



**Modelling *Cryptosporidium* risks in catchments
to protect drinking water supplies**

Alan Smalley

A thesis submitted for the degree of
Doctor of Philosophy

Submitted December 2025

The University of Sheffield
Faculty of Engineering
School of Mechanical, Aerospace and Civil Engineering

Academic Supervisors: Dr. James Shucksmith and Dr. Isabel Douterelo

Industrial Supervisor: Dr. Michael Chipps

*To Mum, Dad
and Clarice*

Declaration

I confirm that the Thesis is my own work and that I am aware of the University's Guidance on Academic Integrity. This work has not previously been presented for an award at this, or any other, university.

The following paper (presented in Chapter 2 of the thesis) has been published previously:
Smalley, A. L., Douterelo, I., Chipps, M. and Shucksmith, J. D. (2025) 'Data-driven prediction of daily *Cryptosporidium* river concentrations for water resource management: Use of catchment-averaged vs spatially distributed features in a Bagging-XGBoost model', *Science of The Total Environment*, 991, pp. 179794. <https://doi.org/10.1016/j.scitotenv.2025.179794>

Abstract

Understanding source water characteristics – and the catchments which influence them – can play a vital role in protecting drinking water supplies. Many water providers collect vast quantities of water quality monitoring data, but this data is often neglected after first use, though it may contain important information for understanding contaminant sources, processes and risks. The research in this thesis examines *Cryptosporidium* – a major waterborne pathogen – by applying machine learning (ML) and Quantitative Microbial Risk Assessment (QMRA) approaches to long-term water quality, hydro-meteorological and geographical datasets. The overall aim is to advance understanding of how catchment conditions influence *Cryptosporidium* river concentrations and the consequent effect on public health risk in drinking water supply systems. The opening work describes a first-of-its-kind ML-based daily *Cryptosporidium* prediction tool for use at a major abstraction site on the River Thames. The principal advance is the comparison and integration of catchment-averaged and spatially distributed environmental and source data to demonstrate predictability at operational timescales and identify potential source areas, conditions and lag times associated with elevated *Cryptosporidium* concentrations. The second study presents a QMRA for London’s drinking water supply, integrating river, reservoir and final treated water data across a complex multi-source system. This enables comparison between loading-sensitive raw water-based and treated water-based risk estimates, improving assessment of seasonal effects and treatment robustness. A third study applies an explainable ML approach to evaluate the impact of static, pseudo-static and dynamic environmental variables across multiple catchments. Higher humidity and recent rainfall events were associated with higher *Cryptosporidium* river concentrations, and 33% of sites were identified as dominated by *Cryptosporidium* shed by human – as opposed to animal – hosts. Taken together, these findings have important implications for drinking water providers and catchment managers, and demonstrate the latent value within large-scale historical monitoring datasets, which can be readily harnessed using ML tools.

Funding

This work was supported by the UK Research and Innovation, Engineering and Physical Sciences Research Council (EPSRC) through their funding of the Water Infrastructure and Resilience (WIRe) Centre for Doctoral Training [EP/S023666/1], with industrial sponsorship by Thames Water.

Collaborative work between the University of Sheffield and the University of São Paulo, which contributed significantly to the contents of Chapter 3, was supported by a Royal Society Grant (IES\R2\232010 - International Exchanges 2023, Round 2).

Acknowledgements

I would like to thank my supervisors, James Shucksmith and Isabel Douterelo, for their help and guidance throughout. I could not have wished for wiser, kinder or more supportive mentors over the past four years.

My industrial supervisor, Michael Chipps, has been a pleasure to work with from day one, and I have been very fortunate to have access to someone with such depth and breadth of knowledge. I greatly appreciate Michael's support and encouragement, and am grateful to him for introducing me to the other Thames Water teams along the way.

Huge thanks are also owed to Dean Khanna, Jo Clint and James Townsend at Thames Water for always being so helpful and providing information whenever I came up with another question or request for data. This thesis would not exist without their help. Thanks also to Ben Martin, Christian King, Vincent Jones, Chiara Coleman, Heide Hardiman, Nadia Begum, Sarah Braceman and Sean Cable at Thames Water, along with many others.

I am also grateful to Vaida Suslovaite for early coding advice and to Ian Guymmer for providing travel time data for the Thames catchment.

Many thanks to Marcelo Laurretto, Maria Tereza Pepe Razzolini, Maria Ines Zanolli Sato and João Carlos Setubal for their time, advice and encouragement. Their support was essential to the QMRA work.

I am very grateful for the friendship and support of Archana Dayal and Manoj Menon – having the chance to be involved in their publications gave me extremely useful experience of the peer-review process early in my PhD studies. I also thank Rob Ashurst for giving me a lab tech job at the Department of Geography a few years ago and for being such a great person to work with. It was my time in that role, supporting students as they undertook their own water quality projects, which ultimately led to me embarking on this PhD.

Thanks are also due to everyone involved in the WIRe Centre for Doctoral Training, including Lindsay, Jackie, Justine, Pete, Luca, Jaime and, of course, all of the WIRe students.

Finally, I would like to give a very big thank you to family and friends, and of course to my mum and dad, who deserve far more thanks than I could ever express here.

Page intentionally left blank

Table of Contents

Declaration.....	iv
Abstract.....	v
Funding	vi
Acknowledgements.....	vii
List of Abbreviations	xiii
List of Figures	xv
List of Tables	xix
Chapter 1: Introduction	1
1.1 Background	1
1.2 Risks of <i>Cryptosporidium</i> to drinking water resources: a brief summary.....	2
1.3 AI tools and concepts relevant to this thesis.....	4
1.3.1 AI and the water sector.....	4
1.3.2 XGBoost.....	5
1.3.3 SHAP.....	8
1.4 Research questions and Thesis structure	9
1.5 Publications and Conference Presentations	11
1.5.1 Journal Papers (published).....	11
1.5.2 Journal Papers (drafted for submission).....	11
1.5.3 Conference Presentations.....	12
Chapter 2: Data-driven prediction of daily <i>Cryptosporidium</i> river concentrations for water resource management: Use of catchment-averaged vs spatially distributed features in a Bagging-XGBoost model	13
2.1 Abstract.....	13
2.2 Introduction	14
2.3 Background	16
2.4 Methods.....	20
2.4.1 Study area and <i>Cryptosporidium</i> monitoring	20
2.4.2 Model type.....	22
2.4.3 Model inputs.....	23
2.4.4 Travel times and time lags	27

2.4.5 Input selection for models	29
2.4.6 Hyperparameter tuning	29
2.4.7 Model validation and performance metrics	31
2.4.8 Local interpretation using SHAP values	32
2.5 Results	33
2.5.1 Lag time and hyperparameter tuning	33
2.5.2 Overall performance and comparison of models	33
2.5.3 SHAP analysis	37
2.5.4 Identifying possible source locations/areas.....	39
2.6 Discussion.....	40
2.6.1 Predicting <i>Cryptosporidium</i> at daily timescales	40
2.6.2 Differential performance of catchment-averaged and spatially distributed models	41
2.6.3 Applying model performance and SHAP analysis to understand <i>Cryptosporidium</i> transport and sources	42
2.6.4 Modelling and feature interpretation limitations.....	45
2.7 Conclusions	47
Chapter 3: Quantitative Microbial Risk Assessment of <i>Cryptosporidium</i> for London's water supply: Analysis of raw and treated water to inform supply management	49
3.1 Abstract.....	49
3.2 Introduction	50
3.3 Method	53
3.3.1 Description of drinking water assets.....	53
3.3.2 <i>Cryptosporidium</i> monitoring.....	55
3.3.3 QMRA workflow	56
3.3.4 Modelling risk reduction under selective abstraction regimes	60
3.3.5 Sensitivity analysis	62
3.4 Results	62
3.4.1 Summary statistics, sampling frequency and <i>Cryptosporidium</i> detections.....	62
3.4.2 Volumetric data and proportional allocation	64
3.4.3 Log-removal values	65
3.4.4 QMRA outputs and selective abstraction modelling	65
3.4.5 Sensitivity analysis	69
3.4.6 Selective abstraction modelling: An example application of the model	70
3.5 Discussion.....	71
3.5.1 Overall infection risk	71

3.5.2 Effectiveness of selective abstraction.....	74
3.5.3 Application of QMRA to complex systems.....	75
3.5.4 Analysis of log-removal modelling.....	76
3.5.5 Key determinants of risk	77
3.5.6 Example applications of RW-QMRA: SARs and other risk mitigation strategies	79
3.6 Conclusions	79
Chapter 4: Environmental factors and sources influencing <i>Cryptosporidium</i> river	
concentrations: A data-rich spatio-temporal analysis using Explainable AI.....	81
4.1 Abstract.....	81
4.2 Introduction	82
4.3 Method	85
4.3.1 <i>Cryptosporidium</i> data acquisition and cleaning.....	85
4.3.2 Summary statistics for spatial and temporal analysis.....	86
4.3.3 Environmental data and GIS processing	86
4.3.4 Multi-catchment <i>Cryptosporidium</i> ML model (XGBoost)	89
4.3.5 Explainable AI (SHAP analysis)	90
4.4 Results.....	92
4.4.1 Spatial and temporal analysis	92
4.4.2 XGBoost model performance.....	97
4.4.3 Feature interpretation using SHAP outputs.....	99
4.4.4 Dominant source types for each site	101
4.5 Discussion.....	102
4.5.1 <i>Cryptosporidium</i> spatial patterns and hotspots.....	103
4.5.2 <i>Cryptosporidium</i> seasonality.....	103
4.5.3 Key environmental variables affecting <i>Cryptosporidium</i> river concentration.....	105
4.5.4 Spatial averaging: considerations and limitations	106
4.5.5 Major sources of <i>Cryptosporidium</i> in rivers.....	107
4.6 Conclusions	109
Chapter 5: Discussion.....	
5.1 Overall synthesis and contribution of the thesis	111
5.2 Immediate impact: trial and implementation of river prediction tool	112
5.2.1 Context.....	112
5.2.2 Operational trial: method	113
5.2.3 Operational trial: results	113
5.2.4 Operational implementation	116

5.3 Scientific implications, limitations and future directions	117
5.3.1 <i>Cryptosporidium</i> sources and environmental drivers.....	117
5.3.2 <i>Cryptosporidium</i> risk assessment.....	119
5.3.3 ML modelling: strengths and limitations	120
5.3.4 Uncertainties in routine <i>Cryptosporidium</i> monitoring data	122
5.3.5 Future work.....	124
Chapter 6: Conclusions	125
6.1 Research Question 1	125
6.2 Research Question 2	126
6.3 Research Question 3	127
6.4 General advances in understanding	127
References	131
Appendix A: Supplementary Material for Chapter 2	157
A.1 Soil Moisture and Soil Temperature Data	157
A.2 Estimation of travel times (spatially distributed model only)	159
A.2.1 Jobson travel time calculation	159
A.2.2 Model Calibration	160
A.2.3 Error Analysis	161
A.3 Results from Hyperparameter Tuning	163
Appendix B: Supplementary Material for Chapter 3	165
B.1 Parametric Modelling to Fix Lower Limits	165
B.2 Pooled Dose-Response Model	167
B.2.1 Overview	167
B.2.2 Isolate-Specific Bootstrap Modelling	167
B.2.3 Hierarchical Bayesian Model	168
B.2.4 Pooled DR QMRA Outputs	169
Appendix C: Supplementary Material for Chapter 4	173

List of Abbreviations

AI – artificial intelligence

Coag-Floc – coagulation and flocculation

CoCo-DAFF – Counter-current dissolved air flotation/filtration

CSO – combined sewer overflow

DR – dose-response

DWI – Drinking Water Inspectorate (for England and Wales)

DWQR – Drinking Water Quality Regulator (for Scotland)

EA – Environment Agency (for England and Wales)

EDM – event duration monitoring

FIB – faecal indicator bacteria

GIS – geographical information system

IMS-IFM – immunomagnetic separation-immunofluorescence microscopy

KNN – K-nearest neighbours

LDFA - linear discriminant function analysis

LRC – log-removal credit

LRV – log-removal value

ML – machine learning

OS – Ordnance Survey

PCR – polymerase chain reaction

QMRA – quantitative microbial risk assessment

qPCR – quantitative polymerase chain reaction

RF – random forest

RGF – rapid gravity filtration

RMSE – root mean square error

SHAP – SHapley Additive exPlanations

SSF – slow sand filtration

STW – sewage treatment works

SVM – support vector machines

SWAT – soil water assessment tool

TW - Thames Water

UNESCO – United Nations Educational, Scientific and Cultural Organization

USEPA – United States Environmental Protection Agency

WHO – World Health Organization

WTW – water treatment works

XAI – eXplainable Artificial Intelligence

XGBoost – eXtreme Gradient Boosting

List of Figures

Fig. 1.1 Simplified structure of an XGBoost regression model, as used in Chapters 2 and 4.	6
Fig. 2.1 Summary of key characteristics and input parameters used in previous <i>Cryptosporidium</i> models. (Notes: *Spatial distribution often varies with input type, hence sometimes two figures are provided; *ML models (all other models are deterministic); **SWAT models; ***New models (presented in this paper). For deterministic models, the number of <i>Cryptosporidium</i> river measurements used for validation (v) is provided only, whereas for ML models, the number of both training (t) and validation (v) data points is provided.)	17
Fig. 2.2 Maps showing the catchment location on the mainland and the towns, cities and river network within the catchment boundary, along with the five major abstraction points on the River Thames which provide raw water for approximately 70% of London’s drinking water. Base map: OS MiniScale (Ordnance Survey, 2025).	21
Fig. 2.3 Estimated travel times for 5x5 km cells under three different flow regimes in the study catchment	28
Fig. 2.4 Diagram of model processes and model outputs, showing dates (T) over which training and validation data is split for walk forward validation, alongside key hyperparameters, outputs and evaluation metrics.	31
Fig. 2.5 (a) Measured and predicted concentrations of <i>Cryptosporidium</i> (generated by Models 1b and 5b), (b) catchment-averaged rainfall and (c) discharge. SHAP values showing the impact on Model 1b predictions of (d) core inputs and (e) hydrological inputs and (f) soil meteorological inputs (presented as cumulative values for all lag times), and (g) the impact on Model 5b of rainfall cells and CSOs as cumulative totals. (Note: for simplicity, the effect of SHAP base values has been omitted.)	36
Fig. 2.6 Maximum positive and negative SHAP impact scores for key individual features and grouped features in (a) Model 1b and (b) Model 5b. (Note: catchment-averaged and spatially distributed features with an absolute maximum impact of <0.25 are grouped together by feature type; RD = River Distance (km) from Walton intake; Ln=lag time, expressed as n days.).....	38
Fig. 2.7 Maximum SHAP impact scores for (a) 5x5 km rainfall cells and (b) CSO binary inputs, with (c) human population density from 2011 Census (Reis et al., 2017), alongside livestock densities for (d) cattle, (e) pigs (f) poultry and (g) sheep from APHA (2022).	39
Fig. 3.1 Schematic map of the major Thames Water assets supporting London’s drinking water supply.....	54

Fig. 3.2 <i>Cryptosporidium</i> detections for the 6-year period of study, plotted by day-of-year, showing (a) river intakes, (b) WTW inlets and (c) WTW outlets.	63
Fig. 3.3 Water volume data for the study period (01.03.2018 to 29.04.2024) showing (a) proportions of total volumes abstracted from intakes A-F, (b) proportions of total output volumes generated by WTWs 1-5 and (c) total intake volumes supplying individual WTWs (based on proportional allocation calculations).	64
Fig. 3.4 Modelled distributions of (a-e) storage stage log-removal values ($LRV_{storage}$) and (f-j) treatment stage log-removal values ($LRV_{treatment}$).	66
Fig. 3.5 Annual infection risk distribution outputs from Iowa DR QMRA models for biannual periods, based on <i>Cryptosporidium</i> concentrations in (a) treated water, and (b-e) river water with (b) no selective abstraction, (c) SAR-1, (d) SAR-2 and (e) SAR-H.	67
Fig. 3.6 Sensitivity analysis based on Spearman correlation coefficient (r), for (a) TW-QMRA (b) RW-QMRA (without SAR), using the Iowa DR model.	69
Fig. 3.7 Annual infection risk (for all biannual periods combined, 2018-2024) plotted against annual abstraction volume reduction for different SARs. (For Iowa DR QMRAs only and idealised SAR implementation, with smoothed mean daily abstraction volumes for 2018-2024.)	71
Fig. 4.1 Locations on the British mainland of (a) sampling sites (with Site Numbers), (b) long-term mean <i>Cryptosporidium</i> concentrations ($\bar{C}_{LT}$), (c) peak <i>Cryptosporidium</i> season and seasonal amplitude, and (d) bar chart of peak seasons.	93
Fig. 4.2 Boxplots of catchment-averaged environmental variables grouped by peak season. Significant Tukey HSD differences ($\alpha = 0.05$) were identified between Summer- and Winter-peak sites only. <i>Cryptosporidium</i> concentration is shown for reference, although no seasonal differences were significant.	96
Fig. 4.3 Correlations between <i>Cryptosporidium</i> concentration and seasonal amplitude with environmental data, based on long-term (≤ 24 years) and full catchment averages. Results are provided for all 52 sites and for sites defined by peak season. Spring-peaking sites were omitted due to low ($n=2$) numbers.	97
Fig. 4.4 Out-of-fold (OoF) predictions from (a) Model 1 and (b) Model 4.	98
Fig. 4.5 Ranked Mean Absolute SHAP (MAS) values for each feature in the four models. (fc= full catchment).	99
Fig. 4.6 Sites characterised by the relative contribution of source types to the modelled <i>Cryptosporidium</i> concentrations, shown by (a) their location and (b) the proportion of total sites.	102
Fig. 5.1 Example outputs from Model A as viewed by end-users during operational trials, showing (a) predicted <i>Cryptosporidium</i> concentrations alongside a prediction range (based on the mean $\pm 1SD$);	

and (b) predicted threshold intervals/less-than values. Actual (measured) <i>Cryptosporidium</i> concentrations are displayed in the right-hand columns in both tables.	114
Fig. A.1 Locations of the 4 COSMOS monitoring stations in the Walton Catchment.	157
Fig. A.2 Mean daily soil moisture (as % volumetric water content at 5cm depth) for the 4 COSMOS monitoring sites in the Walton catchment.	157
Fig. A.3 Mean daily soil temperature (in °C at 5cm depth) for the 4 COSMOS monitoring sites in the Walton catchment.	158
Fig. A.4 Measured and calculated (using Equation A.1) velocities for three sections of the study catchment.	161
Fig. A.5 Empirical cumulative distribution error plots for velocity calculations.	162
Fig. A.6 F-score outputs from hyperparameter tuning using Model 1b (with the second fewest features) and model 5b (with the largest number of features). Vertical yellow lines indicate “optimum” value based on an assessment of combined performance at the three thresholds. Because of the greater number of low exceedances, high F1-score and F2-scores were regarded as more important than high F3-scores. (Note that in each trial the optimum values for the other three elements were used.)	163
Fig. B.1 MLE-fitted models for <i>Cryptosporidium</i> concentration data from all intakes. The “best” model (based on AIC outputs) is given in the shaded box, along with the Q50 value for that model.	165
Fig. B.2 MLE-fitted models for <i>Cryptosporidium</i> concentration data from all WTW inlets. The “best” model (based on AIC outputs) is given in the shaded box, along with the Q50 value for that model.	166
Fig. B.3 MLE-fitted models for <i>Cryptosporidium</i> concentration data from all WTW outlets. The “best” model (based on AIC outputs) is given in the shaded box, along with the Q50 value for that model.	166
Fig. B.4 Bootstrapped dose-response (k) value distributions for three <i>C. parvum</i> isolates and combined distribution generated by hierarchical Bayesian model.	168
Fig. B.5 Annual infection risk distribution outputs from the pooled DR QMRA models for biannual periods, based on <i>Cryptosporidium</i> concentrations in (a) treated water, and (b-e) river water with (b) no selective abstraction, (c) SAR-1, (d) SAR-2 and (e) SAR-H.	170
Fig. B.6 Sensitivity analysis based on Spearman correlation coefficient (r), for (a) TW-QMRA (b) RW-QMRA (without SAR), using the pooled DR model.	171
Fig. C.1 Out-of-fold (OoF) predictions from (a) Model 2 and (b) Model 3.	173

Page intentionally left blank

List of Tables

Table 2.1 Input parameters included in each of the tested models, as indicated by greyed-out cells. (Sp. Dist. = Spatially Distributed)	30
Table 2.2 Optimised maximum lag time and hyperparameter settings following initial trials	33
Table 2.3 Performance metric results for the different models. Blank values arise where True or False Positives were not predicted by the model. Cell shading denotes highest, intermediate and lowest performing models for each metric	34
Table 3.1 Summary of Log-removal Credit (LRC) estimates for WTWs and <i>Cryptosporidium</i> -relevant treatment processes. (Coag-Floc = Coagulation-Flocculation; CoCo-DAFF = Counter-Current Dissolved Air Flotation/Filtration.) LRCs taken from Hijnen and Medema (2007) and Edzwald et al. (2001).....	55
Table 3.2 Trigger thresholds applied in the selective abstraction-based risk management of <i>Cryptosporidium</i> at Thames Water (Thames Water, 2021)	61
Table 3.3 Summary statistics and measurement counts for the intakes, WTW inlets and WTW outlets. (Cen = Censored values (non-detects); Det= Detections (uncensored values). For concentration statistics, censored data points are here treated as 0 oocysts L ⁻¹ . Shading indicates highest, intermediate and lowest concentrations.).....	63
Table 3.4 Annualised infection risk values for QMRA outputs across the full study period (01.03.2018 to 29.02.2024). Highlighted value = above USEPA guideline.....	68
Table 4.1 Environmental datasets used as features in the XGBoost model.....	87
Table 4.2 <i>Cryptosporidium</i> measurement counts, frequency, periods, summary concentration statistics and seasonality for all sites included in the study. Sites with ≥ 1000 data points or ≤10-day mean interval are highlighted green; sites with concentrations/seasonal amplitude ≥ 1 oocysts L ⁻¹ are highlighted pink.	94
Table 4.3 Significant pairwise environmental differences between summer-peaking and winter-peaking catchments, using the Tukey HSD post-hoc test ($p < \alpha$, where $\alpha = 0.05$). Note that no significant differences were found for other season pairs.	95
Table 4.4 Data filtering settings and performance metrics from the four models.	98
Table 4.5 Base features for all four models ranked by Mean Absolute SHAP (MAS), alongside direction of effect and most predictive spatial scale. (Colour ramp indicates high, intermediate and low MAS values; SD=Standard Deviation; POS = Positive; NEG = Negative.)	100
Table 5.1 End-user feedback following operational trials	115

Table A.1 Matrix showing Pearson correlation coefficients for the soil moisture data at the 4 COSMOS monitoring sites in the Walton catchment, and the mean of the 4 sites (for 01.01.2016 – 01.04.2023 data).....	158
Table A.2 Matrix showing Pearson correlation coefficients for the soil temperature data at the 4 COSMOS monitoring sites in the Walton catchment, and the mean of the 4 sites (for 01.01.2016 – 01.04.2023 data).....	158
Table A.3 Curve-fitted constants for Equation A.1 applied to three sections of the study catchment.	160
Table A.4 Estimated error in the predicted travel times for cells at three different river distances from the abstraction point when ± 1 SD in the calculated velocities was applied.....	162
Table B.1 Challenge test data used in pooled DR QMRA model. (from Messner et al., 2001).	167
Table C.1 Top 20 SHAP interaction contributions for humidity features with land use and source-related features.	173

Chapter 1: Introduction

1.1 Background

Drinking water providers around the world are under increasing pressure to meet the challenges posed by population growth, expanding land development and climate change (UNESCO, 2020). In many countries, droughts and intensive rainfall events are expected to become both more frequent and more severe, leading to seasonal and/or year-long reductions in the quantity of water available for abstraction, while at the same time decreasing the quality of raw water entering drinking water treatment works (WTWs), due to the mobilisation of contaminants by overland flow, soil erosion and flooding (Jones, Bierkens and van Vliet, 2024). The need to include catchment and raw water information in drinking water quality risk management is now written into legislation, with the European Union (2020) Drinking Water Directive mandating that water providers take a source-to-tap approach to maintaining safe drinking water.

More informed, data-driven approaches to drinking water management have the potential to mitigate some of these challenges. The increasing availability of remote sensing data, combined with an explosion in Artificial Intelligence (AI)-based modelling tools, provide rich new opportunities for understanding catchments, protecting both drinking water supplies and the environments upon which they depend. Running alongside this are the vast datasets which have been amassed by water providers around the world; in some cases, because of the regularity of sampling and the consistency of methods applied, these constitute the most substantial water quality datasets available in their countries of origin. Unfortunately, such data – which is primarily collected to meet regulatory requirements and/or provide a short-term measure of water quality status – is often permanently archived shortly after collection, wasting a potentially valuable resource which could be interrogated to deliver scientific insight and understanding.

The concept of untapped monitoring data underpins the three research projects presented in this thesis, each of which seeks to exploit historical datasets to generate learning which can aid or inform drinking water providers. The focus is both on adding to a wider understanding of the threat discussed, and to help provide tools which can lead to a more

data-driven approach to decision-making, at timescales which are relevant to abstraction and water quality managers.

1.2 Risks of *Cryptosporidium* to drinking water resources: a brief summary

The work contained in this thesis focuses on *Cryptosporidium*, a waterborne pathogen which can cause acute gastroenteritis in humans and animals (Ryan et al., 2021). Although *Cryptosporidium* has multiple routes of transmission, infection by contaminated drinking water has been associated with the largest outbreaks on record (Efstratiou et al., 2017a). The most severe example of this occurred in Milwaukee, Wisconsin in 1993, when approximately 403,000 people were infected, leading to at least 54 deaths (Mac Kenzie et al., 1994; Hoxie et al., 1997). The severity and scale of this event had an impact on water providers around the world, leading to a much greater awareness of the threat posed by *Cryptosporidium* to drinking water supplies, and the need for appropriate treatment processes and monitoring measures to prevent future outbreaks.

Because of the resistance of the pathogen to chlorine disinfection, drinking water safety typically relies on multi-barrier treatment methods in combination with other risk management strategies (Betancourt and Rose, 2004); these can include catchment management approaches to reduce contamination of raw waters, frequent monitoring of source waters to gauge *Cryptosporidium* loadings, and/or selective abstraction to reduce loadings when raw water concentrations are high. The methods for benchmarking risk vary around the world, with some countries (e.g. the Netherlands and Australia) opting for a health burden/infection risk-based approach, whereas other countries (including the UK) apply concentration-based limits for pathogens in the treated water and then estimate the probability of these limits being exceeded (WHO, 2016).

In the UK, a spate of small outbreaks in the 1980s (Nichols et al., 2006), resulted in the formation of the *Expert Group on Cryptosporidium in Water* in 1989, the findings from which were published as the Badenoch Reports (Badenoch, 1990; 1995), followed by a third report by Bouchier (1998). These have formed the basis for *Cryptosporidium* risk management by water providers in the UK, though legislation regarding monitoring has passed through two distinct phases since that time. In 2000-2007, continuous sampling of treated drinking

water for *Cryptosporidium* analysis was mandated for sites designated as of “significant risk”, and which did not have membrane filters in place. This statute was then removed (UK Government, 2007), after which greater emphasis was placed on targeted monitoring and more detailed risk assessments to demonstrate robustness during rare events – e.g. when raw concentrations are very high, or during treatment failures (DWI, 2021). Monitoring of treated waters remains in place at many sites, however, and between 2005-2022, 168 *Cryptosporidium* detections were recorded in England and Wales (DWI, 2025a). Occasional outbreaks continue to be linked to drinking water contamination, with 231 confirmed infections following an incident in Cwellyn in 2006 (DWI, 2025a), and an estimated 422 infections resulting from the Pitsford incident in 2008 (Puleston et al., 2014).

Cryptosporidium concentrations in raw waters show high spatial and temporal variability (Vermeulen et al., 2019). They are often characterised by low baseline levels interrupted with sudden episodic increases (Bouchier, 1998). Understanding when and why these spikes occur would be an important contribution to *Cryptosporidium* risk management (Ligda et al., 2024). Although the main sources of *Cryptosporidium* are generally understood – i.e. animal waste mobilised from pastureland/manure-fertilised fields, and human waste released by the sewage network – there remains considerable disagreement about the relative importance of these sources at different locations and different times (Bukhari et al., 1997; Sturdee et al., 2007; Swaffer et al., 2014). Rainfall, a key environmental driver, is known to be important in mobilising and transporting the organism from source areas to water bodies, but the effects are non-linear and earlier work by modellers has found that rainfall data alone cannot be used as a reliable predictor of *Cryptosporidium* concentrations (Atherholt et al., 1998; Ligda et al., 2024).

In the absence of other available data sources, previous studies have necessarily relied on small datasets collected for specific research purposes – e.g. Tang et al. (2011), Sato et al. (2013), Bataiero et al. (2019) and Liu et al. (2019). The routine monitoring datasets collected by drinking water providers in the UK and other countries around the world therefore constitute a significant resource which, until now, has remained largely unexplored and underutilised. When combined with recently-developed AI-based data analysis and modelling tools, such datasets are arguably more valuable today than ever before, providing new opportunities to advance understanding on a range of water quality

issues – understanding which can then be applied to protect both the environment and drinking water supplies.

1.3 AI tools and concepts relevant to this thesis

1.3.1 AI and the water sector

The use of AI is expected to revolutionise the water sector in the years ahead (UNESCO, 2025a). Its effects are already being felt, with AI-based approaches playing an increasing role in operational settings, where they have been applied for a wide range of purposes, including water quality assessment, wastewater treatment and control, and distribution network management (International Water Association, 2020). Machine learning (ML) algorithms are an important subcategory of AI, most of which work on the basis of training models to identify (i.e. “learn”) patterns in data which are then applied to generate outputs in the form of either predictions, classifications or decisions. ML models are commonly divided into two types: (i) unsupervised models, which train on unlabelled data, usually to find clusters or reduce dimensionality; and (ii) supervised models, which are trained using labelled input data to predict a target dataset (Hastie et al., 2009). Commonly used examples of supervised models include linear regression, artificial neural networks, tree-based ensemble models (e.g. random forest) and support vector machines (Zhu et al., 2022).

In more recent years, the application of ML models has expanded beyond the generation of primary model outputs (e.g. predictions or classifications), to analysis of the importance of different inputs (or “features”), so as to increase understanding of the systems being modelled. Models differ in their degree of interpretability, ranging from fully transparent (“white-box”) models – e.g. linear regression – in which the relationship between inputs and outputs is explicitly defined, to complex (“black-box”) models, whose internal logic cannot be directly inspected or expressed in a simple, human-readable form. Many modern models fall between these two extremes (“grey-box models”); for these, feature effects can be approximated using post-hoc methods, but the complete decision-making process cannot be exactly characterised. The interpretation of model inputs to understand systems is often referred to by the umbrella term, eXplainable AI (XAI) (Zhu et al., 2023).

XAI is very much an emerging field, and one which promises to transform how systems are understood and managed. The environmental and water quality sectors have historically relied on process-based models, grounded in known physical, chemical and/or biological mechanisms. While these models provide mechanistic insight, they can struggle to represent complex interactions and nonlinearity – situations where ML-based models are often most effective. Where data of sufficient quality and quantity is available to train models, ML and XAI tools therefore offer powerful new alternatives for advancing understanding across all scientific disciplines (Minh et al., 2022).

The work detailed in this thesis exploits this still developing field by making extensive use of a supervised machine learning model, eXtreme Gradient Boosting (XGBoost), in combination with one of the most widely used XAI methods - SHapley Additive exPlanations (SHAP). A brief explanation of both is provided below.

1.3.2 XGBoost

XGBoost is an open source ML library which can be applied to both regression and classification tasks (Chen and Guestrin, 2016). It belongs to the family of ensemble tree-based models, in which multiple decision trees are trained and the outputs from each tree combined into a single model output. Gradient boosted models, such as XGBoost, extend this approach by training trees sequentially in an additive model (“boosting”) and applying gradient-based optimisation to correct the errors (“residuals”) in previous trees (Fig. 1.1).

The “extreme” element relates to a number of additional optimisation steps, including regularisation to reduce overfitting (by removing weak, unstable effects), efficient tree “pruning” (to remove unhelpful detail) and improved split finding (to identify useful decision points). Unlike many similar models, XGBoost can also handle missing data, by incorporating data absence as a learning element (Piraei et al., 2023).

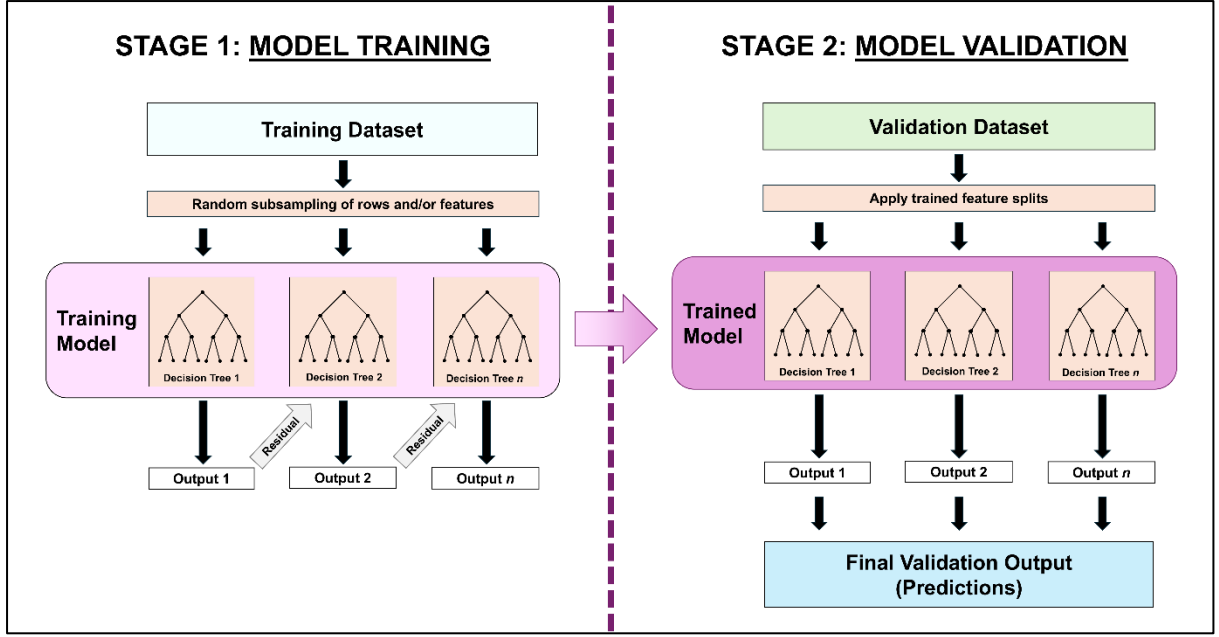


Fig. 1.1 Simplified structure of an XGBoost regression model, as used in Chapters 2 and 4.

The model can be defined by a series of equations provided by Chen and Guestrin (2016), beginning with the shape of the regression trees ($\mathcal{F}$), which for this (and other tree-based models) is often given as:

$$\mathcal{F} = \{f(x) = w_{\{q(x)\}} \mid q : \mathbb{R}^m \rightarrow \{1, \dots, T\}, w \in \mathbb{R}^T\} \quad (1)$$

where $x \in \mathbb{R}^m$ denotes an input data point with m features; $q(x)$ assigns x to one of T leaves; and $w \in \mathbb{R}^T$ stores the corresponding leaf weights.

Given this function, XGBoost models the prediction as an additive ensemble of regression trees:

$$\hat{y}_i = \phi(x) = \sum_{k=1}^K f_k(x_i), f_k \in \mathcal{F} \quad (2)$$

where $\hat{y}_i$ denotes the predicted output for data point i ; x_i is the corresponding input feature vector; $\phi(x_i)$ is the model prediction; and each $f_k \in \mathcal{F}$ is a regression tree whose outputs are summed over K trees.

Model parameters are learned by minimising a regularised objective function:

$$L(\phi) = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \quad (3)$$

where $L(\phi)$ denotes the regularized objective function; $l(y_i, \hat{y}_i)$ measures the loss for data point i ; and $\Omega(f_k)$ penalizes the complexity of each regression tree. The regularisation term is defined as:

$$\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (4)$$

where T is the number of leaves in the tree; w_j denotes the weight of leaf j ; and γ and λ are regularization parameters. This term discourages overly complex trees by penalizing both the number of leaves and the magnitude of leaf weights.

At boosting iteration t , the objective after adding a new tree (f_t) becomes:

$$L^{(t)} = \sum_i l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_k) \quad (5)$$

where $L^{(t)}$ denotes the objective at iteration t ; $\hat{y}_i^{(t-1)}$ is the current prediction for data point i ; and $f_t(x_i)$ is the contribution of the newly added regression tree.

The loss is then approximated using a second-order Taylor expansion:

$$L^{(t)} = \sum_i \left[l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2 \right] + \Omega(f_t) \quad (6)$$

where:

$$g_i = \frac{\partial l(y_i, \hat{y})}{\partial \hat{y}} \quad \text{and} \quad h_i = \frac{\partial^2 l(y_i, \hat{y})}{\partial \hat{y}^2} \Big|_{-\hat{y} = \hat{y}_i^{(t-1)}} \quad (7)$$

Here, g_i and h_i denote the first- and second-order derivatives (gradient and Hessian) of the loss with respect to the prediction. In practice, these derivatives are evaluated at the current prediction, $\hat{y}_i^{(t-1)}$, prior to adding the new tree.

Equations 6 and 7 represent one of the main extreme boosting elements of XGBoost. While standard gradient boosting relies on first-order gradients (g_i), XGBoost additionally incorporates second-order information through the Hessian terms (h_i).

Equations governing tree construction, split selection, and pruning are not reproduced here, as they pertain primarily to implementation details rather than the core conceptual formulation (further details are available from Chen and Guestrin (2016)).

XGBoost has been used to model time series and to make predictions or classifications in many environmental and water quality applications (Wu et al., 2021a; Masood et al., 2023; Rawat et al., 2023). It regularly out-performs other ML models in both its predictive performance and computation time (Ligda et al., 2024), and is noted for its reliability when modelling datasets with high dimensionality, or multiple correlating features (“multicollinearity”) (Drobnič et al., 2020).

1.3.3 SHAP

SHAP is a game theory-based post-hoc method for explaining the outputs of ML models in terms of feature contributions (Lundberg and Lee, 2017). These are compiled by comparing model outputs with selected features in multiple combinations, to generate average estimates of feature contribution (known as “SHAP values”). Although SHAP is model-agnostic (i.e. usable with a wide variety of ML model types), a tree-specific version, known as Tree SHAP, can be applied to models such as XGBoost; this uses the explicit tree structure to efficiently compute SHAP outputs (Lundberg et al., 2018).

SHAP values quantify the effect of a feature on individual predictions (“local interpretability”) and on predictions across the full validation dataset (“global interpretability”). The direction of a feature’s effect on predictions (i.e. positive or negative) is described at the local level by the SHAP value, and at the global level by SHAP “dependency”, in which the relationship is calculated/plotted between all predictions and all SHAP values for a given feature (Molnar, 2022).

There are numerous examples of the combined use of XGBoost and SHAP – e.g. studies seeking to understand spatial effects on landslide susceptibility (Zhang et al., 2023), or interpret time series data to elucidate environmental effects on river sediment loads (Piraei et al., 2023), and in water quality research, the most relevant example of which is the work of Ligda et al. (2024), to connect meteorological, hydrological and water quality parameters with *Cryptosporidium* and *Giardia* river concentrations.

It should be noted that XGBoost, in common with many tree-based models, can generate feature importance metrics independently of SHAP. However, these metrics are limited in that they do not convey the direction of a feature’s effect and they cannot be used to analyse local, instance-level contributions. Model-agnostic approaches such as SHAP

therefore play a crucial role in the interpretive application of ML models, enabling a more detailed understanding of the factors driving individual predictions.

1.4 Research questions and Thesis structure

The research detailed in this thesis focuses on *Cryptosporidium* – a decision which was taken in the initial phase of the project after discussions with the research sponsor, Thames Water (TW). For the reasons outlined above (Section 1.2) and provided in more detail in the individual chapters of this thesis, managing the *Cryptosporidium* risk to drinking waters is a significant challenge for many water providers around the world, and there are multiple knowledge gaps regarding the pathogen. The key research questions addressed in this work are therefore as follows.

1. What is the potential to develop data-led predictive models to provide warning of periods of high *Cryptosporidium* at daily resolution in a large and complex catchment (such as the Thames)? Further, how can such a model be interpreted to understand key sources and transport mechanisms?
2. How can monitoring datasets and asset information be applied to better understand and quantify *Cryptosporidium* risk to drinking water supplies serving large population centres, and how can such understanding be used to inform effective management strategies?
3. How can wider environmental datasets be used to understand the key drivers, sources and environmental factors that influence the spatial and temporal variability in observed source water *Cryptosporidium*?

In this thesis, the questions above are addressed sequentially in three key chapters, formatted as research papers. Each chapter/paper contains a separate literature review to contextualise and justify the identified research question. Hence the thesis consists of six chapters, with **Chapters 2-4** based on published/draft papers, which detail the most significant research contributions of this work.

Chapter 2 describes the development and testing of a data-driven tool for predicting *Cryptosporidium* concentrations at a major river abstraction site serving London's WTWs. This project – which provided the initial impetus for the work as a whole – was motivated by a specific remit from TW to increase the effectiveness of their current selective abstraction-

based risk mitigation strategy by providing advance notice of elevated *Cryptosporidium* river concentrations. It applies a readily adoptable (and adaptable) machine learning (ML) method, using publicly-available environmental data in combination with TW's own water quality monitoring data. Further, it deploys explainable AI methods to attempt to explore model sensitivities and identify key *Cryptosporidium* sources within the Thames catchment.

Chapter 3 details a whole-city *Cryptosporidium*-based Quantitative Microbial Risk Assessment (QMRA) study for London's drinking water. It applies extensive TW monitoring datasets at multiple sites and three stages in the water treatment system (river intake, reservoir, and WTW outlets) to assess both current treatment performance and demonstrate the potential of this approach to scenario-test alternative risk mitigation strategies. The work is seen in part as a call to the UK water sector to adopt QMRA-based approaches to understanding risk, particularly when considering major changes to treatment systems, such as those currently being planned at TW.

Chapter 4 takes a much broader geographical view, collecting together a vast *Cryptosporidium* river concentration dataset (consisting of $\approx 52,000$ measurements) from the past 25 years and covering sites around the British mainland. This data is analysed to understand broad spatial and temporal trends, and an ML model is combined with XAI methods to identify non-linear patterns, revealing the key dynamic and static environmental factors associated with high *Cryptosporidium* concentrations and the dominant sources in different locations.

Chapter 5 provides a discussion of the research project as a whole and presents further details of an operational trial of the *Cryptosporidium* river prediction tool (explained in Chapter 2) at a TW abstraction site in 2024-25. It summarises the insights gleaned from the trial and briefly describes the current roll-out of the tool for full operational use across multiple sites. This chapter then brings together the key scientific implications, limitations and future avenues of research.

Chapter 6 briefly summarises the main conclusions and contributions of the work presented in this thesis.

1.5 Publications and Conference Presentations

Outputs from this research project have been disseminated in several publications and conference presentations, which are summarised below.

1.5.1 Journal Papers (published)

Smalley, A. L., Douterelo, I., Chipps, M. and Shucksmith, J. D. (2025) 'Data-driven prediction of daily *Cryptosporidium* river concentrations for water resource management: Use of catchment-averaged vs spatially distributed features in a Bagging-XGBoost model', *Science of The Total Environment*, 991, pp. 179794.

<https://doi.org/10.1016/j.scitotenv.2025.179794> (Chapter 2 is a reproduction of this paper.)

Author Contributions:

- **Alan Smalley** performed preliminary data analysis, proposed and developed the model, wrote and tested all code, generated and analysed all model outputs, composed the first draft and edited all subsequent drafts following feedback.
- **James Shucksmith** (academic supervisor) provided critical feedback and input into the proposed study, the modelling method and the interpretation of outputs, advised on structural issues and offered feedback at multiple stages in the drafting process.
- **Isabel Douterelo** (academic supervisor) provided critical feedback and input into the modelling method and the interpretation of outputs, advised on chapter structure and offered feedback on later drafts.
- **Michael Chipps** (industrial supervisor) identified the operational relevance of the research topic in line with Thames Water priorities, provided information regarding the real-world application of the model and input into the interpretation of model findings, and offered feedback on later drafts.

1.5.2 Journal Papers (drafted for submission)

Smalley, A. L., Lauretto, M. S., Razzolini, M. T. P., Douterelo, I., Chipps, M. and Shucksmith, J. D. (2026) Quantitative Microbial Risk Assessment of *Cryptosporidium* for London's water supply: Analysis of raw and treated water to inform supply management. *Intended journal: Water Research*. (Chapter 3 is a reproduction of this paper.)

Author Contributions:

- **Alan Smalley** performed preliminary data analysis, proposed and developed the QMRA modelling framework, wrote and tested all code, generated and analysed all model outputs, composed the first draft and edited all subsequent drafts following feedback.
- **James Shucksmith** (academic supervisor) provided critical feedback and input into the proposed study, the modelling method and the interpretation of outputs, advised on structural issues and offered feedback at multiple stages in the drafting process.
- **Isabel Douterelo** (academic supervisor) provided critical feedback and input into the modelling method and the interpretation of outputs and offered feedback on later drafts.
- **Michael Chipps** (industrial supervisor) provided input into the modelling method and the interpretation of outputs, and offered feedback on later drafts.

- **Marcelo Lauretto** and **Maria Tereza Razzolini** (research collaborators) contributed guidance on standard QMRA and statistical methods and provided feedback on later drafts.

Smalley, A. L., Douterelo, I., Chipps, M. and Shucksmith, J. D. (2026) Environmental factors and sources influencing *Cryptosporidium* river concentrations: A data-rich spatio-temporal analysis using Explainable AI. *Intended journal: Water Research*. (Chapter 4 is a reproduction of this paper.)

Author Contributions:

- **Alan Smalley** proposed the idea for the study, obtained the data, performed data analysis, trialled and developed the model, wrote all code, generated and analysed all model outputs, composed the first draft and edited all subsequent drafts following feedback.
- **James Shucksmith** (academic supervisor) provided critical feedback and input into the proposed study, the modelling method and the interpretation of outputs, advised on structural issues and offered feedback at multiple stages in the drafting process.
- **Isabel Douterelo** and **Michael Chipps** (academic and industrial supervisors) provided critical feedback and input into the modelling method and the interpretation of outputs, and offered feedback on later drafts.

1.5.3 Conference Presentations

Smalley, A. L., Shucksmith, J. D., Douterelo, I., Speight, V. and Chipps, M. (2023) 'Understanding *Cryptosporidium* trends in a major UK river system.' *40th IAHR World Congress (Vienna, 2023)*.

Smalley, A. L., Shucksmith, J. D., Douterelo, I. and Chipps, M. (2025) 'Implementing a machine learning approach for predicting *Cryptosporidium* river concentrations to assist operational water resource decisions on the River Thames.' *SimHydro 2025: Which data for water and models? (Nice, 2025)*.

Smalley, A. L., Douterelo, I., Chipps, M. and Shucksmith, J. D. (2025) 'Predicting *Cryptosporidium* river concentrations: The development and testing of a data-driven model for operational use.' *CCWI 2025 - 21st Computing & Control for the Water Industry Conference (Sheffield, 2025)*. doi.org/10.15131/shef.data.29920904

Chapter 2: Data-driven prediction of daily *Cryptosporidium* river concentrations for water resource management: Use of catchment-averaged vs spatially distributed features in a Bagging-XGBoost model

Reproduced from: Smalley, A. L., Douterelo, I., Chipps, M. and Shucksmith, J. D. (2025) 'Data-driven prediction of daily *Cryptosporidium* river concentrations for water resource management: Use of catchment-averaged vs spatially distributed features in a Bagging-XGBoost model', *Science of The Total Environment*, 991, pp. 179794. <https://doi.org/10.1016/j.scitotenv.2025.179794>

2.1 Abstract

Cryptosporidium is a waterborne pathogen which poses a major challenge to water utilities because of its resistance to chlorination and its infectivity at very low concentrations. The ability to make predictions of *Cryptosporidium* concentrations in rivers would aid significantly in abstraction-based risk management of water resources, but current models are inappropriate for making predictions at the temporal resolutions required to inform abstraction decision-making. This study utilises *Cryptosporidium* data collected over 7 years at a major river abstraction site in South East England, alongside publicly-available remote sensing data, to train a Bagging-XGBoost model for predictive applications at daily timescales. Different combinations of catchment-averaged and spatially distributed datasets were trialled as model inputs. The highest-performing models predicted 69-75% of >1 oocysts L⁻¹ exceedances, and they also predicted the timing of 78-89% of higher (>2 oocysts L⁻¹) exceedances. Interpretation of predictions using SHapley Additive exPlanations analysis indicated that sources near (<30 km) to the intake were the most important and identified catchment-averaged rainfall at 1 and 2-day lag time and antecedent *Cryptosporidium* measurements as significant inputs. The study demonstrates the potential of such models when an unparsimonious approach to feature selection is taken, because of their ability to discern non-linear trends and their resistance to multicollinearity and redundancy in the input data. This advances current understanding by showing how spatially distributed catchment information and temporally dynamic predictors can be combined to identify the locations and timescales most relevant to *Cryptosporidium* mobilisation and transport. Such

models could improve the ability of water utilities to predict pathogen peaks and aid abstraction decision-making, thereby reducing the loadings of this pathogen to reservoirs and water treatment works.

2.2 Introduction

Cryptosporidium is a pathogen of major concern to drinking water providers because of its resistance to chlorination and its infectivity at very low concentrations. The organism was identified as the leading cause of 63 % of waterborne outbreaks reported worldwide between 2011 and 2016 (Efstratiou et al., 2017a). *Cryptosporidium* persists in the environment in its transmissible “oocyst” form, with each oocyst measuring approximately 4-6µm in diameter and protected by a robust outer shell (Rose et al., 2002). When using standard methods, it is estimated that *Cryptosporidium* will be detectable in 45.3% of surface water samples globally (Daraei et al., 2020). Infection in humans and animals can lead to cryptosporidiosis, a self-replicating gastrointestinal illness which can be particularly dangerous for immunocompromised patients (Bouzid et al., 2013). Current risk management of *Cryptosporidium* in water systems is often reactive and/or precautionary, largely because of inadequacies in the understanding of key sources, pathways and processes within the water cycle, and the lack of a robust means of quantifying risk. Elevated concentrations of the *Cryptosporidium* organism in raw water can pose significant challenges to water treatment works (WTWs) in terms of treatment costs and potential risks to public health if removal and/or deactivation methods prove inadequate (Betancourt and Rose, 2004). In the UK, detections of *Cryptosporidium* in drinking water must be reported to the regulator (the Drinking Water Inspectorate (DWI)) and can lead to enforcement measures, including regulatory fines for the associated water company (DWI, 2022).

The pathogen is typically of greatest concern to those WTWs which rely on surface waters for their source water, although groundwater supplies can also be contaminated (Bouchier, 1998). The major diffuse sources of *Cryptosporidium* in rivers are believed to be faecal waste from grazing livestock (Sturdee et al., 2007) and animal/human waste-based fertilisers (including manures, slurries and digestates). Pathogens from these waste-associated sources can be mobilised by runoff generated during rainfall events, usually entering rivers as overland flow (Bhattarai et al., 2011; Swaffer et al., 2014). Previous studies have

identified infection prevalence and oocyst loadings as higher in cattle than in other major livestock groups, although poultry, pigs, goats, sheep and other farmed animals also constitute important reservoirs of environmental *Cryptosporidium* (Golomazou et al., 2024). Diffuse runoff containing waste from wild animals can also contribute to *Cryptosporidium* river loadings in many catchments (Sturdee et al., 1999). Point sources of *Cryptosporidium* are regarded as largely human (Medema and Schijven, 2001), entering the river either as treated discharge from sewage treatment works (STWs) (Bukhari et al., 1997), or untreated sewage spilled from combined sewer overflows (CSOs) (Gibson III et al., 1998). It is generally accepted that rainfall is a key driver of *Cryptosporidium* contamination of surface water bodies (Bhattarai et al., 2011), but correlations between rainfall and *Cryptosporidium* concentrations in rivers are not found at all sites (Atherholt et al., 1998), and where they are present, they are often weak and limited to certain times of the year (Coffey et al., 2010). Because laboratory analysis is costly and time consuming, there is an overall lack of published data on *Cryptosporidium* in surface waters, especially at resolutions which would facilitate analysis of sources and pathways in the environment.

Changes in the contaminant/water quality status of the raw surface waters can be managed by selective surface water abstraction and/or the adjustment of water treatment options. For example, daily abstraction volumes from major intakes on the River Thames are currently adjusted in response to the routine monitoring of *Cryptosporidium* concentrations in the raw water, with decisions concerning abstraction suspension, reduction or resumption triggered at threshold concentrations of 1, 2, 4, and 5 oocysts L⁻¹ (depending on the storage times of the reservoirs to which the river water is being abstracted) (Thames Water, 2021). Unfortunately, long *Cryptosporidium* analysis times (typically 2 days) mean that there is a lag between contamination events and operational decision-making, so that in many cases water above key threshold levels is still pumped into the reservoirs. Equally, resumption of abstraction after the *Cryptosporidium*-contaminated water has passed downstream of the abstraction point is also delayed – a particular concern for regions such as South East England, which are under water stress. Furthermore, the effectiveness of such decision-making is heavily reliant on frequent (ideally daily) monitoring, which is not undertaken at all abstraction points because of the high cost and time demands associated with *Cryptosporidium* sampling and analysis. The ability to predict/model *Cryptosporidium*

concentrations in rivers over a short (e.g. daily) timeframe would therefore significantly improve risk management and decision-making strategies (Dobson and Mijic, 2020), as well as assisting in water resource management.

2.3 Background

Catchment hydrological-water quality models have traditionally been classified as either spatially distributed, semi-distributed or catchment-averaged (also known as “aggregated” or “lumped”) models, with the former typically requiring greater complexity, increased computing time and higher data demands than the latter (Tran et al., 2018). Studies comparing distributed and catchment-averaged models have demonstrated that increased spatial-distribution of data and model design does not automatically equate to increased predictive power (Sinha et al., 2022). Some distributed deterministic models have been developed for the specific purpose of predicting raw water contaminant concentrations at sub-daily resolution, with a view to informing abstraction timing (Asfaw et al., 2018; Suslovaite et al., 2024). To date, these approaches have only been applied to a small number of contaminants (including metaldehyde and *E. coli*, but not *Cryptosporidium*) and limited to relatively small ($\approx 300 \text{ km}^2$) test catchments.

Previous attempts to explicitly model *Cryptosporidium* river concentrations are summarised in Fig. 2.1. They are generally limited to lower temporal resolution (monthly and above) models, or snapshot-based models intended to capture the current loadings to, or concentrations within, rivers. Four models did generate daily predictions (Brion et al., 2001; Medema and Schijven, 2001; Dorner et al., 2006; Tang et al., 2011), but these were validated against extremely limited datasets of ≤ 68 *Cryptosporidium* measurements (taken at intervals of many days/weeks). The models are predominantly deterministic and vary in spatial scale and complexity. The partially-distributed Soil Water Assessment Tool (SWAT) has been used for three catchment-scale *Cryptosporidium* models, with two focusing on very small ($<30 \text{ km}^2$) homogenous catchments (Coffey et al., 2010; Tang et al., 2011) and a third (Liu et al., 2019) looking at a larger ($\approx 4000 \text{ km}^2$) catchment and addressing possible mitigation scenarios. The SWAT studies combine a rainfall-runoff element (accounting for mobilisation and transport of *Cryptosporidium*, and dilution effects) with components which quantify oocyst loadings from animal and/or human waste-based sources. The latter can

Authors	Spatial Distribution [†]	Model Scale (km ²)	Prediction Time Steps	No. of <i>Cryptosporidium</i> river measurements used for training (t) & validation (v)											Animal Data	Human Data	Hydrological Data	Land Data	Meteorological Data	Oocyst Data	Source Data	Water Quality Data	Microbiological Data	Time-Based Data
					Livestock	Livestock	Livestock	Animal	Human	Hydrological	Land	Meteorological	Oocyst	Source	Water Quality	Microbiological	Time-Based							
Walker Jr and Stedinger (1999)	By subcatchment	4,142	N/A	Not stated	X																			
Brion et al. (2001)*	Catchment-Averaged	17,527	Daily	51 (t) 17 (v)																				
Medema and Schijven (2001)	By subcatchment	ca. 40,000	Daily	68 (v)																				
Ferguson et al. (2005)	By subcatchment	ca. 10,000	N/A	None	X	X	X																	
Donner et al. (2006)	1 x 1 km	130	Daily	30 (v)		X																		
Coffey et al. (2010)**	By subcatchment	29	Monthly	None	X	X																		
Tang et al. (2011)**	By subcatchment	4.0 - 4.2	Daily	40 (v)	X	X																		
Hofstra and Vermeulen (2016)	ca. 50 x 50 km	Global	N/A	None																				
Liu et al. (2019)**	30 x 30 m to 1 x 1 km	4,166	Monthly	None	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Vermeulen et al. (2019)	ca. 50 x 50 km	Global	Monthly	> 4000 (v)	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Ligda et al. (2020)*	By subcatchment	ca. 5,000	Monthly	136 (t+v)																				
Ligda et al. (2024)*	By subcatchment	ca. 5,000	Monthly	136 (t+v)																				
New Catchment-Averaged Model (1b)***	Catchment-averaged	9,300	Daily	≥ 1,165 (t) 683 (v)																				
New Spatially Distributed Model (5b)***	5 x 5 km to catchment-averaged	9,300	Daily	≥ 1,165 (t) 683 (v)																				

Fig. 2.1 Summary of key characteristics and input parameters used in previous *Cryptosporidium* models. (Notes: [†]Spatial distribution often varies with input type, hence sometimes two figures are provided; *ML models (all other models are deterministic); **SWAT models; ***New models (presented in this paper). For deterministic models, the number of *Cryptosporidium* river measurements used for validation (v) is provided only, whereas for ML models, the number of both training (t) and validation (v) data points is provided.)

include parameters which are indicative of initial oocyst quantities at the source (e.g. infection rates, shedding estimates, total faecal matter from grazing animals and wastewater discharge volumes), alongside factors which affect either oocyst survival (e.g. temperature and sunlight hours), or which might impede oocyst transport (e.g. adsorption to soil particles and settling processes). Other models working at either the catchment scale (Walker Jr and Stedinger, 1999) or the global scale (Hofstra and Vermeulen, 2016; Vermeulen et al., 2019) predict loadings to water courses averaged over months, years or the very long-term (i.e. not in real-time).

In recent years, machine learning (ML) modelling tools have been increasingly adopted for water quality applications due to their ability to consider complex non-linear relationships and their applicability to large, complex catchments in which deterministic approaches would be computationally expensive and challenging to calibrate. Brion et al. (2001) produced the first data-driven ML model for predicting *Cryptosporidium* river concentrations in the form of a neural network model of a large (17,527 km²) mixed urban-rural catchment in the USA, with inputs of river discharge, rainfall and water quality (including microbiological) parameters. Similarly, Ligda et al. (2020) applied a Linear Discriminant Function Analysis (LDFA) model, with inputs of mean monthly air temperature, total weekly rainfall and faecal indicator bacteria (FIB) data to produce mean monthly predictions of broad categories of *Cryptosporidium* concentrations. Both models were trained and validated on relatively small datasets (containing a total of 68 and 136 data points, respectively). The dataset used by Ligda et al. (2020) was later re-applied in a new study (Ligda et al., 2024), which trialled and compared the performance of different ML models in combination with additional physicochemical water quality inputs for the prediction of *Cryptosporidium* and *Giardia* river concentrations. The models trialled were Support Vector Machines (SVM), K-Nearest Neighbours (KNN), Random Forest (RF) and eXtreme Gradient Boosting (XGBoost), with RF and XGBoost found to be the best-performing models for *Cryptosporidium* and *Giardia*, respectively. All three studies showed the potential advantages of combining meteorological and/or hydrological data with river water quality data to generate predictions, but it is also worth noting that the use of FIB data (which usually requires analysis times of ≥ 24 hours) would be a significant limitation on

the applicability of such models for making real-time daily predictions for water resource management.

The choice of appropriate inputs (or “feature selection”) is a key element of model design and it is common practice to perform model trials using different combinations of inputs, then to compare the performance metrics of the validated models in order to identify the most important features (Dhal and Azad, 2022). For time series-based ML models which contain environmental data, such trials can also be used to decide on the spatial and temporal resolution of the input data and the use of data at different time lags to capture delayed effects – e.g. due to contaminant travel times (Choubin et al., 2018). Many ML models incorporate methods for calculating feature importance, but in recent years model-agnostic methods have also been developed, such as SHapley Additive exPlanations (SHAP) – a game theory-based approach which quantifies the scale and direction of impact (i.e. additive or subtractive) for each feature, along with local interpretation – i.e. the assessment of feature importance for each prediction in isolation (Lundberg and Lee, 2017). SHAP was deployed to assess the *Cryptosporidium* ML models developed by Ligda et al. (2024), and such approaches have huge potential for enhancing the use of ML tools as exploratory methods in order to advance understanding of system processes (Piraei et al., 2023).

In summary, previous attempts to model *Cryptosporidium* highlight a number of ongoing challenges, including: (i) a lack of robust understanding concerning key catchment sources and transport mechanisms, and their relative significance and required level of detail when modelling at different scales and resolutions; and (ii) the need to access larger *Cryptosporidium* and environmental datasets at higher temporal resolution (e.g. with daily instead of monthly intervals), either for the purpose of training more effective ML models, or for providing the information required to develop a process-based understanding of *Cryptosporidium* fluxes in response to specific rainfall events and conditions. To date there are no published models which have been proven to provide an operational understanding of the risk posed by *Cryptosporidium* to water abstraction systems at daily or sub-daily resolution; addressing the above challenges would be a significant step towards developing such a model. Recent advances in the availability of catchment sensing data – much of which is publicly available – combined with a parallel increase in open source ML tools, do

however provide new opportunities for modelling and predicting catchment water quality (Chen et al., 2022). For example, in the UK an extensive network of sensors has been installed on the vast majority of CSOs, providing Event Duration Monitoring (EDM) of spills of untreated wastewater into rivers, lakes and coastal waters (Giakoumis and Voulvoulis, 2023; Suslovaite et al., 2024).

The present study will present the development and testing of a data-driven tool for predicting *Cryptosporidium* concentrations in a river within a large and complex catchment. Making use of a new dataset comprising 7 years of *Cryptosporidium* measurements (much of which is at daily resolution) and open source environmental data, together with freely-available ML software, the aim is to produce a predictive tool which can be used operationally to inform abstraction management decisions. The novelty of this study lies in evaluating whether catchment-averaged and spatially distributed environmental and source data can be integrated with antecedent *Cryptosporidium* measurements within an interpretable machine-learning framework to improve scientific understanding of the locations, conditions and timescales associated with elevated *Cryptosporidium* risk. Specific objectives of the study are to (1) carry out trials using an appropriate ML model to demonstrate for the first time that well-validated *Cryptosporidium* predictions can be made at daily timescales; (2) compare models using different inputs (including catchment-averaged and spatially distributed data) to determine how alternative spatial and temporal representations of catchment processes influence model performance and interpretation; and (3) use model performance and feature importance metrics to identify the key features for predicting elevated *Cryptosporidium* concentrations, including drawing conclusions about key source types and locations.

2.4 Methods

2.4.1 Study area and *Cryptosporidium* monitoring

The catchment area is defined by a downstream monitoring and abstraction point at Walton WTW, located to the west of London (Fig. 2.2). The catchment covers an area of approximately 9,300 km² and contains both extensive rural areas and major urban centres. The Walton raw water intake, along with three intakes within 30 km (river distance) upstream, and one intake approximately 2 km downstream, all abstract to a series of

bankside reservoirs. As a collective, these five intakes are the source for approximately 70% of London's drinking water, serving approximately 6.5 million people, while the river system as a whole provides drinking water for a number of large towns and cities, including Oxford, Swindon, Reading, Slough, Maidenhead and Guildford. 127 sewage treatment works (STWs) and 285 CSOs discharge into the catchment, as defined by the Walton intake. For the purpose of this study, the catchment has been subdivided into 443 5x5 km cells, corresponding to the cells for which rainfall radar data is available.

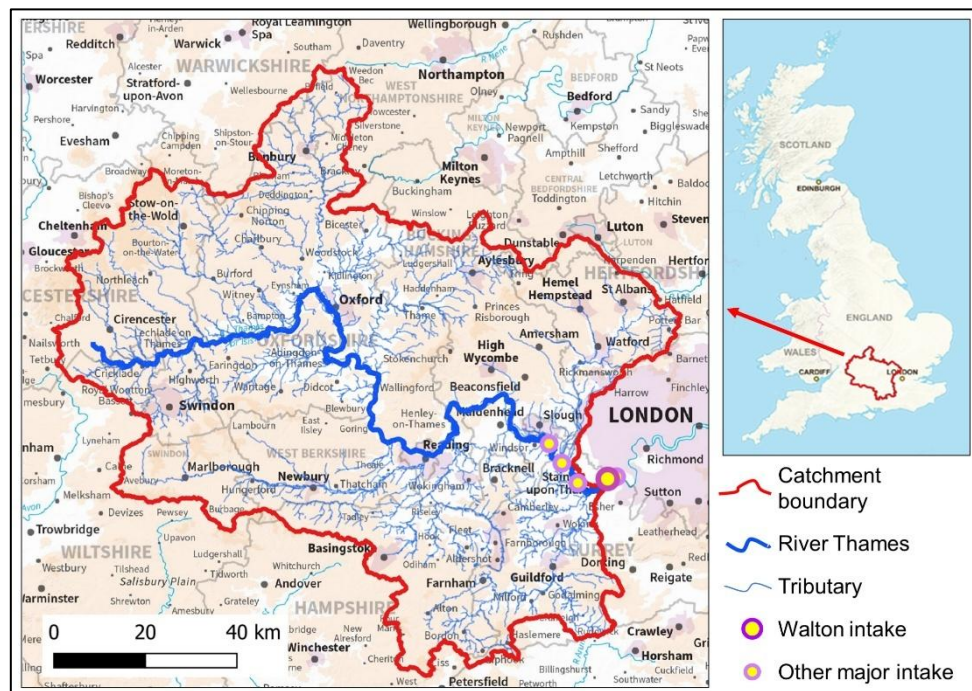


Fig. 2.2 Maps showing the catchment location on the mainland and the towns, cities and river network within the catchment boundary, along with the five major abstraction points on the River Thames which provide raw water for approximately 70% of London's drinking water. Base map: OS MiniScale (Ordnance Survey, 2025).

Cryptosporidium monitoring data for the River Thames (sampled at the Walton intake) was made available by Thames Water for the period between Jan 2000 and Mar 2023. From 2016 onwards, most of this data is at approximately daily temporal resolution, especially during the months where *Cryptosporidium* is at its highest. *Cryptosporidium* analysis was carried out by a UKAS-accredited laboratory using the standard Blue Book immunomagnetic separation-immunofluorescence microscopy (IMS-IFM) method (Environment Agency, 2010). Concentration data was not adjusted to compensate for analytical recovery because the available recovery data was not sample-specific, and therefore would not account for

variations in sample matrix. Furthermore, it was deemed appropriate to remain consistent with current Thames Water selective abstraction protocols, which apply concentration data without any post-analysis correction factor.

2.4.2 Model type

The model described below was designed to predict daily *Cryptosporidium* concentrations at the Walton intake on the River Thames. The predicted concentration outputs were then converted into positive or negative classification outputs, representing exceedances of different threshold concentrations (1, 2, 3, 4 and 5 oocysts L⁻¹). These thresholds were selected in accordance with the abstraction management trigger levels currently applied operationally by Thames Water.

A Bagging-XGBoost model was used, in combination with an unparsimonious approach to feature selection. XGBoost is a supervised learning algorithm, which resembles earlier ensemble tree-based models (such as Random Forest) by generating multiple decision trees to reduce overfitting and improve model performance (Chen and Guestrin, 2016; Bisong, 2019). XGBoost also incorporates a gradient descent function to correct the errors of previous trees constructed in the algorithm (Friedman, 2001). Although not specifically designed for working with time series data, it has been used extensively for time series modelling in water resources engineering (Niazkar et al., 2024), and specifically applied to water quality modelling studies aimed at estimating water quality indices (Masood et al., 2023) and predicting persistent organic pollutant concentrations in the Great Lakes of the USA and Canada (Wu et al., 2021a). Numerous studies have demonstrated the superior predictive performance of XGBoost for a variety of water quality applications when compared with other ML algorithms (e.g. Random Forest and neural network-based models) (Wu et al., 2021b; Masood et al., 2023; Rawat et al., 2023). XGBoost is effective at modelling non-linear trends (Bisong, 2019) and it has the capacity to handle datasets with a large number of features and different data types (e.g. both numerical and categorical data). Furthermore, its performance is not generally impaired by the presence of redundant inputs or multicollinearity (i.e. the use of multiple correlating features) (Pham et al., 2021).

Prior to adopting an XGBoost-based approach, preliminary trials were conducted using a range of alternative regression models, including Multiple Linear Regression, Random

Forest, Support Vector Machine and Multilayer Perceptron models, with default hyperparameter settings throughout and with the inputs matching those used in Models 1b and 5b (as described in Section 2.4.5). XGBoost-based models were found to give the best overall predictive performance (based on the metrics outlined in Section 2.4.7), which further justifies this choice of ML approach for the present study.

The term “Bagging” involves running models multiple times on different subsets of data and aggregating the model outputs. Bagging can be incorporated into an individual XGBoost model by setting the *subsample* hyperparameter (which defines the proportion of data randomly sampled for each tree) to a value of <1. When bagging across multiple models (as in the present study), a different random state setting is applied with each run of the model in order to ensure that the data is subsampled differently (and reproducibly) each time. Bagging across multiple iterations of a model generally reduces overfitting and is particularly effective where the input data is noisy or highly variable (Hastie et al., 2009). Aggregation of model outputs is typically performed using weighted averages as part of a Bayesian or frequentist statistical approach (Dormann et al., 2018), or by simple averaging (Deng et al., 2022), which was the method adopted in this study.

Although it is considered good practice to remove autocorrelating and redundant features prior to running ML models (Zhu et al., 2023), such feature engineering can be laborious and runs the risk of removing inputs which contain useful data – a particular concern given the current knowledge gaps regarding the most important inputs for predicting *Cryptosporidium*. For this reason, an unparsimonious approach to feature selection was adopted and the models were run with a high (>90) number of features, including highly autocorrelating inputs at different time lags.

Various forms of the model (trialling different combinations of inputs as described below) were executed using the XGBoost library (v. 1.5) installed in a Python environment (v. 3.8.8) and using the Anaconda distribution (available from www.anaconda.com).

2.4.3 Model inputs

The choice of input parameters was informed by existing literature detailing *Cryptosporidium* sources, mobilisation mechanisms and factors affecting the pathogen’s longevity in catchments (see Fig. 2.1). The parameters used can be divided into three

categories: (i) core inputs (common to all models presented here); (ii) environmental inputs (e.g. meteorological and hydrological data); and (iii) CSO inputs (specifically the EDM datasets). A full description and justification of parameters used is provided below. It should be noted that a number of water quality parameters – including turbidity, nitrate, ammonium, phosphate and electrical conductivity, measured in grab samples at the intake, or upstream sonde stations – were also considered as possible inputs, but were dismissed due to substantial gaps in the available data.

2.4.3.1 Core inputs

Three core inputs were deployed uniformly across all models. The first of these was *Crypto_Rolling_Mean*, a 3-week rolling mean *Cryptosporidium* concentration, based on historical measurements from T_{-7} to T_{-28} days (where T_0 is the current day on which *Cryptosporidium* concentrations are to be predicted by the model). It was necessary to include this feature as a scaling input to allow the model to distinguish between the severity of *Cryptosporidium* seasons. A temporal input – *Day_of_Year* – was included to capture cyclical annual trends that may not be covered by the other input data; for example, events in the farming calendar, or time windows during which heavy rainfall might be more likely to mobilise *Cryptosporidium*. This input was defined by a numerical value from 1-366. A third input – *High_Season* – was included to group together the commonly “high” *Cryptosporidium* months (Oct-Feb) into individual seasons, therefore allowing the model to learn to delineate temporally between more and less severe *Cryptosporidium* seasons (which were evident in the data). Seasons are here defined as from 1st Oct to 28th/29th Feb, and were delineated in the input by starting year number (e.g. 2016 for the Oct 2016-Feb 2017 season). Days in the period from Mar-Sept were designated a zero value in all years.

2.4.3.2 Environmental inputs

Environmental inputs include: rainfall, number of consecutive antecedent dry days, soil moisture, soil temperature, river discharge and river discharge rate-of-change. These are outlined in greater detail below.

Previous models have applied rainfall as a catchment-averaged parameter (Brion et al., 2001), or as a spatially distributed (location-specific) value (Dorner et al., 2006). In the latter case, routing is required to impose appropriate lag times which are reflective of the time-of-

transport from the source to the abstraction/monitoring point. In the present study, rainfall was used in different forms in both catchment-averaged and spatially distributed versions of the models.

Rainfall radar data (NIMROD) for 5x5km cells and 15-minute resolution was obtained from the Centre for Environmental Data Analysis (CEDA) (Met Office, 2003) and converted to mean daily rainfall values. Gaps in the radar data were filled using the HadUKGrid 5x5km daily rainfall measurements (Met Office et al., 2021). The rainfall data in this form was used in the spatially distributed versions of the model by imposing location-specific time lags (see Section 2.4.4.2).

Catchment-averaged daily rainfall was obtained by calculating the daily mean for all cells within the catchment. The number of consecutive dry days is important for capturing first-flush effects, when relatively large amounts of faecal waste may be washed from fields or from CSO pipes following an extended dry period (Mamun et al., 2020). This parameter was calculated from the catchment-averaged rainfall, with a dry day defined as one with <1 mm of rainfall.

Soil moisture was included because of its positive impact on runoff generation during rainfall events (Singh et al., 2021) and its effect on *Cryptosporidium* survival, with desiccation regarded as one of the most important mechanisms of oocyst deactivation (Robertson et al., 1992). Daily soil moisture values compiled by the COSMOS-UK network were obtained from the UK Centre for Ecology & Hydrology (UKCEH) (Smith et al., 2024). Data at the four measurement sites in the catchment was retrieved (Chimney Meadows, Chobham Common, Sheepdrove and Waddesdon) and averaged to give a mean soil moisture value as % volumetric content at 5 cm depth.

Temperature can be a significant factor controlling oocyst survival, with very low (<-10 °C) or very high (>37 °C) temperatures known to dramatically increase rates of oocyst deactivation (King et al., 2005). Because a major source of *Cryptosporidium* in the environment is believed to be animal waste on agricultural land, soil temperature (as opposed to air temperature) was taken as an input, with values measured by gauging stations at the four COSMOS-UK sites and available from the same website. Hourly measurements were averaged to obtain a mean daily soil temperature in °C.

The locations of the four monitoring sites for soil moisture and soil temperature are spread across the catchment (see Appendix A for locations and variability over time). Whilst the correlation between sites is very high (>0.84 for soil moisture; >0.97 for soil temperature) and overall variability is low, it is acknowledged that the average of these values is unlikely to fully represent the full spatial variability of catchment conditions. Future studies could consider the value of populating these inputs with spatially distributed satellite-based data; however, for the present study priority was given to reliability of the data and its availability in real-time.

River discharge has been a key input in previous *Cryptosporidium* river models (e.g. Liu et al., 2019). Increases in flow will typically follow rainfall-runoff events and may coincide with increased spills from CSOs. Abrupt changes in discharge can also liberate entrained oocysts from bed and bankside sediment (Crockett, 2004). Conversely, high flow rates can have a diluting effect, leading to lower *Cryptosporidium* concentrations (Knapp et al., 2022). Daily discharge data (in $\text{m}^3 \text{ day}^{-1}$) for the nearest downstream gauging station (Kingston-upon-Thames) was accessed via the Hydrology Data Explorer, managed by Defra (<https://environment.data.gov.uk/hydrology>).

2.4.3.3 CSO inputs

The major human sources of *Cryptosporidium* released to rivers in the UK are believed to be discharges of treated sewage effluent (from STWs) and spills of untreated sewage (from CSOs) (Medema and Schijven, 2001). Despite the large number (120) of STWs in the study catchment and the sizeable contribution of treated wastewater to the Thames river network, the daily regularity of effluent discharge volumes means that these would not generally be expected to generate abrupt increases in *Cryptosporidium* in the river. By contrast, CSO spills of untreated sewage are highly intermittent and therefore present a more obvious cause of sudden spikes, particularly given that they are also likely to contain higher concentrations of oocysts than treated discharges (Nasser, 2016). EDM data recording CSO spills in the catchment was made available by Thames Water for the period from January 2018 to August 2022. This data also includes discharges from storm tanks, which are used to store excess influent during high flow conditions, usually in response to rainfall events. For simplicity, henceforth in Chapter 2 the term "CSOs" will refer to both CSOs and storm tanks, except where the text explicitly distinguishes between the two

sources. Note that storm tank spills are likely to coincide with periods of high effluent discharge from the main STW outfalls, but this latter source is not represented in the model input data.

EDM data from post-August 2022 was obtained via the publicly-accessible Thames Water API Portal (<https://data.thameswater.co.uk>). EDM sensors record the time and date at which spills from CSOs start and stop (at 1-minute resolution). Of the 285 CSOs which fall within the study catchment, EDM data for 201 CSOs has been acquired, although monitoring at each site commenced at different times. At the start of 2019, 44 of the CSOs were being monitored; by the start of 2020, this figure had risen to 161, rising further to 189 monitored CSOs by September 2020. All EDM data was converted into daily binary form – i.e. the timestamped spill start and spill stop records were converted into continuous daily binary time series, with 0 denoting no spills in a day, and 1 denoting a day in which one or more spills occurred. A catchment-averaged version of the CSO spill data, *Catchment-Wide CSO Spills*, was also calculated by summing all of the spilling CSOs for each day.

2.4.4 Travel times and time lags

2.4.4.1 Multi-lagged inputs: time lag trials for catchment-averaged inputs and discharge inputs

The selection of appropriate lag times for the catchment-averaged inputs (rain, consecutive dry days, soil moisture and soil temperature), along with river discharge, discharge rate-of-change and catchment-wide CSO spills, was essential because events resulting in the mobilisation and transport of *Cryptosporidium* can affect in-river concentrations of the pathogen many days later (Atherholt, 1998). The simplest (catchment-averaged) version of the model (see Table 2.1) was trialled using a minimum lag time of 1 day and maximum lag times (*MaxLag*) ranging from 2 to 30 days, to determine the most suitable maximum time lag (in days), based on model performance. For example, in the first test the lag times were 1 and 2 days; in the second test they were 1, 2 and 3 days; in the third test they were 1, 2, 3 and 4 days (and so on). The minimum lag time was fixed at 1 throughout because recent conditions were presumed to exert a strong influence over *Cryptosporidium* concentrations. In each trial, the lag times were applied uniformly to all lagged inputs. 30 days was deemed an appropriate upper limit, since this exceeded the estimated river travel times for all of the

5x5 km cells of the catchment under a Q50 flow regime (see Fig. 2.3). Lag times for features are denoted by an “L” followed by the number of days (e.g. L02 refers to a 2-day lag time).

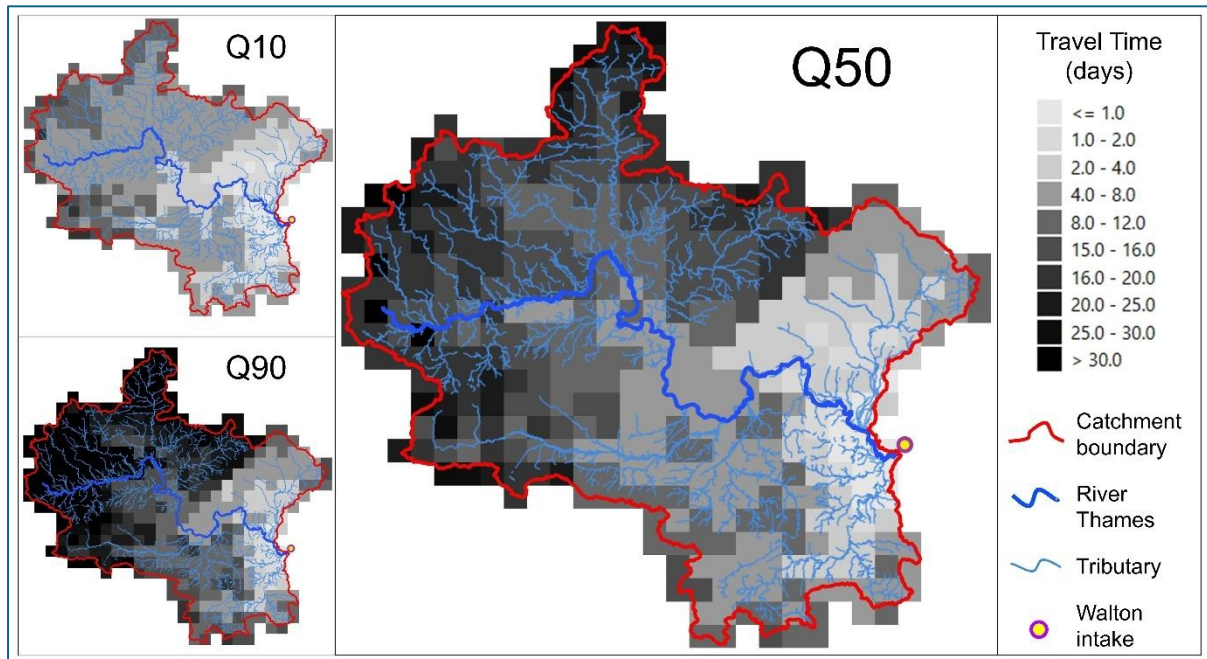


Fig. 2.3 Estimated travel times for 5x5 km cells under three different flow regimes in the study catchment

2.4.4.2 Single-lagged inputs: calculating flow regime-dependent travel times for spatially distributed data

Single-lagged inputs were restricted to the spatially distributed data – i.e. the 5x5 km cell rainfall and CSO spill data – which were lagged and averaged according to estimated travel times under different flow regimes. Note that these two types of spatially distributed data were chosen because they represented the primary mechanisms/sources which could lead to the sudden increases in *Cryptosporidium* in the river (Medema and Schijven, 2001; Bhattarai et al., 2011; Swaffer et al., 2014). Because of the size and the complexity of upstream river system, it was necessary to make broad estimates of in-river travel time ranges from each cell within the catchment to the Walton intake. An empirical approach described in Jobson (1997) was adopted, in which catchment area and discharge values are used to estimate the transport velocity of a contaminant peak through a river network.

The present study applied the Jobson (1997) method, using 110 catchment travel time data points obtained through a combination of tracer tests carried out by Environmental Tracing Systems Ltd (1998) and a spill time-of-transport model developed by Wallingford

HydroSolutions (2016). Both sources provided data for limited areas of the catchment, including sections of the Thames and a number of its tributaries. Curve-fitting was applied separately to the tributaries, to the upper Thames (with catchment areas of $<1600 \text{ km}^2$), and to the lower Thames (with catchment areas of $>1600 \text{ km}^2$). QGIS (v. 3.32.3) combined with OS Open Rivers Data and OS 1:50,000 terrain data (Ordnance Survey, 2025) was used to obtain the catchment area and river distances for the most downstream point in each 5x5 km catchment cell, and mean velocity (and subsequently travel time) was calculated under different flow regimes (Q01 to Q99) for all cells within the catchment. The quantile of each day's discharge was calculated and rounded to the nearest Q10 to give a quantile interval that could then be used to look up appropriate travel times. When broken down by flow regime, error at 68% confidence interval for example cells in the catchment was shown to be <1 day when distances between cell and abstraction point were <150 km at high (Q10) flow, <60 km at medium (Q50) flow, and <20 km at low (Q90) flow. The error at low flow is not regarded as a critical limitation to the model because 70% of >1 oocyst L^{-1} exceedances occur at flows of Q50 and above, and because inputs for more distant cells were averaged over multiple days. Estimated travel times in the catchment for the three flow regimes are shown in Fig. 2.3. Further information concerning curve-fitted equations and error analysis is provided in Appendix A.

2.4.5 Input selection for models

Table 2.1 shows the inputs included in each tested model. All models contained the three core features, along with the following time-lagged inputs: discharge, discharge rate-of-change, catchment-averaged soil moisture, catchment-averaged soil temperature and catchment-averaged consecutive antecedent dry days. Models labelled “a” omitted catchment-averaged rainfall, whereas those labelled “b” included this input. Model 2 also contained catchment-wide CSO Spills, whereas Models 3 and 4 contained spatially distributed cell rainfall data and CSO data, respectively. Model 5 contained *both* spatially distributed cell rainfall data and CSO data.

2.4.6 Hyperparameter tuning

Hyperparameter tuning was initially carried out using Scikit-Learn's GridsearchCV (Pedregosa et al., 2011) but the tuned parameters did not differ markedly from the default

settings and produced no improvement in model predictions. Automatic hyperparameter tuning was therefore omitted entirely from future models and – with two exceptions – the default hyperparameter settings were adopted throughout. The exceptions were

Table 2.1 Input parameters included in each of the tested models, as indicated by greyed-out cells. (Sp. Dist. = Spatially Distributed)

	Input parameters	Input Name	Units	Models									
				1 a	1 b	2 a	2 b	3 a	3 b	4 a	4 b	5 a	5 b
Core	Day of year	<i>Day_of_Year</i>	Day										
	High Season Oct-Feb	<i>High_Season</i>	-										
	<i>Cryptosporidium</i> 3-week mean	<i>Crypto_Rolling_Mean</i>	oocysts L ⁻¹										
Catchment-Averaged	Catchment-averaged rainfall	<i>Catchment_Rain</i>	mm										
	Continuous antecedent dry days	<i>Dry_Days</i>	Days										
	Discharge	<i>Discharge</i>	m ³ s ⁻¹										
	Discharge daily-rate-of-change	<i>Discharge_ROC</i>	m ³ s ⁻¹										
	Soil moisture	<i>Soil_Moisture</i>	% volume										
	Soil temperature	<i>Soil_Temp</i>	°C										
	Catchment-wide CSO spills	<i>CSO_Sum</i>	Spills										
Sp. Dist.	Cell-based rainfall	<i>Rainfall_Cell</i>	mm										
	CSO spills (individual CSOs)	<i>CSO</i>	Spills										

n_estimators (number of decision trees in the model) and *subsample* (percentage of the training data randomly sampled and used in each decision tree). Preliminary model trials were performed using values between 5 and 500 for the *n_estimators* hyperparameter and values of 0.1 to 1.0 for the *subsample* hyperparameter. The minimum number of “bagging” repeats (*No of iterations*) was established by training and validating models through different iteration totals, ranging from 5 to 200, in order to establish at which point no further improvement in performance was achieved. Model performance was significantly improved by shuffling the input data columns before each iteration, an approach which reduces biases in the XGBoost random sampling method (Alsahaf et al., 2022). Random seeds were used for both the subsampling and column shuffling so that the results could be reproduced. All of the manual model trials described above were carried out using Model 1b (which had the smallest number of features, with the exception of 1a) and Model 5b

(which had the highest number of features) to see if hyperparameter tuning was feature-dependent.

2.4.7 Model validation and performance metrics

Walk-forward validation was applied so as to replicate the way in which the developed model would be deployed operationally (Fig. 2.4). This is an appropriate alternative to K-fold validation where data has a time-determined order. In walk-forward validation, the model is trained using the latest available data in a sliding or extending window, to predict the next time step (Suradhaniwar, 2021). Because of the minimum 2-day analysis time for *Cryptosporidium* and the need to incorporate mean *Cryptosporidium* data from days T_{-7} to T_{-28} , the latest data included in the training model was at T_{-4} . The trained model was then used to predict the current day's *Cryptosporidium* concentrations (i.e. on day T_0). The training data commenced on 01.01.2016 and initially ended on 28.02.2021 (containing a minimum of 1165 *Cryptosporidium* measurements).

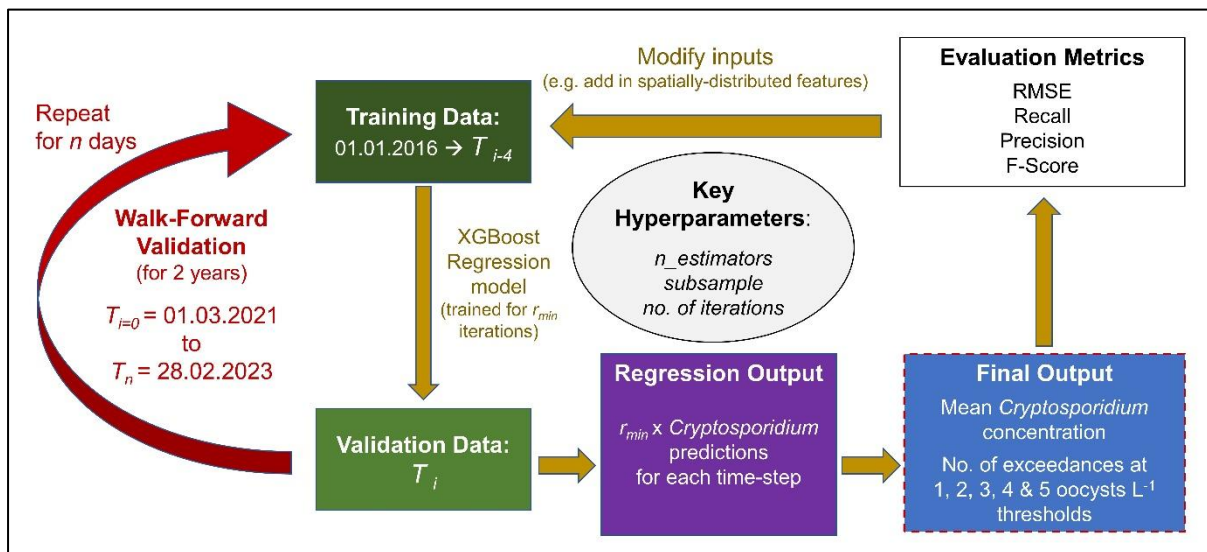


Fig. 2.4 Diagram of model processes and model outputs, showing dates (T) over which training and validation data is split for walk forward validation, alongside key hyperparameters, outputs and evaluation metrics.

The initial validation start date was 01.03.2021, with a final validation date of 28.02.2023, generating two full years of validation data (containing a total of 683 measurements). Days (i.e. rows) in the training data for which *Cryptosporidium* measurements were unavailable were removed prior to training and validating the model. The validation outputs from all

iterations of a single model were averaged to give a final predicted concentration, in accordance with the bagging method described in Section 2.4.2.

Model outputs were assessed using four performance metrics, beginning with (i) the Root Mean Square Error (RMSE), which was applied to the concentration output and represents the standard deviation of residuals (i.e. the error in the concentration predictions). The remaining three metrics evaluated the model's performance when concentrations were converted into classification outputs – i.e. positive or negative exceedances of the 1, 2, 3, 4 and 5 oocysts L⁻¹ thresholds. These metrics were: (ii) Recall (R), which is the proportion of predicted exceedances – i.e. True Positives (TPs) – relative to the total number of measured exceedances; (iii) Precision (P), representing the proportion of predicted exceedances which were correct; and (iv) F-score (F), which provides a combined evaluation of R and P (calculated from the harmonic mean of the two metrics) (Alpaydin, 2014). An increase in the number of TPs results in higher R, P and F scores, but likewise, an increase in the number of False Positives (FPs) (i.e. incorrectly predicted exceedances), will produce lower P and F scores. Because reliable prediction of exceedances is central to the use of the model in abstraction decision-making, the exceedance-based metrics (R, P and F) were regarded as the most important for assessing model performance.

2.4.8 Local interpretation using SHAP values

SHAP analysis was carried out on the models to quantify the relative contribution of individual features to model predictions above and below threshold levels. The SHAP approach provides local interpretability – i.e. the quantification of feature importance at the level of the individual prediction – revealing both the magnitude of a feature's impact the direction of that impact (i.e. whether it increases or decreases the predicted value) (Lundberg and Lee, 2017). Analysis was performed on the validation data and averaged across all iterations within each model. The results were then applied in three different ways: (i) as time series plots of the cumulative totals of the core and lagged discharge and catchment-averaged values; (ii) as bar plots of maximum impact scores for the most important features and for features grouped by type; and (iii) to produce maps of the maximum impact of the spatially distributed inputs (rainfall cells and CSOs) for comparison with livestock and human population distributions. Livestock density data (for cattle, sheep,

pigs and poultry) was obtained from the livestock surveys of Great Britain (APHA, 2022), and the residential human population data was taken from the 2011 UK census (Reis et al., 2017).

2.5 Results

2.5.1 Lag time and hyperparameter tuning

Final results from the maximum lag time and hyperparameter trials using Models 1b and 5b are shown in Table 2.2. Note that trials were carried out multiple times to zero-in on the settings which produced the highest performing models, with the results for each trialled element obtained using the “optimum” settings for the other three elements. Performance was reduced when values lower than those stated in Table 2.2 were used, whereas higher values either reduced performance (in the case of *MaxLag* and *subsample*), or increased computation time without improving performance (in the case of *n_estimators* and *No. of iterations*). (See Appendix A for more detailed presentation of the hyperparameter tuning outputs.)

Table 2.2 Optimised maximum lag time and hyperparameter settings following initial trials

Model setting / hyperparameter	Optimum value	Explanation / Significance
<i>MaxLag</i> (days)	18	All catchment-averaged / discharge-related inputs in models 1-5 are applied with lag times from 1-18 days
<i>n_estimators</i>	25	Each iteration of the XGBoost model generates 25 decision trees (i.e. estimators) from the training data
<i>Subsample</i>	0.5	The XGBoost model randomly subsamples 50% of the training data for each decision tree
<i>No. of iterations</i>	20	Each model reruns the XGBoost algorithm 20 times, producing mean <i>Cryptosporidium</i> concentration predictions and probabilities of threshold exceedance from the combined outputs of each iteration

2.5.2 Overall performance and comparison of models

Performance metric results for the different models are shown in Table 2.3. RMSE showed minimal variation (<0.04 oocysts L⁻¹) between the models, indicating that when predicting absolute concentrations, the models differed only marginally. The exceedance-based

metrics (R, P and F) do show more pronounced variation, however. All of the models performed well when predicting lower (>1 oocysts L^{-1}) threshold exceedances, but were less effective when predicting moderate (>2 oocysts L^{-1}) exceedances (with R_1 scores more than $2 \times R_2$ scores in every case). A maximum of two higher exceedances (>3 oocysts L^{-1}) were

Table 2.3 Performance metric results for the different models. Blank values arise where True or False Positives were not predicted by the model. Cell shading denotes **highest**, **intermediate** and **lowest** performing models for each metric

Metric Type	Threshold (oocysts L^{-1})	Metric	Model									
			Catchment-Averaged & Discharge Inputs				Catchment-Averaged, Discharge + Spatially Distributed Inputs					
			Without CSO Data		With CSO Data		With Cell Rain		With CSOs		With Cell Rain + CSOs	
			1a	1b	2a	2b	3a	3b	4a	4b	5a	5b
True Positives	1	TP ₁	40	44	44	43	45	45	47	46	47	48
	2	TP ₂	6	8	4	8	8	9	8	9	9	10
	3	TP ₃	1	1	1	1	1	0	0	1	2	0
False Positives	1	FP ₁	33	20	32	20	25	24	30	21	26	25
	2	FP ₂	2	5	2	5	11	13	4	10	14	12
	3	FP ₃	1	1	1	1	1	2	0	2	1	2
Recall	1	R ₁	0.625	0.680	0.688	0.672	0.703	0.703	0.734	0.719	0.734	0.750
	2	R ₂	0.261	0.348	0.174	0.348	0.348	0.391	0.348	0.391	0.391	0.435
	3	R ₃	0.083	0.083	0.083	0.083	0.083	0.000	0.000	0.083	0.167	0.000
Precision	1	P ₁	0.548	0.688	0.579	0.683	0.643	0.652	0.610	0.687	0.644	0.658
	2	P ₂	0.750	0.615	0.667	0.615	0.421	0.409	0.667	0.474	0.391	0.455
	3	P ₃	0.500	0.500	0.500	0.500	0.500	0.000		0.333	0.667	0.000
F-score	1	F ₁	0.584	0.688	0.629	0.677	0.672	0.677	0.667	0.702	0.686	0.701
	2	F ₂	0.387	0.444	0.276	0.444	0.381	0.400	0.457	0.429	0.391	0.444
	3	F ₃	0.143	0.143	0.143	0.143	0.143			0.133	0.267	
Root Mean Square Error		RMSE	0.590	0.564	0.587	0.559	0.558	0.564	0.562	0.564	0.560	0.552
Total Number of Features			93	111	111	129	522	540	279	297	708	726
Total Number of Exceedances	1	64 (9.4%)										
	2	23 (3.4%)										
	3	12 (1.8%)										
	4	6 (0.9%)										
	5	4 (0.6%)										
Total Measurements		683										

predicted by the models and there were no predictions higher than 4 oocysts L^{-1} , hence metrics for these levels have been omitted from the results.

On the whole, catchment-averaged models (Models 1 and 2) performed similarly to the spatially distributed models (Models 3-5), but there were subtle differences. Model 1b performed better than the other three catchment-averaged models, including those with catchment-wide CSO inputs (Models 2a and 2b). The inclusion of spatially distributed data (Models 3-5) led to a slight increase in Recall (because of a corresponding increase in TPs), but this came at the cost of Precision, with Models 3-5 suffering from higher numbers of FPs, compared with Model 1b. Spatially distributed models based on CSO data (4a and 4b) performed slightly better than those based on cell rainfall data (3a and 3b). Although the differences are modest, Model 5b – which incorporated all of the catchment-averaged parameters (except *CSO_Sum*) along with all available spatially distributed parameters – was the highest performing model, with the lowest RMSE, the highest R_1 and R_2 scores and the second highest F_1 and F_2 scores.

It is notable that including rainfall data in some form was shown to have a positive effect on the predictive performance of the models at the 1 oocysts L^{-1} threshold levels. Models which incorporated this input (i.e. those subscripted with “b”, or containing cell rainfall inputs) had F_1 scores of 0.677-0.702, whereas those models which omitted rainfall entirely had lower F_1 scores of 0.584-0.667.

On many occasions, the reduced model performance at higher exceedances was a question of prediction *magnitude*, not prediction *timing* – i.e. although the models did not predict a majority of the moderate-to-high exceedances, they frequently predicted a lower exceedance on the same day, with Models 1b and 5b forecasting a >1 oocyst L^{-1} exceedance on over 67-100% of moderate-to-high (>2 to >5 oocysts L^{-1}) exceedances. Taking all of the >1 oocysts L^{-1} exceedances together, the models correctly predicted 69-75% of exceedances, whereas for all >2 oocysts L^{-1} exceedances, the models predicted a >1 oocysts L^{-1} on 78-89% of occasions.

Fig. 2.5a shows predicted outputs from the highest performing catchment-averaged model (Model 1b) and the highest performing model incorporating spatially distributed data (Model 5b), alongside measured *Cryptosporidium* values. The differences between the two model outputs are subtle, with Model 5b correctly identifying seven more exceedances than Model 1b, although it did this at the cost of reduced precision (see Table 2.3). Measured *Cryptosporidium* peaks were higher and more frequent in Oct 2022-Feb 2023 than they

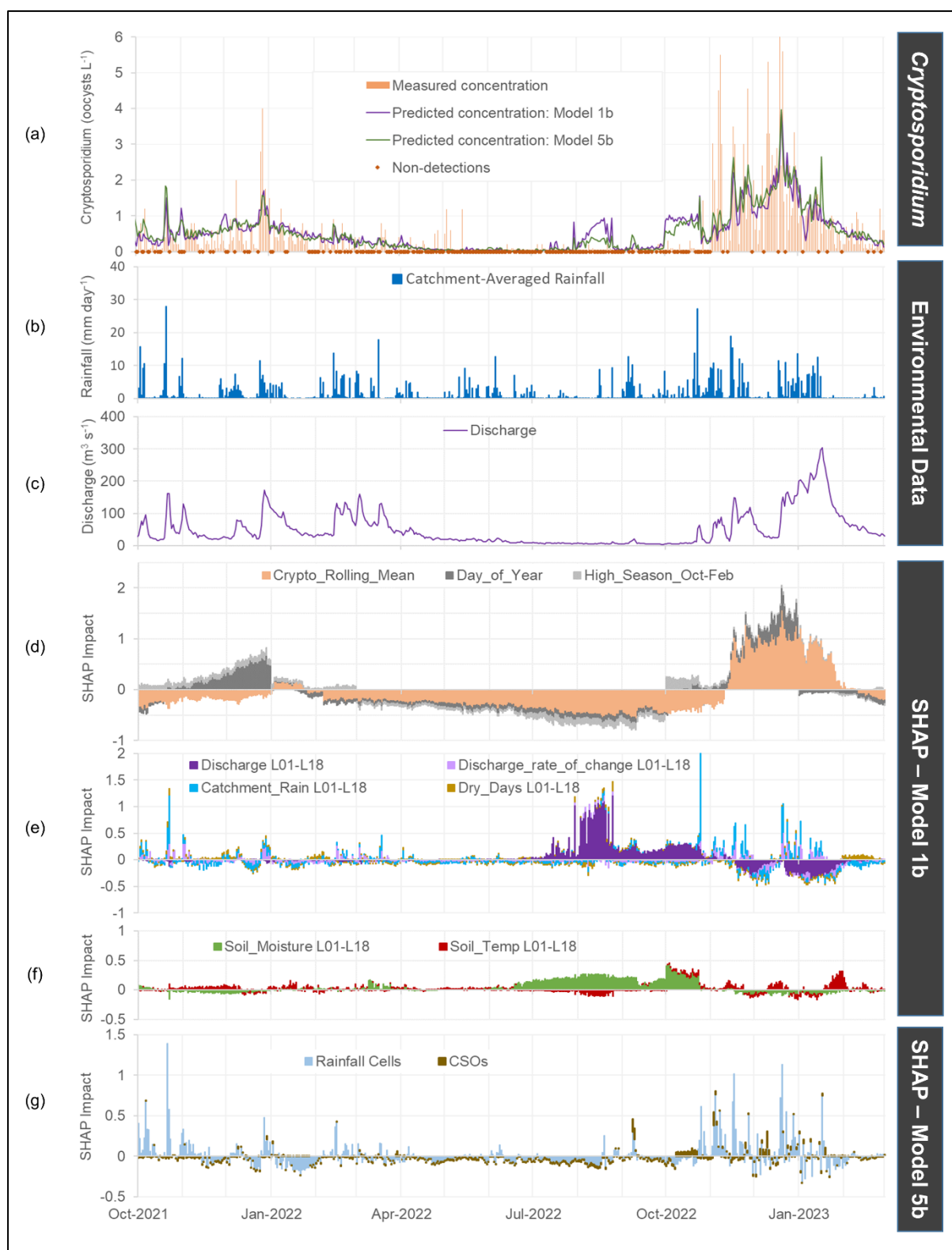


Fig. 2.5 (a) Measured and predicted concentrations of *Cryptosporidium* (generated by Models 1b and 5b), (b) catchment-averaged rainfall and (c) discharge. SHAP values showing the impact on Model 1b predictions of (d) core inputs and (e) hydrological inputs and (f) soil meteorological inputs (presented as cumulative values for all lag times), and (g) the impact on Model 5b of rainfall cells and CSOs as cumulative totals. (Note: for simplicity, the effect of SHAP base values has been omitted.)

were in the preceding year (Oct 2021-Feb 2022); the outputs of both models reflected this, with notably fewer predicted exceedances in the earlier period, indicating that the models are capable of distinguishing between more and less severe *Cryptosporidium* seasons. Fig. 2.5b-c show two key environmental variables (catchment-averaged rainfall and discharge), illustrating that the highest peaks in *Cryptosporidium* sometimes coincided with, or lagged slightly behind, rainfall events and increases in discharge.

For the most part the models underestimated the highest measured *Cryptosporidium* peaks, but for two months (Aug and Oct 2022) the predictions consistently exceeded observed values. The environmental variables do not offer any clear explanation for these incorrectly high predictions, although the occurrence of several small rainfall events at a time when discharge was low may provide a partial explanation.

2.5.3 SHAP analysis

Fig. 2.5d-f displays mean SHAP impact scores for the three core inputs and catchment-averaged inputs (expressed as cumulative values across all lag times) for Model 1b.

Crypto_Rolling_Mean was the most positively impactful feature (i.e. increasing the predicted concentration) for the more severe *Cryptosporidium* season (Oct 2022-Feb 2023), but it had very little impact on the less severe season (Oct 2021-Feb 2022), where instead *Day_of_Year* was more impactful. During the months in which observed *Cryptosporidium* is typically low (defined here as Mar-Sept), predictions are shown to be partially dominated by the negative impact (i.e. acting to reduce the predicted concentration) of *Crypto_Rolling_Mean*.

Of the non-core inputs, *Discharge*, *Catchment_Rain*, *Soil_Moisture* and *Soil_Temp* have the most influence on predictions, although these impacts vary with the time of year. *Discharge* has the most striking positive impact in Aug-Oct 2022 and a negative impact in Dec 2022 and Jan 2023. *Catchment_Rain*'s impact is predominantly positive, but also much more intermittent than the other catchment-averaged/discharge-based inputs, and often of higher magnitude. *Soil_Moisture*'s impact is also largely positive and (like *Discharge*) occurs most prominently in Aug-Sep 2022, whereas *Soil_Temp* switches frequently from imposing a small positive to small negative impact, with the highest-magnitude impact occurring in late Jan 2023.

The combined SHAP values for spatially-distributed inputs (Fig. 2.5g) show that the impact of *Rainfall Cells* is of much greater magnitude than the *CSOs*, although *CSOs* do occasionally have a higher impact (e.g. in early Sept, mid-Oct and early Dec 2022). As with *Catchment_Rain*, the spatially-distributed inputs have a very intermittent, high-magnitude (relative to other catchment-averaged inputs) impact on predictions. This impact is generally positive, although negative impacts of moderate magnitude do occur in Dec-Jan of both winter seasons.

Ranking the features by their highest maximum positive impact (Fig. 2.6) shows that in both Model 1b and 5b, catchment-averaged rainfall has the greatest effect when applied with a 2-day lag time (*Catchment_Rain_L02*), although a 1-day lag time is also important for Model 1b. *Rainfall Cells* within 30 km river distance of the intake provided the highest positive impact on Model 5b, although *Rainfall Cells* ≥ 30 km from the intake also had a high maximum collective impact. *Crypto_Rolling_Mean* and *Day_of_Year* had roughly equal impact on both models, but *Soil_Moisture* and *Discharge* were more impactful in Model 1b than 5b. *Discharge* had the most impact at a 6-day lag time, whereas *Discharge_ROC* has most effect on Model 1b at a 1-day lag time. The maximum impact of *CSOs* on Model 5b was relatively low.

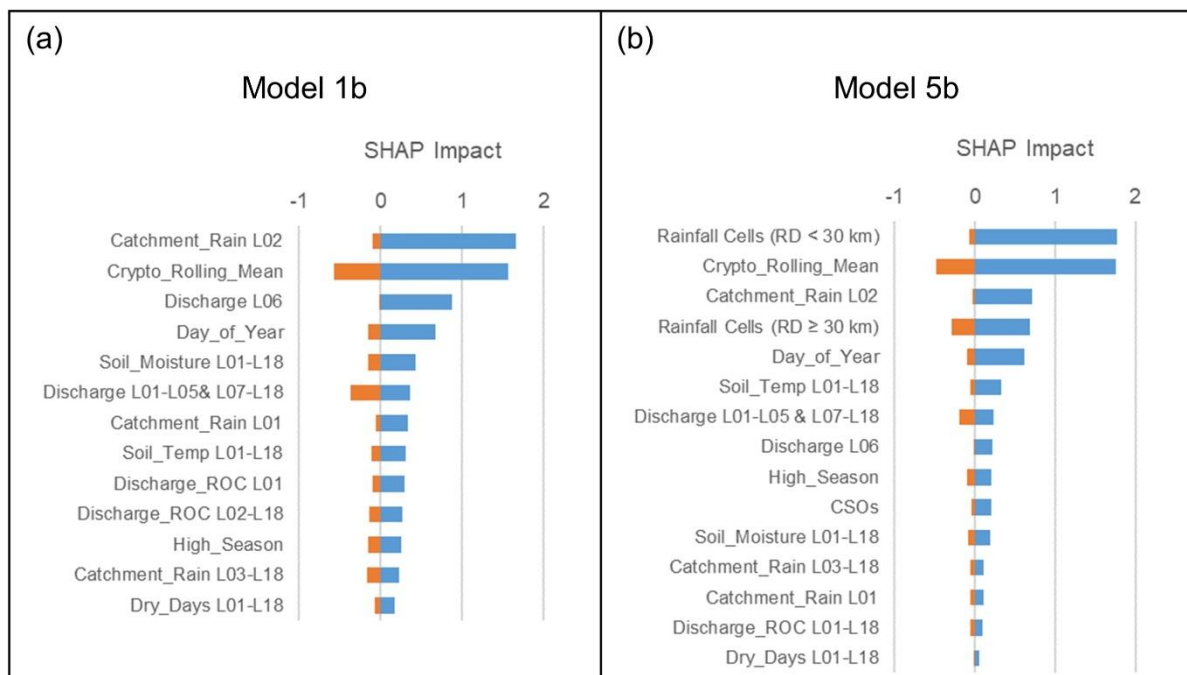


Fig. 2.6 Maximum positive and negative SHAP impact scores for key individual features and grouped features in (a) Model 1b and (b) Model 5b. (Note: catchment-averaged and spatially distributed features with an absolute maximum impact of <0.25 are grouped

together by feature type; RD = River Distance (km) from Walton intake; Ln=lag time, expressed as n days.)

2.5.4 Identifying possible source locations/areas

Fig. 2.7a-b shows the locations of spatially distributed features (rainfall cells and CSOs), along with their maximum positive SHAP impact scores (from Model 5b). Here, high-scoring features can be treated as an indicator of possible *Cryptosporidium* source areas/locations. The most impactful rainfall cells are 387, 444, 498 and 583, all of which are within 30 km river distance of the intake, which is also an area of relatively high (human) population (Fig. 2.7b). The majority of the moderately impactful cells are also located near to the intake, although there are exceptions in the north-east of the catchment (245), central areas (296 and 349), and seven cells in the far west of the catchment (e.g. 310 and 426). There isn't a clear overlap between rainfall cell impact and

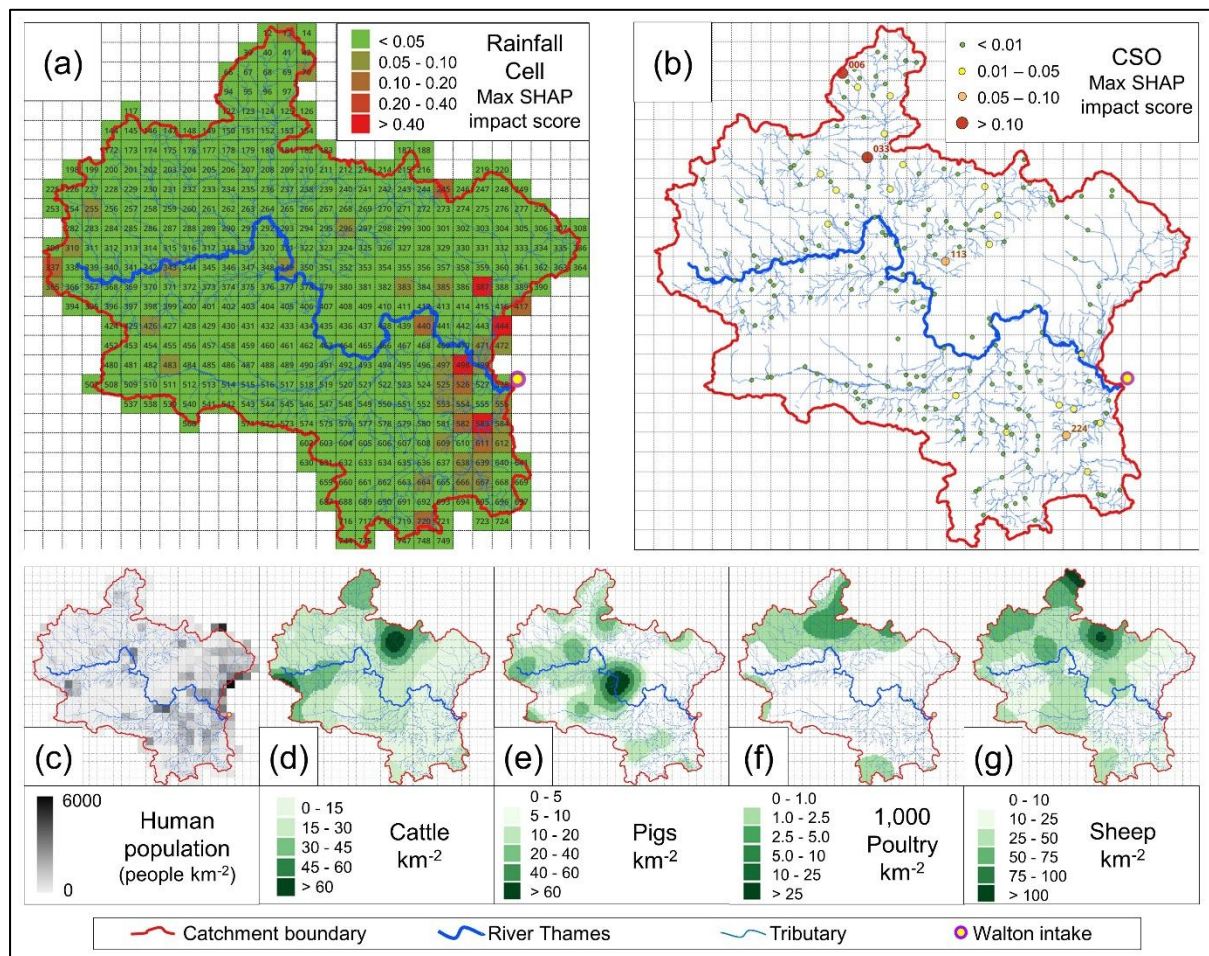


Fig. 2.7 Maximum SHAP impact scores for (a) 5x5 km rainfall cells and (b) CSO binary inputs, with (c) human population density from 2011 Census (Reis et al., 2017), alongside livestock densities for (d) cattle, (e) pigs (f) poultry and (g) sheep from APHA (2022).

livestock distributions, although the higher-impact cells in the west do coincide with higher cattle densities (Fig. 2.7c), and those positioned towards the centre and north-east of the catchment coincide with higher sheep, pig, poultry and cattle densities.

The impact of the CSO inputs is lower in magnitude than that of the rainfall cells. The most impactful CSO inputs (006 and 033) are far (>100 km river distance) from the intake and in relatively low population areas, although one moderately impactful CSO (224) is much closer to the intake and in a higher population area.

2.6 Discussion

2.6.1 Predicting *Cryptosporidium* at daily timescales

This study has demonstrated that predictions of *Cryptosporidium* can be made at daily timescales using an ML model. Although four previous studies have used either ML or deterministic models to generate daily predictions (Brion et al., 2001; Medema and Schijven, 2001; Dorner et al., 2006; Tang et al., 2011), each of these was validated against sparse data (with ≤ 68 *Cryptosporidium* measurements), whereas in this study models were trained using a minimum of 1165 *Cryptosporidium* measurements over approximately 5 years and validated with 683 measurements spread across 2 years. Furthermore, unlike several of the cited studies, this model does not require FIB data as an input. Such a well-validated demonstration of the predictability of *Cryptosporidium* at daily timescales constitutes an important finding, given the variety and uncertainty of pathogen sources, the daily, seasonal and annual variability in *Cryptosporidium* concentrations in the river, and the size and complexity of the catchment concerned. Predictions of the type presented here are very likely to have operational utility, since water supply managers do not normally require absolute *Cryptosporidium* values in order to make abstraction decisions, but only a reliable indicator of whether or not concentrations are likely to exceed certain thresholds – and the highest-performing models correctly predicted 69-75% of >1 oocysts L^{-1} exceedances, and predicted elevated concentrations on 78-89% of >2 oocysts L^{-1} exceedances. The approach taken – namely, the application of a Bagging-XGBoost model with an unparsimonious approach to feature selection – also has the benefit of simplicity in

that it does not require extensive preliminary data analysis in order to reduce dimensionality.

2.6.2 Differential performance of catchment-averaged and spatially distributed models

The highest performing spatially distributed model (Model 5b) slightly outperformed the highest performing catchment-averaged model (Model 1b) on the majority of metrics. More generally, those models incorporating spatially distributed features predicted more exceedances (i.e. had higher Recall values), but they also suffered from lower Precision values than Model 1b. This may be due to the far larger number of features used in the spatially distributed models, which can be a source of feature noise or overfitting (Dhal and Azad, 2022) – an effect which could be amplified by the high levels of multicollinearity known to be present in the input data (Chan et al., 2022). The fact that the spatially distributed models with the fewest features (Models 4a and 4b) had higher precision than those models with >500 features, adds further credence to this interpretation.

The higher Recall of Model 5b compared with Model 1b indicates that cell rainfall and CSO data provided relevant extra (source-related) information to the model. A major component of this is the more distant (≥ 30 km) rainfall cell inputs (see Fig. 2.6b), which will differ greatly from catchment-averaged rainfall because of the combination of extended time lags and averaging over multiple days. Catchment-averaged rainfall will serve as a reasonable proxy for the more impactful nearby (< 30 km) rainfall cells, hence the reason why catchment-averaged models are still relatively effective. However, cell rainfall data for more distant locations in the catchment (and therefore more distant *Cryptosporidium* sources) plays a dominant role in increasing the number of exceedances predicted by Model 5b relative to the catchment-averaged models.

Based on these results, it is arguable that the simpler catchment-averaged modelling approach provides a more time-efficient and robust option for many water abstraction operators, since for large catchments it would require a much smaller number of input features and it would also remove the need to obtain travel time estimates for discrete points within the catchment.

2.6.3 Applying model performance and SHAP analysis to understand *Cryptosporidium* transport and sources

Rainfall was shown to be a key input, with comparison of the models consistently demonstrating that the inclusion of rainfall features in some form resulted in improved performance. This was further supported by the SHAP analysis, which showed that rainfall inputs taken as a collective were the most impactful features in the two highest performing models (1b and 5b). This is an important finding, and one which highlights the centrality of rainfall as a mechanism for transferring *Cryptosporidium* into rivers, by mobilising and transporting (as rainfall-runoff) oocysts which are present in faecal matter on fields and in urban areas, as well as flushing oocysts from drainage networks and increasing the volumes of untreated sewage entering rivers (via CSO spills). This interpretation is well-supported by previous studies (Atherholt et al., 1998; Bhattarai et al., 2011) and established theory regarding *Cryptosporidium* transport in catchments (FWR, 2011).

The importance of catchment-averaged rainfall at 2-day lag time to both Models 1b and 5b indicates that rainfall over nearby sources (i.e. within 2 days' travel time) influences *Cryptosporidium* concentrations in the river more than rainfall over more distant areas – and therefore that nearby sources are more dominant. This is backed-up by the Model 5b SHAP output, in which cells that were near (<30 km river distance) to the intake had the largest impact on predictions. Existing theory states that *Cryptosporidium* from more distant sources will be subject to greater levels of dispersion and die-off during transport (Medema and Schijven, 2001), and hence will generally have a less pronounced effect on pathogen concentrations than sources which are near to the monitoring point. However, more distant sources are still relevant, as shown by the high impact of rainfall cells ≥30 km upstream of the intake, and by the improved performance of the models when extended lag times (up to 18 days) were applied to the catchment-averaged data, suggesting that less recent rainfall events – and hence more distant upstream sources – do have an important secondary effect on *Cryptosporidium* concentrations at the intake.

In Model 5b, the higher overall impact of rainfall cells relative to CSOs lends support to the theory that mobilisation of livestock-shed oocysts by rainfall-runoff is more important than mobilisation of human-shed oocysts in CSO spills. On the other hand, the correspondence of the most impactful rainfall cells with areas of moderate-to-high human population and

relatively low livestock population (e.g. cells 387 and 444) contradicts this and points to the significance of human sources. It is important, however, to draw a distinction between human and urban inputs; CSO spills can contain both untreated sewage and urban storm water, and are therefore likely to be dominated by human-shed oocysts in many cases. However, rainfall-runoff over higher (human) population areas will also generate discharges of urban storm water into the river system, which can contain oocysts shed by domestic animals and wild animals. Indeed, previous studies show that oocysts shed by non-human animals (e.g. dogs and pigeons) can be the primary source of surges in *Cryptosporidium* river concentrations after rainfall over built-up areas (Müller et al., 2020).

The superior performance of spatially distributed models which included CSO data (i.e. Models 4a, 4b, 5a and 5b) relative to those which excluded CSOs (Models 3a and 3b) also adds support to the significance of human sources. This could be particularly true at certain times of the year, such as in autumn or early winter, when river levels are often lower than in mid-late winter, and so potential dilution effects are reduced; this effect appears to be in evidence for periods in Oct and early Dec 2022, when the CSOs have a greater impact than the rainfall cells (Fig. 2.5g). It should be noted however that the binary form of the CSO inputs may serve to inhibit their effectiveness, since the data in this form provides no information to the model concerning the scale of untreated sewage spills. Overcoming these deficiencies in the CSO input data is challenging and would require access to information which was unavailable for the present study, including dimension data for each individual asset, which could then be used to estimate/scale flow volumes (Suslovaite et al., 2024). Furthermore, although spills from STW storm tanks are included in the dataset, discharges from the main STW effluent outfalls are not represented within the CSO inputs, even though such discharges may contain inadequately treated sewage during periods of high hydraulic loading.

Viewed as a whole, the SHAP analysis of the spatially distributed inputs can be taken as evidence that livestock, human and urban sources of *Cryptosporidium* are all important in determining river concentrations of the pathogen (with urban sources made up of a mixture of human and non-human oocysts). The impact of rainfall cells is of higher magnitude than the CSO inputs, but the CSOs add an important “polishing” effect to the final prediction (i.e.

by raising predictions above threshold concentrations when their lower-magnitude impacts are applied).

Soil moisture was shown to have the most (positive) impact in Aug-Oct – this corresponds to the months in which surface moisture levels often undergo significant change as the drier conditions of late summer transition into the wetter periods of autumn. This observation matches well with theory, since high moisture values should reduce the rate of oocyst desiccation on the land (Robertson et al., 1992) and increase runoff volumes (and hence the potential for *Cryptosporidium* mobilisation) during rainfall events (Singh et al., 2021). It should also be noted that basing the average soil moisture (and soil temperature) on data taken from four sites in the catchment is a potential model limitation. In particular, the use of spatially distributed (satellite-based) soil moisture data could complement (and increase the predictive utility of) the spatially distributed rainfall data.

Discharge also had the most positive impact in Aug-Oct, whereas the negative (i.e. subtractive) impacts of high discharge were shown to play an important role later in the season, as illustrated by the SHAP outputs for Nov 2022-Feb 2023 (Fig. 2.5e). The effects of this input parameter are expected to be non-linear. For example, high discharges may sometimes coincide with higher loadings of *Cryptosporidium* from rainfall-runoff or CSO spills, and with more rapid transport of mobilised oocysts in the catchment. However, high discharge could also lead to greater dilution, and hence reduced concentrations of the pathogen (Knapp et al., 2022). The impact of discharge rate-of-change at 1-day time lag on Model 1b (Fig. 2.6a) could represent the influx of *Cryptosporidium* oocysts liberated by re-suspended bed and bankside sediment in the wake of abrupt changes in flow (Crockett, 2004; Drummond et al., 2018).

Two of the most impactful features – *Crypto_Rolling_Mean* and *Day_of_Year* – serve as surrogates for other processes for which we lack appropriate data. *Day_of_Year* represents seasonal elements, such as the annual farming calendar (e.g. when manure-based fertiliser might be applied to the fields, or when livestock may be transferred from grazing fields to winter holds), or the timing of peak infection rates in the UK in both human and livestock populations – typically in the months of Sep and Apr, respectively, although the monitoring of infections in livestock is arguably too patchy to provide a reliable picture of seasonal trends (Public Health England, 2024; APHA, 2024). *Crypto_Rolling_Mean* provides a

surrogate for seasonal and longer-scale (e.g. annual) variations in *Cryptosporidium*. There are many factors which may result in longer-term variation, including sporadic outbreaks in the animal or human populations, as well as differential rates of oocyst survival and/or release into the environment as a result of meteorological or farming effects which have not been represented elsewhere by the input data. This feature's centrality is consistent with the model design, since it was included precisely for the purpose of providing a scaling factor to inform the model of the severity of the current *Cryptosporidium* season.

2.6.4 Modelling and feature interpretation limitations

All models were ineffective at predicting the magnitude of exceedances above 3 oocysts L⁻¹, although they were much more effective at predicting the timing of such exceedances. This is to be expected, given that higher exceedances are much rarer in the dataset and hence the models have fewer such instances on which to train. The performance of the models may also be limited by the quality of the *Cryptosporidium* data itself, which is subject to issues of sampling representativeness, analytical recovery and subjectivity of analysis method (Efstratiou et al., 2017b; Hassan et al., 2021). Although Thames Water's selective abstraction protocols apply *Cryptosporidium* measurements without correction for analytical recovery, a refined version of the model could still incorporate an estimated recovery component to improve predictive performance and interpretability. Direct adjustment of concentrations based on laboratory recovery data is problematic because routine quality control tests rarely match the matrices of the samples, but a continuously monitored proxy for recovery – e.g. turbidity – could be considered, particularly where sufficient site-specific laboratory data is available to establish a reliable relationship. This should be treated as an exploratory approach, however, because previous studies have found the relationship between turbidity and oocyst recovery to be weak, complex, or context-dependent (Feng et al., 2003; Francy et al., 2004). Collectively, these data quality challenges underline the necessity to work with large *Cryptosporidium* datasets when training models.

A one-day time-step was adopted because this was the maximum temporal resolution of the *Cryptosporidium* data, but sub-daily time-steps could reveal subtleties that are not visible in the current models. In particular, the discrete daily, sub-weekly or sub-monthly

sampling approach associated with routine monitoring by water companies provides only a snapshot of river conditions and may fail to capture short-duration contamination events. Episodic inputs associated with CSO spills, elevated discharges from STW outfalls, or localised rainfall-runoff events occurring close to the sampling locations could therefore be underrepresented even within the approximately daily dataset used here, potentially obscuring some short-timescale relationships between the source-related features (i.e. CSO spills and rainfall) and *Cryptosporidium* concentrations in the resulting models.

Unfortunately, although sub-daily environmental datasets are widely available, sub-daily *Cryptosporidium* data is not collected as part of routine water sector monitoring.

It should also be noted that travel time calculations for such a complex river network are challenging, and it is therefore probable that inaccurate lag times have been applied to some of the spatially distributed features. If this is the case, it will of course compromise the effectiveness of the spatially distributed models and the associated SHAP analysis. Interpretation of the spatially distributed rainfall data is likely to be further impaired by the homogeneity of rainfall events over what is a region of relatively low relief. In particular, the high correlation between adjacent/proximate rainfall cells will reduce the ability of an ML model to discern the effects of rainfall over different parts of catchment. The higher-impact cells highlighted by the SHAP analysis may in fact represent effects produced by nearby cells, which could complicate interpretation of likely sources when SHAP outputs are compared with livestock and human population data. Furthermore, a number of factors not considered here are also likely to influence the potential for animal waste to reach surface water bodies; these include the storage locations of waste-based fertilisers, soil type, land gradient and the proximity of grazing animals to watercourses.

Finally, it is important to bear in mind that although XGBoost model performance has a high tolerance for multicollinearity in the input data, this is less true for model interpretability by feature importance analysis (including SHAP), with the potential generation of outputs which over/under-estimate individual feature impacts (Drobnič et al., 2020). This is one drawback of the unparsimonious approach used here. The spatially distributed features are a particular concern in this context because of the low magnitude of their associated SHAP impacts; findings concerning the most important sources/source areas and their relative significance should therefore be treated with caution. Other studies have analysed model

inputs beforehand and removed highly-correlating features (Takefuji, 2025), but even when such approaches are adopted, the question regarding which features to retain/remove remains problematic, and can result in the loss of valuable input data (O'Brien, 2017). However, there is undoubted scope for future researchers to engage with the challenging problems posed by feature importance and multicollinearity in the context of data-rich ML-based pathogen river models; such work could also seek to understand the effect of feature interactions (Alomari and Andó, 2024).

2.7 Conclusions

This study advances the modelling of *Cryptosporidium* in river systems by evaluating how catchment-averaged and spatially distributed environmental and source data can be integrated with antecedent *Cryptosporidium* measurements within an interpretable ML framework. It is the first published attempt to make daily predictions of *Cryptosporidium* concentrations in a complex river system by training and validating a model using approximately daily monitoring data in a long-term (7-year) dataset. It successfully demonstrates that such predictions – which are of great relevance to water resource management – can be practically made utilising commonly available environmental datasets alongside existing raw water monitoring data. The contribution is therefore both predictive and explanatory: the modelling framework provides a means of assessing which spatial representations, environmental drivers, source indicators and lag times are most informative for understanding short-term *Cryptosporidium* risk. XGBoost was used because of its reputation for high predictive performance and its ability to deal with multicollinearity in the input data – a key quality, given the use of input parameters at multiple lag times.

The most successful model (containing both catchment-averaged and spatially distributed data) achieved F-scores of 0.701 and 0.444 when predicting exceedances above 1 and 2 oocysts L⁻¹, respectively. A simpler model (using catchment-averaged inputs) performed almost as well; it may therefore be preferable in an operational setting to opt for a catchment-averaged model, particularly if spatially distributed data and associated travel time estimates are unavailable, or challenging to obtain. Models which excluded rainfall entirely from the input data performed less well, demonstrating the centrality of rainfall as a mobilisation mechanism.

All of the models tended to underestimate moderate-to-high exceedances (>2 oocysts L^{-1}), although the models were more successful at predicting the *timing* of such exceedances, predicting a >1 oocysts L^{-1} exceedance on 78-89 % of days when *Cryptosporidium* exceeded 2 oocysts L^{-1} .

Sources near to the intake were shown to have the greatest effect on *Cryptosporidium* concentrations in the river, as demonstrated by the importance of catchment-averaged rainfall at 2-day lag time and by the impact of spatially distributed rainfall features <30 km upstream. SHAP analysis of spatially distributed model inputs supports the theory that livestock, human and mixed urban sources of *Cryptosporidium* are all important in determining the pathogen's concentrations in the river. However, the effectiveness of spatially distributed features may have been partially impeded by potential errors in travel time estimation and (in the case of CSO inputs) by the binary form of the data. The SHAP analysis results should be considered in the light of the high number of features and multicollinearity in the input data, which is particularly high in the case of the spatially distributed rainfall inputs.

Discharge antecedent to rainfall events was also shown to be an important feature, with discharge rate-of-change, soil temperature and soil moisture adding further useful information to the model, although the vital contribution of two proxy inputs representing antecedent *Cryptosporidium* concentrations and day of the year show that there are still major knowledge and data gaps which, if addressed, could lead to an improved model.

The approach outlined here could be readily adopted at many abstraction sites, at little or no cost. Although such ML models rely on extensive training data, it should be noted that many water providers in the UK and around the world have already collected substantial *Cryptosporidium* datasets as part of their routine monitoring regimes (Smeets et al., 2007). When combined with remote sensing environmental data and powerful ML software tools (both of which are often freely available), these existing datasets contain significant untapped potential, which could be harnessed not only to support future modelling, but also to advance scientific understanding of *Cryptosporidium* source contributions, transport pathways and temporal variability in complex catchments.

Chapter 3: Quantitative Microbial Risk Assessment of *Cryptosporidium* for London's water supply: Analysis of raw and treated water to inform supply management

Reproduced from draft journal paper: Smalley, A. L., Lauretto, M. S., Razzolini, M. T. P., Douterelo, I., Chipps, M. and Shucksmith, J. D. (2026) Quantitative Microbial Risk Assessment of *Cryptosporidium* for London's water supply: Analysis of raw and treated water to inform supply management. Intended journal: *Water Research*.

3.1 Abstract

Quantitative Microbial Risk Assessment (QMRA) is a widely used approach for understanding and managing *Cryptosporidium* risk in drinking water systems, but is often constrained by reliance on raw/source water concentrations in combination with treatment-performance assumptions or log-removal credits. This study integrates source water, reservoir and treated water data within a novel parallel QMRA framework to evaluate *Cryptosporidium* risk in final treated supplies for a complex multi-source drinking water system. A whole-city approach for London was adopted, accounting for complex blending of source waters in the reservoir storage and treatment system, and in the drinking water supply network. Parallel raw water (RW-QMRA) and treated water (TW-QMRA) risk outputs were generated for biannual periods to quantify seasonal effects and model selective abstraction-based risk management strategies. Log-removal values were computed using additional reservoir datasets and a novel, hybrid data imputation method with parametrically-defined lower limits to deal with the high proportion (>98%) of left-censored data points in the treated water data.

TW-QMRA results revealed no clear seasonality, with median biannual risk values over 6 years ranging from 0.0037% to 0.0057% - below the USEPA "tolerable" risk guideline. RW-QMRA outputs showed clear seasonality (reflecting high *Cryptosporidium* river concentrations in winter), with above-guideline median risk for some Sep-Feb periods (up to 0.012%). Selective suspension/reduction in river abstraction volumes was shown to lower median risk to below the guideline level for all periods, but the overall risk reduction was modest (<0.3 log units) compared with potential reductions from alternative risk mitigation options (e.g. the 2-4 log reduction achievable with UV treatment). The absence of

seasonality in the TW-QMRA outputs provided partial evidence of system robustness, while the RW-QMRA outputs – which are less vulnerable to left-censored data uncertainties – offered a more conservative and loading-sensitive assessment of risk variation.

3.2 Introduction

Cryptosporidium is a protozoan pathogen which can cause gastrointestinal illness in infected humans and animals. The largest *Cryptosporidium* outbreaks on record have been caused by the consumption of contaminated drinking water (Mac Kenzie et al., 1994; Widerström et al., 2014; Pignata et al., 2019). Risk management of *Cryptosporidium* by drinking water providers typically relies on multi-barrier treatment options, often involving physical removal by coagulation-flocculation and/or filtration (Betancourt and Rose, 2004). The pathogen is resistant to chlorination at standard treatment concentrations, but deactivation can be achieved either by ozonation or UV treatment, though at additional financial cost per unit of treated water (USEPA, 2025). Hence, when applying risk mitigation strategies to protect public health, it is essential to have some means of calculating the effectiveness of different approaches. Such calculations are often challenging, however, especially for large and complex systems where risk is a function of many factors, including pathogen loadings in raw water supplies (e.g. river systems, which are influenced by site-specific and highly variable seasonal factors, such as rainfall-runoff), complex and interlinked treatment systems, and population characteristics (e.g. differential immunity to infection). The need for risk quantification becomes even more pressing when prioritising investment, making asset management decisions or considering the effective implementation of other risk mitigation strategies, such as selective surface water abstraction to reduce loadings of contaminants to water treatment works (WTW).

Quantitative Microbial Risk Assessment (QMRA) is an established approach advocated by the World Health Organisation (WHO) which has been used in a drinking water context for an array of pathogens, including *Cryptosporidium* (Medema et al., 2009). The two key inputs usually required are (i) the exposure of the population to the pathogen and (ii) the dose-response relationship exhibited by members of the population. Exposure typically breaks into two further components: the concentration of the pathogen in the ingested water (taken from water quality data) and the volume of ingested water. Although point

estimates of inputs can be used, the preferred QMRA approach is to probabilistically model each input separately, generating distributions which can be randomly sampled in a Monte Carlo analysis. Final risk outputs are then in the form of daily and annual infection risk distributions, which reflect both the uncertainty and natural variability of the inputs (WHO, 2016).

Each of the inputs presents its own challenges (Medema et al., 2009). For example, analytical datasets for pathogens (including *Cryptosporidium*) often contain high numbers of non-detects – i.e. left-censored values below the analytical limit of detection (LOD). In many recent studies, the method adopted for dealing with non-detects is parametric modelling, using censored Maximum Likelihood Estimation (MLE), in which a likelihood function is applied to determine the most probable fitted distribution (Helsel, 2012). Two problematic aspects of the MLE method are that it relies on assumptions regarding the shape of the concentration distribution (e.g. gamma or log-normal), and that the curve-fitting process is biased towards the uncensored data points (Antweiler and Taylor, 2008). It is also regarded as non-ideal for modelling data which is highly left-censored, with one study recommending 80% as an upper limit (Hewett and Ganser, 2007). Alternative methods have been used – e.g. the semi-parametric Regression on Order of Statistics (ROS) (Helsel and Cohn, 1988) and the non-parametric Kaplan-Meier (KM) and multiple-imputation (MI) methods) (Helsel, 2012). MI methods – which are based on repeated random sampling from distributions – have also been shown to outperform the alternatives (Canales et al., 2018) and are less reliant on assumptions regarding underlying distributions.

The majority of published QMRA studies are also based on either raw (untreated) or treated water datasets in isolation – and where both data streams are available, they are rarely symmetrical, with one containing many more data points than the other (Owens et al., 2020). QMRAs based on raw water data (RW-QMRA) must incorporate a log-removal element to calculate the final treated concentrations after water has passed through the WTW(s). The most common approach is to assign log removal credits (LRCs) appropriate to the on-site treatment processes. LRCs are usually taken from the literature, but these may not reflect true, site-specific performance of the treatment processes (Schmidt et al., 2020). Use of LRCs also fails to capture the inherent variability/uncertainty in the system, which is a central aim of the QMRA method. An alternative approach is to model log-removal values

(LRVs) directly, using real data, but this requires access to both raw and treated concentration measurements. A potential advantage of RW-QMRAs is that they can incorporate changes in risk due to observed differences in pathogen loadings (e.g. seasonal changes in river concentrations), but data availability often precludes this – i.e. if concentration monitoring data is sparse, it cannot be divided into seasonal subsets while maintaining statistical robustness.

QMRA studies based on treated water data (TW-QMRA) do not require assumptions/calculations for log-removal, since the concentration data represents the status of the water, post-treatment. In cases where treated water frequently contains detectable pathogens, TW-QMRA can be the preferred approach, because its source data directly reflects the condition of the water being ingested (Guzmán Barragán et al., 2022). However, in robust and effective water treatment systems, where pathogens are almost always removed to levels below the LOD, the concentration data will be highly left-censored, meaning that (compared with the raw data) there is considerable uncertainty regarding true pathogen levels in the water. In such cases, TW-QMRAs also have limited scope for assessing changing environmental loadings (e.g. due to seasonal effects), because unlike RW-QMRA, such loadings are not inherently incorporated into the input data. In summary, both RW-QMRA and TW-QMRA can be subject to high levels of uncertainty in their exposure assessments; for RW-QMRA, this uncertainty resides chiefly in the log-removal estimates, whereas in TW-QMRA the bulk of the uncertainty lies in the left-censored concentration data.

In a number of countries (including the UK, Netherlands, Australia, USA and Canada) monitoring of *Cryptosporidium* concentration in raw source waters is undertaken at many sites as part of the standard risk management strategy (Clancy and Hunter, 2004; Schets et al., 2008; Cook et al., 2013; Petterson et al., 2021). Regulatory requirements in the UK also require water providers to collect extensive treated water data for WTWs considered at higher risk; when combined with the large raw water datasets, this constitutes a significant resource which – from a research perspective – has not been fully tapped (Smeets et al., 2007). Rich datasets for both raw and treated waters provide researchers with an opportunity to perform parallel RW-QMRA and TW-QMRA at the same sites to assess the effect of different environmental loadings over time (e.g. seasonal effects), quantify

treatment efficacy by calculating LRVs, and evaluate the robustness of the treatment systems by comparing the response to different loadings seen in both the RW-QMRA and TW-QMRA. If treatment systems are robust, changes in environmental loadings seen in the RW-QMRA outputs should be either less apparent or indiscernible in the TW-QMRA (i.e. because the processes will remove the extra influx of pathogens). Furthermore, the RW-QMRA can be used to scenario-test hypothetical risk management strategies to estimate improvements in relative risk.

The aim of this study is to develop and evaluate a dual RW-QMRA and TW-QMRA approach for assessing pathogen risk across a complex multi-source drinking water system, using the first city-wide QMRA of *Cryptosporidium* for London's drinking water as a case study. The study utilises novel and extensive datasets to explore and refine QMRA approaches (e.g. log removal calculations) in the context of large and complex water supply systems, and applies a hybrid data imputation method to model *Cryptosporidium* concentrations – a step which is required because of the reliance on exclusively left-censored data in some of the seasonal TW-QMRA input data. Further, we demonstrate the applicability of QMRA to inform management strategies by considering the effectiveness of a currently-active selective abstraction protocol (described below), and scenario-test alternate approaches to show the potential risk reduction and explore trade-offs with abstraction volumes.

3.3 Method

3.3.1 Description of drinking water assets

Thames Water-managed assets provide drinking water to approximately 76 % of the population of Greater London (≈ 7.4 million people). Surface water provides the source for 87-89 % of this water (based on 2016-2025 Thames Water output data) and this is obtained from 6 abstraction points on the River Thames (Intakes A-D₃), and 2 abstraction points on the River Lee (Intakes E & F) (see Fig. 3.1). Water from each of the intakes is transferred to one or more of the multiple bankside reservoirs which flank the rivers. The reservoirs supply water to 5 large WTWs, with each WTW receiving water from at least 2 reservoirs, the proportional blend of which changes over time due to a variety of circumstances (e.g. water quality considerations, reservoir levels and maintenance). Transfer of water between reservoirs (for maintenance or water storage management) is also routine. The WTWs

pump water both to local distribution networks and to the Thames Water Ring Main (TWRM). Hence water produced at any of the WTWs can be consumed in any part of Greater London connected to the TWRM. For simplicity, this study will focus exclusively on water blended in the TWRM, which is the supply accessed by the majority ($\approx 76\%$) of consumers in Greater London (Dickens and Bensted, 1988).

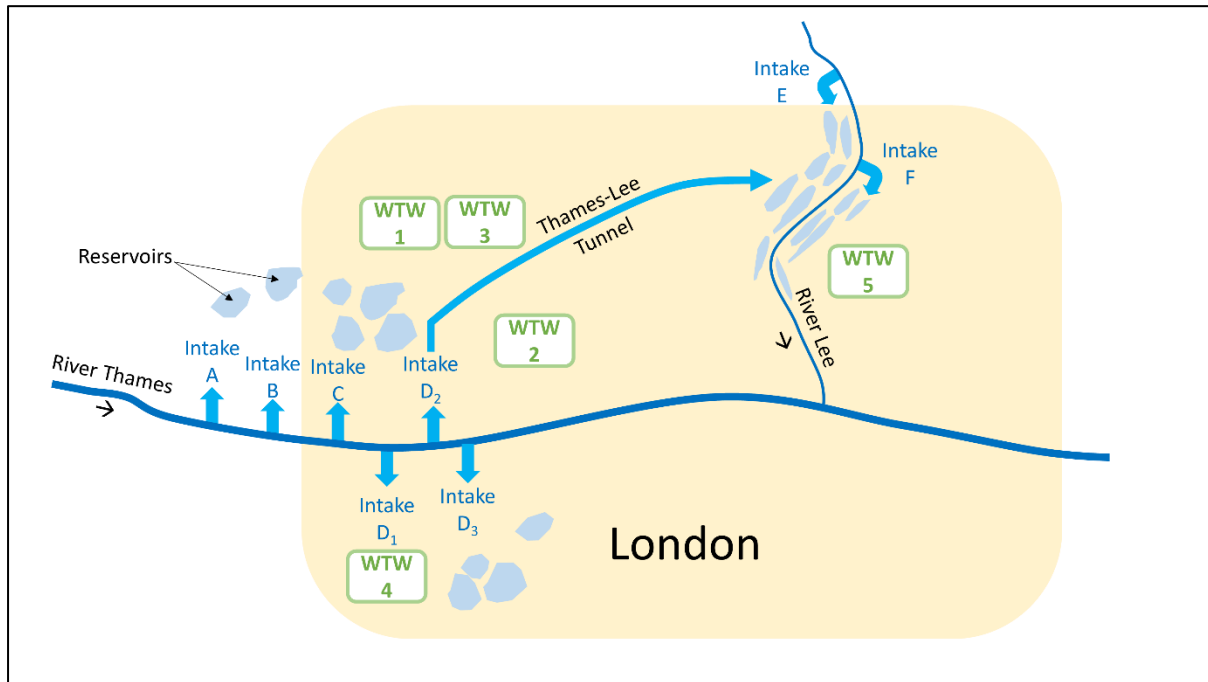


Fig. 3.1 Schematic map of the major Thames Water assets supporting London's drinking water supply.

The WTWs mitigate the *Cryptosporidium* risk primarily by physical removal of oocysts during the treatment process. This is either by Rapid Gravity Filtration (RGF) followed by Slow Sand Filtration (SSF), as in WTWs 1, 2, 3 and 5, or by Coagulation-Flocculation, combined with Counter-Current Dissolved Air Flotation/Filtration (CoCo-DAFF), as in WTW 4. Estimates of log-removal credits (LRCs) for the five water treatment works are given in Table 3.1, based on literature values (Hijnen and Medema, 2007; Edzwald et al., 2001). There is disagreement on the effectiveness of RGF for *Cryptosporidium* removal, with previous UK guidance conservatively assuming an LRC of zero (Badenoch, 1990), although the consensus opinion is that RGF can achieve ≈ 1 -log removal with effective use of coagulants (Hijnen and Medema, 2007). Overall LRCs both with and without the literature values for RGF are therefore provided in Table 3.1. All sites also apply ozone at the tertiary stage, and two

sites (WTWs 1 and 4) apply pre-ozone at the head of the treatment train, but the system and Ct (concentration-time) levels are not designed to achieve inactivation of oocysts. For this reason, ozone log-removal credits have not been applied to the LRCs quote in this study.

Table 3.1 Summary of Log-removal Credit (LRC) estimates for WTWs and *Cryptosporidium*-relevant treatment processes. (Coag-Floc = Coagulation-Flocculation; CoCo-DAFF = Counter-Current Dissolved Air Flotation/Filtration.) LRCs taken from Hijnen and Medema (2007) and Edzwald et al. (2001).

LRC		Rapid Gravity Filtration (RGF)	Slow Sand Filtration (SSF)	Coag-Floc	CoCo-DAFF	Overall LRC			
						With RGF LRC		Without RGF LRC	
						Mean	Range	Mean	Range
	Mean:	1.1	4.8	1.9	1.7				
	Range:	0.0-2.3	2.7-6.5	0.4-3.8	1.4-2.0				
Asset									
WTW 1, 2, 3 & 5		✓	✓			5.9	2.7 - 8.8	4.8	2.7-6.5
WTW 4				✓	✓	-	-	3.6	1.8 - 5.8

3.3.2 *Cryptosporidium* monitoring

Routine monitoring of *Cryptosporidium* concentrations is undertaken at the river abstraction points, WTW inlets and WTW outlets, with all analysis carried out according to the standard immunofluorescence and enumeration-based Blue Book Method (Environment Agency, 2010). Intakes D₁ and D₂ are located ≈1 km apart, but monitoring is only carried out at Intake D₁, hence for this study these will be treated as a single monitoring site. Intake D₃, just downstream of D₂, supplies only small volumes of water intermittently, therefore its volumes have been combined with those of Intake D₁.

Cryptosporidium river concentrations are highly seasonal in the Thames catchment, with most years registering their highest concentrations in Nov-Jan, and moderately high values in Sep, Oct and Feb (Smalley et al., 2025). The data was therefore split into biannual sets so as to consider the “low” (Mar-Aug) and “high” (Sep-Feb) concentration periods separately. These will be referred to henceforth as Summer and Winter periods, respectively. WTW outlet *Cryptosporidium* data is available from Nov 2017 onwards, but to ensure that each biannual period contains data across all months, the analysis has been limited to the 6 years from 1 Mar 2018 to 29 Feb 2024.

3.3.3 QMRA workflow

Parallel quantitative microbial risk analysis is based on *Cryptosporidium* concentrations measured in the river water (RW-QMRA) and the treated water (TW-QMRA). The modelling steps can be broken down as follows: (1) proportional allocation of intake water to WTWs; (2) concentration modelling; (3) log-removal modelling; (4) ingestion rate modelling; (5) dose-response modelling; and (6) the final QMRA models. All data processing was performed in Python (version 3.13.5) using the Anaconda distribution (available from www.anaconda.com).

3.3.3.1 Proportional allocation

Thames Water provided daily inflow volumes from the intakes to specific reservoirs, and daily outflow volumes from the reservoirs to specific WTWs. This data was compiled and used to calculate the relative proportion of water from individual intakes to individual WTWs (along the intake → reservoir(s) → WTW pathway) for each biannual period.

3.3.3.2 Concentration modelling

A Hybrid Multiple-Imputation (HMI) approach with bootstrapping and parametrically defined lower limits has been adopted for the concentration modelling. For simplicity and consistency, the method has been applied uniformly to all three of the different concentration datasets (i.e. river intake, WTW inlet and WTW outlet data). The advantage of this approach over MLE or ROS are that it relies on fewer prior assumptions regarding the distributions of the parent data (Canales et al., 2018).

The “hybrid” element relates to the differential treatment of the censored and uncensored values. Censored values are substituted with new values sampled from a log-uniform distribution, with an upper limit (C_{LOD}), determined by the LOD (i.e. the censored value) and a lower limit, (C_{LL}) defined by 0.1x the 50th percentile (Q_{50}) of a parametric model, which was fitted to the entirety of the data for each of the data types (see Appendix B for MLE modelling). This is a similar approach to that applied by Canales et al. (2018) to confine MI boundaries.

Uncensored values are treated using an empirical simulation method in which the original uncensored data is divided into bins and then randomly sampled from a uniform distribution within each bin (with sampling proportional to the frequency of the original

data points). Bootstrapping was implemented to generate 10,000 samples per biannual time period per WTW. For RW-QMRA, the number of bootstrapped values drawn from each relevant intake was determined by the proportional allocation data.

Note that analytical recovery has not been factored into the analysis. Although the correction of concentration data by a recovery factor is recommended in QMRA where suitable recovery data is available (Medema et al., 2009; WHO, 2016), some previous studies have avoided doing so on the basis that incomplete recovery is partly compensated for by the inclusion of non-viable oocysts and non-infectious *Cryptosporidium* species in standard IMS-IFM oocyst counts (Smeets et al., 2007; Sato et al., 2013). Furthermore, the available recovery data was limited to routine quality control tests that did not reflect the matrices of the environmental samples and therefore did not constitute true sample-specific recovery estimates; applying such correction factors therefore risked introducing additional error (Hamilton et al., 2018).

3.3.3.3 Log-removal modelling

Log-removal values (LRVs) were required for the RW-QMRA to simulate the two-stage reduction in *Cryptosporidium* concentrations during (i) storage in reservoirs and (ii) treatment in WTWs. Given access to both treated and untreated water datasets, these were calculated as in equations 8 and 9:

$$LRV_{storage} = LOG_{10} \left(\frac{C_i}{C_{wi}} \right) \quad (8)$$

$$LRV_{treatment} = LOG_{10} \left(\frac{C_{wi}}{C_{wo}} \right) \quad (9)$$

where the concentrations (in oocysts L⁻¹) are given by C_i (at the intake), C_{wi} (at the WTW inlet) and C_{wo} (at the WTW outlet). Data for each asset pair (i.e. intake/inlet or inlet/outlet) was first ranked by concentration across the full six-year period (Smeets et al., 2008), then independently divided into quantile bins. Within each bin, upstream and downstream values were randomly paired to compute LRVs. The process was repeated over 1,000 iterations to yield a distributional estimate of removal performance over the study period. Censored concentrations were substituted with values randomly sampled from a log-uniform

distribution between the previously defined lower limit (C_{LL}) and the censored value (C_{LOD}). Uncensored values were applied directly without substitution. Note that rank-based pairing was preferred over time-based pairing because of the difficulty in apply appropriate lag times for storage that would be appropriate for multiple reservoirs with different (and dynamic) retention times, and because time-based pairing was found to generate extreme and unrealistic LRVs.

3.3.3.4 Ingestion rate and dose-response model

Ingestion rate (IR , in mL day⁻¹) was modelled using a log-normal distribution fitted to the regulator's survey of unboiled tap water consumption by adults in the UK (DWI, 2008). The fitted natural-log parameters were log-mean (μ) = 5.63 and log-standard deviation (σ) = 1.198, with the distribution capped at the 99th percentile (4,523 mL day⁻¹).

The daily ingested dose (d) (in oocysts) was calculated as in Equation 10:

$$d = C \cdot IR / 1000 \quad (10)$$

where C is the concentration (oocysts L⁻¹) in the ingested water. Dose-response was based on data published in DuPont et al. (1995) and fitted to an exponential model, as shown in Equation 11:

$$P_{inf_d} = 1 - e^{-k \cdot d} \quad (11)$$

where P_{inf_d} is the daily probability of infection at *Cryptosporidium* dose d (in oocysts) and k is the dose-response constant (also known as the infectivity coefficient), which represents the probability that a single organism will survive the host's defences to initiate an infection. This is commonly based on challenge studies involving the infection of human volunteers.

However, such studies have demonstrated substantial variation in infectivity between different *Cryptosporidium* isolates. The Iowa isolate was the focus of the earliest human challenge study for *Cryptosporidium*, incorporating 29 volunteers and 8 different doses in the 30-1,000,000 oocysts range (DuPont, 1995). Subsequent challenge studies working with different *C. parvum* isolates have identified isolates with lower infectivity (UCP) and higher infectivity (TAMU), using four dose levels and involving 17 and 14 volunteers, respectively (Messner et al., 2001). The DuPont (1995) data was applied in this study for the primary Iowa dose-response (DR) model, on the basis that it constitutes the most substantial single-

isolate dataset available, and because it represents a middle ground between higher- and lower-infectivity isolates. Uncertainty in k was modelled using a bootstrap method described in Haas, Rose and Gerba (1999) and Guzmán Barragán et al. (2022), to generate an empirical distribution of k consisting of 20,000 values.

To evaluate the effects of higher- and lower-infectivity isolates on risk estimates, a secondary pooled DR model was applied, in which k -values were sampled from a broader distribution representing the dose-response data from UCP, Iowa and TAMU isolates. The distribution of 50,000 k -values was obtained by applying the hierarchical Bayesian framework described by Messner et al. (2001). This approach treats isolate-specific dose-response parameters as samples from a wider population of human-infective *Cryptosporidium* strains and propagates between-strain variability through the Monte Carlo simulation (see Appendix B.2 for more information). The QMRA outputs from the two dose-response models are referred to henceforth as Iowa DR QMRA and pooled DR QMRA.

3.3.3.5 Final QMRA model

Infection risk was computed according to a standard QMRA method (WHO, 2016), with daily probability of infection (P_{inf_d}) as shown in Equation 11. Infection probability during each biannual period (P_{inf_period}), which is the likelihood of at least one infection during that time, is then given by Equation 12:

$$P_{inf_period} = 1 - \prod_{i=1}^n (1 - P_{inf_d,i}) \quad (12)$$

where n is the number of days in a biannual period (181-184) and $(1 - P_{inf_d,i})$ is the probability of *not* being infected on day i . To enable comparison with the USEPA annual infection risk guideline value, biannual infection probabilities were converted to annualised form using Equation 13:

$$P_{inf_y} = 1 - (1 - P_{inf_period})^{365/n} \quad (13)$$

A Monte Carlo modelling approach was adopted to capture variability and uncertainty in the input parameters. Monte Carlo simulation involved sampling from the previously described bootstrapped or parametric distributions to obtain daily values for concentration (C), log removal values ($LRV_{storage}$ and $LRV_{treatment}$), ingestion rate (IR) and dose-response constant

(k). (Note that for the TW-QMRA, the LRV values defaulted to zero.) Equation 11 was then applied to compute the daily infection probability (P_{inf_d}). This process was repeated for n days, appropriate to the relevant biannual (Mar-Aug or Sep-Feb) period, after which the resulting P_{inf_d} values were used to calculate P_{inf_period} and P_{inf_y} (Equations 12 and 13). This sequence constitutes a single Monte Carlo simulation, and the process was repeated across 20,000 iterations for all time periods and for assets in the RW-QMRA and TW-QMRA models. (Note also that, in line with the conventional QMRA practice, k varied between iterations but remained constant within each iteration.)

To represent the blending of water from all modelled WTWs supplying London, daily infection probabilities from WTWs were weighted by their relative WTW output volumes and combined into a single blended infection probability distribution for each period. The standard QMRA approach is to model risk for children and adults simultaneously (WHO, 2016), but in this study risk values are calculated for adults only. Age-related risk outputs for *Cryptosporidium* differ only as a consequence of differential consumption rates, as all other model inputs are usually identical; therefore, risk outputs for adults consistently exceed those for children because of higher adult ingestion rates. (Razzolini, Santos and Bastos, 2010; Sato et al., 2013; Guzmán Barragán et al., 2022). A QMRA for adults-only therefore provides the most conservative (i.e. highest-risk) output.

3.3.4 Modelling risk reduction under selective abstraction regimes

Selective abstraction of river water can be an effective way of reducing contaminant loadings to reservoirs or water treatment works. Typically, this involves reducing or suspending abstraction when concentrations rise above “trigger” thresholds – a practice used by Thames Water for managing *Cryptosporidium* risk. Table 3.2 shows the two Selective Abstraction Regimes (SARs) currently in place.

Because of its use of raw water data, the RW-QMRA can be applied to model the effect of applying Thames Water’s SARs on infection risk (this is not true for the TW-QMRA, because of its use of treated water data as an input). The effect of SARs was assessed by re-running the RW-QMRA model with Suspension Threshold (ST) and Reduction Threshold (RT) conditions. For RT, two types were applied to represent the suspension protocols described in Table 3.2; these were reduction of the original abstraction volume by 50% (⁵⁰RT) and by

25% (^{25}RT). Because abstraction volumes do not feed directly into the QMRA model, changes in abstraction volumes were modelled by removing and reducing values from the bootstrapped data. Specifically, concentrations which exceeded the ST were removed altogether, whereas concentrations which exceeded ^{50}RT and ^{25}RT were reduced by 50% and 25%, respectively. Threshold concentrations were applied to the data in the ST $\rightarrow$ ^{50}RT $\rightarrow$ ^{25}RT order and reductions were only applied once (i.e. values reduced because they exceeded ^{50}RT were not further reduced if the adjusted value also exceeded ^{25}RT).

Table 3.2 Trigger thresholds applied in the selective abstraction-based risk management of *Cryptosporidium* at Thames Water (Thames Water, 2021)

<i>Cryptosporidium</i> river concentration (oocysts L ⁻¹)	Selective Abstraction Regime 1 (SAR-1) Applied at long retention time reservoirs	Selective Abstraction Regime 2 (SAR-2) Applied at short retention time reservoirs
>1	Maintain pumping	Reduce pumping to maintain current volume of storage (or as close as possible)
>2	Reduce pumping to maintain current volume of storage (or as close as possible)	Reduce pumping to 50% of current outflow (or as close as possible)
>4	Reduce pumping to 50% of current outflow (or as close as possible)	Stop all pumping
>5	Stop all pumping	Stop all pumping

This approach was carried out for SAR-1 and SAR-2 ($\text{ST}_5^{50}\text{RT}_4^{25}\text{RT}_2$ and $\text{ST}_4^{50}\text{RT}_2^{25}\text{RT}_1$, respectively, where the subscript denotes the trigger threshold concentration). To demonstrate the use of the model to scenario test new risk management strategies, a hypothetical abstraction regime (SAR-H) was modelled, based on a more stringent set of thresholds ($\text{ST}_{1.5}^{50}\text{RT}_{0.5}$).

Selective abstraction inevitably reduces overall abstraction volumes. To quantify the effect of this alongside risk reduction, the percentage abstraction volume reduction was calculated for each SAR. Abstraction deficit was defined as the % reduction from the mean annual abstraction volume (where the latter value was constructed from smoothed daily abstraction datasets, corrected for existing SAR suspensions/reductions).

3.3.5 Sensitivity analysis

To evaluate the relative impact of the different inputs, sensitivity analysis was performed by calculating the Spearman rank correlation coefficients for the annualised risk values with model inputs (note that the latter were averaged for each iteration within the Monte Carlo simulation) (Vásquez et al., 2014).

3.4 Results

3.4.1 Summary statistics, sampling frequency and *Cryptosporidium* detections

The summary statistics in Table 3.3 show that for each asset type, mean *Cryptosporidium* (C_{mean}) was highest at Intake E, WTW Inlet 2 and WTW Outlet 3 – three assets which are not directly linked (although WTWs 2 and 3 do draw water from shared reservoir sources).

Lowest C_{mean} values were at intake F, and the linked assets WTW Inlet 4 and WTW Outlet 4. However, WTW Outlet 4 also had the highest mean LOD – an important consideration when modelling the left-censored data, since high LODs will produce higher imputed concentration values. Sampling frequency varied across the sites, with least frequent sampling at Intakes B followed by Intake C. Fortunately, Intakes B and C are only ≈ 5 km apart on River Thames, so the values can be legitimately merged (the option taken in this study).

Across the three data types, there was a stark contrast in the number of *Cryptosporidium* detections (i.e. uncensored values), which made up 36-73% of the river intake data, 3-9% of the WTW inlet data, and 0.7-2.0% of the WTW outlet data. When the detections from all 6 years are plotted by day-of-year (Fig. 3.2), the intakes follow the seasonal pattern previously observed for this catchment (Smalley et al., 2025), with more detections and higher concentrations occurring in Sep/Oct to Feb/Mar. This seasonal trend was reproduced in the WTW inlet data, although the differences between the months were less stark. The seasonality in the WTW outlet data was lagged by comparison, with far more detections in Jan-Mar than in Sep-Dec, and with detections also occurring (albeit less frequently) in the late spring and summer months.

Table 3.3 Summary statistics and measurement counts for the intakes, WTW inlets and WTW outlets. (Cen = Censored values (non-detects); Det= Detections (uncensored values). For concentration statistics, censored data points are here treated as 0 oocysts L⁻¹. Shading indicates highest, intermediate and lowest concentrations.)

		Concentration (oocysts L ⁻¹)				% Det	% Cen	Measurements
Site		C _{mean}	C ₉₅	C _{max}	LOD _{mean}			
Intakes	A	0.28	1.2	5.5	0.1	50.1	49.9	803
	B	0.38	1.5	4.4	0.1	55.9	44.1	102
	C	0.30	1.5	4.4	0.1	50.5	49.5	289
	D	0.45	2.0	13.4	0.1	54.9	45.1	1738
	E	0.53	2.0	6.5	0.1	73.1	26.9	468
	F	0.18	1.0	7.5	0.1	35.7	64.3	856
WTW Inlets	1	0.0075	0.0	0.5	0.1	4.9	95.1	815
	2	0.0131	0.1	1.0	0.1	8.8	91.2	872
	3	0.0062	0.1	0.4	0.1	5.3	94.7	795
	4	0.0035	0.0	0.2	0.1	3.0	97.0	1336
	5	0.0079	0.1	0.6	0.1	5.4	94.6	3403
WTW Outlets	1	1.6E-6	0	3.0E-4	3.0E-4	0.7	99.3	1145
	2	2.9E-6	0	3.5E-4	3.4E-4	1.1	98.9	622
	3	9.1E-6	0	1.5E-3	3.5E-4	2.0	98.0	539
	4	1.0E-6	0	6.2E-4	6.2E-4	0.2	99.8	875
	5	3.0E-6	0	7.9E-4	3.4E-4	1.0	99.0	1233

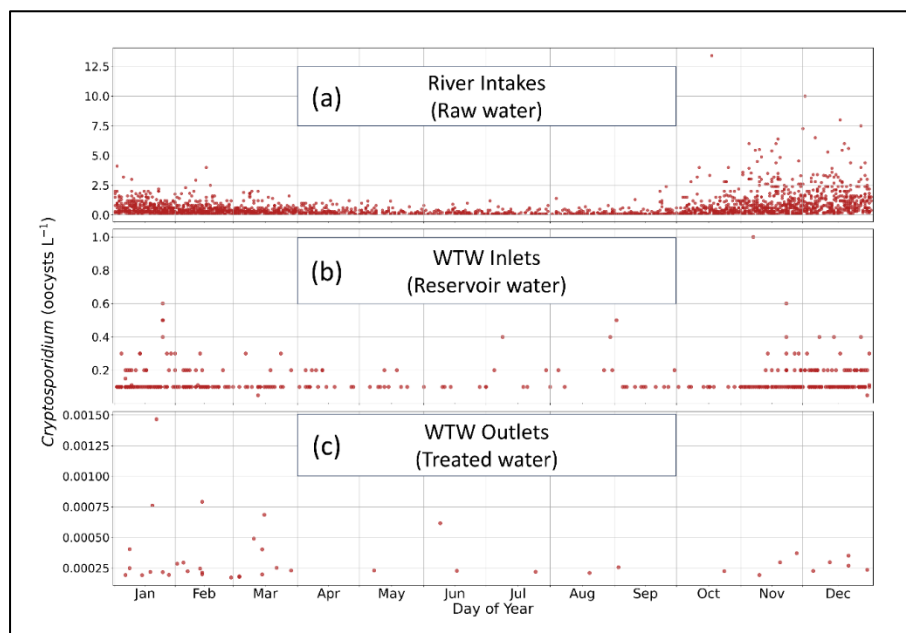


Fig. 3.2 *Cryptosporidium* detections for the 6-year period of study, plotted by day-of-year, showing (a) river intakes, (b) WTW inlets and (c) WTW outlets.

3.4.2 Volumetric data and proportional allocation

Cumulative intake volumes over the study period (Fig. 3.3a) show that the largest quantities of water were abstracted at Intakes A and (collectively) at Intakes D₁ and D₂. The lowest

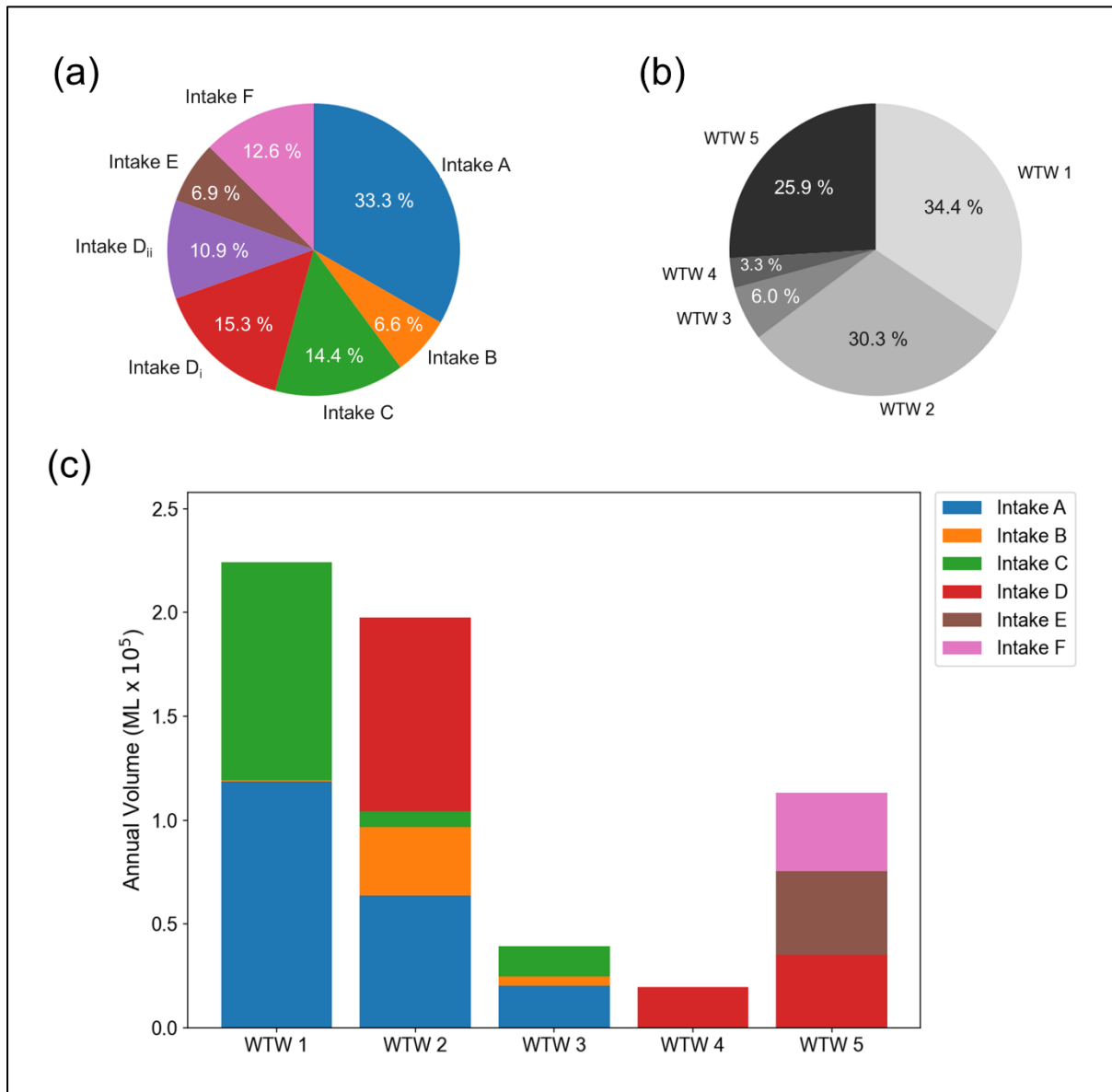


Fig. 3.3 Water volume data for the study period (01.03.2018 to 29.04.2024) showing (a) proportions of total volumes abstracted from intakes A-F, (b) proportions of total output volumes generated by WTWs 1-5 and (c) total intake volumes supplying individual WTWs (based on proportional allocation calculations).

abstraction volumes were taken at Intakes E and B, but collectively these still account for 13.5% of the total abstracted volume. Volumetric data for treated water (Fig. 3.3b) shows that WTWs 1, 2 and 5 are the largest works, collectively generating > 90 % of the output from London's large WTWs. Although WTWs 3 and 4 are much smaller, collectively these

provide 9.3 % of the total volume emerging from the five large WTWs serving London. Proportional allocation calculations reveal which intakes serve as the primary sources for the WTWs (Fig. 3.3c). Intake A is a major source for WTWs 1, 2 and 3, whereas Intake C is a major source for WTWs 1 and 3. Intake D is the exclusive source for WTW 4 and a major source for WTWs 2 and 5. Intakes B, E and F predominantly serve just one WTW. Note that the water supplied from very small works (constituting $\approx 10\%$ of the overall supply and largely based on groundwater sources) has been omitted from these proportional calculations and all subsequent risk evaluations.

3.4.3 Log-removal values

Although based on different paired datasets, Modelled LRV_{storage} distributions (Fig. 3.4a-e) show a consistent log removal effect due to reservoir storage, with mean values in the 0.83-0.95 range. This is slightly lower than the LRV_{storage} value of 1-log commonly deployed by Thames Water for modelling and risk assessment (Thames Water, 2025). Modelled $LRV_{\text{treatment}}$ distributions (Fig. 3.4f-j) are also very similar across the WTWs, with means in the 2.97-3.21 range. WTW 4 has the highest spread, with an interquartile range (Q_{25} - Q_{75}) of 0.65. For WTW 4, modelled $LRV_{\text{treatment}}$ means are approximately 0.6 log lower than mean literature values for Coag-Floc treatment with CoCo-DAFF (Table 3.1). For WTWs 1, 2, 3 and 5 the discrepancy is much greater, with modelled $LRV_{\text{treatment}}$ means ≈ 1.6 -1.7 lower than literature values (when potential RGF effects are excluded) and 2.7-2.8 lower when RGF is included.

3.4.4 QMRA outputs and selective abstraction modelling

Except where stated, results presented here will focus on the lowa DR QMRAs. TW-QMRA risk outputs (Fig. 3.5a) show no obvious summer-winter seasonality, and all but one of the distributions falls below the USEPA guideline value of 10^{-4} infections person⁻¹ yr⁻¹. (Note that although median and/or mean risk values have been used in the past when measuring against guideline values, the preferred and more conservative

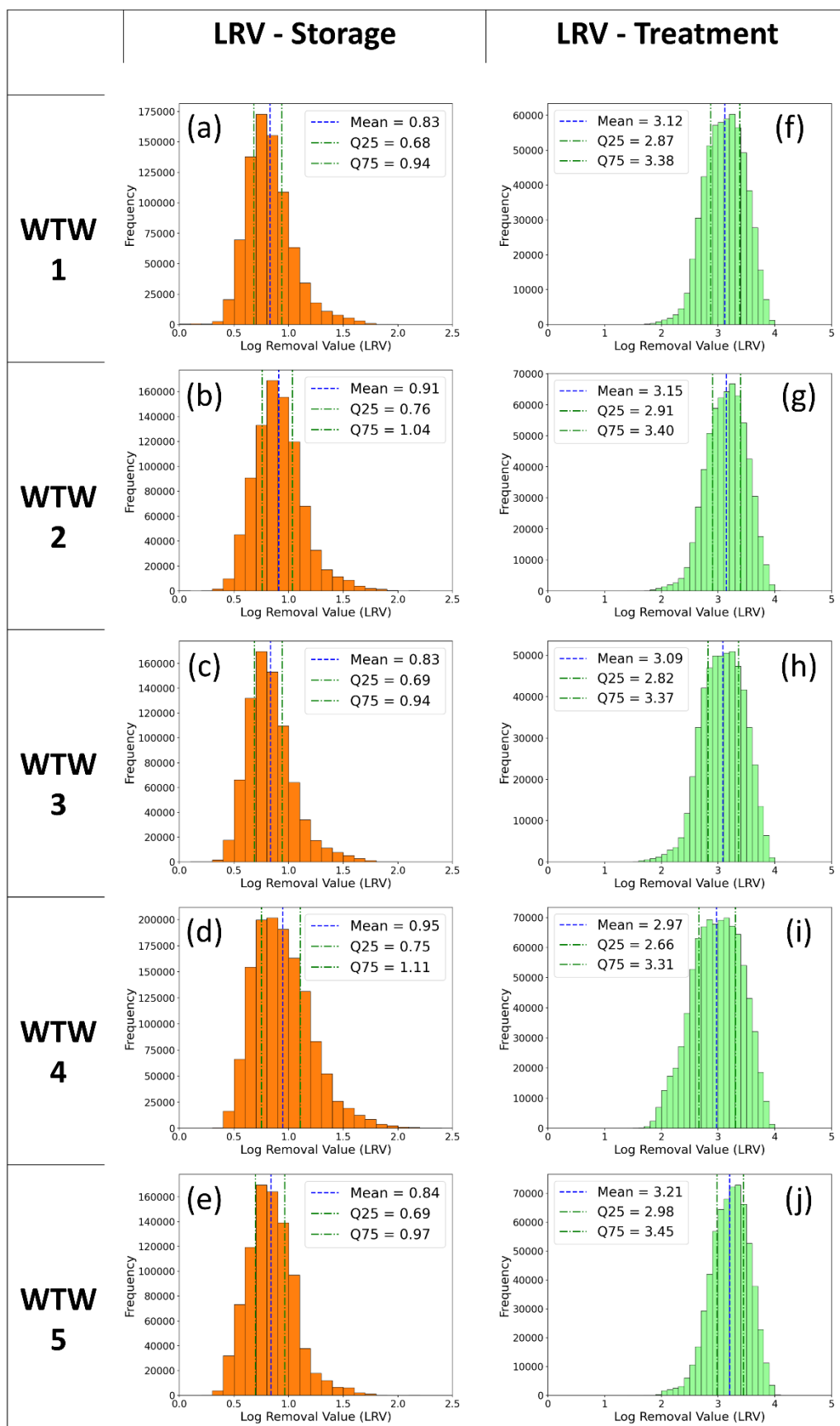


Fig. 3.4 Modelled distributions of (a-e) storage stage log-removal values (LRV_{storage}) and (f-j) treatment stage log-removal values ($LRV_{\text{treatment}}$).

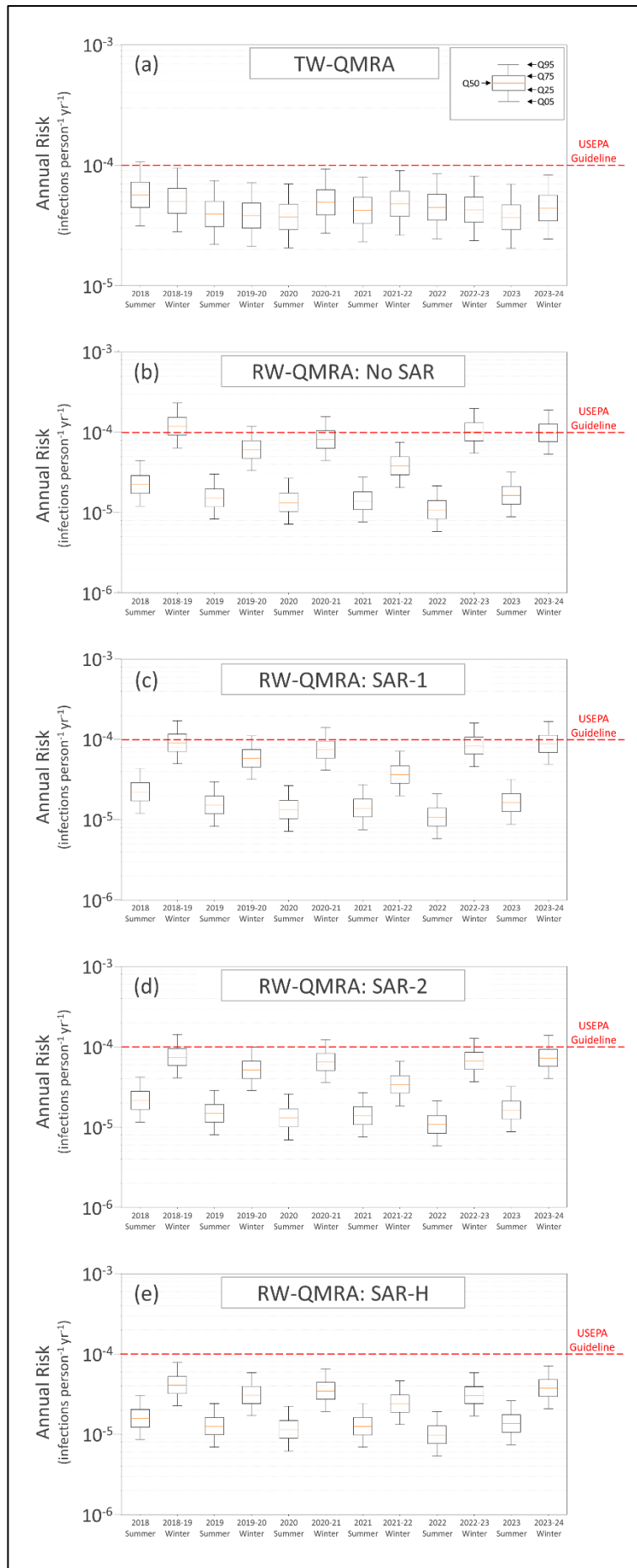


Fig. 3.5 Annual infection risk distribution outputs from Iowa DR QMRA models for biannual periods, based on *Cryptosporidium* concentrations in (a) treated water, and (b-e) river water with (b) no selective abstraction, (c) SAR-1, (d) SAR-2 and (e) SAR-H.

approach in recent years has been to use the 95th quantile (Q95) values (Schijven et al., 2011). Both median and Q95 values will therefore be considered in this study.) The period with the highest risk is 2018 Summer (underlining the lack of clear seasonality), which exceeds the USEPA guideline at Q95. The range within each distribution and the temporal variation between each biannual period (based on Q5-Q95 values) falls within 1 log unit, between 2×10^{-5} and 2×10^{-4} infections person⁻¹ yr⁻¹ (i.e. 0.002% to 0.02%).

RW-QMRA outputs without selective abstraction (Fig. 3.5b) show strong seasonality, with higher risk distributions for all Winter periods compared with Summer periods. In three periods, the Winter median falls above the USEPA guideline, and for all but one Winter period the Q95 values exceed the guideline. The range (Q5-Q95) of risk across all periods is greater than for the treated outputs, ranging from 5×10^{-6} to 3×10^{-4} infections person⁻¹ yr⁻¹ (0.0005% to 0.03%), although when grouped separately, the Q5-Q95 range is ≈ 1 log unit for the Winter periods and ≈ 1 log unit for the Summer. Summer periods are extremely consistent across the 6-year study period. Note also that lower Winter period distributions reflect periods of reduced *Cryptosporidium* river concentrations in the winter months of 2019-20 and 2021-22 (data not presented).

Annual risk statistics across the full six-year study period are also provided (Table 3.4), showing that at the median level both TW-QMRA and RW-QMRA fall well below the USEPA guideline, whereas Q95 for RW-QMRA (No SAR) falls very slightly above the guideline. This illustrates the innately conservative nature of biannual modelling (i.e. inclined to generate higher risk outputs for some periods) when compared to models which take in one or more

Table 3.4 Annualised infection risk values for QMRA outputs across the full study period (01.03.2018 to 29.02.2024). Highlighted value = above USEPA guideline.

Dose-Response (DR) Model →		Annual Infection Risk (infections person ⁻¹ yr ⁻¹)			
		Iowa DR		Pooled DR	
Model	Scenario	Median	Q95	Median	Q95
TW-QMRA	-	4.39×10^{-5}	8.39×10^{-5}	4.17×10^{-5}	1.12×10^{-3}
RW-QMRA	No SAR	4.79×10^{-5}	1.05×10^{-4}	4.43×10^{-5}	1.29×10^{-3}
	SAR-1	4.32×10^{-5}	8.83×10^{-5}	4.15×10^{-5}	1.12×10^{-3}
	SAR-2	3.73×10^{-5}	7.58×10^{-5}	3.59×10^{-5}	9.71×10^{-4}
	SAR-H	2.26×10^{-5}	4.42×10^{-5}	2.19×10^{-5}	5.99×10^{-4}

complete years. Applying the pooled DR QMRA resulted in similar median values across all models, but raised Q95 values by at least an order of magnitude. (Full outputs from the pooled DR QMRA are presented in Appendix B.2).

3.4.5 Sensitivity analysis

Sensitivity analysis (Fig. 3.6) reveals that the treated and river water models respond differently to the two types of concentration data, with TW-QMRA most sensitive to censored data and RW-QMRA most sensitive to the uncensored data. This pattern reflects

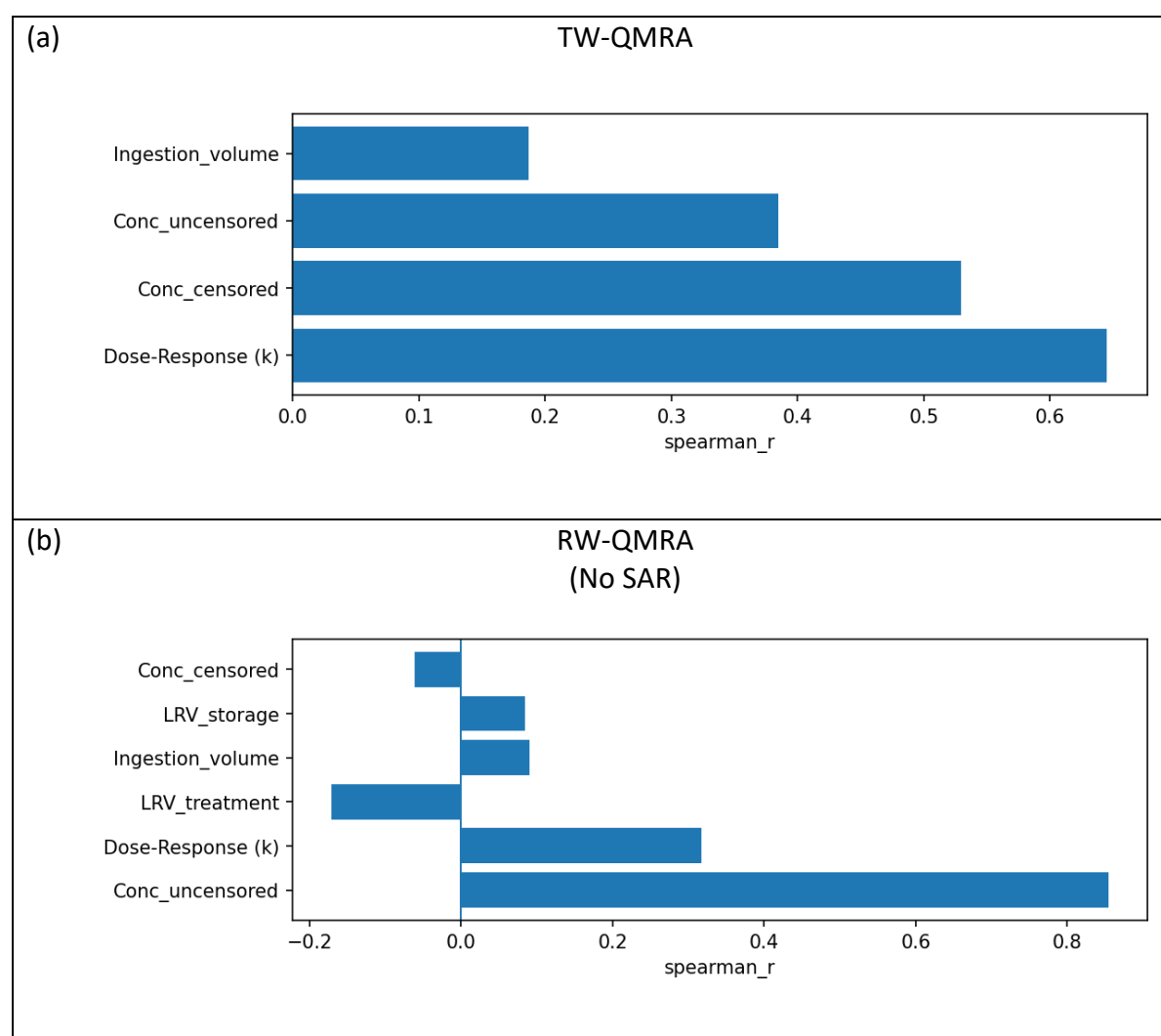


Fig. 3.6 Sensitivity analysis based on Spearman correlation coefficient (r), for (a) TW-QMRA (b) RW-QMRA (without SAR), using the Iowa DR model.

the relative abundance of each data type in the two models (i.e. for TW-QMRA, censored data >> uncensored data; for RW-QMRA, uncensored data > censored data). However, though the uncensored data is secondary in the TW-QMRA, it nevertheless has a disproportionately strong effect (given that there are only 40 detections in $\approx 4,400$ data points, across all WTW outlets). The TW-QMRA is most sensitive to dose-response (k) overall – this is most likely a consequence of the lower variance in the TW-QMRA concentration data.

RW-QMRA is also sensitive to dose-response, albeit to a lesser extent. $LRV_{treatment}$ has a greater (negative) impact on risk than $LRV_{storage}$. Somewhat surprisingly, $LRV_{storage}$ shows a positive correlation with risk outputs; this is most likely because the log-removal effects due to storage are being obscured by the larger log-removal effects of treatment. Both TW-QMRA and RW-QMRA are only moderately sensitive to ingestion volume.

3.4.6 Selective abstraction modelling: An example application of the model

Additional RW-QMRA models (for different SARs) are provided to demonstrate an application of the model which is of direct relevance for understanding both current and future risk levels. SAR-1 and SAR-2 – which represent current abstraction regimes employed by Thames Water – both result in decreased risk (relative to the baseline “No SAR” model), reducing Winter period medians to less than the USEPA guideline (Fig. 3.5c-d). Q95 values for SAR-1 and SAR-2 remain above or on the guideline level for the higher Winter periods, however. The overall reduction in Winter median and Q95 risk values achieved by selective abstraction is, however, modest at <0.3 log units.

When modelling SAR-H – a much more stringent hypothetical strategy – all Q95 values are brought within the USEPA recommendation (Fig. 3.5e). Median and Q95 values in Winter seasons are reduced by ≈ 0.4 - 0.6 log units. As would be expected, given the low *Cryptosporidium* concentrations found in most Mar-Sep periods, the Summer risk distributions show almost no change in response to abstraction regime.

Fig. 3.7 shows the effect of SAR risk reduction in the context of water abstraction volume reduction, modelled for the entire study period (with exact risk values shown in Table 3.4). Median annual infection risk drops from 4.8×10^{-5} for No SAR to 3.7×10^{-5} infections person⁻¹ yr⁻¹ for SAR-2 – equivalent to 1 fewer person in 100,000 becoming infected per year – but

this comes at the cost of an abstraction reduction of 1.6%. Use of the hypothetical SAR-H reduces median risk to 2.3×10^{-5} infections person⁻¹ yr⁻¹, in return for ≈6% volume reduction. It is notable however that more dramatic risk reductions are observed if the more conservative Q95 risk statistic is used instead of the median – a more conservative risk mitigation strategy may therefore lead to the acceptance of greater reductions in abstraction volume.

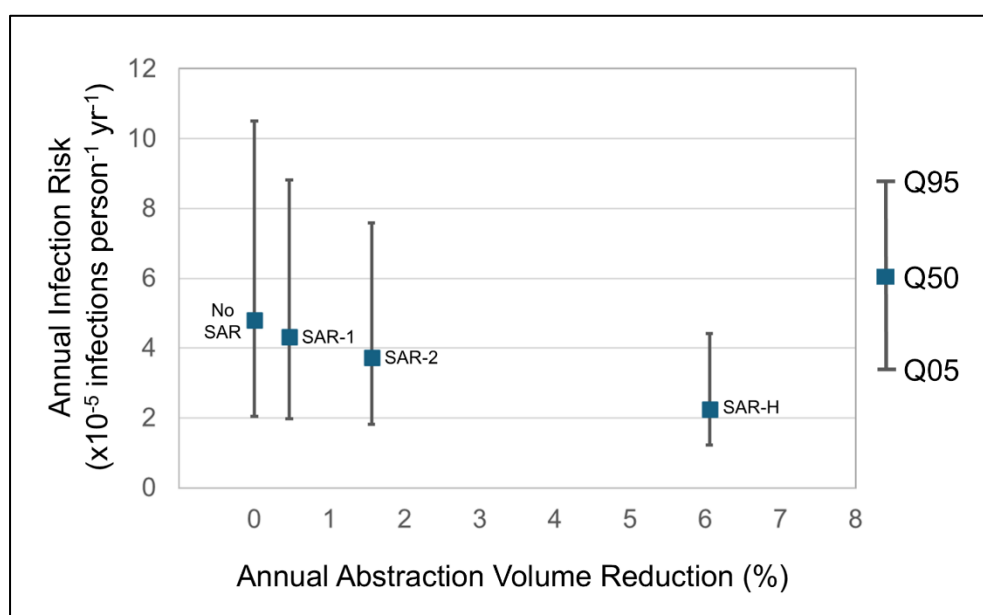


Fig. 3.7 Annual infection risk (for all biannual periods combined, 2018-2024) plotted against annual abstraction volume reduction for different SARs. (For Iowa DR QMRAs only and idealised SAR implementation, with smoothed mean daily abstraction volumes for 2018-2024.)

3.5 Discussion

3.5.1 Overall infection risk

3.5.1.1 Treated water QMRA (TW-QMRA)

TW-QMRA outputs show consistent median risk levels below the USEPA guideline of 10^{-4} infections person⁻¹ yr⁻¹ (0.01%) (Regli et al., 1991); this indicates that current risk management practices and water treatment processes are effective at reducing the risk of *Cryptosporidium* infection to tolerable levels. The HMI modelling method adopted here is appropriate given the preponderance of left-censored data and, in the case of some

biannual periods, data subsets which consist exclusively of left-censored values, and which are therefore unamenable to standard MLE modelling.

The annual increases in *Cryptosporidium* concentrations found in Oct-Feb in both the raw (river intake) and reservoir (WTW inlet) data did not result in pronounced seasonality in the TW-QMRA outputs. This could be taken as evidence that the combination of reservoir storage and WTW processes are sufficiently robust to deal with the increased loadings of oocysts in the cold season. However, it could also indicate a weakness in the TW-QMRA method, given that a slight increase in frequency of detections was observed in the treated water in Jan-Mar (Fig. 3.2c). The fact that this overlaps with both Summer and Winter seasons may be the reason why these detections are not reflected in the biannual risk outputs; in which case, it is arguable that the treated water's slightly delayed seasonality would be better represented by Winter seasons defined as Oct-Mar, instead of Sep-Feb. An explanation for the lagged seasonality in treated water detections might be the negative effect of low temperatures on SSF biofilms (Schmutzdecke), which have been found to be pronounced when temperatures drop below $\approx 8^{\circ}\text{C}$ (Maiyo et al., 2023). Given that SSFs are the primary mechanism of oocyst removal for 96.3% of the water considered in this study (Table 3.1 and Fig. 3.3), temperature-related treatment performance is a critical consideration for protecting London's drinking water supply.

Absence of seasonality in TW-QMRA outputs could also be due to the dominance of left-censored concentration data over the uncensored data, but the high sensitivity of the TW-QMRA risk outputs to uncensored data (Fig. 3.6) shows that detections are significantly affecting the outputs (i.e. they are not completely masked by the non-detects).

3.5.1.2 Raw water QMRA (RW-QMRA)

RW-QMRA model outputs show a strong seasonal effect, with median risk for three "Winter" periods exceeding the USEPA tolerable risk guideline, whereas all "Summer" period median and Q95 values were within the tolerable risk band. The RW-QMRA model is based on more assumptions than the TW-QMRA model, and it requires more data processing steps. The assumptions relate to the proportional allocation of river intake water to WTWs and log-removal rates at both the storage and treatment stages. For reasons discussed below (Section 3.5.4), $\text{LRV}_{\text{treatment}}$ outputs (Fig. 3.4f-j) appear conservative

when compared with LRCs taken from the literature (Table 3.1), and will therefore generate correspondingly higher risk outputs.

The higher number of uncensored concentration values in the river intake data means that the RW-QMRA is better able to represent variations in seasonal risk – variations which will never be fully discernible in the TW-QMRA unless the system frequently fails (i.e. producing more detections). However, the reduction in risk achieved by the application of SAR-1 and SAR-2 showed that the current abstraction strategy could be playing a role in bringing all median values to tolerable risk levels (although Q95 values still remain above the threshold).

Biannual modelling (as opposed to annual or multiple-year modelling) is inherently conservative – i.e. inclined to generate higher risk outputs for certain periods – because the modelled concentration distributions during the higher-risk months are not merged with concentration data from the lower-risk months. When risk is assessed across a full year, median risk for the city-wide treatment system falls well within tolerable levels, irrespective of modelling method applied (Table 3.4), and Q95 values are also below the guideline when currently active abstraction scenarios are applied.

The models do not consider analytical recovery, which, based on previous studies using the same or closely comparable methods, has been shown to vary between 6–89% for the raw-water data and 16–60% for the treated-water data (Environment Agency, 2010). This could lead to significant underestimates in both RW-QMRA and TW-QMRA risk outputs, though these may be balanced in part by the inclusion of non-infectious oocysts in the concentration data – i.e. oocysts which are non-viable, or which belong to species which are not regarded as pathogenic to humans (Medema et al., 2009; Ryan et al., 2021). The inclusion of both recovery rates and estimates of the viable and infectious fractions is recommended in QMRA (Owens et al., 2020). Where possible, recovery data should be sample-specific because recovery can vary widely in response to water conditions (Schijven et al., 2011), but non-sample-specific recovery data can still be incorporated probabilistically where suitable study-specific data is available; for example, Daniels et al. (2018) adjusted exposure estimates by applying modelled recovery distributions derived from spiking trials and literature-informed viability and infectious-fraction assumptions. It should be noted, however, that attempting to model these correction factors without adequate data risks introducing additional bias (Hamilton et al., 2018).

3.5.1.3 Interpreting parallel QMRA outputs

Taken together, the RW-QMRA and TW-QMRA outputs provide complementary estimates of *Cryptosporidium* risk in the final blended treated water. Both models estimate risk at the same endpoint and share common ingestion and dose-response components, but they reconstruct the exposure concentration using different input data and model design. The RW-QMRA is driven by raw water concentration data and therefore retains the pronounced seasonal variation in *Cryptosporidium* loadings observed at the river intakes, with proportional allocation and aggregated storage/treatment LRV distributions used to translate those loadings to the final treated supply. Importantly, the seasonal pattern in the RW-QMRA outputs is not a consequence of applying season-specific LRVs: modelled LRV distributions were generated using data from the full six-year study period to avoid circularity, an approach which would be expected, if anything, to dampen rather than amplify seasonal differences in risk. By contrast, the TW-QMRA uses WTW outlet concentration data directly, meaning that its outputs are more strongly shaped by the high proportion of left-censored treated water measurements and by the small number of detections. Differences between the two sets of outputs are therefore not simply a contrast between “raw water” and “treated water” QMRA, but arise from the interaction between substantively different concentration datasets and the modelling assumptions required to estimate risk in the final blended treated water.

3.5.2 Effectiveness of selective abstraction

In the RW-QMRA model presented here, the quantifiable seasonal risk reduction achieved through selective abstraction is modest. The SAR-2 strategy delivers a modelled reduction of <0.3 log units in return for a relatively small ($\approx 1.6\%$) reduction in mean annual abstraction volume. Greater seasonal risk reductions (0.4-0.6 log units) can be achieved using the more stringent SAR-H strategy, but only at the cost of a substantially larger ($\approx 6\%$) reduction in mean annual abstraction volume – a level of supply loss that may be operationally and politically difficult to justify in the water-stressed south-east of the UK.

It should also be noted that the examples presented here represent “idealised” abstraction management strategies in which abstraction is always reduced or suspended sufficiently early to prevent elevated *Cryptosporidium* loadings entering reservoirs once river

concentrations exceed threshold values. In practice, such operational perfection is unlikely to be achievable because of monitoring uncertainty, forecasting limitations, response delays and competing supply demands. Consequently, real-world implementation of any SAR is likely to yield smaller reductions in infection risk, while simultaneously incurring greater losses in deployable output than those shown in Fig. 3.7.

Importantly, the reductions in infection risk achieved through selective abstraction are also small relative to the ≈ 3 -log reduction already achieved by existing treatment processes (Fig. 3.4). More substantial reductions in risk could instead be achieved through the addition of further treatment barriers. For example, UV disinfection can typically provide 2-4 log inactivation of *Cryptosporidium*. A 2-log UV treatment barrier would reduce the seasonal Iowa DR RW-QMRA median outputs to approximately 10^{-7} to 10^{-6} infections person⁻¹ year⁻¹ – at least two orders of magnitude below the USEPA guideline – while avoiding any reduction in water supply volume. The installation of UV contactors at the major London WTWs is currently in the planning stages (Thames Water, 2023). Although implementation at this scale will involve considerable capital and operational expenditure, the model outputs presented here demonstrate the potential public health benefits of this investment. Selective abstraction remains, however, a useful risk reduction method which can be continually practised without fresh investment and at significantly lower running costs. (Note that the largest associated costs are staff time for sampling and analysis of grab samples.)

3.5.3 Application of QMRA to complex systems

The QMRA models described here provide an example of how the complexities in a major water supply system can be modelled using a proportional allocation method to represent mixing of multiple sources of raw water and storage reservoirs. Note that this complexity resides in the RW-QMRA model – proportional allocation is not required for the TW-QMRA model, although the weighting of risk values, using relative WTW output volumes, is required in both models to account for the final blending of treated water. Applying the same method for other sites requires a substantial hydrological dataset in which all flows to and from assets are routinely recorded (and ideally at daily resolution, to reflect daily operational decision-making). Where such data is not collected at all locations, estimates

can be applied instead, such as standard operational pumping rates and pumping times, but the final model will be subject to greater error.

3.5.4 Analysis of log-removal modelling

The high consistency in the log-removal distributions (Fig. 3.4) suggests that performance across the system is relatively uniform, with no obviously failing or vulnerable assets. Although sites vary in their frequency of detections (with WTW 3 having the most detections, at 2%), this is not reflected in the LRV distributions. Another notable trend in the LRV data is the consistent skewing of the distributions. The right skew seen in the $LRV_{storage}$ distributions (Fig. 3.4a-e) can be explained by the potential for occasional high log-removal effects during extended reservoir retention times, whereas the very slight left-skew in the $LRV_{treatment}$ distributions (Fig. 3.4f-j) could be reflective of occasional dips in treatment performance (von Sperling, Verbyla and Oliveira, 2020).

Mean values for $LRV_{treatment}$ are substantially lower than their equivalents in the literature, particularly for the SSF-based sites (see Table 3.1), but the modelled distributions do overlap with previously published ranges, and SSF performance is expected to be in the lower range in temperate climates (Maiyo et al., 2023). The use of a parametrically-defined lower limit for censored data imputation in the log removal model may also have led to overly conservative LRVs, since subtle changes in both upstream and downstream concentrations can lead to substantial differences in final LRV values (von Sperling, Verbyla and Oliveira, 2020). The method of pairing upstream and downstream data in the intake → reservoir(s) → WTW pathway can also have a dramatic effect on LRV estimates. Several pairing methods were tested in initial trials, including date-based pairing (with site-specific lag time distributions) and simple rank-based pairing. Quartile-ranking – the method presented here – resulted in the most realistic outputs (i.e. the absence of unrealistic negative log-removal rates, which are sometimes generated by other pairing methods), and it also produced the closest match between RW-QMRA and TW-QMRA outputs, which is one method of validating log-removal modelling/pairing methods (Smeets et al., 2008).

It should also be noted that averaging and combining literature-based LRCs for multiple treatment processes can result in overestimation of log-removal rates (Schmidt, Anderson and Emelko, 2020). While the use of literature-based LRCs is a viable option, the wide

variety in specific treatment conditions and source water characteristics makes site-specific LRVs the preferred choice where suitable data is available (von Sperling, Verbyla and Oliveira, 2020). In the RW-QMRA model, the LRV distributions fulfil their intended function of providing approximate log-removal values for the individual assets, while also representing uncertainty and natural variability. An obvious modification which could be made would be to model LRVs by season, potentially capturing the effect of both seasonally-different loadings and treatment performance (e.g. reduced SSF performance in colder months); this option was not taken however because of the risk of circularity (i.e. applying LRV distributions to exactly the same concentration datasets from which they were generated) and because trials showed that LRV calculations were highly sensitive to seasonal loadings, leading to exaggerated LRVs for Winter and low LRVs in Summer. Aggregating LRVs across all time periods is an effective way of removing these effects.

3.5.5 Key determinants of risk

Both models are sensitive to concentration data inputs (Fig. 3.6), which is desirable from a modelling perspective, since these are the inputs for which we have an abundance of “real” and site-specific monitoring data (albeit much of it left-censored). By contrast, the high sensitivity of both TW-QMRA and RW-QMRA to dose-response is less ideal, because of the uncertainties inherent in this model input. This issue is made even more apparent by the order-of-magnitude increases in Q95 values which are observed when the pooled DR model is used instead of the Iowa DR model (Table 3.4); this illustrates the potential for more infectious species/genotypes and strains to raise infection risk well above the “tolerable” levels. The pooled DR QMRA is still derived from only a small number of *C. parvum* isolates (Iowa, TAMU and UCP), meaning that it cannot fully capture the likely diversity of infectivity present among environmental *Cryptosporidium* populations; however, this is an inherent limitation of *Cryptosporidium* QMRA studies generally, owing to the extremely limited number of experimentally derived human dose-response datasets currently available.

Determining which *Cryptosporidium* isolate data is most appropriate for QMRA is problematic, since in almost all cases risk assessors will have little or no information concerning the relative infectivity of detected oocysts at their sites. What’s more, existing dose-response datasets were gathered from infection challenge studies of healthy human

volunteers, meaning that there is a notable lack of controlled data for very young, elderly or immunocompromised subjects, and no controlled data showing the effect of ingesting very low (<10) numbers of oocysts, which are more appropriate to exposures via drinking water following treatment/network failures (Messner and Berger, 2016). Epidemiological data collected during outbreaks offers another means of generating dose-response models (Eisenberg et al., 1998), but this is not currently the preferred option for *Cryptosporidium* QMRA studies (Medema et al., 2009). Consequently, uncertainty associated with dose-response selection and low-dose extrapolation remains one of the dominant structural uncertainties within *Cryptosporidium* QMRA (Hamilton et al., 2024).

In previous studies, ingestion volume has been identified as more impactful than dose-response (Owens et al., 2020), but the opposite trend was found in this study. This may be due to the nature of the ingestion model used here, which has a lower mean than those used in many published studies. The low mean may be a consequence of greater hot drink consumption in the UK (i.e. boiled tap water, as opposed to unboiled tap water represented by the ingestion model), especially when compared to the warmer locations considered in some previous studies. Nevertheless, more up-to-date consumption data would be preferable, given the changes in drinking habits over the previous decade (e.g. an increased trend for using refillable water bottles and greater access to drinking water in schools, workplaces and community centres), but this concern is mitigated somewhat by the moderate sensitivity of both models to ingestion volume.

An important limitation of the study is that the models do not consider contamination risks arising within the drinking water distribution system (DWDS). Such risks are difficult to quantify within QMRA frameworks because they are often associated with sporadic infrastructure failures, ingress events or operational incidents rather than routine treatment performance and monitored source water quality. Nevertheless, major *Cryptosporidium* incidents in the UK have in some cases originated within the DWDS, rather than in source waters or treatment works; for example, the 2024 Brixham outbreak was linked to contamination entering the network through a flooded air valve chamber (BBC, 2024; DWI, 2025). Consequently, the risk estimates presented here should be interpreted as risks associated with treated water production and WTW outlet quality, rather than total end-consumer risk across the full drinking water supply and distribution system.

3.5.6 Example applications of RW-QMRA: SARs and other risk mitigation strategies

By modelling different abstraction-based risk mitigation methods alongside abstraction volume reductions, the optimum combination of permissible risk and permissible reduction can be identified (Fig. 3.7). The analysis demonstrates that SARs are potentially effective in reducing risk, though the cost in terms of reduced abstraction volume becomes increasingly severe as the thresholds are lowered. Water resource pressures are expected to become more acute for SE England, therefore understanding this trade-off will become increasingly important.

Other uses of the RW-QMRA model could include estimating risk reductions resulting from additional treatment measures – e.g. the introduction of UV contactors to deactivate oocysts. Modelled reductions in risk can be tied directly to installation and operating cost estimates as part of a cost-benefit analysis, leading to more informed and targeted allocation of funds so as to maximise the positive the impact of investment in public health. The QMRA-based approach is not widely adopted in the UK, and in the case of *Cryptosporidium*, water companies usually adhere to concentration-based limits, which are commonly 1.5 to 2 orders-of-magnitude below the analytical LOD (e.g. the Thames Water (2023) limit is 10^{-5} oocysts L^{-1}). Hence QMRA-based approaches could introduce more nuance into risk discussions and decision-making, rather than appraising risk in binary terms, which focus on detection (i.e. failure) vs absence (i.e. perceived pass).

3.6 Conclusions

This study has demonstrated the application of parallel raw water and treated water QMRA within a complex multi-source drinking water system, integrating data from multiple stages of abstraction, storage and treatment to evaluate *Cryptosporidium* risk in the final treated water supply. Observed variations between the different QMRA outputs reflect the different areas of uncertainty in the two models – specifically, the log-removal modelling in RW-QMRA, and the heavy reliance on left-censored data in TW-QMRA. Although TW-QMRA is based on ingested water concentrations, high LODs (relative to the unachievably low LOD required to confirm tolerable risk) mean that left-censored modelling assumptions will always be critical where detections are infrequent. TW-QMRA is also unsuitable for exploring changes in risk in response to differing environmental conditions, pathogen

loadings or risk management strategies. RW-QMRA can model these changing risks effectively because it considers the full water treatment pathway, from intake to WTW output. Such approaches are more in line with recent legislation, advising water providers to adopt a “source-to-tap” approach to risk management, which directly incorporates risk imposed by catchment conditions (European Union, 2020).

Both TW-QMRA and RW-QMRA found that the current overall risk to London’s drinking water from *Cryptosporidium* is within the tolerable risk levels defined by the USEPA (Regli et al., 1991). RW-QMRA shows there is a seasonal increase in risk due to high winter river concentrations of the pathogen. By incorporating different abstraction scenarios into the model, risk variation can be quantified alongside estimated reductions in abstraction volume. Selective abstraction results in a <0.3-log reduction in risk, which is modest compared with the 2-4 log reductions (based on potential oocyst inactivation rates) that could be achieved by installing UV treatment at the major London WTWs – work which is currently planned for 2025-2035.

QMRA is not the preferred approach for risk assessment in the UK water sector, but this study can be seen as a test case for where it could be adopted to assess risk beyond individual treatment works to encompass complex water supply networks. Severe water scarcity has been projected for the South East of England by 2050, due to increased population, increased development and climate change (Environment Agency, 2025), whereas climate-driven increases in the frequency of intense rainfall events (Chan et al., 2023) are expected to generate greater wash-off of animal waste from fields, potentially leading to higher pathogen loadings to surface waters. Such conditions – which are also projected for many other parts of the world – inevitably increase the tension between achieving water quality targets and meeting water demand (Jones et al., 2024). The QMRA-based analysis demonstrated here can provide a method of addressing these competing priorities by extracting value from extensive monitoring data held in water company databases and applying it to better understand existing and future risk, under a range of management scenarios.

Chapter 4: Environmental factors and sources influencing *Cryptosporidium* river concentrations: A data-rich spatio-temporal analysis using Explainable AI

Reproduced from draft journal paper: Smalley, A. L., Douterelo, I., Chipps, M. and Shucksmith, J. D. (2026) Environmental factors and sources influencing *Cryptosporidium* river concentrations: A data-rich spatio-temporal analysis using Explainable AI. *Intended journal: Water Research*.

4.1 Abstract

Cryptosporidium is a waterborne parasite which poses significant risks to drinking water safety. Despite its importance, considerable uncertainty remains regarding the effect of environmental factors on the scale and timing of *Cryptosporidium* concentrations in source waters. This paper performs a spatio-temporal analysis of the environmental factors influencing *Cryptosporidium* river concentrations by applying eXplainable AI (XAI) tools to a large-scale, long-term nationwide dataset. The study brings together extensive land use, topographical, hydro-meteorological and human/animal density data to analyse the seasonality and magnitude of *Cryptosporidium* in rivers and advance understanding of how static catchment characteristics interact with dynamic environmental conditions to influence *Cryptosporidium* concentrations across heterogeneous river systems.

A machine learning (ML) model – XGBoost – was trained with environmental inputs at different spatial scales. Interpretation of feature effects using SHapley Additive exPlanations (SHAP) found that humidity and rainfall had a strong positive effect on *Cryptosporidium* concentrations. Human population density was shown to be more highly predictive of *Cryptosporidium* concentrations than either grassland (grazing) areas or livestock density, indicating that human sources of *Cryptosporidium* are more important than animal sources in many UK rivers. Density of combined sewer overflows (CSOs) and sewage treatment works (STW) were found to be more effective predictors at the local (≤ 25 km upstream) scale, pointing to the possible impact of local point sources. The study revealed strong seasonality in *Cryptosporidium* at many sites, with the timing of peak concentrations varying significantly with catchment characteristics, such as human population, STW density, urban area and prevailing meteorological conditions. The principal

advance of this work lies in demonstrating how an XAI-based modelling framework can combine static, pseudo-static and dynamic environmental datasets across multiple catchments to identify drivers of pathogen occurrence and seasonality in river systems, yielding more generalisable insights than previous single-catchment or short-term studies.

4.2 Introduction

Cryptosporidium is a pathogen capable of causing severe gastro-intestinal illness in humans and animals. It exists in the environment in its transmissible “oocysts” form, large numbers of which are shed in the faeces of infected hosts (Rose et al., 2002). Once in the environment, oocysts can be transported to surface waters and groundwaters, potentially contaminating vital sources for drinking water. *Cryptosporidium* poses a major drinking water treatment challenge because of its resistance to chlorine disinfection, its infectivity at very low doses, and long (1-2 day) laboratory analysis times, which prevent real-time monitoring (Okhuysen et al., 1999; Betancourt and Rose, 2004; Hassan et al., 2021). Contamination of drinking waters can lead directly to outbreaks, threatening public health and undermining confidence in drinking water safety (Efstratiou et al., 2017a), with the worst cases often making the news, as seen in the recent events in the South West of the UK (Horton, 2025).

Grazing livestock (Sturdee et al., 2007; Bhattarai et al., 2011) and the application of manure-based fertilisers (Swaffer et al., 2014) are generally regarded as the most significant animal sources, with mobilisation from fields to water courses usually driven by rainfall events, leading to transport of oocysts in rainfall-runoff (Atherholt et al., 1998). Most human sources of *Cryptosporidium* enter rivers either in the regular daily discharges from sewage treatment works (STWs) (Bukhari et al., 1997), or in irregular spills from combined sewer overflows (CSOs) (Gibson III et al., 1998), with the latter generally occurring in the wake of rainfall events, or due to storm drainage/sewer network issues. Other factors which have been considered to influence concentration levels in river systems include temperature (King et al., 2005), sunlight hours (Dorner et al., 2006), soil moisture (Samadder et al., 2010) and river discharge (Swaffer et al., 2014).

Cryptosporidium infections (in humans and animals) and detections in the raw waters are often highly seasonal, although seasonality has been shown to vary widely between

countries and climate types (Ikiroma and Pollock, 2021; Wang et al., 2023). Such seasonality is likely the consequence of multiple factors, including extended oocysts survival times in the wetter, cooler conditions, increased rainfall-runoff volumes over moist and/or saturated soils and increased loadings from human and/or animal sources during seasonal infection peaks (Wang et al., 2023).

A greater understanding of the relative importance of different *Cryptosporidium* sources and mobilisation factors under differing environmental conditions would provide vital information to water supply operatives and could feed into the development of new predictive models and catchment management strategies. Previous research on *Cryptosporidium* in raw waters has generally focussed on single or adjoining catchments (Sturdee et al., 2007), with a notable absence of any study comparing multiple catchments to explore the influence of environmental variables. Previous catchments studies are also often based on small datasets (e.g. <20 water sample measurements) over relatively brief (e.g. 1-2 years) timescales and/or infrequent sampling (Robinson et al., 2011; Wells et al., 2015); they are therefore ill-suited for discerning long-term seasonal and/or annual trends. Globally-based *Cryptosporidium* loadings and river concentrations have been estimated using models based on land use, population data, STW removal efficiencies, rivers flows and weather data (Vermeulen et al., 2019), but these studies were not designed to represent or interpret important temporal (i.e. seasonal) trends and influences. To date, no national or international studies exist which are based on consistent *Cryptosporidium* sampling protocols and consistent analytical procedures over long durations and at regular (e.g. sub-monthly or sub-weekly) sampling frequency. Of particular need is a detailed exploration of the intertwined influence of static or pseudo-static catchment characteristics (e.g. slope or population density), and dynamic factors, such as rainfall, temperature and discharge, on observed *Cryptosporidium* concentrations.

In the UK, source waters at high risk of *Cryptosporidium* contamination are routinely monitored as part of standard risk management practice. A number of sites have been monitored in this way for 20+ years, but the data is fragmented between databases managed by multiple water companies. Collectively, these datasets have considerable potential for investigating spatial and temporal patterns in *Cryptosporidium* source water concentrations, which could inform understanding of related drinking water risks (Smeets et

al., 2007). In addition to this, the combined explosion in freely available AI-based modelling tools and remote sensing data has opened new avenues for data-driven research (Zhu et al., 2023). The interrogation of machine learning (ML) models using eXplainable AI (XAI) tools is an increasingly common practice, which has been deployed in a number of river catchment studies. Wu et al. (2021b) used several ML models (including XGBoost – a tree-based ensemble model) to track sources of river pathogens in a watershed in California, while Zhang et al. (2022) applied a different tree-based model (Random Forest) along with SHapley Additive exPlanations (SHAP) to connect land use with river water quality (e.g. nutrients) in the Potomac River Basin. Both studies used catchment-specific meteorological/environmental data in their models and applied XAI methods to analyse for non-linear/interactive trends that would not have been discernible using conventional statistical analysis.

The aim of this study is to develop an integrated spatio-temporal analytical framework for understanding environmental controls on *Cryptosporidium* concentrations in rivers at a national scale. To achieve this, a multi-catchment *Cryptosporidium* monitoring dataset covering the entirety of the British mainland is analysed alongside static, pseudo-static and dynamic environmental datasets using XAI methods. Specific objectives are to: (i) provide an overview of how *Cryptosporidium* river concentrations vary spatially and temporally (e.g. peak timing and strength of seasonality) on the British mainland; (ii) identify the most important environmental factors influencing *Cryptosporidium* river concentrations over time and space; and (iii) determine the relative importance of animal and human sources in the study catchment. An XAI approach will be adopted to meet these objectives, making use of freely available modelling software and open-source environmental data. Taken as a whole, this work advances understanding of how static catchment characteristics interact with dynamic hydro-meteorological conditions to influence *Cryptosporidium* concentrations across heterogeneous river systems. The study also demonstrates the utility of XAI as a transferable framework for analysing long-term pathogen monitoring data across multiple catchments.

4.3 Method

In carrying out this research, the following steps were taken: (i) initial *Cryptosporidium* data acquisition and cleaning to compile the study dataset; (ii) calculation of summary statistics for spatial and temporal analysis; (iii) geographical processing of environmental data; (iv) development of a multi-catchment *Cryptosporidium* ML model; and (v) coupling the ML model with XAI to analyse the potential impact of environmental factors on *Cryptosporidium* concentrations. All coding was performed in Python (v. 3.13.5), using the Anaconda distribution (v. 6.0.7), and all geographical information systems (GIS) analysis was performed in QGIS (v. 3.40.10-Bratislava).

4.3.1 *Cryptosporidium* data acquisition and cleaning

Environmental Information Regulation (EIR) requests for *Cryptosporidium* concentration data in untreated source (i.e. raw) waters were submitted to the major water companies in the UK, all of which responded positively, providing full data retrievals. The resulting dataset comprised approximately 210,000 *Cryptosporidium* measurements. All samples were analysed in accredited laboratories using the standard Blue Book method employed by all water companies in the UK (Environment Agency, 2010). Recovery efficiencies for this method are typically $\leq 60\%$ (Environment Agency, 2010); results should therefore be considered underestimates of true concentrations, although oocysts counts are likely to include non-viable oocysts or species/genotypes which are not infectious to humans (Daniels et al., 2018; Ryan et al., 2021). In the absence of sample-specific recovery data, we have excluded recovery considerations from the analysis of *Cryptosporidium* datasets, in line with the approach taken by previous studies (e.g. Ligda et al., 2020; Guzmán Barragán et al., 2022; Ligda et al., 2024).

Frequency and periods of measurement varied widely across sampling sites, with >4000 data points spanning the years 2000-2024 available for some sites, whereas other sites only had a single measurement. The type of sampling sites also varied widely, with some water company datasets dominated by either river abstraction points, impoundment reservoirs, bankside reservoirs, or boreholes, reflecting the prevailing source waters across the different regions.

The dataset was filtered to consider only river sites (the focus of this study). Measurements taken outside of the 24-year study period (01.01.2000 to 31.12.2023) were removed (note that this is the period for which gridded meteorological data was available – see next section). River sites were excluded if they contained fewer than 100 data points in total or fewer than 2 years of sampling data. 52 sites satisfied these criteria (constituting 52,181 *Cryptosporidium* measurements), all of which were at abstraction sites – i.e. intakes where water is transferred (by gravity or by pumping) to a reservoir or water treatment works (WTW). For security reasons, the exact position of abstraction sites is rarely disclosed by water companies, therefore sampling sites had to be inferred from the site names/descriptions. Final site coordinates were confirmed by identifying the associated abstraction point on Google Maps (satellite view), usually by observation of an intake channel, eel screen, adjacent WTW or reservoir.

4.3.2 Summary statistics for spatial and temporal analysis

Summary statistics, based on long-term (≤ 24 years) *Cryptosporidium* data, were calculated for all sites, with left-censored values (i.e. values below the detection limit) treated as zeros. Mean monthly *Cryptosporidium* concentration ($\bar{C}_{monthly}$) aggregated across all sampling years was calculated for each site, along with seasonal amplitude (A_s), a commonly used metric for seasonal change (Zhao and Zeng, 2014), calculated as shown in Equation 14:

$$A_s = \max(\bar{C}_{monthly}) - \min(\bar{C}_{monthly}) \quad (14)$$

Sites were grouped by the timing of their *Cryptosporidium* seasonal peak and differences in environmental variables between these groups were assessed using one-way ANOVA with Tukey's Honestly Significant Difference (HSD) post-hoc tests (Zuur, Ieno and Smith, 2007). Spearman's rank correlation analysis was carried out for long-term mean *Cryptosporidium* ($\bar{C}_{LT}$), seasonal amplitude and long-term catchment-averaged environmental data (see below) to discern which factors may influence seasonality.

4.3.3 Environmental data and GIS processing

A summary of the environmental data used is provided in Table 4.1. The data is divided into static and dynamic types, with static datasets including both factors which are truly unchanging during the study's timescale – i.e. catchment-averaged elevation, aspect (as N:S

Table 4.1 Environmental datasets used as features in the XGBoost model.

	Data	Units	Resolution	Notes / Source	Base Feature
Static	OSGB coordinates (of sampling site)	m	1m	Compiled from site info / Google Maps	Eastings Northings
	Soil Percentage Runoff (SPR)	%	1x1km	Based on Hydrology of Soil Types classification 1995 (Boorman, 1995)	SPR
	Aspect - North:South ratio	Decimal	50m	Based on OS Terrain 50 (Ordnance Survey, 2025)	N:S_Aspect
	Elevation	m			Elevation
	Slope angle (mean)	°			Slope
	Human population density	people km ⁻²	1x1km	2011 Census data (Reis et al., 2017)	Human_Pop
	Livestock density (cattle, pigs, poultry & sheep)	head km ⁻²	1x1km	Animal and Plant Health Agency (APHA, 2022)	Cattle_Pop Poultry_Pop Sheep_Pop
	Land use (Aggregated classifications)	% area	1x1km	UKCEH Land Cover Map 2023 (Morton et al., 2024)	Land_Arable Land_Grassland Land_Forested Land_Heath_&_Bog Land_Urban
	Sewage Treatment Works (STWs)	count km ⁻²	N/A		STWs
	Combined Sewer Overflows (CSOs)	count km ⁻²	N/A		CSOs
Dynamic	River discharge	m ³ day ⁻¹	Daily	Hydrology Data Explorer (Defra, 2025)	Discharge
	Rainfall	mm	Daily, 5x5km	HadUKGrid 2023 (Met Office et al., 2021)	Rainfall
	Antecedent Precipitation Index (API)				API
	Air Temperature	°C	Monthly, 5x5km		Air_Temp
	Relative Humidity	%			Humidity
	Sunlight (monthly mean)	hours			Sunlight_Hours
	Wind speed	m s ⁻¹			Wind_Speed

ratio), slope angle, soil type and soil potential runoff (SPR) – and characteristics which are pseudo-static for the purpose of this study (because they are based on a single snapshot dataset), but which in reality will be subject to variation over the 24-year study window – i.e. livestock density, human population density, land use and (potentially) CSO and STW density (although the latter two are only likely to change minimally). Dynamic datasets include meteorological and hydrological (river discharge) data, all at either mean daily or mean monthly resolution.

The environmental data was chosen to represent different known factors affecting *Cryptosporidium* concentrations in rivers. Rainfall has been shown in numerous studies to be the main driver of oocyst mobilisation, either as rainfall-runoff (Atherholt et al., 1998), CSO spills (Gibson III et al., 1998), or urban stormwater discharges (Müller et al., 2020). Soil potential runoff (SPR) and mean slope angle are also included to capture possible rainfall-runoff effects, which would be expected to be more dominant in areas of higher gradient and with less permeable (i.e. higher SPR) soils. Slope aspect and sunlight hours are included because oocyst longevity has been shown to be reduced by exposure to UV light (Jagai et al., 2009). Air temperature, relative humidity, slope aspect and antecedent precipitation index (or API, which serves here as a proxy for soil moisture (Schoener and Stone, 2020)) are included because reduced oocyst survival rates are most associated with desiccation (Robertson et al., 1992), and with high (>37°C) or low (<-10°C) temperatures (King et al., 2005).

Human population density, urban area, and STW and CSO densities are included to represent potential human sources/source areas of *Cryptosporidium*, whereas grassland area and livestock density represent potential animal sources/source areas. Other land use data is included to account for livestock on heaths and bogs, or wild animals in forests and on arable land.

OS Terrain 50 (Ordnance Survey, 2025) elevation data was used to generate full catchment areas for each site. Truncated catchments, representing drainage areas 5 km and 25 km linear distance upstream of each intake, were also generated. Catchment-averaged values for all environmental datasets described above were obtained for the full and the truncated catchments, with the latter included to capture effects that might otherwise be missed when data is averaged over the entire catchment area (Wu et al., 2021b) – e.g. a high density of grazing animals adjacent to the site.

River discharge data was included because previous studies show correlations between flow rates and increased contaminant loads (Ezzati et al., 2023). Data from the nearest gauging station (GS) was retrieved and scaled using the ratio between the GS and intake catchment areas. When GS data was not available for the specific river required, the nearest GS in the wider river system was used. Three further datasets – Eastings, Northings and elevation –

were included to add extra detail to our understanding of the geographical and topographical spread of river-borne *Cryptosporidium*.

4.3.4 Multi-catchment *Cryptosporidium* ML model (XGBoost)

An XGBoost model was developed to predict mean monthly *Cryptosporidium* concentrations at each site, using the catchment-specific environmental variables as input features.

XGBoost – a gradient-boosted, tree-based ensemble model (Chen and Guestrin, 2016) – was selected for its strong predictive performance, its robustness to high-dimensional feature sets, and its ability to handle substantial multicollinearity (Pham and Ho, 2021). It has been applied in many fields of research, including spatio-temporal water-quality modelling (Niazkar et al., 2024), to predict *Cryptosporidium* concentrations from meteorological, hydrological and water-quality variables (Ligda et al., 2024; Smalley et al., 2025), and to link catchment land use with river-water quality (Zhang et al., 2022).

Initial tests used a fully static dataset in which all dynamic variables were reduced to single summary statistics; this approach performed poorly, indicating that both static land-use/geographical characteristics and time-varying meteorological/hydrological variables were required to train an effective model. Given substantial variation in sampling frequency across sites, some temporal aggregation was necessary to ensure target data quality. Because several meteorological datasets were only available at monthly resolution, the model was trained using mean monthly data, which was adequate for identifying temporal patterns.

Prior to modelling, the data was filtered according to two criteria: (i) minimum monthly *Cryptosporidium* sample count (with thresholds of 4, 6, 8, and 10 samples), and (ii) minimum temporal coverage within each month (50% or 75% of days sampled). Months within each site dataset which did not meet these criteria were removed. The model was re-run using four threshold combinations (Models 1-4) and the outputs were compared for both predictive performance and XAI stability (see below).

The modelling pipeline first filtered the date-specific *Cryptosporidium* data, then merged it with daily discharge, rainfall and antecedent precipitation index (API). For rainfall, the lagged rolling mean over days D1–D7 (where D1 is the day prior to sampling) was used, reflecting earlier findings that rainfall is most predictive of *Cryptosporidium* concentrations

when lagged or averaged over the preceding week (Ligda et al., 2020; Smalley et al., 2025). For API, the 7-day lag (D7) was used as a proxy for pre-rainfall event soil moisture. The resulting daily dataset was averaged by calendar month and merged with other monthly dynamic variables and static catchment attributes into a single modelling data frame. Left-censored *Cryptosporidium* values were included as zeros in the monthly averages. These accounted for 20–95% of observations depending on the site, but this reduced to a maximum of 50% in the final monthly averages. Target values (mean monthly *Cryptosporidium*) were log-transformed, a standard approach when concentrations span multiple orders of magnitude (Sokolova et al., 2022). Monthly means equal to zero were removed prior to log transformation. XGBoost hyperparameters were tuned manually, with the following final settings: $n_estimators = 100$, $num_parallel_tree = 10$, $subsample = 0.5$, and $colsample_bytree = 0.5$, with the latter two parameters included to reduce overfitting. Manual k-fold cross-validation used 5 splits and 20 repeated iterations. Mean out-of-fold (OoF) predictions were inverse-log transformed to obtain linear-scale predictions. Model performance was assessed using Root Mean Square Error (RMSE), Normalised Root Mean Square Error (NRMSE) and Explained Variance Score (EVS). Metrics were provided for both linear and log-scale predictions (Hastie et al., 2009).

4.3.5 Explainable AI (SHAP analysis)

XAI analysis of monthly time series *Cryptosporidium* data and catchment data was performed to detect the drivers of more short-term surges in *Cryptosporidium* which may not be apparent in the long-term averaged or monthly-aggregated data. Unlike the statistical tests/correlations carried out for the spatial and temporal analysis, the XAI approach can detect non-linear or complex feature effects, and interactions between features.

SHAP is a model-agnostic XAI package for interpreting ML models (Lundberg and Lee, 2017). It improves on standard feature analysis tools by quantifying both the magnitude and the direction (i.e. positive or negative) of feature effects. Tree SHAP – a version of SHAP designed to analyse ensemble tree-based models (Lundberg et al., 2018) – was applied to the multi-catchment *Cryptosporidium* XGBoost model to quantify the impact, linear

dependency and non-linear dependency of features, each of which is described in detail below.

SHAP values are the amounts by which a feature increases or decreases an individual prediction – e.g. in our model, a SHAP value of -1 for air temperature would mean that this feature decreased the log-transformed *Cryptosporidium* concentration prediction by 1 unit from the SHAP baseline prediction. To evaluate overall feature impact, the “Global” (i.e. across all OoF predictions) Mean Absolute SHAP values (MAS_f) for each feature (f) were calculated according to Equation 15:

$$MAS_f = \frac{1}{N} \sum_{k=1}^N |S_f^k| \quad (15)$$

where N is the number of predictions, k is the index of a specific prediction, and S_f^k is the SHAP value for feature f and prediction k . Interactions between model predictors were assessed using Tree SHAP interaction values and summarised as the mean absolute interaction contribution across all OoF predictions.

SHAP **linear dependency** (S_{LD}) and **non-linear dependency** (S_{NLD}) were calculated to determine the strength and direction (positive or negative) of the relationship between the features and predictions. S_{LD} is given by the Pearson correlation coefficient for SHAP impact scores with raw feature values. S_{NLD} is then calculated as in Equation 16:

$$S_{NLD} = 1 - |S_{LD}| \quad (16)$$

For the final SHAP results, outputs for features applied at three spatial scales (i.e. full catchment, and 5km and 25km catchments) were aggregated to obtain base feature values. Features were then ranked by the magnitude of their effect (i.e. by MAS values).

To determine the relative contributions of animal and human sources of *Cryptosporidium* at each site, SHAP Positive Contribution Scores (PCSs) – i.e. the means of all positive site-specific SHAP values for each feature – were computed. Total PCSs for human-associated and animal-associated features were then summed and converted into % Animal and % Human values. Note that human source features were human population, urban area, STWs and CSOs; animal source features were grassland area and livestock densities.

4.4 Results

4.4.1 Spatial and temporal analysis

The spatial distribution of the sites is displayed in Fig. 4.1a, with summary statistics for each provided in Table 4.2. Sites are spread widely across the British mainland and therefore cover a range of catchment types, including upland and lowland areas, rural and highly urbanised catchments.

Long-term mean *Cryptosporidium* concentrations (Fig. 4.1b) are generally higher in the South East and South West of England and in Wales – these are the warmer and (to a lesser extent) wetter parts of the British mainland. The cooler regions of the north and the drier regions to the east do not have any sites with $\bar{C}_{LT} > 0.4$ oocysts L⁻¹. The collection of higher-*Cryptosporidium* sites around the London area suggest that rivers in heavily urbanised/high population areas may be more susceptible to contamination, but this is offset slightly by the high values at Sites 21, 51 and 52, which are not in heavily urbanised areas (although the latter two are in locations which have high summer influxes of tourists). There are also several sites in highly urbanised areas which have low $\bar{C}_{LT}$ values (e.g. Sites 7 and 36).

Temporal analysis at monthly resolution revealed the timing of peaks (here categorised by season) and the seasonal amplitude (Fig. 4.1c-d). Winter peaks are the most common (22 sites), followed by autumn (19), summer (9) and then spring (2). Both spring sites are found in Yorkshire and The Humberside (although on different rivers). Autumn- and winter-peaking sites are widely spread, whereas summer-peaking sites are found near the south coast, the northern parts of England and Wales, and in Scotland. Sites with high $\bar{C}_{LT}$ tend to peak in autumn or winter (e.g. Sites 21, 34, 37, 46 and 51), although Site 52 (in the SW) is an exception, peaking in summer. Sites with high seasonal amplitude usually also have high mean *Cryptosporidium* concentration. Sites 30 and 38 (in the East and SE) are unusual in having relatively low mean *Cryptosporidium* values ($\bar{C}_{LT} \approx 0.4$ oocysts L⁻¹) but high seasonal amplitudes; hence these sites have very low $\bar{C}_{monthly}$ concentrations for most of the year which are interrupted by high autumn spikes.

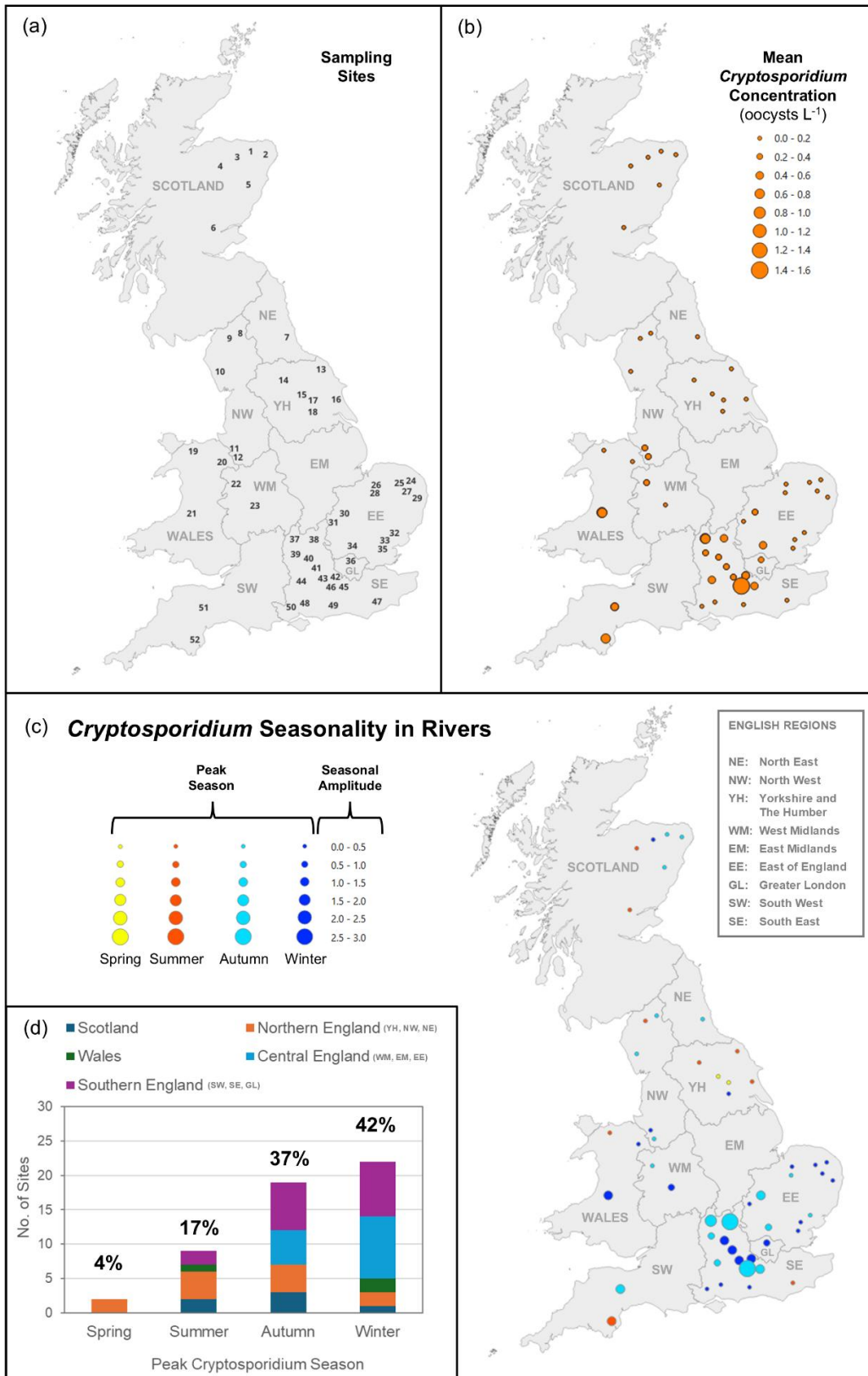


Fig. 4.1 Locations on the British mainland of (a) sampling sites (with Site Numbers), (b) long-term mean *Cryptosporidium* concentrations ($\bar{C}_{LT}$), (c) peak *Cryptosporidium* season and seasonal amplitude, and (d) bar chart of peak seasons.

Table 4.2 *Cryptosporidium* measurement counts, frequency, periods, summary concentration statistics and seasonality for all sites included in the study. Sites with ≥ 1000 data points or ≤ 10 -day mean interval are highlighted green; sites with concentrations/seasonal amplitude ≥ 1 oocysts L⁻¹ are highlighted pink.

		Measurements						Cryptosporidium Concentration			Seasonal	
Site No.	Region	Site Count	Earliest Year	Latest Year	Time Span (years)	Yearly Count	Mean Interval (days)	Mean ($\bar{C}_{LT}$)	Std Dev	Q95	Peak	Amplitude A_s
								(oocysts L ⁻¹)				
1	Scotland	688	2012	2025	13	51	7	0.15	0.46	0.63	AUT	0.49
2		685	2012	2025	13	51	7	0.13	0.48	0.48	AUT	0.30
3		502	2012	2024	13	39	9	0.08	0.33	0.52	WIN	0.28
4		1305	2013	2025	12	106	3	0.09	0.58	0.34	SUM	0.24
5		1609	2012	2025	13	119	3	0.04	0.11	0.18	AUT	0.08
6		687	2012	2025	13	51	7	0.04	0.17	0.23	SUM	0.10
7	North East	475	2013	2024	12	40	9	0.06	0.15	0.30	AUT	0.13
8	North West	518	2009	2025	16	32	12	0.01	0.03	0.10	AUT	0.02
9		598	2000	2025	25	23	16	0.11	0.24	0.52	SUM	0.13
10		544	2009	2024	16	35	11	0.04	0.14	0.20	AUT	0.10
11		889	2000	2025	25	35	10	0.38	0.74	1.40	WIN	0.40
12		550	2009	2024	16	34	11	0.26	0.44	1.20	AUT	0.38
13	Yorkshire and The Humber	1118	2001	2024	24	47	8	0.13	0.52	0.50	SUM	0.38
14		231	2003	2024	21	11	34	0.05	0.13	0.27	SUM	0.10
15		916	2001	2024	24	39	9	0.15	0.32	0.68	SPR	0.16
16		1077	2001	2024	24	45	8	0.12	0.56	0.50	SUM	0.22
17		3582	2001	2024	24	152	2	0.12	0.57	0.50	SPR	0.24
18		1386	2001	2024	24	58	6	0.11	0.26	0.59	WIN	0.14
19	Wales	1083	2008	2024	16	67	5	0.05	0.18	0.18	SUM	0.19
20		552	2009	2024	16	35	11	0.07	0.18	0.40	WIN	0.21
21		428	2009	2024	16	27	13	0.68	1.85	2.97	WIN	1.07
22	West Midlands	253	2020	2024	5	52	7	0.29	0.41	1.14	AUT	0.47
23	East of England	118	2020	2024	5	26	14	0.16	0.22	0.60	WIN	0.58
24		336	2018	2024	7	51	7	0.03	0.09	0.20	WIN	0.08
25		512	2000	2024	25	21	18	0.08	0.54	0.30	WIN	0.26
26		1207	2000	2024	25	49	8	0.02	0.10	0.20	WIN	0.05
27		598	2000	2024	24	25	15	0.03	0.15	0.10	WIN	0.11
28		1280	2000	2024	25	52	7	0.03	0.20	0.20	AUT	0.13
29		407	2015	2024	10	42	9	0.03	0.09	0.20	WIN	0.09
30		253	2000	2009	9	28	13	0.33	0.85	1.54	AUT	1.36
31		1123	2000	2024	25	45	8	0.06	0.32	0.39	WIN	0.21
32		344	2018	2024	7	52	7	0.06	0.13	0.30	AUT	0.19
33		1074	2018	2024	7	158	2	0.17	0.28	0.70	WIN	0.33
34		1248	2000	2024	24	51	7	0.53	1.09	2.00	AUT	0.93
35		1073	2018	2024	7	158	2	0.12	0.22	0.60	WIN	0.31
36	London	1773	2000	2024	24	73	5	0.21	0.50	1.00	WIN	0.58
37	South East	1568	2000	2019	19	81	4	0.79	2.56	3.60	AUT	1.88
38		101	2013	2014	2	51	7	0.41	1.42	1.90	AUT	2.62
39		3157	2000	2024	24	130	3	0.39	1.28	1.60	AUT	0.69
40		1739	2000	2024	24	71	5	0.39	0.81	1.76	WIN	1.22
41		209	2009	2024	15	14	27	0.37	0.60	1.50	WIN	1.18
42		3411	2000	2024	24	140	3	0.55	0.99	2.41	WIN	1.47
43		425	2009	2024	15	28	13	0.31	0.59	1.48	WIN	1.02
44		2418	2000	2024	24	99	4	0.42	0.90	1.60	AUT	0.87
45		2256	2000	2024	24	93	4	0.47	1.39	1.82	AUT	1.13
46		2169	2000	2024	24	89	4	1.57	2.88	6.92	AUT	2.95
47		565	2020	2024	5	117	3	0.08	0.23	0.39	SUM	0.22
48		991	2020	2024	5	204	2	0.01	0.03	0.00	WIN	0.01
49		621	2021	2024	4	162	2	0.02	0.07	0.12	WIN	0.04
50		1032	2020	2024	5	213	2	0.02	0.07	0.13	WIN	0.04
51	South West	251	2000	2015	14	18	21	0.44	1.18	2.35	AUT	1.13
52		246	2000	2011	10	24	15	0.69	0.99	2.68	SUM	1.16

Total Count	52,181
-------------	--------

When testing if site-specific environmental variables differed between sites grouped by their peak season, significant differences ($p < \alpha$, where $\alpha = 0.05$) were identified between summer-peaking and winter-peaking sites only (Table 4.3). Spring-peaking sites were excluded from the analysis because of the low sample number ($n=2$). Box-plots for significant pairs (Fig. 4.2) reveal that winter-peaking sites are in areas with significantly higher mean air temperature, sunlight hours, human population, STW density, and arable and urban land areas, whereas summer-peaking sites are generally in more upland areas with significantly higher mean slope, humidity, rainfall, wind speed and heath/bog areas.

Table 4.3 Significant pairwise environmental differences between summer-peaking and winter-peaking catchments, using the Tukey HSD post-hoc test ($p < \alpha$, where $\alpha = 0.05$). Note that no significant differences were found for other season pairs.

Variable	Season 1	Season 2	Adjusted p-value
Slope	Summer	Winter	0.0041
Land_Urban	Summer	Winter	0.0044
Air_Temp	Summer	Winter	0.0058
Land_Arable	Summer	Winter	0.0096
Rainfall	Summer	Winter	0.0118
Land_Heath_&_Bog	Summer	Winter	0.0204
Human_Pop	Summer	Winter	0.0315
Eastings	Summer	Winter	0.0373
STWs	Summer	Winter	0.0389
Humidity	Summer	Winter	0.0416
Sunlight_Hours	Summer	Winter	0.0444
Wind_Speed	Summer	Winter	0.0449

Cryptosporidium concentration and seasonal amplitude are highly correlated (Fig. 4.3) – this is to be expected, since higher concentration sites are more likely to have high measured range (given that the detection limit for analysis is usually fixed at 0.1 oocysts L^{-1}). When sites are analysed as a whole, moderate positive correlations show that sites with higher *Cryptosporidium* and seasonal amplitude tend to be larger catchments with higher average elevation, higher human population density, higher density of STWs and CSOs, higher proportions of urban areas and grassland and (to a lesser extent, based on weaker correlations) higher cattle density, sheep density and discharge. Summer-peaking sites show higher positive correlation between *Cryptosporidium*/seasonal amplitude and

elevation, air temperature, humidity and sunlight hours, whereas autumn-peaking sites showed much higher correlation with urban area, human population, air temperature and sunlight hours. In winter-peaking sites, positive correlations were with STW and CSO density, sheep density, grassland area and overall catchment area.

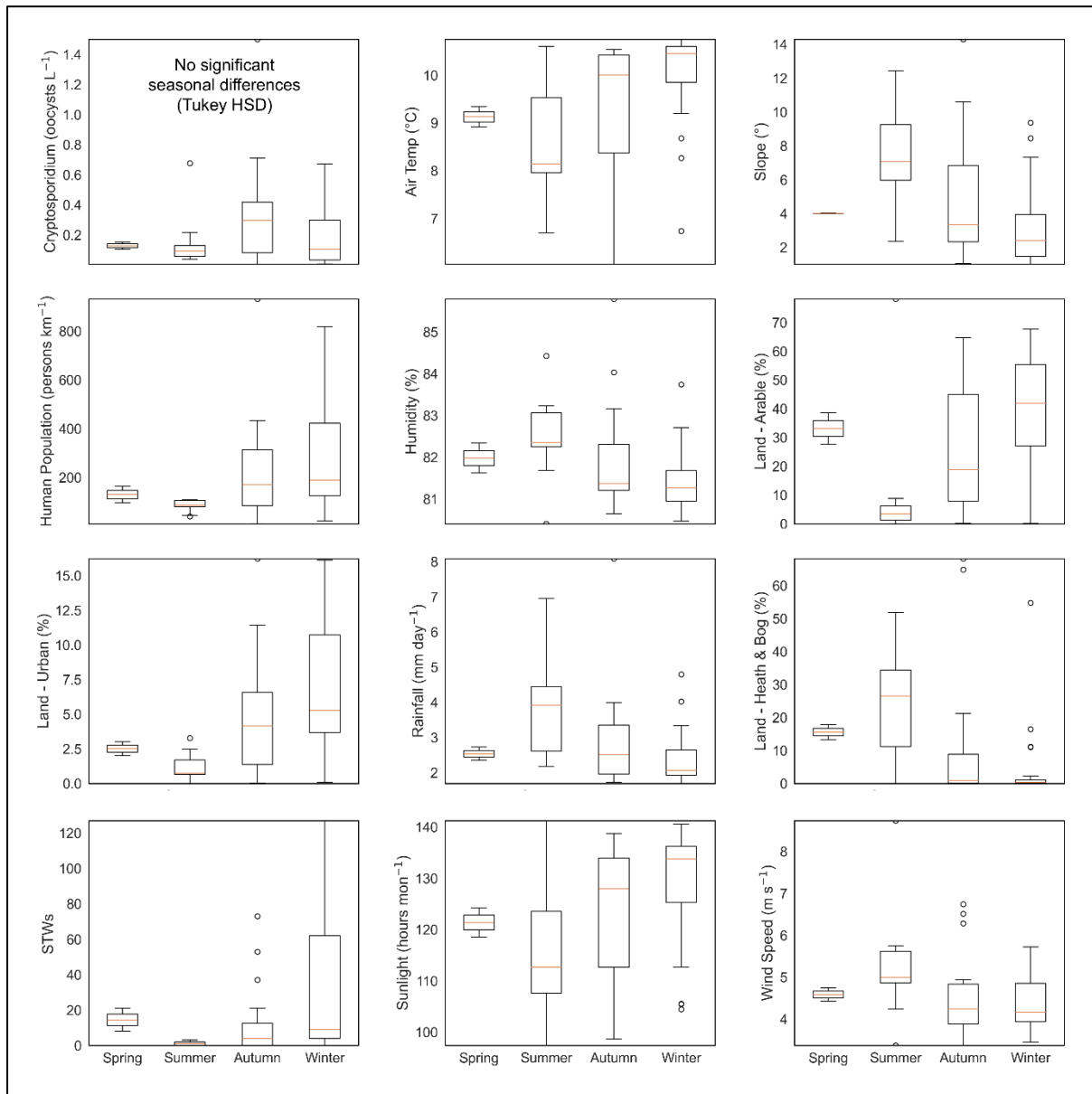


Fig. 4.2 Boxplots of catchment-averaged environmental variables grouped by peak season. Significant Tukey HSD differences ($\alpha = 0.05$) were identified between Summer- and Winter-peak sites only. *Cryptosporidium* concentration is shown for reference, although no seasonal differences were significant.

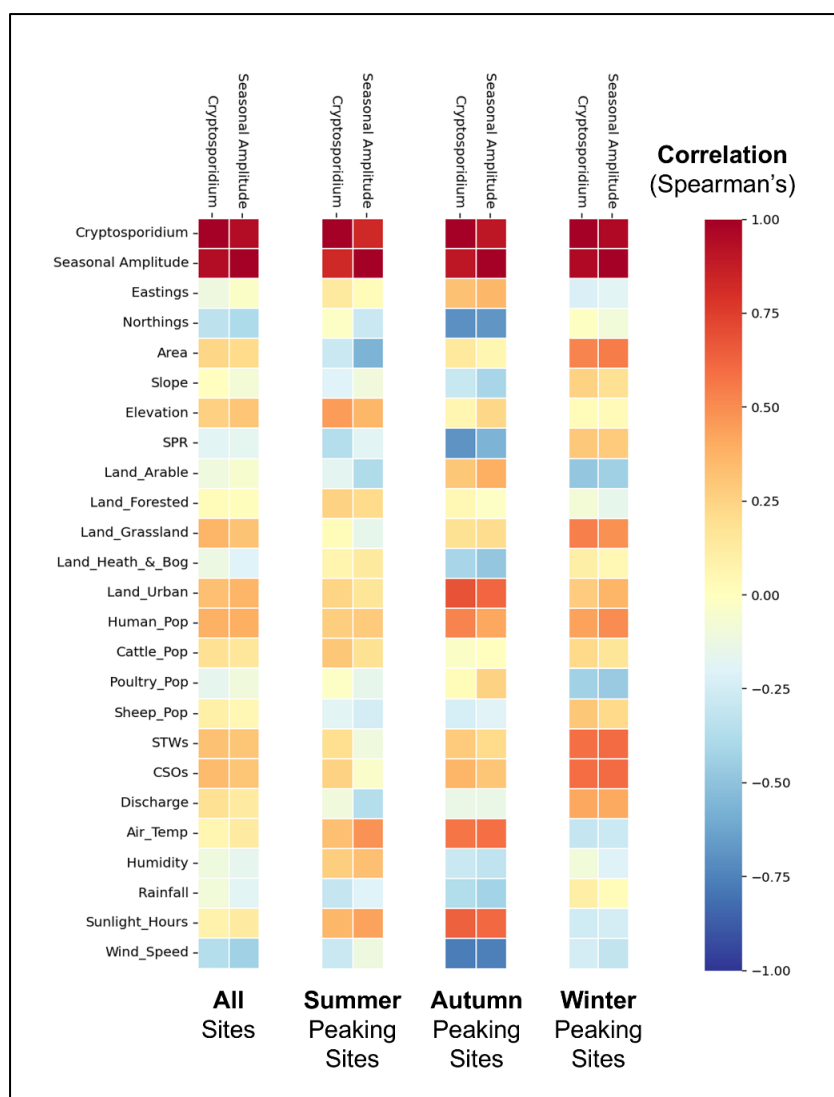


Fig. 4.3 Correlations between *Cryptosporidium* concentration and seasonal amplitude with environmental data, based on long-term (≤ 24 years) and full catchment averages. Results are provided for all 52 sites and for sites defined by peak season. Spring-peaking sites were omitted due to low ($n=2$) numbers.

4.4.2 XGBoost model performance

The four XGBoost models differed only in the amount of *Cryptosporidium* data filtering and the effect this had on the size and quality of the final datasets (see Table 4.4). Model 1 contained the largest ($N=4409$) and most diverse dataset, whereas the dataset used in Model 4 was the smallest ($N=1110$) and least diverse, but also the most statistically-robust (with each monthly mean based on ≥ 10 values).

Table 4.4 Data filtering settings and performance metrics from the four models.

	Model Filtering		Total No. of Sites	Total No. Data Points (N)	Linear Scale Predictions			Log-Scale Predictions	
	Min Monthly Values	Min Monthly Spread			RMSE	NRMSE	EVS	RMSE	EVS
Model 1	4	50	51	4409	0.69	0.04	0.27	0.41	0.55
Model 2	6	50	43	2065	0.90	0.05	0.23	0.42	0.61
Model 3	8	75	34	1548	0.80	0.05	0.26	0.42	0.64
Model 4	10	75	32	1110	0.65	0.07	0.35	0.42	0.64

OoF plots (Fig. 4.4) show predicted data from the four separate k-fold splits for Models 1 and 4. The only notable difference between the two plots is the number of data points, which is much lower in Model 4. (Plots for Models 2 and 3 aren't presented because of their high similarity with the ones shown here – see Appendix C.) The four models were compared in terms of both predictive performance (Table 4.4) and XAI outputs (Fig. 4.5). Despite differing substantially in sample size and data-quality filtering, the models showed very similar behaviour. On the linear scale, Models 2 and 3 exhibited slightly higher error

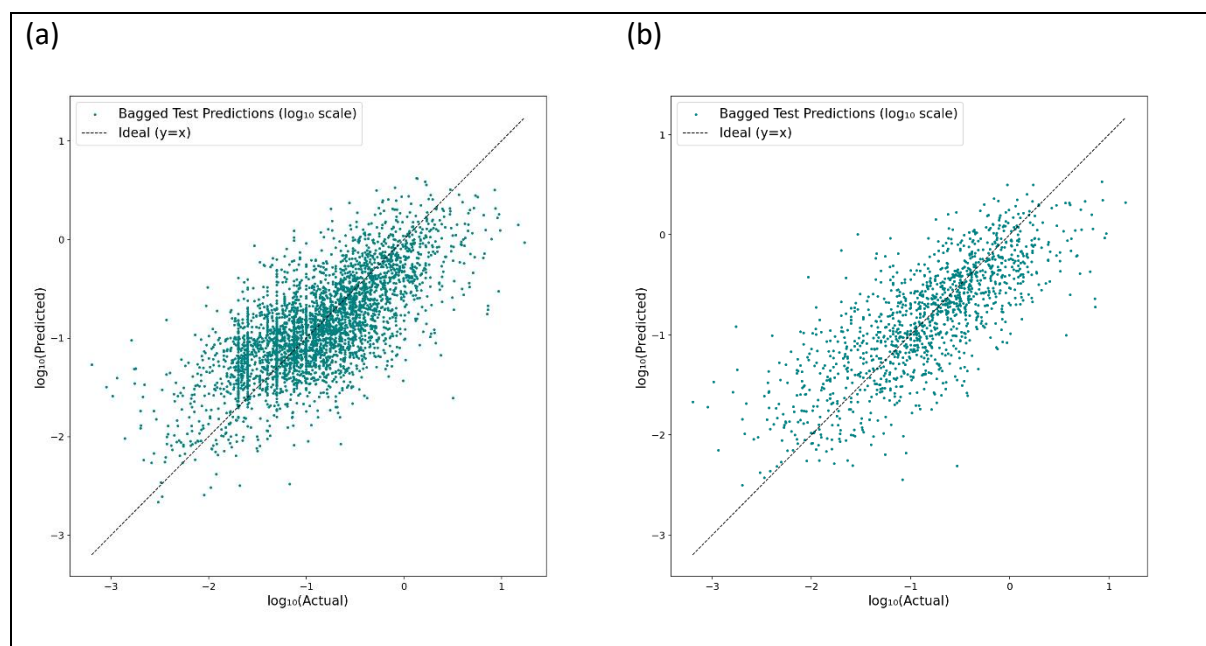


Fig. 4.4 Out-of-fold (OoF) predictions from (a) Model 1 and (b) Model 4.

(RMSE and NRMSE) than Models 1 and 4. On the log-transformed scale, RMSE values were nearly identical across all four models, and explained variance (EVS) exceeded 0.55 (i.e. more than half) in every case. For environmental models with highly skewed targets, EVS values of 0.5-0.6 are generally considered strong indicators that the model is capturing meaningful structure in the data (Hastie et al., 2009). Because no single model clearly outperformed the others, their XAI outputs were examined in parallel.

4.4.3 Feature interpretation using SHAP outputs

Consistent with other applied studies using SHAP (e.g. Ligda et al., 2024), SHAP outputs were not validated against a formal standard. However, the XAI literature emphasises consistency across data subsets as a key indicator of reliability (Yuan et al., 2023; Gan et al., 2025), therefore SHAP stability was assessed by comparing outputs across the four different models, each of which is based on different data subsets. Global SHAP values were highly consistent across all models, with SHAP feature rankings converging early within each model's iterative process, and each generating identical final rankings (Fig. 4.5). Such agreement indicates that the models learned the same underlying predictor-response relationships, despite differences in training sample size and quality. When viewed alongside the consistent predictive performance of the XGBoost models, this provides good evidence for the robustness and reliability of the SHAP outputs.

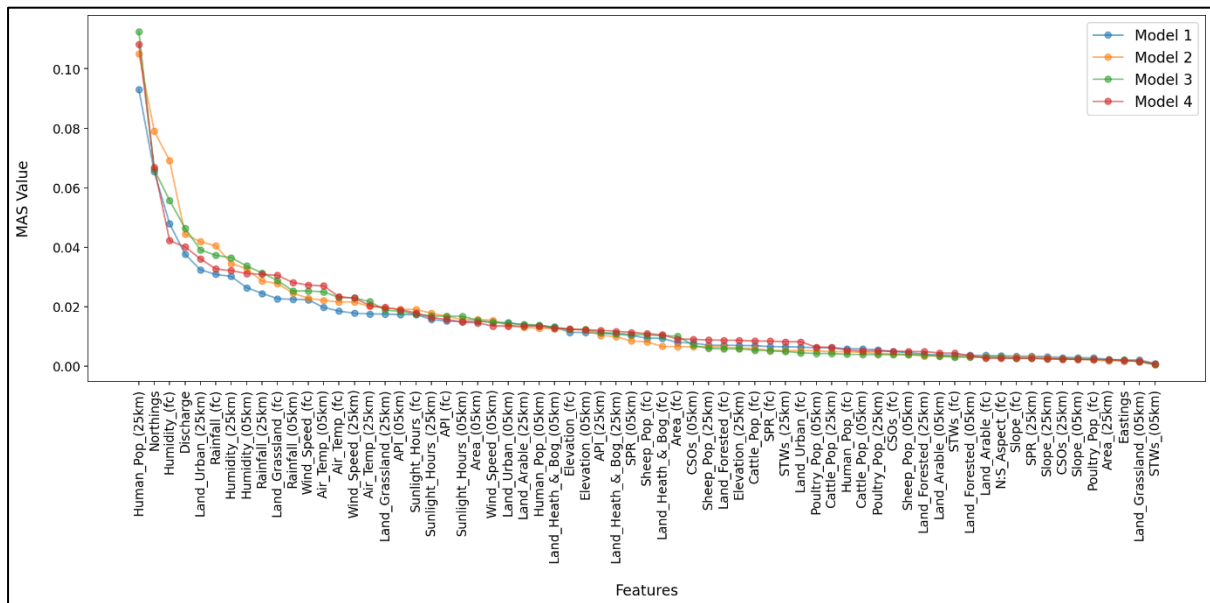


Fig. 4.5 Ranked Mean Absolute SHAP (MAS) values for each feature in the four models. (fc= full catchment).

Feature values ranked by MAS are presented in Table 4.5. For simplicity, spatially-defined features (i.e. those based on full catchment, 25km and 5km upstream areas) have been summed to give a single base feature value. With the exception of those labelled “complex”, all of the features were consistent across the models in the (global) direction of their effect – i.e. either having a positive (increasing) or negative (decreasing) effect on predictions.

Table 4.5 Base features for all four models ranked by Mean Absolute SHAP (MAS), alongside direction of effect and most predictive spatial scale. (Colour ramp indicates high, intermediate and low MAS values; SD=Standard Deviation; POS = Positive; NEG = Negative.)

Base Feature	Mean Absolute SHAP (MAS)						Direction of Effect	Most Predictive Spatial Scale
	Model 1	Model 2	Model 3	Model 4	Overall Mean	SD		
Humidity	0.105	0.155	0.139	0.130	0.132	0.018	POS	Full catchment
Human_Pop	0.113	0.115	0.120	0.127	0.119	0.006	POS	25km
Rainfall	0.078	0.076	0.080	0.083	0.079	0.002	POS	25km
Land_Grassland	0.042	0.096	0.077	0.047	0.066	0.022	POS	25km
Sunlight_Hours	0.048	0.048	0.058	0.076	0.058	0.012	NEG	05km
Air_Temp	0.056	0.059	0.057	0.057	0.057	0.001	NEG	05km
Wind_Speed	0.055	0.055	0.046	0.046	0.050	0.005	NEG	05km
Land_Heath_&_Bog	0.033	0.048	0.063	0.041	0.046	0.011	NEG	05km
Northings	0.065	0.035	0.037	0.036	0.043	0.013	NEG	-
Land_Urban	0.054	0.039	0.044	0.031	0.042	0.008	POS	25km
Discharge	0.038	0.040	0.046	0.042	0.042	0.003	POS	-
API	0.044	0.041	0.042	0.038	0.041	0.002	Complex	05km
Elevation	0.030	0.029	0.030	0.044	0.033	0.006	POS	Full catchment
Area	0.025	0.034	0.032	0.027	0.030	0.004	Complex	05km
CSOs	0.016	0.038	0.030	0.025	0.027	0.008	POS	05km
SPR	0.020	0.018	0.031	0.031	0.025	0.006	Complex	Full catchment
Land_Forested	0.015	0.014	0.018	0.031	0.019	0.007	Complex	Full catchment
Poultry_Pop	0.015	0.016	0.013	0.016	0.015	0.001	Complex	05km
Land_Arable	0.021	0.016	0.010	0.010	0.014	0.005	NEG	25km
Slope	0.010	0.011	0.010	0.019	0.013	0.004	NEG	05km
Eastings	0.002	0.008	0.004	0.033	0.012	0.012	Complex	-
Sheep_Pop	0.021	0.007	0.007	0.007	0.011	0.006	NEG	25km
Cattle_Pop	0.019	0.008	0.007	0.008	0.011	0.005	POS	Full catchment
STWs	0.011	0.009	0.010	0.012	0.010	0.001	POS	25km
N:S_Aspect	0.004	0.008	0.013	0.011	0.009	0.004	POS	-

Humidity had the greatest positive effect in all of models, followed by human population and rainfall. Air temperature and sunlight hours had relatively strong negative effects. Grassland area had the most positive effect of the land use features. Heath and bog areas and arable areas had negative effects, while the effect of forested areas was complex.

Of the features directly quantifying the density of potential *Cryptosporidium* hosts, only human population ranked high, with all of the livestock features having a low (and in the case of sheep and poultry, a negative) effect. Two of the three human source-related features – urban area and CSOs – had modest positive effects, while the effect of STWs, although also positive, was much lower.

N:S aspect had a very weak positive effect (indicating that north-facing slopes are associated with higher *Cryptosporidium* concentrations). Slope has a negative effect, suggesting that more rapid rainfall-runoff and storm responses do not necessarily lead to higher concentrations of the pathogen, although mean higher elevations overall do appear to favour higher concentrations. Two runoff-related features – API and SPR – both showed complex patterns of association.

The most predictive spatial scale also varied between features. A majority of the meteorological features were most predictive at the 5 km scale, the exceptions being humidity (full catchment scale) and rainfall (25 km scale). Land use impacts for grassland, urban and arable areas were greatest at 25 km scale, in contrast to heath/bog (5 km) and forested areas (full catchment). CSOs and STWs had the most effect at the 5 and 25 km scale respectively. Topographical features were variable, with slope having the most effect at the 5 km scale, and elevation at the full catchment scale.

4.4.4 Dominant source types for each site

Fig. 4.6a shows that sites dominated by human sources are located predominantly in the South East, Greater London and the East of England. Animal source-dominated sites are more prevalent in the western and northern regions, and are totally dominant in Scotland. When sites are counted using a 75% dominance threshold (Fig 4.6b), human-dominated sites are slightly more common than animal dominated sites, but mixed sites (in which neither site is more than 75% dominant) are the most abundant.

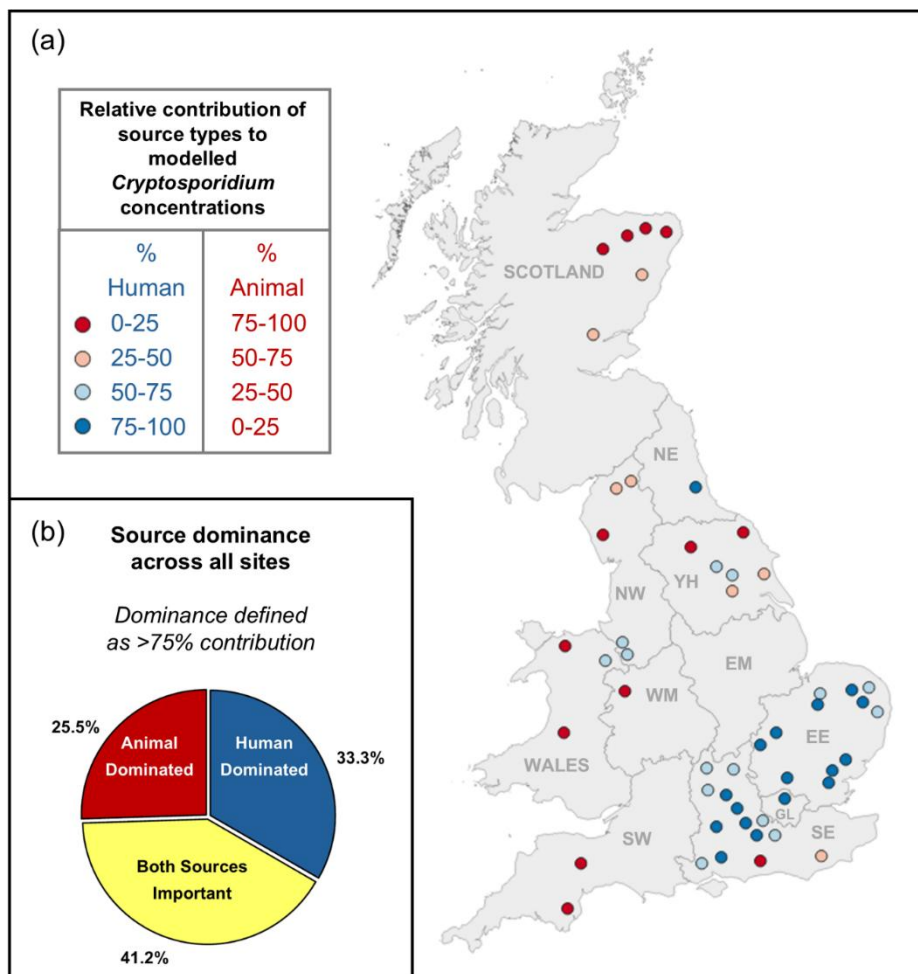


Fig. 4.6 Sites characterised by the relative contribution of source types to the modelled *Cryptosporidium* concentrations, shown by (a) their location and (b) the proportion of total sites.

4.5 Discussion

A spatio-temporal analysis has been performed 52 sites widely distributed across the British mainland. This dataset – which is unprecedented in the literature – has been integrated with spatial environmental data, daily discharge data and meteorological data. A key novelty of the study is the parallel analysis of static, pseudo-static and dynamic catchment datasets at three distinct temporal resolutions: (i) a long-term scale, represented by site-averaged values across the 24-year record, used to identify spatial patterns; (ii) a seasonal scale, based on monthly values aggregated across all years, used to characterise recurrent seasonal behaviour; and (iii) a short-term scale, using monthly time-series data within an XAI modelling framework, to identify spatial and temporal drivers of *Cryptosporidium* variability that may not be discernible at coarser resolutions.

4.5.1 *Cryptosporidium* spatial patterns and hotspots

Cryptosporidium river concentrations vary widely across the British mainland (Fig. 4.1), with all of the highest concentration sites in the southern half of England and in Wales, lower values in the English midlands and northern areas, and very low values in Scotland. The spatial patterns are broadly consistent with much earlier Water Industry-based monitoring (DWI, 1998), which found fewer detections in rural upland catchment waters compared with lowland mixed urban-rural catchments. The picture is also consistent with output maps from the GloWPa-Crypto C1 model (Vermeulen et al., 2019), in which *Cryptosporidium* concentrations were predicted to be highest in the southern parts of the British mainland.

Correlation analysis (Fig. 4.3) across all sites shows that mean *Cryptosporidium* levels tend to be higher in sites with higher human population, higher proportions of grassland and/or urban land, higher density of STWs or CSOs, and (to a lesser extend) higher cattle and sheep densities – a finding which ties in well with existing understanding of *Cryptosporidium* sources and high-risk water bodies (Medema et al., 2009). Larger catchment areas are also associated higher *Cryptosporidium* concentrations, possibly because large catchments are more likely to include one or more major sources. Areas with high humidity or rainfall are not associated with higher mean *Cryptosporidium*, reflecting previous observations that climatic variables averaged over the long term are often poor predictors of *Cryptosporidium* in water bodies; this is particularly true for rainfall, which – due to the episodic nature of precipitation events and their direct effect on driving oocyst transport – is usually most predictive at approximately daily resolution (Wilkes et al., 2009; Wilkes et al., 2013).

Although higher temperatures usually lead to higher rates of oocyst deactivation (King and Monis, 2007), the positive correlation between *Cryptosporidium* concentration with air temperature found here is most likely due to generally higher mean temperatures in lowland southern sites, which have high human and livestock populations.

4.5.2 *Cryptosporidium* seasonality

Previous studies have observed seasonality in *Cryptosporidium* river concentrations, although these have been limited to consideration of single or adjoining catchments. Both autumn and winter-spring peaks have been recorded in catchments in the North West of England (Sturdee et al., 2007; Robinson et al., 2011), with oocysts shed by both livestock

and wild animals. Concentrations were found to peak in late spring in a catchment in the Scottish Cairngorms (Wells et al., 2015), most likely as a consequence of the calving/lambing season. This study further illustrates the seasonality of *Cryptosporidium*, but it does so over a much larger spatial scale, while performing a broader analysis of the key drivers of seasonal variations.

Summer-peaking seasons were found to be in upland catchments containing heath/bog areas, with higher mean slope, humidity and rainfall (Fig. 4.2). They were also sites with low overall *Cryptosporidium* concentrations (with one exception – site 52, in the South West). Summer peaks are therefore likely due to high-intensity summer storms, with grazing sheep a contender as the major source host (although no significant difference in sheep numbers was found between summer-peaking sites and sites peaking in other seasons). It is notable that some of sites (including Site 52) are major tourist centres, therefore summer peaks may also be due to human-shed oocysts entering water courses, either from CSOs during rainfall events and/or when sewer systems are overburdened at peak tourist months (Pasciucco et al., 2023).

Autumn/winter-peaking sites are the most common, accounting for 79% of all sites. All but one of the highest *Cryptosporidium* sites ($\bar{C}_{LT} > 0.6$ oocysts L⁻¹) peaked in the cold-seasons, but a majority of autumn/winter-peaking sites did not have significantly higher mean *Cryptosporidium* concentrations than spring/summer-peaking sites. Correlation analysis suggests that autumn peak concentrations tend to be higher in more urban catchments, perhaps reflecting the September peak in human *cryptosporidiosis* cases (Public Health England, 2024), whereas higher winter peaks occur in the larger mixed urban-pastureland catchments, with higher proportions of grassland, and higher densities of STWs and CSOs (Fig. 4.3).

A contributing factor in the autumn-peaking sites could be river discharge, which is usually lower in autumn than in winter, and therefore has less capacity to dilute the influx of oocysts from STWs and CSOs. This interpretation is supported by previous studies (Knapp et al., 2022) and by the negative correlation between discharge and *Cryptosporidium* concentrations in autumn-peaking sites (Fig. 4.3). First-flush effects (particularly from urban areas) may also be relevant, mobilising human-shed oocysts from CSOs which have lain

dormant for many weeks/months, and oocysts shed by wild/domesticated urban animals (Müller et al., 2020).

4.5.3 Key environmental variables affecting *Cryptosporidium* river concentration

In the XAI analysis, humidity was shown to have the greatest positive effect on mean monthly *Cryptosporidium* predictions (Table 4.5). This is a striking result given that out of approximately 13 published *Cryptosporidium* river modelling studies, only one (Liu et al., 2019) included humidity as a model input. A major component of humidity's impact is likely as a seasonal trainer for the model – an important consideration given that the highest peaks are in the cooler and more humid autumn/winter seasons. Air temperature and sunlight hours also serve as seasonal trainers, but the direct association between humidity and rainfall may account for why humidity is the most predictive feature, given that rainfall is known to be a short-term driver of increased *Cryptosporidium* concentrations (Atherholt et al., 1998). This result also supports previous work proposing that desiccation is often the main cause of oocyst degradation in the environment (Robertson et al., 1992); in which case, longer oocyst survival times would be expected in months with higher humidity.

Reduced desiccation could be important in both rural areas – which, though not dominant at all sites, are still a major source of *Cryptosporidium* (see Section 4.5.5) – and in urban areas. In the latter, the longevity of oocysts in animal waste (e.g. from birds) deposited on relatively impermeable urban roof and ground surfaces may be particularly sensitive to humidity. This interpretation is supported in part by the mean absolute SHAP interaction contributions between humidity features and land use or source-related features (Appendix C, Table C.1), which show that humidity-human population and humidity-urban area interactions rank among the strongest feature pairs, followed by humidity-grassland area. In this context, human population may be acting partly as a proxy for urbanisation.

The model outputs provide clear evidence of the importance of rainfall as a mobilisation and transport mechanism for *Cryptosporidium*. The way that rainfall data is used is key to its predictive utility, however. The authors' previous work (Smalley et al., 2025) identified 2 days as the most predictive lag time, and early trials for this model (not presented) found a similar result, with rainfall lagged by 1-7 days all serving as useful training data. This is supported by the earlier correlation analysis findings and previous studies (Wilkes et al.,

2009; Wilkes et al., 2011) which have shown that rainfall has a greater short-term influence on *Cryptosporidium* river concentrations (i.e. through mobilisation and transport), but that average rainfall often shows no association.

It should be recognised that the routine *Cryptosporidium* monitoring data used here represents discrete samples collected at frequencies ranging from approximately daily to sub-monthly, depending on site and monitoring period. Furthermore, the temporal aggregation of rainfall, discharge and API data prior to modelling may smooth or obscure short-duration contamination events and transient mobilisation processes, potentially limiting the ability of the models to fully resolve short-timescale relationships between environmental drivers and *Cryptosporidium* concentrations.

Models which include both rainfall and soil moisture are more likely to capture the impact of heightened rainfall-runoff when soils are saturated (Singh et al., 2021). API – which has been included as a proxy for soil moisture – has a moderate MAS, although its association with predictions is complex. API can be a poor surrogate at times of high evapotranspiration or extreme (e.g. poorly drained or very sandy) soil types, therefore actual soil moisture values could improve the model further.

4.5.4 Spatial averaging: considerations and limitations

All but three of the features included in the model were represented using spatially averaged values (the exceptions being discharge, northings and eastings). The implications of spatial averaging are unlikely to be equal across all variables, however. Some may be subject to greater spatial heterogeneity (e.g. CSOs, which tend to cluster in urban areas), while others may not be well-represented by "lumped" catchment-averaged values because their influence depends strongly on location and hydrological connectivity (Lei et al., 2021). In particular, the impact of source-related variables (human population, urban and grassland land cover, STWs, CSOs and livestock populations) may be poorly captured by the model if used in catchment-averaged form only, since this would not account for the potential influence of sources concentrated close to river channels or the monitoring points. This effect is likely to be particularly important in smaller rivers, where there is reduced opportunity for mixing and dilution of contamination inputs over short downstream distances.

The use of truncated upstream catchments (25 km and 5 km) was intended partly to mitigate this issue by increasing model sensitivity to contamination sources located closer to the monitoring points. Nested spatial discretisation approaches of this kind have been used in previous models to represent landscape heterogeneity and scale-dependent hydrological processes (Dehotin and Braud, 2008), and the stronger predictive performance of several source-related variables at the truncated catchment-scale supports the interpretation that proximity to upstream sources is an important control on observed concentrations.

However, defining an appropriate degree of truncation is challenging, not least because the study rivers differ substantially in discharge, catchment size and topography, and this is likely to affect the spatial scale across which local effects manifest.

Lateral proximity of sources to the river network may in some cases be more important than upstream distance (Sliva and Williams, 2001), in part because of the increased likelihood that oocysts adjacent to river channels will enter the river system. Future refinement of the truncated catchment approach could therefore include spatial averaging over a riparian buffer zone, with buffer width potentially adjusted according to local slope, hydrological connectivity or distance from the monitoring point, so as to better represent areas most likely to contribute to rapid mobilisation and transport of *Cryptosporidium* to river systems (Lei et al., 2021).

4.5.5 Major sources of *Cryptosporidium* in rivers

The high MAS for human population (Table 4.5) relative to the other source-related features (urban area, STWs and CSOs for human sources; grassland area and livestock densities for animal sources) supports the argument that high *Cryptosporidium* concentrations in rivers are more commonly caused by oocysts from human – as opposed to animal – waste. This is partially supported by the earlier correlation analysis for all sites (Fig. 4.3) in which features representing human sources (population, urban area, STWs and CSOs) showed moderate-to-high correlations with *Cryptosporidium*, whereas of those features representing animal sources, only grassland area showed a similar level of correlation.

However, the low MAS values for STWs and CSOs demonstrate the difficulty of interpreting relative importance of sources from ML models – if human waste is the major source of *Cryptosporidium* in many cases, almost all of that waste would be entering river systems via

either STWs or CSOs, therefore a stronger signal for either feature might be expected. The issue here may be that simple asset density values (i.e. STWs or CSOs per km²) are a crude way to try to capture the effects of these assets, since STWs can vary in their daily discharge volumes by orders of magnitude, and CSOs can exhibit very different patterns of spilling. It is also noteworthy that CSOs, STWs and urban area were most predictive at the smaller spatial scales (5 km and/or 25 km) – in part, this may reflect the reduced mixing and dilution of contamination that is observed when samples are taken a short distance downstream of point sources (Sterk et al., 2016).

For animal sources of *Cryptosporidium*, grassland area was a more predictive feature than any of the livestock density features. Gaps/anomalies in the livestock data may be responsible for this, with the sheep and poultry data open to the most doubt because the surveys used to gather the data are voluntary. The submission of survey forms is a compulsory requirement for beef and dairy farmers, therefore the cattle data, though imperfect, is generally regarded as more reflective of true numbers. This may be the reason why cattle population is the only livestock feature with a consistent positive effect.

The pattern of more human-dominated sites in the South East and East (Fig. 4.6a) partially reflects the general spread of population across the British mainland, but it also captures other effects – e.g. the relative position of urban areas, CSOs and STWs in relation to the sites. Some river intakes are in areas of high human population density, but they are upstream of STWs, CSOs and may therefore be less impacted by human faecal pollution than other sites in urban areas. Likewise, a number of human-dominated sites are in areas with relatively low human population. This seems counter-intuitive, but previous studies have shown that in some cases CSO spills can have a significant effect on pathogen river concentrations in largely rural catchments (Suslovaite et al., 2024).

Overall, the findings of the study add further support to the growing body of evidence that human-shed oocysts are often dominant in rivers found in highly urbanised and mixed urban-rural catchments (Vermeulen et al., 2019). Although catchment partnership programmes – in which water companies work with farmers to reduce agricultural loadings – should continue to be supported and expanded, addressing human sources must become the central priority in most mixed and urbanised catchments. Nationwide investment has already been proposed (Ofwat, 2025a) to upgrade STWs and sewage networks so as to

improve the quality of treated discharge waters and reduce CSOs spill volumes. The evidence of this study is that such upgrades have the potential to significantly reduce the risk posed by *Cryptosporidium* to drinking water quality and public health.

4.6 Conclusions

This study presents a spatial and temporal overview of *Cryptosporidium* river concentrations across the British mainland, based on long-term (20+ years) monitoring data from 52 sites. It provides the first multi-catchment analysis of spatial and seasonal variations in *Cryptosporidium* river concentrations and their environmental drivers, and the compiled dataset constitutes the largest and highest quality set of raw water concentration measurements yet assembled for this pathogen. Novel integration of static, pseudo-static and dynamic data enabled the identification and interpretation of relationships between temporally dynamic drivers (e.g. rainfall and humidity) and more spatially fixed catchment characteristics (e.g. urbanisation and population density), allowing patterns and trends to be identified that were either previously unknown, only supported by isolated studies, or difficult to resolve using conventional statistical or single-catchment approaches.

Distinct spatial differences were found in nationwide *Cryptosporidium* levels, with the highest *Cryptosporidium* sites in the south of England and Wales, and the lowest *Cryptosporidium* sites in Scotland. Seasonal amplitude varied greatly between sites and showed high correlation with long-term mean *Cryptosporidium* concentrations. A majority (79%) of sites peaked in autumn/winter, with only 17% peaking in summer and 4% in spring.

Humidity, human population and rainfall are all positive drivers of *Cryptosporidium* concentrations; rainfall events in high population areas during humid periods are likely to generate higher *Cryptosporidium* river concentrations than the opposite conditions. Both human and animal *Cryptosporidium* sources are important, but the model outputs indicate that human sources have the greatest influence over spatial and temporal variations in *Cryptosporidium*, and are the dominant source in at least 33% of sites. These findings contribute to a broader conceptual understanding of how anthropogenic and environmental drivers interact to influence pathogen transport and persistence across heterogeneous catchments.

To mitigate future *Cryptosporidium* risks to river water quality (and hence human health), the results presented here provide a strong case for addressing urban sources of contamination in mixed and urbanised catchments. This could include improving STW treatment efficiencies (to reduce oocyst counts in STW discharges) and improving network infrastructure (to reduce CSO spills).

This study demonstrates the strong research potential of long-term water quality monitoring datasets and the value of integrating static and dynamic environmental information within an XAI framework to investigate complex spatial and temporal patterns in river pathogen concentrations across multiple catchments. The approach described could be readily applied to other long-term water quality monitoring datasets to advance understanding of contamination sources, pathways and environmental drivers.

Chapter 5: Discussion

This chapter draws together the main scientific and practical contributions of the thesis. It first synthesises the findings of the three research chapters, focusing on the use of routine monitoring data to predict *Cryptosporidium* concentrations, quantify drinking water risk, and identify environmental and source-related drivers of river-borne *Cryptosporidium* across multiple catchments. It then describes the operational trial and implementation of the river prediction tool developed in Chapter 2, before discussing specific findings in greater detail and outlining recommendations for future work.

5.1 Overall synthesis and contribution of the thesis

This thesis has examined how routine water industry monitoring data, when combined with environmental datasets and modelling approaches, can be used to improve understanding and management of *Cryptosporidium* risk in catchments and drinking water systems. Across the three research chapters, the work addresses three connected challenges: predicting short-term elevations in *Cryptosporidium* river concentrations, quantifying the implications of those concentrations for drinking water risk, and identifying the environmental and source-related factors that explain spatial and temporal variation in *Cryptosporidium* occurrence.

The central contribution is therefore not only the development of individual models, but the demonstration that long-term operational datasets can be repurposed to generate both predictive tools and new scientific understanding. Chapter 2 developed a data-driven machine learning model to support selective abstraction decisions at Thames Water, providing the first model of its kind capable of predicting daily *Cryptosporidium* river concentrations. The chapter showed how antecedent *Cryptosporidium* measurements can be combined with either catchment-averaged or spatially distributed hydro-meteorological and source data, and processed through an interpretable machine learning framework, to predict daily river concentrations and identify influential source areas and lag times.

Chapter 3 applied *Cryptosporidium* monitoring data for raw, reservoir and treated waters to develop and apply a dual QMRA framework for a complex multi-source drinking water

system. This was motivated by the potential for UK water company monitoring data to support risk modelling, and by the limited adoption of QMRA-based approaches within the UK water sector. The London-wide QMRA therefore provides a case study showing how operational monitoring datasets can be used to quantify pathogen risk, evaluate selective abstraction, and explore wider questions relating to drinking water risk management. Omission of the drinking water distribution system from the framework, however, means that it represents a source-to-WTW outlet – not a full source-to-tap – risk assessment.

Chapter 4 then broadened the analysis across multiple catchments by bringing together raw water *Cryptosporidium* data collected by water companies across the British mainland. The chapter showed how static catchment characteristics and dynamic hydro-meteorological conditions interact to influence *Cryptosporidium* concentrations at national scale, and demonstrated how XAI methods can be used to identify environmental drivers, source indicators and seasonal patterns across heterogeneous catchments.

Taken together, these studies show that routine monitoring data can support a catchment-to-WTW outlet understanding of *Cryptosporidium* risk, while also highlighting the limitations of empirical models and the uncertainties associated with discrete sampling, a lack of clear analytical recovery data and potential variation in pathogen infectivity. This work sits within the wider framework set out by the Drinking Water Directive (European Union, 2020), which emphasises catchment-level risk assessment and source-informed approaches to maintaining safe drinking water.

5.2 Immediate impact: trial and implementation of river prediction tool

5.2.1 Context

Between Oct 2024 and Feb 2025, the *Cryptosporidium* river prediction tool (described in Chapter 2) was tested at Walton abstraction point, a major source of raw water for London's drinking water supply. The aim was to trial the recently developed model in an operational setting. Specific objectives were: (i) identify and address issues encountered when using the model operationally; and (ii) obtain feedback from users on how the model outputs might be presented and used in a future roll-out.

5.2.2 Operational trial: method

A working model was made available to Thames Water managers and research staff. Staff accessed the model on an approximately daily basis to view predictions and provide the research team with informal feedback concerning the model's utility, user-friendliness and reliability, and any suggestions for areas of improvement.

The model was initially developed and tested in Python (v. 3.8.8), but prior to the trial the code was re-platformed to run on Google Colab, with input and output files accessed in Google Drive. This option was taken to make the code accessible to multiple users without the need for a desktop Python installation, and to facilitate real-time editing of the code should issues arise.

A data update module was inserted into the beginning of the code so that with each (daily) run, data would be automatically retrieved from several online platforms, and input files would be repopulated accordingly, prior to running the Bagging-XGBoost model. Only the catchment-averaged model (Model A – see Chapter 2) was tested operationally.

Model outputs were presented to users in two different ways: (i) as absolute concentrations, alongside 68% confidence intervals (± 1 standard deviation); and (ii) exceedance/non-exceedance classes of threshold levels (i.e. ≤ 1 , >1 , >2 , >3 oocysts L^{-1}). These thresholds were based partially on those currently in operational use at the site (Thames Water, 2021). Examples of the two user outputs are provided in Fig. 5.1.

The two types of user data gathered during the operational trials were: (i) feedback collected from external end-users of the code (i.e. members of staff at Thames Water); and (ii) learning notes compiled by the researchers during daily running and maintenance of the code.

5.2.3 Operational trial: results

End-user feedback from the trials is presented in Table 5.1. General comments pointed to the potential utility of the trialled model, but also emphasised the need for automation of tasks at all levels, including automatic retrieval and updating of input data, automated running of the model on a daily basis and, finally, the automated emailing of model outputs (alongside other relevant decision-making parameters) to end-users. Concentration-based

prediction outputs (as opposed to interval-based outputs) were preferred, and the flagging of data input gaps alongside predictions was recommended.

(a)

	Date	Day	<i>Cryptosporidium</i> Concentration (oocysts L ⁻¹)		
			Predicted Value	Predicted Range	Actual Value
	26/11/2024	Tue	2	1.5 - 2.6	1.2
	27/11/2024	Wed	1.4	0.9 - 2.0	3
	28/11/2024	Thu	1.7	1.0 - 2.3	2
	29/11/2024	Fri	1.9	1.2 - 2.6	1.6
	30/11/2024	Sat	1.4	0.9 - 1.8	0.5
	01/12/2024	Sun	0.9	0.6 - 1.3	0.4
	02/12/2024	Mon	0.9	0.5 - 1.3	0.6
	03/12/2024	Tue	1.2	0.7 - 1.6	1.2
	04/12/2024	Wed	1	0.7 - 1.3	1
	05/12/2024	Thu	0.8	0.5 - 1.2	Data Not Yet Available
Today --->	06/12/2024	Fri	1	0.5 - 1.5	Data Not Yet Available

(b)

	Date	Day	<i>Cryptosporidium</i> Concentration (oocysts L ⁻¹)	
			Predicted Value	Actual Value
	26/11/2024	Tue	1 - 2	1.2
	27/11/2024	Wed	1 - 2	3
	28/11/2024	Thu	1 - 2	2
	29/11/2024	Fri	1 - 2	1.6
	30/11/2024	Sat	1 - 2	0.5
	01/12/2024	Sun	≤ 1	0.4
	02/12/2024	Mon	≤ 1	0.6
	03/12/2024	Tue	1 - 2	1.2
	04/12/2024	Wed	≤ 1	1
	05/12/2024	Thu	≤ 1	Data Not Yet Available
Today --->	06/12/2024	Fri	≤ 1	Data Not Yet Available

Fig. 5.1 Example outputs from Model A as viewed by end-users during operational trials, showing (a) predicted *Cryptosporidium* concentrations alongside a prediction range (based on the mean ± 1 SD); and (b) predicted threshold intervals/less-than values. Actual (measured) *Cryptosporidium* concentrations are displayed in the right-hand columns in both tables.

The need to adapt the model to other abstraction sites was also pointed out as an important consideration, and the potential advantages of being able to make predictions beyond the current day (up to 3 days in advance, particularly over weekends) were also identified.

Predictions beyond today (i.e. as far as 1-3 days in the future) were identified as a desirable feature. Although not included in the original design of the model, it would be feasible to adapt the model to include both nowcast and forecast outputs. To achieve this, the model would need to incorporate weather forecast data in addition to the existing weather monitoring data. Forecast outputs from such models would be particularly sensitive to the reliability of the rainfall predictions, but the extra information provided by the adapted models would likely enhance their overall functionality and hence their potential contribution to the risk evaluation and risk management process.

Table 5.1 End-user feedback following operational trials

Category	End-user comments
Modelling Approach	Adaptation of the model for deployment at other major abstraction points would be desirable.
	Nowcasting (i.e. predicting today's <i>Cryptosporidium</i> concentration) is the priority for informing abstraction decision-making, but the ability to predict 1-3 days in the future would also be helpful, particularly before weekends and bank holidays.
	The coding language (Python) is suitable for integration with existing company systems.
Output Formats	Predictions showing absolute concentrations are more useful because they allow the end-user to see how close to the next threshold the prediction is. This information is lost when predictions are presented in interval form.
	It would be useful to flag those days on which the full input data was unavailable for a given prediction, since model performance has been shown to be affected negatively if the full set of model inputs are not available.
	Model outputs could be combined with other parameters currently employed in abstraction-based risk management of <i>Cryptosporidium</i> (e.g. upstream sampling site data). A suitable risk matrix could then be used to aid decision-making.
	Automated daily emails of the combined outputs would be ideal.

Learning notes compiled by the researchers during the model trials recorded a number of challenges, the majority of which related to data download/availability issues, which only became apparent when the model was used in real-time. In particular, data platform downtimes (which affected the soil moisture and temperature datasets most dramatically) drew attention to the need to build a degree of redundancy into the model.

These adjustments fell into two types: (i) the addition of backup data sources for parameters wherever possible and (ii) the exclusion of features from the model entirely if data remained unavailable for a sustained period (3+ days).

Back-up data sources were added for rainfall and temperature (from multiple weather stations within the catchment), and discharge (based on the same gauging station data, but

accessed via a different online platform), and these platforms were accessed to populate the data frames where data gaps occurred. Finally, an alternative model was used which excluded soil moisture entirely (to be used during those periods when the latest soil moisture was unavailable for 3+ days). These adjustments were shown to improve daily predictions when data gaps occurred.

5.2.4 Operational implementation

Shortly after the trial, the DWI served Thames Water with a notice requesting that they complete the following measures by 31 Mar 2026 (DWI, 2025b):

*Design, develop and implement a system to predict *Cryptosporidium* oocyst concentrations to minimise the risk of *Cryptosporidium* oocysts being abstracted into the raw water storage reservoirs.*

The newly tested *Cryptosporidium* river forecasting tool provided a timely option for fulfilling this requirement, hence the task of implementing the tool began in August 2025, when an annotated version of the working code (for Model A) was shared with a computer science team contracted by Thames Water. This team have now adapted the model for use at two further sites, including one site ≈ 20 km upstream of the Walton intake, and another site in north London (on the River Lee).

In its final form, the model outputs will be combined with upstream sampling data and emailed to water quality and abstraction managers on a daily basis. Decision-making is to be risk matrix-based – i.e. where a certain number of indicators suggest the risk of elevated *Cryptosporidium* is high, the decision to suspend/reduce abstraction is given. The predicted data is presented as a results table similar to that shown in Fig. 5.1a, with absolute concentration values. Predictions from the previous ten days are provided alongside measured values, so that a judgement can be made regarding the recent performance of the model; this will then be factored into the risk matrix-based decision-making process.

The intention is to roll-out the tool for use at further sites in future. The three sites where it is currently being implemented account for $\approx 72\%$ of the total surface water abstracted to supply London's drinking water (equating to $\approx 55\%$ of London's overall supply).

5.3 Scientific implications, limitations and future directions

5.3.1 *Cryptosporidium* sources and environmental drivers

Of the three projects presented here, the nationwide *Cryptosporidium* study (Chapter 4) provides the best introduction to the *Cryptosporidium* challenge facing water companies in the UK and, potentially, other countries around the world. The data shows clearly that although *Cryptosporidium* is ubiquitous, its abundance in the environment can vary dramatically over space and time. The very low river concentrations found in low-population, highly rural areas of Scotland and some parts of the north of England, contrasted with the high concentrations found in the densely populated areas of the south, indicate that human sources must be given more attention in future. A number of previous studies have been unambiguous in pointing to humans as a major source of *Cryptosporidium* in raw waters (Medema and Schijven, 2001; Montemayor et al., 2005; Phillip et al., 2007; Nasser, 2016) – and in some cases, as the *primary* source (Vermeulen et al., 2019). Water companies which are responsible for wastewater and drinking water treatment are in a powerful position to effect change, since by upgrading their own STWs and drainage systems (reducing CSO spills), they can improve raw water quality and thereby reduce the need to also upgrade drinking WTWs in the same catchments. There are already wide-ranging and long-term plans to implement major upgrades in the UK and many of these could help reduce pathogen loadings and hence protect drinking water supplies (Defra, 2022).

The findings from Chapter 4 show that higher human population density is more associated with human-shed oocysts in rivers, whereas high percentage grassland area is more associated animal-shed oocysts. Catchment managers could use this information to assess whether oocysts found in rivers are likely to have a human or animal source, and apply that inference when allocating resources to improve river conditions. Given the strength of human population as a model predictor, the conservative option may be to assume that human sources are dominant in even moderately populated catchments. Seasonal patterns, such as the apparent overlap between autumn-peaking seasons and human-shed oocysts, could also be useful in enabling catchment managers to make judgements about the most dominant sources. Source proximity also appears to be important, as represented by the spatial analyses in both Chapters 2 and 4, with sources only a few km upstream likely to be

more impactful because the oocysts they release are less affected by dispersion and die-off than oocysts from sources further upstream (Medema and Schijven, 2001); this trend will not be universally true, however, and more distant sources have also been shown to be potentially important.

The ML models presented here represented the potential influence of CSOs on *Cryptosporidium* river concentrations in two ways: in Chapter 2, CSO EDM records were converted into daily binary spill inputs, while in Chapter 4, CSO density within each full or truncated catchment was included as a spatial input. Future models could incorporate CSO information in formats more likely to reflect the magnitude of the spill events. Possible options include using daily spill duration (e.g. spill minutes per day), or using local rainfall radar data in a secondary model to weight or scale spill days and spill durations according to rainfall depth or intensity (Bizer and Kirchhoff, 2022). The preferred solution would be to estimate spill volumes directly, but this usually requires knowledge of CSO asset dimensions and access to the level or pressure measurements upon which EDM records are based (Suslovaite et al., 2024). Although this is feasible in principle, such data is unavailable for most sites, making this approach unsuitable at present for catchment-scale or nationwide modelling.

The positive association of humidity with *Cryptosporidium* may reflect the role of moist conditions in extending oocyst persistence, but it will also be a consequence of broader seasonal effects, and could represent soil moisture or urban runoff-related impacts. This is the first time humidity has been identified as a powerful predictor for *Cryptosporidium*, and this knowledge could inform other models (including the daily prediction model, described in Chapter 2). Rainfall was confirmed to be an important driver of modelled *Cryptosporidium*, although this is only true if rainfall in the seven days prior to *Cryptosporidium* measurements is used to train models (monthly rainfall, for example, was a far weaker predictor). This result was observed in both Chapters 2 and 4. Unlike humidity, rainfall is a variable which acts most powerfully over the short-term, mobilising and transporting oocysts from source areas to water bodies.

Overall, the findings of the models presented in both Chapter 2 and Chapter 4 could be used to develop models in other locations around the world. They show the predictive utility of meteorological and land use data for estimating current *Cryptosporidium* concentrations

(i.e. “nowcasting”) – provided that the data is of sufficient resolution and quality. Moving beyond nowcasting, to predict *Cryptosporidium* one or more days in the future, may be feasible in some areas, but it is likely to depend on the complexity of catchment. The daily model presented here predicts *Cryptosporidium* in one of the most complex catchments in the UK; there would be much value and scope in repeating a similar approach in smaller settings, with less diversity of potential sources.

5.3.2 *Cryptosporidium* risk assessment

The QMRA study featured here presents just one type of scenario testing – namely, quantifying the effects of different abstraction suspension regimes on infection risk. The findings showed that, judged on mean annual risk, the median and Q95 *Cryptosporidium* risk to London’s drinking water was below the USEPA guideline (i.e. in the “tolerable risk” band). Seasonal risk was shown to be slightly above the guideline (at Q95) for some biannual periods, showing the value of assessing temporal variation in risk in situations where pathogen loadings are highly variable. A single hypothetical scenario (SAR-H) was presented here, but QMRA analysis could be applied more flexibly to address changing priorities. For instance, in preparation for times of drought, less stringent SARs could be modelled using hypothetical river concentration data (reflecting seasons of different severity) to determine how high thresholds can be raised before the change in relative risk becomes unacceptable.

Note that the risk estimates presented in Chapter 3 should be interpreted in the light of two important limitations. Firstly, the absolute magnitude of the QMRA outputs is sensitive to the dose-response data and model applied. Although the Iowa DR QMRA was retained as the primary analysis, the pooled DR QMRA analysis showed that the inclusion of dose-response data obtained using different isolates can increase upper-percentile risk estimates substantially. Consequently, the comparison with the USEPA guideline should be interpreted as conditional on dose-response assumptions, rather than as a fixed boundary between acceptable and unacceptable risk.

Secondly, the QMRA framework estimates risk associated with source water quality, storage, treatment and final treated water production (considered as a blended risk from multiple WTW outlets); it does not quantify contamination risks arising within the drinking water distribution system. The outputs should therefore be understood as estimates of risk

at the WTW outlets, rather than total end-consumer risk across the full source-to-tap system.

Thames Water is currently embarking on a major programme to add UV contactors to four of the five major WTWs discussed in this thesis. The project, which is funding-dependent, is intended to be completed over 10 years, beginning with installation of UV treatment at two WTWs in 2025-2030, followed by installation at the remaining two in 2030-2035 (Thames Water, 2023). The UV technology will play a role in mitigating risk in the years ahead, and could also increase water supply by justifying a relaxation in selective abstraction during periods of drought. The QMRA model could therefore inform the timing and location of the rollout by determining where the maximum risk reduction will occur in response to installation of UV technology. It could also be used to test other scenarios regarding the technology's design/operational use, such as the minimum UV inactivation rate required (e.g. 2-log or 4-log), or when in the year it should be applied (e.g. just when loadings begin to increase in the river, or when temperatures drop below a threshold believed to impact SSF performance).

5.3.3 ML modelling: strengths and limitations

A broader consideration arising from this work concerns the role of human judgement in ML-based modelling. The feature selection in the models presented in Chapters 2 and 4 was guided primarily by prior understanding of catchment processes and by data availability. This theory-informed approach has the benefit of constraining models to plausible inputs, but it can also lead to the inclusion of less predictive features and the potential exclusion of relevant features. While manual feature selection methods were applied in Chapter 2 to determine the most predictive lag times for catchment inputs (e.g. rainfall and discharge), an automated feature selection approach could have been taken in both the catchment-specific and nationwide *Cryptosporidium* models. Typically, such methods begin with a larger and more diverse set of initial inputs and then use recursive feature elimination (RFE) to obtain a more predictive subset of features (Bulut et al., 2025). A number of such tools exist which have been specifically applied to XGBoost, and which would therefore be suitable for use with the models described here – e.g. Recursive Feature Elimination with CV (Scikit-learn, 2026) and Boruta (Kursa and Rudnicki, 2011).

Because of the sensitivity of SHAP to multicollinearity in inputs, the use of automated RFE has the potential to improve model interpretation by removing redundant correlating features, and can support the identification of important effects which may not have been considered theoretically important (Aas et al., 2021). However, aggregation of SHAP outputs across related variables, as undertaken in both Chapters 2 and 4, partially addressed this issue by enabling broader environmental processes to be interpreted collectively.

The work also highlights both the strengths and limitations of ML relative to mechanistic and stochastic modelling approaches. The ML methods used here were well suited to integrating large heterogeneous datasets and identifying non-linear relationships without requiring mechanistic representation of pathogen transport processes (Razavi et al., 2022). However, the models depended on extensive historical training datasets collected over many years and were therefore fundamentally empirical. Such models perform more reliably when predicting within previously observed conditions than when extrapolating to novel scenarios, such as extreme climatic events, major infrastructure changes or future increases in urbanisation (Cuddington et al., 2013).

Mechanistic and stochastic approaches can be more effective when extrapolating beyond the range of the training data, since they are based, respectively, on process representation and uncertainty propagation, rather than on empirical pattern recognition (An et al., 2015). However, mechanistic/stochastic *Cryptosporidium* models capable of supporting the type of near real-time operational decision-making attempted in Chapter 2, or interpretation of environmental effects provided in Chapter 4, are currently unavailable (Smalley et al., 2025). ML models therefore provide a powerful method of generating operationally useful outputs directly from routinely collected monitoring and environmental data. Furthermore, for the intended applications, accurate prediction and system understanding of relatively common elevated river concentrations were more important than precise estimation of rare extreme events. This differs somewhat from applications such as hydrological flood forecasting, where prediction of extreme values may itself be the primary objective (Anik et al., 2025).

In this sense, the ML approaches presented here should not necessarily be viewed as replacements for mechanistic or stochastic models, but rather as complementary tools for exploiting high volumes of data in operational environments. Although the models are data-driven, domain knowledge has strongly influenced variable selection, interpretation of

environmental relationships and operational implementation. Future hybrid approaches, such as physics-informed neural networks (PINNs), which combine physical constraints with data-driven learning, may therefore offer a useful compromise between mechanistic realism, interpretability and predictive performance (Raissi et al., 2019; Karpatne et al., 2017).

5.3.4 Uncertainties in routine *Cryptosporidium* monitoring data

A central strength of this thesis is the use of routine water industry monitoring data at a scale rarely available for environmental *Cryptosporidium* research. The datasets provide long temporal and wide spatial coverage, and are of direct relevance to operational drinking water risk management. However, the same data carries uncertainties which affect the interpretation of model outputs. Firstly, concentration measurements are subject to variable analytical recovery, which can differ between laboratories, sites and samples over time. This uncertainty may not be random: samples influenced by sewage, runoff or high-particulate events could plausibly have different recovery efficiencies from cleaner samples, meaning that recovery bias may vary in response to the contamination processes being modelled (Feng et al., 2003). Similarly, individual laboratories or sites may exhibit consistent patterns of recovery, which could support exploratory correction or uncertainty modelling in future work, even where true sample-specific recovery data is unavailable. However, such corrections would need to be applied cautiously, since poorly constrained adjustments could introduce additional bias. In addition to recovery issues, different sites and samples will contain different proportions of viable oocysts and human-infectious *Cryptosporidium* species or genotypes (Swaffer et al., 2018).

Addressing these uncertainties at the modelling stage is challenging, although there is scope to modify future monitoring procedures to reduce uncertainty. For example, sample-specific recovery can be estimated by spiking each sample with labelled oocysts that can be differentiated from environmental oocysts. Products such as ColorSeed have been developed for this purpose (Warnecke et al., 2003; Franczy et al., 2004), but have yet to be widely adopted by the UK water sector. Likewise, regular molecular characterisation of a subset of positive samples can provide useful information on species/genotype composition at key abstraction points. Polymerase chain reaction (PCR)-based methods are already used

for *Cryptosporidium* genotyping in the water industry, and these can help assess the likely human-health relevance of detected oocysts by determining whether they belong to species/genotypes associated with human infection (Di Giovanni et al., 2010; UKWIR, 2020). PCR can also be combined with propidium monoazide staining to perform PMA-PCR-based viability assays (Brescia et al., 2009). Such methods do not definitively establish oocyst infectivity, and in most cases it may only be feasible to gather such data intermittently from positive samples or slides containing multiple oocysts, but catchment-specific information of this type would nevertheless reduce current levels of uncertainty associated with standard enumeration-based monitoring datasets (UKWIR, 2020).

A further limitation is that the monitoring data represents discrete samples rather than continuous observations, meaning that short-duration contamination events associated with CSO spills, elevated STW discharges or localised rainfall-runoff events may have been missed, particularly where such events occurred close to monitoring points. To some extent this issue is inherent to any grab sampling-based monitoring regime, particularly in the context of the long analysis times required for *Cryptosporidium*, which are a barrier to sub-daily sampling. A number of DNA-based methods, including quantitative PCR (qPCR), have been developed to detect and quantify *Cryptosporidium* more rapidly than the standard IMS-IFM procedure currently deployed by UK water companies (Hassan et al., 2021), but these methods are also subject to recovery uncertainties, and the conversion of a qPCR signal into a reliable oocyst count is non-trivial (Chalmers et al., 2021). Research is also underway to combine continuous flow set-ups with a fluorescent micro-bead technology which can bind selectively to oocysts (Ofwat, 2025b), but such methods are primarily intended to make current monitoring regimes more rapid and practicable, not to facilitate sub-daily monitoring.

The modelling techniques presented in Chapters 2 and 4 are not limited to *Cryptosporidium*, however – they could be applied to other contaminants, including chemical contaminants for which sub-daily monitoring may be available either by direct or proxy measurement. In the years ahead, the substantial expansion in the use of continuous monitoring (sonde) stations in UK river systems (Environment Act 2021, Section 82, 2021; see also Section 5.3.5) will equip modellers with multiple time-resolved datasets for each major catchment, increasing the potential to reliably capture, model and understand contamination events.

5.3.5 Future work

Several suggestions for specific future work concerning *Cryptosporidium* modelling and risk mitigation have already been made in the preceding discussion, but many of the methods described in this thesis could be readily applied to other locations and/or repeated for other contaminants of concern (including, but not limited to, other pathogens). Given the easy accessibility of remote sensing data and modelling software, the main limiting factor for future studies of this type would be the availability of the water quality monitoring datasets required to train models. However, as this work demonstrates, water companies possess considerable archives of potentially valuable data, the vast majority of which has not been applied to any purpose beyond its original monitoring remit. In the UK there has been an increasing move towards data-sharing as the default state – a change which is being pushed for by the regulator (Ofwat, 2021) – combined with the formation of the Open Data Framework, adopted by a group of water companies and other stakeholders (Stream, 2024), which is already making large operational/monitoring datasets available online. In many cases, the full value of this data will only be realised when water quality data from multiple water companies is made public by default, so that researchers have access to all that is available.

In addition to historical water quality data, access to continuous monitoring data is constantly expanding. At the catchment level, this includes the spill monitoring of $\approx 16,800$ CSOs in the UK (The Rivers Trust, 2025). New legislation (Environment Act 2021, Section 82, 2021) mandates the installation of water quality monitoring stations upstream and downstream of every CSO and STW discharge point in England, recording multiple parameters (e.g. dissolved oxygen and ammonium) at least every 15 mins. Full rollout is intended to be completed by 2035, with high priority sites addressed first. The overall effect could be transformative for the environmental and drinking water sectors, reducing the reliance on spot-sampling and transitioning to a continuous environmental intelligence model, which would allow for real-time responses to events. Such opportunities will only be realised, however, if automated data processing, analysis and modelling methods keep pace with the data deluge. Data-driven, AI-based approaches to catchment modelling of the kind presented here are therefore likely to play an increasingly important role in both research and management decision-making.

Chapter 6: Conclusions

The work contained within this thesis has applied long-term water quality monitoring data, environmental data and a variety of modelling methods to increase current understanding of *Cryptosporidium* in rivers and provide useful tools for (i) predicting *Cryptosporidium* river concentrations, (ii) identifying high-risk areas, (iii) understanding environmental sources and drivers of the pathogen, and (iv) quantifying the risk the pathogen poses to populations via drinking water.

The main research questions have been addressed as described below.

6.1 Research Question 1

*What is the potential to develop data-led predictive models to provide warning of periods of high *Cryptosporidium* at daily resolution in a large and complex catchment (such as the Thames)? Further, how can such a model be interpreted to understand key sources and transport mechanisms?*

A first-of-its-kind ML-based model has been developed, providing operational predictions of *Cryptosporidium* at daily resolution in the large and complex Thames Catchment, and thereby demonstrating the potential of ML approaches to predict and explain previously intractable processes, such as *Cryptosporidium* mobilisation and transport. The model is now in the process of being implemented for operational use at three sites serving London, with the possibility of expanding its application to other sites in the Thames valley, where it would contribute to protecting public water supplies serving Oxford, Reading and other urban centres to the west of London.

The study advances current understanding by showing the dominance of rainfall-driven sources within ≈ 30 km upstream distance of monitoring points, as evidenced by both the primacy of 2-day lagged rainfall as a predictor of peaks and by the SHAP analysis of the most impactful rainfall locations. Sources from more distant areas of the catchment (> 100 km river distance) were also shown to be important, however, albeit to a lesser extent.

Provided monitoring data is available to train the models, the approaches used here could be adapted for other pathogens and/or chemical contaminants wherever there is a need to anticipate peaks in river concentrations and/or identify source areas or environmental factors. In this study, homogeneity of rainfall events was a significant challenge when trying to connect source areas to *Cryptosporidium* peaks in the river, but in catchments containing more distinct microclimate effects, the spatial analysis method applied would be a powerful tool for source attribution.

6.2 Research Question 2

How can monitoring datasets and asset information be applied to better understand and quantify Cryptosporidium risk to drinking water supplies serving large population centres, and how can such understanding be used to inform effective management strategies?

A dual QMRA approach was developed to analyse risk in the final treated supply of a complex multi-source system in a major city, and to consider the impact of management strategies. Novel data treatment processes have been applied to generate concentration distributions from values below LODs, using a hybrid multiple-imputation method bounded by a parametric lower limit – an approach which allowed for periods containing only left-censored data to be modelled. When analysed across the full study period, and applying current management practices, both treated and raw water model outputs showed that the Q95 risk levels remained below the USEPA guideline of 1 infection per 10,000 people per year under the primary Iowa DR QMRA. However, assessment across biannual periods revealed that the Q95 *Cryptosporidium* risk to London's drinking water was marginally above the USEPA guideline during some winters, and the pooled DR QMRA outputs showed that Q95 risk could rise to 1 infection per 1,000 people per year if a broader range of infectivity rates was considered.

This analysis demonstrates the impact of catchment processes on risk and underlines the importance of including raw water quality data in risk assessments. It also demonstrates that selective abstraction can provide a measurable but modest reduction in modelled risk, while the method provides a readily adaptable tool for comparing future abstraction management and treatment scenarios, including higher-impact interventions, such as UV disinfection.

6.3 Research Question 3

*How can wider environmental datasets be used to understand the key drivers, sources and environmental factors that influence the spatial and temporal variability in observed source water *Cryptosporidium*?*

A multi-catchment analysis has been conducted for the British mainland, applying monitoring data collected over 20+ years, using standardised sampling and analysis methods, to study geographical spread, seasonality and the major environmental factors affecting *Cryptosporidium* river concentrations. A key novelty of this work was the use of both static (land use/land characteristics) and dynamic (meteorological/hydrological) datasets in an ML model trained with mean monthly *Cryptosporidium* data, allowing for the recognition of patterns which – because of the highly sporadic nature of both *Cryptosporidium* and rainfall, and the potential for complex interactions of environmental factors – would be hidden or obscured when only assessing data averaged over the long-term.

The study identified clear findings about the geographical spread of *Cryptosporidium*, the relative importance of human vs animal sources and the key environmental drivers (with humidity and rainfall found to be the most important). A significant finding of the XAI model was that 33% of sites are human source-dominated (based on specific thresholds applied in the study) and that high human population is a stronger predictor of high *Cryptosporidium* concentrations than any of the livestock density/rural land use features. Findings such as these underline the need for investments in sewage treatment and storm/sewer network capacity to reduce overall pathogen loadings to rivers.

The approach taken could be readily applied to other contaminants for which similar monitoring datasets are available, with the potential for significant gains in our understanding of contaminant sources and environmental influences.

6.4 General advances in understanding

Further to the above, in carrying out this work, our general understanding of *Cryptosporidium* risk and behaviour in catchment systems has been advanced over the current state of the art. Key contributions are:

- The most important input for predicting *Cryptosporidium* peak timing at daily resolution is rainfall, but models need to be tested to find the most predictive lag time, which (based on observations in models presented here, and in other trial models, not presented) will be within a window of 0-7 days (with 1-4 days the most likely). Many studies (e.g. Ligda et al., 2020), have tested mean weekly/monthly rainfall as a predictor, often because temporal resolution in the data is low (e.g. monthly sampling). This is the first study to apply large daily and sub-weekly *Cryptosporidium* datasets to this issue, and hence the first study able to bring this level of rigour to analysing *Cryptosporidium*-rainfall associations
- Source inputs at shorter upstream distances were generally more predictive of elevated *Cryptosporidium* than more distant source inputs or whole-catchment averages, suggesting that source proximity and hydrological connectivity are important controls (supported by both the daily *Cryptosporidium* model in Chapter 2 and the nationwide model in Chapter 4).
- Many studies incorporate livestock density data into their estimates/models for *Cryptosporidium* – this remains a valid approach, but where the source data is questionable, more easily quantifiable alternatives (such as grassland area) may generate more accurate outputs.
- While the seasonality of *Cryptosporidium* in rivers has been widely reported, we now have evidence that this seasonality may be driven by a complex array of factors, including land use, land characteristics and the prevailing weather. Winter-peaking sites are more often in urban or arable areas, with lower humidity, lower rainfall and higher air temperature, whereas summer-peaking sites are in more upland and heathland areas. Autumn-peaking sites tend to occupy a midpoint between these two extremes.

Overall, these findings and the methods used in this thesis demonstrate the latent value within historical monitoring datasets – value which can be exploited by researchers to advance knowledge of water quality issues and challenges, particularly when combined with remote sensing data and the ever-expanding array of AI-based analysis and modelling tools. These conclusions should be tempered by an awareness of the potential limitations of

routine monitoring data, including variable analytical recovery, uncertainty in viability and human infectivity (when dealing with pathogen data), and the fact that discrete sampling can miss short-duration contamination events.

A key priority of the water sector must be to consider how data can better be used and exploited, not just for regulatory purposes, but also to develop greater understanding of processes, leading to improved decision-making and risk management strategies. The availability of catchment data is expected to explode in the coming years as the industry continues to expand its monitoring networks, and as other remote sensing technologies advance and become more accessible. Continuing research is therefore required to determine how AI-based tools can be applied to analyse this data. Although *Cryptosporidium* has been the focus of the research described in this thesis, the potential to exploit large datasets to enhance our understanding of a whole range of water quality issues, including contamination causes and sources, has never been greater, and will only increase in the years to come.

Page intentionally left blank

References

- Aas, K., Jullum, M. and Løland, A. (2021) 'Explaining individual predictions when features are dependent: More accurate approximations to Shapley values', *Artificial Intelligence*, 298, pp. 103502. <https://doi.org/10.1016/j.artint.2021.103502>
- Alomari, Y. and Andó, M., (2024) 'SHAP-based insights for aerospace PHM: Temporal feature importance, dependencies, robustness, and interaction analysis', *Results in Engineering*, 21, 101834. <https://doi.org/10.1016/j.rineng.2024.101834>
- Alpaydin, E., (2014) *Introduction to Machine Learning*. MIT Press, 547-591.
<http://ieeexplore.ieee.org/sheffield.idm.oclc.org/document/6917149>
- Alsahaf, A., Petkov, N., Shenoy, V. and Azzopardi, G. (2022) 'A framework for feature selection through boosting', *Expert Systems With Applications*, 187, 115895.
<https://doi.org/10.1016/j.eswa.2021.115895>
- An, D., Kim, N. H. and Choi, J.-H. (2015) 'Practical options for selecting data-driven or physics-based prognostics algorithms with reviews', *Reliability Engineering & System Safety*, 133, pp. 223–236. <https://doi.org/10.1016/j.ress.2014.09.014>
- Anik, M. S. B. M., An, C. and Li, S. S. (2025) 'Evolution from the physical process-based approaches to machine learning approaches to predicting urban floods: a literature review', *Environmental Systems Research*, 14(1), pp. 15.
<https://doi.org/10.1186/s40068-025-00409-3>
- Antweiler, R. C. and Taylor, H. E. (2008) 'Evaluation of Statistical Treatments of Left-Censored Environmental Data using Coincident Uncensored Data Sets: I. Summary Statistics', *Environmental Science & Technology*, 42(10), pp. 3732–3738.
<https://doi.org/10.1021/es071301c>
- APHA (2022) *Livestock Demographic Data Group*. Animal & Plant Health Agency. Available at: <https://www.gov.uk/guidance/view-apha-surveillance-reports-publications-and-data#disease-surveillance-dashboards> (Accessed: 8 February 2025)

- APHA (2024) *APHA animal surveillance reports, publications and dashboards*: Animal and Plant Health Agency. Available at: <https://www.gov.uk/guidance/view-apha-surveillance-reports-publications-and-data#disease-surveillance-dashboards> (Accessed: 8 February 2025)
- Asfaw, A., Maher, K. and Shucksmith, J. D. (2018) 'Modelling of metaldehyde concentrations in surface waters: A travel time based approach', *Journal of Hydrology*, 562, 397-410. <https://doi.org/10.1016/j.jhydrol.2018.04.074>
- Atherholt, T. B., LeChevallier, M. W., Norton, W. D. and Rosen, J. S., 1(998) 'Effect of rainfall on Giardia and Crypto', *Journal of the American Water Works Association*, 90 (9), 66–80. <http://dx.doi.org/10.1002/j.1551-8833.1998.tb08499.x>
- Badenoch, J. (1990) *Cryptosporidium in water supplies. Report of the group of experts*. HMSO.
- Badenoch, J. (1995) *Cryptosporidium in water supplies: second report of the group of experts*. HMSO.
- Bataiero, M. O., Araujo, R., Nardocci, A. C., Matté, M. H., Sato, M., Lauretto, M. d. S. and Razzolini, M. T. P. (2019) 'Quantification of Giardia and Cryptosporidium in surface water: A risk assessment and molecular characterization', *Water Supply*, 19(6), pp. 1823–1830. <https://doi.org/10.2166/ws.2019.059>
- BBC News (2024) 'Water company confirms 'reduction' in water parasite', 2024/05/29. Available at: <https://www.bbc.co.uk/news/articles/c1994ww203ro> (Accessed: 25 May 2026).
- Betancourt, W. Q. and Rose, J. B. (2004) 'Drinking water treatment processes for removal of Cryptosporidium and Giardia', *Veterinary Parasitology*, 126(1), pp. 219–234. <https://doi.org/10.1016/j.vetpar.2004.09.002>
- Bhattarai, R., Kalita, P., Trask, J. and Kuhlenschmidt, M. S. (2011) 'Development of a physically-based model for transport of Cryptosporidium parvum in overland flow', *Environmental Modelling & Software: with Environment Data News*, 26(11), pp. 1289–1297. <https://doi.org/10.1016/j.envsoft.2011.05.011>

- Bisong, E. (2019) *Building Machine Learning and Deep Learning Models on Google Cloud Platform*. Apress Berkeley, CA. Available at: <https://doi.org/10.1007/978-1-4842-4470-8>
- Bizer, M. A. and Kirchhoff, C. J. (2022) 'Regression Modeling of Combined Sewer Overflows to Assess System Performance', *Water Science & Technology*, 86(11), pp. 2848–2860. <https://doi.org/10.2166/wst.2022.362>
- Boorman, D. B., Hollis, J M and Lilly, A (1995) *Hydrology of soil types: a hydrologically based classification of the soils of the United Kingdom*: Institute of Hydrology Report No. 126. 137pp. Available at: https://nora.nerc.ac.uk/id/eprint/7369/1/IH_126.pdf (Accessed: 5 Feb 2025).
- Bouchier, I. (1998) *Cryptosporidium in water supplies*. Third Report of the Group of Experts to Department of the Environment, Transport and the Regions and Department of Health.
- Bouزيد, M., Hunter, P. R., Chalmers, R. M. and Tyler, K. M. (2013) Cryptosporidium Pathogenicity and Virulence. *Clinical Microbiology Reviews*, 26 (1), 115-134. <https://doi.org/10.1128/cmr.00076-12>
- Brescia, C. C., Griffin, S. M., Ware, M. W., Varughese, E. A., Egorov, A. I. and Villegas, E. N. (2009) 'Cryptosporidium propidium monoazide-PCR, a molecular biology-based technique for genotyping of viable Cryptosporidium oocysts', *Applied and Environmental Microbiology*, 75(21), pp. 6856–6863. <https://doi.org/10.1128/AEM.00540-09>
- Brion, G. M., Neelakantan, T. R. and Lingireddy, S. (2001) USING NEURAL NETWORKS to predict peak *Cryptosporidium* concentrations. *Journal of the American Water Works Association*, 93 (1), 99-105. <https://doi.org/10.1002/j.1551-8833.2001.tb09103.x>
- Bukhari, Z., Smith, H. V., Sykes, N., Humphreys, S. W., Paton, C. A., GirdwoOd, R. W. A. and Fricker, C. R. (1997) 'Occurrence of Cryptosporidium spp oocysts and Giardia spp cysts in sewage influents and effluents from treatment plants in England', *Water Science & Technology*, 35(11-12), pp. 385–390. [https://doi.org/10.1016/S0273-1223\(97\)00290-4](https://doi.org/10.1016/S0273-1223(97)00290-4)

- Bulut, O., Tan, B., Mazzullo, E. and Syed, A. (2025) 'Benchmarking variants of recursive feature elimination: insights from predictive tasks in education and healthcare', *Information*, 16(6), pp. 476. <https://doi.org/10.3390/info16060476>
- Canales, R. A., Wilson, A. M., Pearce-Walker, J. I., Verhougstraete, M. P. and Reynolds, K. A. (2018) 'Methods for handling left-censored data in quantitative microbial risk assessment', *Applied and Environmental Microbiology*, 84(20), pp. e01203–18. <https://doi.org/10.1128/AEM.01203-18>
- Chan, J. Y.-L., Leow, S. M. H., Bea, K. T., Cheng, W. K., Phoong, S. W., Hong, Z.-W. and Chen, Y.-L. (2022) 'Mitigating the multicollinearity problem and its machine learning approach: a review', *Mathematics* 10(8), 1283. <https://doi.org/10.3390/math10081283>
- Chan, S. C., Kendon, E. J., Fowler, H. J., Youngman, B. D., Dale, M. and Short, C. (2023) 'New extreme rainfall projections for improved climate resilience of urban drainage systems', *Climate Services*, 30, pp. 100375.
- Chalmers, R. M., Elwin, K. and Robinson, G. (2021) *Use of the Polymerase Chain Reaction (PCR) for the Analysis and Enumeration of Cryptosporidium Oocysts in Drinking Water* [Final report], United Kingdom: Drinking Water Inspectorate (DWI 70/2/332. Available at: https://dwi-production-files.s3.eu-west-2.amazonaws.com/wp-content/uploads/2021/10/20121525/DWI70_2_332.pdf (Accessed: 4 May 2026).
- Chen, J., Chen, S., Fu, R., Li, D., Jiang, H., Wang, C., Peng, Y., Jia, K. and Hicks, B. J. (2022) 'Remote Sensing Big Data for Water Environment Monitoring: Current Status, Challenges, and Future Prospects', *Earth's Future* 10(2), e2021EF002289. <https://doi.org/10.1029/2021EF002289>
- Chen, T. and Guestrin, C. (2016) 'XGBoost: A Scalable Tree Boosting System', *arXiv.org*. <https://doi.org/10.48550/arXiv.1603.02754>
- Choubin, B., Darabi, H., Rahmati, O., Sajedi-Hosseini, F. and Kløve, B. (2018) 'River suspended sediment modelling using the CART model: A comparative study of machine learning techniques', *Science of the Total Environment*, 615, 272-281. <https://doi.org/10.1016/j.scitotenv.2017.09.293>

- Clancy, J. L. and Hunter, P. R. (2004) 'Monitoring of Giardia and Cryptosporidium in Water in the UK and US', in Sterling, C.R. and Adam, R.D. (eds.) *The Pathogenic Enteric Protozoa: Giardia, Entamoeba, Cryptosporidium and Cyclospora*. Boston, MA: Springer US, pp. 129–139. https://doi.org/10.1007/1-4020-7878-1_10
- Coffey, R., Cummins, E., Flaherty, V. O. and Cormican, M. (2010) 'Analysis of the soil and water assessment tool (SWAT) to model *Cryptosporidium* in surface water sources', *Biosystems Engineering*, 106 (3), 303-314.
<http://dx.doi.org/10.1016/j.biosystemseng.2010.04.003>
- Cook, C., Prystajec, N., Ngueng Feze, I., Joly, Y., Dunn, G., Kirby, E., Özdemir, V. and Isaac-Renton, J. (2013) 'A comparison of the regulatory frameworks governing microbial testing of drinking water in three Canadian provinces', *Canadian Water Resources Journal*, 38(3), pp. 185–195. <https://doi.org/10.1080/07011784.2013.822186>
- Crockett, C. S. (2004) *The significance of streambed sediments as a reservoir of Cryptosporidium oocysts*. PhD Thesis. Drexel University.
- Cuddington, K., Fortin, M. J., Gerber, L. R., Hastings, A., Liebhold, A., O'Connor, M. and Ray, C. (2013) 'Process-based models are required to manage ecological systems in a changing world', *Ecosphere*, 4(2), pp. art20. <https://doi.org/10.1890/ES12-00178.1>
- Daniels, M. E., Smith, W. A. and Jenkins, M. W. (2018) 'Estimating Cryptosporidium and Giardia Disease Burdens for Children Drinking Untreated Groundwater in a Rural Population in India', *PLOS Neglected Tropical Diseases*, 12(1), pp. e0006231.
<https://doi.org/10.1371/journal.pntd.0006231>
- Daraei, H., Oliveri Conti, G., Sahlabadi, F., Thai, V. N., Gholipour, S., Turki, H., Fakhri, Y., Ferrante, M., Moradi, A. and Mousavi Khaneghah, A. (2020) 'Prevalence of Cryptosporidium spp. in water: a global systematic review and meta-analysis', *Environmental Science and Pollution Research*, 28 (8), 9498-9507.
<https://doi.org/10.1007/s11356-020-11261-6>
- Defra (2022) *Storm Overflows Discharge Reduction Plan*, London, England: UK Government. Available at: <https://www.gov.uk/government/publications/storm-overflows-discharge-reduction-plan>

Defra (2025) *Hydrology Data Explorer*: UK Government. Available at:

<https://environment.data.gov.uk/hydrology/explore>

Dehotin, J. and Braud, I. (2008) 'Which spatial discretization for distributed hydrological models? Proposition of a methodology and illustration for medium to large-scale catchments', *Hydrology and Earth System Sciences*, 12(3), pp. 769–796.

<https://doi.org/10.5194/hess-12-769-2008>

Deng, X., Ye, A., Zhong, J., Xu, D., Yang, W., Song, Z., Zhang, Z., Guo, J., Wang, T., Tian, Y., Pan, H., Zhang, Z., Wang, H., Wu, C., Shao, J. and Chen, X. (2022) 'Bagging–XGBoost algorithm based extreme weather identification and short-term load forecasting model', *Energy Reports*, 8, 8661–8674. <https://doi.org/10.1016/j.egy.2022.06.072>

Dhal, P. and Azad, C. (2022) 'A comprehensive survey on feature selection in the various fields of machine learning', *Applied Intelligence*, 52 (4), 4543–4581.

<https://doi.org/10.1007/s10489-021-02550-9>

Dickens, W. and Bensted, I. (1988) 'LONDON WATER RING MAIN', *Proceedings of the Institution of Civil Engineers*, 84(3), pp. 445–474.

<https://doi.org/10.1680/iicep.1988.1463>

Di Giovanni, G. D., Hoffman, R. M. and Sturbaum, G. D. (2010) *Cryptosporidium Genotyping Method for Regulatory Microscope Slides*. Water Research Foundation. Available at:

<https://dwi-production-files.s3.eu-west-2.amazonaws.com/wp-content/uploads/2020/10/27111105/DWI70-2-247exsum.pdf>

Dobson, B. and Mijic, A. (2020) 'Protecting rivers by integrating supply-wastewater infrastructure planning and coordinating operational decisions', *Environmental Research Letters*, 15 (11), 114025. <http://dx.doi.org/10.1088/1748-9326/abb050>

Dormann, C. F., Calabrese, J. M., Guillera-Aroita, G., Matechou, E., Bahn, V., Bartoń, K., Beale, C. M., Ciuti, S., Elith, J., Gerstner, K., Guelat, J., Keil, P., Lahoz-Monfort, J. J., Pollock, L. J., Reineking, B., Roberts, D. R., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Wood, S. N., Wüest, R. O. and Hartig, F. (2018) 'Model averaging in ecology: a review of Bayesian, information-theoretic, and tactical approaches for predictive inference', *Ecological Monographs*, 88(4r.), 485–504.

<https://doi.org/10.1002/ecm.1309>

- Dorner, S. M., Anderson, W. B., Slawson, R. M., Kouwen, N. and Huck, P. M. (2006) 'Hydrologic Modeling of Pathogen Fate and Transport', *Environmental Science & Technology*, 40(15), pp. 4746–4753. <http://dx.doi.org/10.1021/es060426z>
- Drobnič, F., Kos, A. and Pustišek, M. (2020) 'On the interpretability of machine learning models and experimental feature selection in case of multicollinear data', *Electronics (Basel)*, 9 (5), 761. <http://dx.doi.org/10.3390/electronics9050761>
- Drummond, J. D., Boano, F., Atwill, E., Li, X., Harter, T. and Packman, A. (2018) 'Cryptosporidium oocyst persistence in agricultural streams—a mobile-immobile model framework assessment', *Scientific Reports*, 8 (1), 4603. <https://doi.org/10.1038/s41598-018-22784-x>
- DuPont, H. L., Chappell, C. L., Sterling, C. R., Okhuysen, P. C., Rose, J. B. and Jakubowski, W. (1995) 'The Infectivity of *Cryptosporidium parvum* in Healthy Volunteers', *The New England Journal of Medicine*, 332(13), pp. 855–859. <https://doi.org/10.1056/nejm199503303321304>
- DWI (1998) *Cryptosporidium: Development of a Risk Assessment*, London: UK Drinking Water Inspectorate. Available at: <https://dwi-production-files.s3.eu-west-2.amazonaws.com/wp-content/uploads/2020/10/27110712/dwi0807.pdf>
- DWI (2008) *National Tap Water Consumption Study: Phase Two Final Report [Report No. DWI 70/2/217]*, London: Drinking Water Inspectorate. Available at: <https://www.dwi.gov.uk/research/completed-research/consumer/national-tap-water-consumption-survey-for-england-and-wales/>
- DWI (2021) 'Guidance on implementing the Water Supply (Water Quality) Regulations'.
- DWI (2022) *Drinking Water 2021. The Chief Inspector's report for drinking water in England*. Available at: https://dwi-content.s3.eu-west-2.amazonaws.com/wp-content/uploads/2022/08/19153555/E02750078_DWI-Public-water_England_V07.pdf (Accessed: 5 February 2025).
- DWI (2025a) *Cryptosporidium – Review of the Reports of the Group of Experts, Literature and Events*, London: Drinking Water Inspectorate. Available at: <https://www.dwi.gov.uk/research/completed-research/cryptosporidium->

[giardia/cryptosporidium-review-of-the-reports-of-the-group-of-experts-literature-and-events/](#)

DWI (2025b) *Thames Water Utilities Limited – AMP8 Hampton Works Cryptosporidium Notice (TMS-2023-00011 v3)*, UK: Drinking Water Inspectorate. Available at: <https://www.dwi.gov.uk/water-companies/improvement-programmes/thames-water-improvement-programmes/tms-2023-00011-v1/>

Edzwald, J. K., Tobiason, J. E., Dunn, H., Kaminski, G. and Galant, P. (2001) 'Removal and fate of Cryptosporidium in dissolved air drinking water treatment plants', *Water Science & Technology*, 43(8), pp. 51–7. <https://doi.org/10.2166/wst.2001.0463>

Efstratiou, A., Ongerth, J. E. and Karanis, P. (2017a) 'Waterborne transmission of protozoan parasites: Review of worldwide outbreaks - An update 2011–2016', *Water Research*, 114, 14-22. <https://doi.org/10.1016/j.watres.2017.01.036>

Efstratiou, A., Ongerth, J. E. and Karanis, P. (2017b) 'Evolution of monitoring for Giardia and Cryptosporidium in water', *Water Research*, 123, 96-112. <https://doi.org/10.1016/j.watres.2017.06.042>

Eisenberg, J. N., Seto, E. Y., Colford Jr, J. M., Olivieri, A. and Spear, R. C. (1998) 'An analysis of the Milwaukee cryptosporidiosis outbreak based on a dynamic model of the infection process', *Epidemiology*, 9 (3), 255–263. <https://www.jstor.org/stable/3703054>

Environment Act 2021, Section 82 (2021) [legislation.gov.uk](https://www.legislation.gov.uk). Available at: <https://www.legislation.gov.uk/ukpga/2021/30/section/82> (Accessed: 5 February 2025).

Environment Agency (2010) *The Microbiology of Drinking Water (2010) - Part 14 - Methods for the isolation, identification and enumeration of Cryptosporidium oocysts and Giardia cysts*, Bristol: Environment Agency. Available at: https://assets.publishing.service.gov.uk/media/5be9956aed915d6a0d6f6fe8/Part_14-oct20-234.pdf (Accessed: 5 February 2025)

Environment Agency (2025) *National Framework for Water Resources 2025: Water for Growth, Nature and a Resilient Future*, Bristol, UK: Environment Agency. Available at:

<https://www.gov.uk/government/publications/national-framework-for-water-resources-2025-water-for-growth-nature-and-a-resilient-future>

Environmental Tracing Systems Ltd (1998) *River Thames Catchment Time-of-Travel Studies*. [Unpublished.]

European Union (2020) *Directive (EU) 2020/2184 of the European Parliament and of the Council of 16 December 2020 on the quality of water intended for human consumption (recast)*. Brussels: European Union.

Ezzati, G., Kyllmar, K. and Barron, J. (2023) 'Long-term water quality monitoring in agricultural catchments in Sweden: Impact of climatic drivers on diffuse nutrient loads', *Science of The Total Environment*, 864, pp. 160978.

<https://doi.org/10.1016/j.scitotenv.2022.160978>

Feng, Y. Y., Ong, S. L., Hu, J. Y., Song, L. F., Tan, X. L. and Ng, W. J. (2003) 'Effect of particles on the recovery of *Cryptosporidium* oocysts from source water samples of various turbidities', *Applied and Environmental Microbiology*, 69(4), pp. 1898–1903.

<https://doi.org/10.1128/AEM.69.4.1898-1903.2003>

Ferguson, C. M., Croke, B., Ashbolt, N. J. and Deere, D. A. (2005) 'A deterministic model to quantify pathogen loads in drinking water catchments: Pathogen budget for the Wingecarribee', *Water Science & Technology*, 52 (8), 191-197.

<http://dx.doi.org/10.2166/wst.2005.0262>

Francy, D. S., Simmons III, O. D., Ware, M. W., Granger, E. J., Sobsey, M. D. and Schaefer III, F. W. (2004) 'Effects of seeding procedures and water quality on recovery of *Cryptosporidium* oocysts from stream water by using US Environmental Protection Agency Method 1623', *Applied and Environmental Microbiology*, 70(7), pp. 4118–4128.

<https://doi.org/10.1128/AEM.70.7.4118-4128.2004>

FWR (2011) *Cryptosporidium in Water Supplies*. Ref. No. FR/R0005. Foundation for Water Research. Available at: <https://fwr.org/publication/cryptosporidium-in-water-supplies> (Accessed: 5 February 2025).

- Gan, L., Zikry, T. M. and Allen, G. I. (2025) 'Are machine learning interpretations reliable? A stability study on global interpretations', *arXiv preprint arXiv:2505.15728*.
<https://doi.org/10.48550/arXiv.2505.15728>
- Giakoumis, T. and Voulvoulis, N. (2023) 'Combined sewer overflows: relating event duration monitoring data to wastewater systems' capacity in England', *Environmental Science: Water Research & Technology*, 9 (3), 707-722. <http://dx.doi.org/10.1039/D2EW00637E>
- Gibson III, C. J., Stadterman, K. L., States, S. and Sykora, J. (1998) 'Combined sewer overflows: A source of Cryptosporidium and Giardia?', *Water Science & Technology*, 38(12), pp. 67–72. [https://doi.org/10.1016/S0273-1223\(98\)00802-6](https://doi.org/10.1016/S0273-1223(98)00802-6)
- Golomazou, E., Mamedova, S., Eslahi, A. V. and Karanis, P. (2024) 'Cryptosporidium and agriculture: A review', *Science of the Total Environment*, 916, 170057.
<https://doi.org/10.1016/j.scitotenv.2024.170057>
- Guzmán Barragán, B. L., de Souza Lauretto, M., Razzolini, M. T. P., Nardocci, A. C. and Marimón Sibaja, K. V. (2022) 'Assessment microbial risks for the presence of Cryptosporidium spp. and Giardia spp. based on the surveillance of the water supply systems in Colombia, 2014–2018', *Water Environment Research*, 94(8), pp. e10776.
<https://doi.org/10.1002/wer.10776>
- Haas, C. N., Crockett, C. S., Rose, J. B., Gerba, C. P. and Fazil, A. M. (1996) 'Assessing the risk posed by oocysts in drinking water', *American Water Works Association Journal*, 88(9), pp. 131. <https://doi.org/10.1002/j.1551-8833.1996.tb06619.x>
- Haas, C. N., Rose, J. B. and Gerba, C. P. (1999) *Quantitative Microbial Risk Assessment*. New York: John Wiley & Sons. <https://doi.org/10.1002/9781118910030>
- Hamilton, K. A., Waso, M., Reyneke, B., Saeidi, N., Levine, A., Lalancette, C., Besner, M. C., Khan, W. and Ahmed, W. (2018) 'Cryptosporidium and Giardia in Wastewater and Surface Water Environments', *Journal of Environmental Quality*, 47(5), pp. 1006–1023.
<https://doi.org/10.2134/jeq2018.04.0132>
- Hamilton, K. A., Ciol Harrison, J., Mitchell, J., Weir, M., Verhougstraete, M., Haas, C. N., Nejadhashemi, A. P., Libarkin, J., Gim Aw, T., Bibby, K., Bivins, A., Brown, J., Dean, K., Dunbar, G., Eisenberg, J. N. S., Emelko, M., Gerrity, D., Gurian, P. L., Hartnett, E., Jahne,

- M., Jones, R. M., Julian, T. R., Li, H., Li, Y., Gibson, J. M., Medema, G., Meschke, J. S., Mraz, A., Murphy, H., Oryang, D., Owusu-Ansah, E. d.-G. J., Pasek, E., Pradhan, A. K., Razzolini, M. T. P., Ryan, M. O., Schoen, M., Smeets, P. W. M. H., Soller, J., Solo-Gabriele, H., Williams, C., Wilson, A. M., Zimmer-Faust, A., Alja'fari, J. and Rose, J. B. (2024) 'Research gaps and priorities for quantitative microbial risk assessment (QMRA)', *Risk Analysis*, 44(11), pp. 2521–2536.
<https://onlinelibrary.wiley.com/doi/abs/10.1111/risa.14318>
- Hassan, E. M., Örmeci, B., Derosa, M. C., Dixon, B. R., Sattar, S. A. and Iqbal, A. (2021) 'A review of *Cryptosporidium* spp. and their detection in water', *Water Science & Technology*, 83(1), pp. 1–25. <http://dx.doi.org/10.2166/wst.2020.515>
- Hastie, T., Friedman, J. H. and Tibshirani, R. (2009) *The elements of statistical learning : data mining, inference, and prediction*. 2nd edition. New York: New York : Springer.
<https://doi.org/10.1007/978-0-387-84858-7>
- Helsel, D. R. (2012) *Statistics for censored environmental data using Minitab and R*. Wiley series in statistics in practice, 2nd edn. Hoboken, N.J: Wiley.
<https://doi.org/10.1002/9781118162729>
- Helsel, D. R. and Cohn, T. A. (1988) 'Estimation of descriptive statistics for multiply censored water quality data', *Water Resources Research*, 24(12), pp. 1997–2004.
<https://doi.org/10.1029/WR024i012p01997>
- Hewett, P. and Ganser, G. H. (2007) 'A Comparison of Several Methods for Analyzing Censored Data', *The Annals of Occupational Hygiene*, 51(7), pp. 611–632.
<https://doi.org/10.1093/annhyg/mem045>
- Hijnen, W. and Medema, G. (2007) 'Elimination of Micro-organisms by Drinking Water Treatment Processes: A Review', *Water Intelligence Online*, 9.
<https://doi.org/10.2166/9781780401584>
- Hofstra, N. and Vermeulen, L. C. (2016) 'Impacts of population growth, urbanisation and sanitation changes on global human *Cryptosporidium* emissions to surface water', *International Journal of Hygiene and Environmental Health*, 219 (7), 599-605.
<https://doi.org/10.1016/j.ijheh.2016.06.005>

- Hoxie, N. J., Davis, J. P., Vergeront, J. M., Nashold, R. D. and Blair, K. A. (1997) 'Cryptosporidiosis-associated mortality following a massive waterborne outbreak in Milwaukee, Wisconsin', *American Journal of Public Health*, 87(12), pp. 2032–2035.
<https://doi.org/10.2105/ajph.87.12.2032>
- Ikiroma, I. A. and Pollock, K. G. (2021) 'Influence of weather and climate on cryptosporidiosis – A review', *Zoonoses Public Health*, 68(4), pp. 285–298.
<https://doi.org/10.1111/zph.12785>
- International Water Association (2020) *Digital Water: Artificial Intelligence — Solutions for the Water Sector*, London: International Water Association. Available at: https://iwa-website-assets.s3.eu-west-2.amazonaws.com/IWA_2020_Artificial_Intelligence_SCREEN_5c537747ff.pdf
- Jagai, J. S., Castronovo, D. A., Monchak, J. and Naumova, E. N. (2009) 'Seasonality of cryptosporidiosis: A meta-analysis approach', *Environmental Research*, 109(4), pp. 465–478. <https://doi.org/10.1016/j.envres.2009.02.008>
- Jobson, H. E. (1997) 'Predicting travel time and dispersion in rivers and streams', *Journal of Hydraulic Engineering*, 123 (11), 971–978. [https://doi.org/10.1061/\(ASCE\)0733-9429\(1997\)123:11\(971\)](https://doi.org/10.1061/(ASCE)0733-9429(1997)123:11(971))
- Jones, E. R., Bierkens, M. F. P. and van Vliet, M. T. H. (2024) 'Current and future global water scarcity intensifies when accounting for surface water quality', *Nature Climate Change*, 14(6), pp. 629–635. <https://doi.org/10.1038/s41558-024-02007-0>
- Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N. and Kumar, V. (2017) 'Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data', *IEEE Transactions on Knowledge and Data Engineering*, 29(10), pp. 2318–2331. <https://doi.org/10.1109/TKDE.2017.2720168>
- King, B. J. and Monis, P. T. (2007) 'Critical processes affecting *Cryptosporidium* oocyst survival in the environment', *Parasitology*, 134(3), pp. 309–323.
<https://doi.org/10.1017/s0031182006001491>
- King, B. J., Keegan, A. R., Monis, P. T. and Saint, C. P. (2005) 'Environmental Temperature Controls *Cryptosporidium* Oocyst Metabolic Rate and Associated Retention of

- Infectivity', *Applied and Environmental Microbiology*, 71(7), pp. 3848–3857.
<https://doi.org/10.1128/aem.71.7.3848-3857.2005>
- Knapp, J. L. A., Li, L. and Musolff, A. (2022) 'Hydrologic connectivity and source heterogeneity control concentration–discharge relationships', *Hydrological Processes*, 36(9), e14683. <https://doi.org/10.1002/hyp.14683>
- Kursa, M. B. and Rudnicki, W. R. (2010) 'Feature Selection with the Boruta Package', *Journal of Statistical Software*, 36(11), pp. 1 – 13. <https://doi.org/10.18637/jss.v036.i11>
- Lei, C., Wagner, P. D. and Fohrer, N. (2021) 'Effects of land cover, topography, and soil on stream water quality at multiple spatial and seasonal scales in a German lowland catchment', *Ecological Indicators*, 120, pp. 106940.
<https://doi.org/10.1016/j.ecolind.2020.106940>
- Ligda, P., Claerebout, E., Kostopoulou, D., Zdragas, A., Casaert, S., Robertson, L. J. and Sotiraki, S. (2020) 'Cryptosporidium and Giardia in surface water and drinking water: Animal sources and towards the use of a machine-learning approach as a tool for predicting contamination', *Environmental Pollution*, 264, pp. 114766.
<https://doi.org/10.1016/j.envpol.2020.114766>
- Ligda, P., Mittas, N., Kyzas, G. Z., Claerebout, E. and Sotiraki, S. (2024) 'Machine learning and explainable artificial intelligence for the prevention of waterborne cryptosporidiosis and giardiasis', *Water Research*, 262, pp. 122110.
<https://doi.org/10.1016/j.watres.2024.122110>
- Liu, W., An, W., Jeppesen, E., Ma, J., Yang, M. and Trolle, D. (2019) 'Modelling the fate and transport of Cryptosporidium, a zoonotic and waterborne pathogen, in the Daning River watershed of the Three Gorges Reservoir Region, China', *Journal of Environmental Management*, 232, pp. 462–474. <https://doi.org/10.1016/j.jenvman.2018.10.064>
- Lundberg, S. and Lee, S.-I. (2017) 'A Unified Approach to Interpreting Model Predictions', *arXiv.org*. <https://doi.org/10.48550/arXiv.1705.07874>
- Lundberg, S. M., Erion, G. G. and Lee, S.-I. (2018) 'Consistent individualized feature attribution for tree ensembles', *arXiv preprint arXiv:1802.03888*.
<https://doi.org/10.48550/arXiv.1802.03888>

- Mac Kenzie, W. R., Hoxie, N. J., Proctor, M. E., Gradus, M. S., Blair, K. A., Peterson, D. E., Kazmierczak, J. J., Addiss, D. G., Fox, K. R. and Rose, J. B. (1994) 'A massive outbreak in Milwaukee of *Cryptosporidium* infection transmitted through the public water supply', *New England Journal of Medicine*, 331(3), pp. 161–167.
<https://doi.org/10.1056/nejm199407213310304>
- Maiyo, J. K., Dasika, S. and Jafvert, C. T. (2023) 'Slow Sand Filters for the 21st Century: A Review', *International Journal of Environmental Research and Public Health*, 20(2).
<https://doi.org/10.3390/ijerph20021019>
- Mamun, A. A., Shams, S. and Nuruzzaman, M. (2020) 'Review on uncertainty of the first-flush phenomenon in diffuse pollution control', *Applied Water Science*, 10 (1), 53.
<https://doi.org/10.1007/s13201-019-1127-1>
- Masood, A., Niazkar, M., Zakwan, M. and Piraei, R. (2023) 'A Machine Learning-Based Framework for Water Quality Index Estimation in the Southern Bug River', *Water* 15(20), 3543. <https://doi.org/10.3390/w15203543>
- Medema, G. J. and Schijven, J. F. (2001) 'Modelling the sewage discharge and dispersion of cryptosporidium and giardia in surface water', *Water Research*, 35(18), pp. 4307–4316.
[https://doi.org/10.1016/S0043-1354\(01\)00161-0](https://doi.org/10.1016/S0043-1354(01)00161-0)
- Medema, G., Teunis, P., Blokker, M., Deere, D., Davison, A., Charles, P. and Loret, J. (2009) 'Risk assessment of *Cryptosporidium* in drinking water', *World Health Organization*, 143. Available at: <https://www.who.int/publications/i/item/WHO-HSE-WSH-09.04> (Accessed: 8 December 2025).
- Messner, M. J. and Berger, P. (2016) 'Cryptosporidium Infection Risk: Results of New Dose-Response Modeling', *Risk Analysis*, 36(10), pp. 1969–1982.
<https://doi.org/10.1111/risa.12541>
- Messner, M. J., Chappell, C. L. and Okhuysen, P. C. (2001) 'Risk Assessment for *Cryptosporidium*: A Hierarchical Bayesian Analysis of Human Dose Response Data', *Water Research*, 35(16), pp. 3934–3940. [https://doi.org/10.1016/s0043-1354\(01\)00119-1](https://doi.org/10.1016/s0043-1354(01)00119-1)

- Met Office (2003) '5 km Resolution UK Composite Rainfall Data from the Met Office Nimrod System', NCAS British Atmospheric Data Centre. Available at: <https://catalogue.ceda.ac.uk/uuid/f91b2c5399c5bf689e29bb15ab45da8a> (Accessed: 5 February 2025).
- Met Office, Hollis, D., Carlisle, E., Kendon, M., Packman, S., Doherty, A. (2024) 'HadUK-Grid Gridded Climate Observations on a 5km grid over the UK, v1.3.0.ceda (1836-2023)', NERC EDS Centre for Environmental Data Analysis. Available at: <https://dx.doi.org/10.5285/5ba67d62cdc249a3bc5b1c38b339beb3> (Accessed: 8 December 2025).
- Minh, D., Wang, H. X., Li, Y. F. and Nguyen, T. N. (2022) 'Explainable artificial intelligence: a comprehensive review', *The Artificial Intelligence Review*, 55(5), pp. 3503–3568. <https://doi.org/10.1007/s10462-021-10088-y>
- Molnar, C. (2022) *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Christoph Molnar. Available at: https://originalstatic.aminer.cn/misc/pdf/Molnar-interpretable-machine-learning_compressed.pdf (Accessed: 8 December 2025).
- Montemayor, M., Valero, F., Jofre, J. and Lucena, F. (2005) 'Occurrence of *Cryptosporidium* spp. oocysts in raw and treated sewage and river water in north-eastern Spain', *Journal of Applied Microbiology*, 99(6), pp. 1455–1462. <https://doi.org/10.1111/j.1365-2672.2005.02737.x>
- Morton, R. D., Marston, C. G., Neil, A. W. and Rowland, C. S. (2024) 'Land Cover Map 2023 (1km summary rasters, GB and N. Ireland)'. Available at: <https://doi.org/10.5285/96bc980a-31b4-4d1b-87e9-007d4932a56b> (Accessed: 5 February 2025).
- Müller, A., Österlund, H., Marsalek, J. and Viklander, M. (2020) 'The pollution conveyed by urban runoff: A review of sources', *Science of The Total Environment*, 709, pp. 136125. <https://doi.org/10.1016/j.scitotenv.2019.136125>
- Mutlu, İ., Hackermüller, J. and Schor, J. (2025) 'Automated curation of spatial metadata in environmental monitoring data', *Ecological Informatics*, 86, pp. 103038. <https://doi.org/10.1016/j.ecoinf.2025.103038>

- Nasser, A. M. (2016) 'Removal of Cryptosporidium by wastewater treatment processes: A review', *Journal of Water & Health*, 14(1), pp. 1–13.
<https://doi.org/10.2166/wh.2015.131>
- Niazkar, M., Menapace, A., Brentan, B., Piraei, R., Jimenez, D., Dhawan, P. and Righetti, M. (2024) 'Applications of XGBoost in water resources engineering: A systematic literature review (Dec 2018–May 2023)', *Environmental Modelling & Software : with environment data news*, 174. <https://doi.org/10.1016/j.envsoft.2024.105971>
- Nichols, G., Chalmers, R., Lake, I., Sopwith, W., Regan, M., Hunter, P., Grenfell, P., Harrison, F. and Lane, C. (2006) *Cryptosporidiosis: A report on the surveillance and epidemiology of Cryptosporidium infection in England and Wales*: Drinking Water Inspectorate (DWI) / Department for Environment, Food and Rural Affairs (Defra) (DWI 70/2/201. Available at: <https://www.dwi.gov.uk/research/completed-research/cryptosporidium-giardia/cryptosporidiosis-a-report-on-the-surveillance-and-epidemiology-of-cryptosporidium-infection-in-england-and-wales/> (Accessed: 8 December 2025).
- O'Brien, R. M. (2017) 'Dropping Highly Collinear Variables from a Model: Why it Typically is Not a Good Idea', *Social Science Quarterly*, 98(1), 360-375.
<https://doi.org/10.1111/ssqu.12273>
- Ofwat (2021) *H2Open – Open data in the water industry: A case for change*, Birmingham, England: Ofwat. Available at: <https://www.ofwat.gov.uk/publication/h2open-open-data-in-the-water-industry-a-case-for-change> (Accessed: 8 December 2025).
- Ofwat (2025a) PR24 *FD sector summary (Revised)*. Available at: <https://www.ofwat.gov.uk/wp-content/uploads/2025/04/PR24-FD-sector-summary-revised.pdf> (Accessed: 8 December 2025).
- Ofwat (2025b) *Rapid Cryptosporidium detection to improve water safety and protect public health*: Water Innovation Fund. Available at: <https://waterinnovation.challenges.org/winners/rapid-cryptosporidium> (Accessed: 8 December 2025).
- Okhuysen, P. C., Chappell, C. L., Crabb, J. H., Sterling, C. R. and DuPont, H. L. (1999) 'Virulence of Three Distinct *Cryptosporidium parvum* Isolates for Healthy Adults', *Journal of Infectious Diseases*, 180(4), pp. 1275–1281. <https://doi.org/10.1086/315033>

- Ordnance Survey, 2025. OS data download and API products. Available at:
<https://www.ordnancesurvey.co.uk/products> (Accessed: 8 February 2025)
- Owens, C. E., Angles, M. L., Cox, P. T., Byleveld, P. M., Osborne, N. J. and Rahman, M. B. (2020) 'Implementation of quantitative microbial risk assessment (QMRA) for public drinking water supplies: systematic review', *Water Research*, 174, pp. 115614.
<https://doi.org/10.1016/j.watres.2020.115614>
- Pasciucco, F., Pecorini, I. and Iannelli, R. (2023) 'A comparative LCA of three WWTPs in a tourist area: Effects of seasonal loading rate variations', *Science of The Total Environment*, 863, pp. 160841. <https://doi.org/10.1016/j.scitotenv.2022.160841>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R. and Dubourg, V. (2011) 'Scikit-learn: Machine learning in Python', *Journal Machine Learning Research*, 12, 2825-2830. Available at:
<https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf>
- Petterson, S., Bradford-Hartke, Z., Leask, S., Jarvis, L., Wall, K. and Byleveld, P. (2021) 'Application of QMRA to prioritise water supplies for Cryptosporidium risk in New South Wales, Australia', *The Science of the Total Environment*, 784, pp. 147107–147107.
<https://doi.org/10.1016/j.scitotenv.2021.147107>
- Pham, X. T. T. and Ho, T. H. (2021) 'Using boosting algorithms to predict bank failure: An untold story', *International Review of Economics & Finance*, 76, pp. 40–54.
<http://dx.doi.org/10.1016/j.iref.2021.05.005>
- Phillip, D. A. T., Rawlins, S. C., Baboolal, S., Gosein, R., Goddard, C., Legall, G. and Chinchamee, A. (2007) 'Relative importance of the various environmental sources of Cryptosporidium oocysts in three watersheds', *Journal of Water & Health*, 6(1), pp. 23–34. <https://doi.org/10.2166/wh.2007.016>
- Pignata, C., Bonetta, S., Bonetta, S., Cacciò, S. M., Sannella, A. R., Gilli, G. and Carraro, E. (2019) 'Cryptosporidium oocyst contamination in drinking water: A case study in Italy', *International Journal of Environmental Research and Public Health*, 16(11), pp. 2055.
<https://doi.org/10.3390/ijerph16112055>

Piraei, R., Afzali, S. H. and Niazkar, M. (2023) 'Assessment of XGBoost to Estimate Total Sediment Loads in Rivers', *Water Resources Management*, 37 (13), 5289-5306.

<http://dx.doi.org/10.1007/s11269-023-03606-w>

Public Health England (2024) *Research and analysis: Cryptosporidium: national laboratory data*. Available at: <https://www.gov.uk/government/publications/cryptosporidium-national-laboratory-data> (Accessed: 8 December 2025).

Puleston, R. L., Mallaghan, C. M., Modha, D. E., Hunter, P. R., Nguyen-Van-Tam, J. S., Regan, C. M., Nichols, G. L. and Chalmers, R. M. (2014) 'The first recorded outbreak of cryptosporidiosis due to *Cryptosporidium cuniculus* (formerly rabbit genotype), following a water quality incident', *Journal of Water & Health*, 12(1), pp. 41–50.

<https://doi.org/10.2166/wh.2013.097>

Raissi, M., Perdikaris, P. and Karniadakis, G. E. (2019) 'Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations', *Journal of Computational Physics*, 378, pp. 686–707.

<https://doi.org/10.1016/j.jcp.2018.10.045>

Rawat, P., Bajaj, M., Sharma, V., Vats, S. (2023) 'A comprehensive analysis of the effectiveness of machine learning algorithms for predicting water quality'. *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)* 14-16, 1108–1114.

<https://doi.org/10.1109/ICIDCA56705.2023.10099968>

Razavi, S., Hannah, D. M., Elshorbagy, A., Kumar, S., Marshall, L., Solomatine, D. P., Dezfuli, A., Sadegh, M. and Famiglietti, J. (2022) 'Coevolution of machine learning and process-based modelling to revolutionize Earth and environmental sciences: A perspective', *Hydrological Processes*, 36(6), pp. e14596. <https://doi.org/10.1002/hyp.14596>

Razzolini, M. T. P., Santos, T. F. D. S. and Bastos, V. K. (2010) 'Detection of *Giardia* and *Cryptosporidium* cysts/oocysts in watersheds and drinking water sources in Brazil urban areas', *Journal of Water & Health*, 8(2), pp. 399–404.

<https://doi.org/10.2166/wh.2009.172>

- Regli, S., Rose, J. B., Haas, C. N. and Gerba, C. P. (1991) 'Modeling the Risk From Giardia and Viruses in Drinking Water', *Journal of the American Water Works Association*, 83(11), pp. 76–84. <https://doi.org/10.1002/j.1551-8833.1991.tb07252.x>
- Reis, S., Liska, T., Steinle, S., Carnell, E., Leaver, D., Roberts, E., Vieno, M., Beck, R., and Dragosits, U. (2017) 'UK gridded population 2011 based on Census 2011 and Land Cover Map 2015', NERC Environmental Information Data Centre. Available at: <https://doi.org/10.5285/0995e94d-6d42-40c1-8ed4-5090d82471e1> (Accessed: 5 February 2025).
- Robertson, L. J., Campbell, A. T. and Smith, H. V. (1992) 'Survival of *Cryptosporidium parvum* oocysts under various environmental pressures', *Applied and Environmental Microbiology*, 58(11), pp. 3494–3500. <https://doi.org/10.1128/aem.58.11.3494-3500.1992>
- Robinson, G., Chalmers, R. M., Stapleton, C., Palmer, S. R., Watkins, J., Francis, C. and Kay, D. (2011) 'A whole water catchment approach to investigating the origin and distribution of *Cryptosporidium* species', *Journal Applied Microbiology*, 111(3), pp. 717–730. <https://doi.org/10.1111/j.1365-2672.2011.05068.x>
- Rose, J. B., Huffman, D. E. and Gennaccaro, A. (2002) 'Risk and control of waterborne cryptosporidiosis', *FEMS Microbiology Reviews*, 26(2), pp. 113–123. [https://doi.org/10.1016/S0168-6445\(02\)00090-6](https://doi.org/10.1016/S0168-6445(02)00090-6)
- Ryan, U. M., Feng, Y., Fayer, R. and Xiao, L. (2021) 'Taxonomy and molecular epidemiology of *Cryptosporidium* and *Giardia* – a 50 year perspective (1971–2021)', *International Journal for Parasitology*, 51(13-14), pp. 1099–1119. <https://doi.org/10.1016/j.ijpara.2021.08.007>
- Samadder, S. R., Ziegler, P., Murphy, T. M. and Holden, N. M. (2010) 'Spatial Distribution of Risk Factors for *Cryptosporidium* Spp. Transport in an Irish Catchment', *Water Environment Research*, 82(8), pp. 750–758. <https://doi.org/10.2175/106143010x12609736966649>
- Sato, M. I. Z., Galvani, A. T., Padula, J. A., Nardocci, A. C., Lauretto, M. d. S., Razzolini, M. T. P. and Hachich, E. M. (2013) 'Assessing the infection risk of *Giardia* and *Cryptosporidium* in public drinking water delivered by surface water systems in Sao Paulo State, Brazil',

Science of the Total Environment, 442, pp. 389–396.

<https://doi.org/10.1016/j.scitotenv.2012.09.077>

Schets, F., Van Wijnen, J., Schijven, J., Schoon, H. and de Roda Husman, A. (2008)

'Monitoring of waterborne pathogens in surface waters in Amsterdam, The Netherlands, and the potential health risk associated with exposure to *Cryptosporidium* and *Giardia* in these waters', *Applied and Environmental Microbiology*, 74(7), pp. 2069–2078. <https://doi.org/10.1128/aem.01609-07>

Schijven, J. F., Teunis, P. F. M., Rutjes, S. A., Bouwknegt, M. and de Roda Husman, A. M.

(2011) 'QMRAspot: A tool for Quantitative Microbial Risk Assessment from surface water to potable water', *Water Research*, 45(17), pp. 5564–5576.

<https://doi.org/10.1016/j.watres.2011.08.024>

Schmidt, P. J., Anderson, W. B. and Emelko, M. B. (2020) 'Describing water treatment

process performance: Why average log-reduction can be a misleading statistic', *Water Research*, 176, pp. 115702. <https://doi.org/10.1016/j.watres.2020.115702>

Schoener, G. and Stone, M. C. (2020) 'Monitoring soil moisture at the catchment scale – A novel approach combining antecedent precipitation index and radar-derived rainfall data', *Journal of Hydrology (Amsterdam)*, 589, pp. 125155.

<https://doi.org/10.1016/j.jhydrol.2020.125155>

Scikit-learn (2026) 'Feature selection — scikit-learn documentation'. https://scikit-learn.org/stable/modules/feature_selection.html (Accessed: 20 April 2026).

Sliva, L. and Williams, D. D. (2001) 'Buffer zone versus whole catchment approaches to studying land use impact on river water quality', *Water Research*, 35(14), pp. 3462–3472. [https://doi.org/10.1016/S0043-1354\(01\)00062-8](https://doi.org/10.1016/S0043-1354(01)00062-8)

Singh, N. K., Emanuel, R. E., McGlynn, B. L. and Miniat, C. F. (2021) 'Soil Moisture Responses to Rainfall: Implications for Runoff Generation', *Water Resources Research*, 57(9), pp. e2020WR028827. <https://doi.org/10.1029/2020WR028827>

Sinha, S., Hammond, A. and Smith, H. (2022) 'A comprehensive intercomparison study between a lumped and a fully distributed hydrological model across a set of 50

catchments in the United Kingdom', *Hydrological Processes*, 36 (3), e14544.

<http://dx.doi.org/10.1002/hyp.14544>

Smalley, A. L., Douterelo, I., Chipps, M. and Shucksmith, J. D. (2025) 'Data-driven prediction of daily Cryptosporidium river concentrations for water resource management: Use of catchment-averaged vs spatially distributed features in a Bagging-XGBoost model', *Science of The Total Environment*, 991, pp. 179794.

<https://doi.org/10.1016/j.scitotenv.2025.179794>

Smeets, P. W. M. H., Dullemont, Y. J., Van Gelder, P. H. A. J. M., Van Dijk, J. C. and Medema, G. J. (2008) 'Improved methods for modelling drinking water treatment in quantitative microbial risk assessment; a case study of Campylobacter reduction by filtration and ozonation', *Journal of Water & Health*, 6(3), pp. 301–314.

<https://doi.org/10.2166/wh.2008.066>

Smeets, P. W. M. H., van Dijk, J. C., Stanfield, G., Rietveld, L. C. and Medema, G. J. (2007) 'How can the UK statutory Cryptosporidium monitoring be used for quantitative risk assessment of cryptosporidium in drinking water?', *Journal of Water & Health*, 5(1), pp. 107–118. <https://doi.org/10.2166/wh.2007.140>

Smith, R.; Antoniou, V.; Askquith-Ellis, A.; Ball, L.A.; Bennett, E.S.; Blake, J.R.; Boorman, D.B.; Brooks, M.; Clarke, M.; Cooper, H.M.; Cowan, N.; Cumming, A.; Evans, J.G.; Farrand, P.; Fry, M.; Harvey, D.; Houghton-Carr, H.; Howson, T.; Jiménez-Arranz, G.; Keen, Y.; Khamis, D.; Leeson, S.; Lord, W.D.; Morrison, R.; Nash, G.V.; O'Callaghan, F.; Rylett, D.; Scarlett, P.M.; St Quintin, P.; Stanley, S.; Swain, O.D.; Szczykulska, M.; Teagle, S.; Thornton, J.L.; Trill, E.J.; Vincent, P.; Ward, H.C.; Warwick, A.C.; Winterbourn, J.B. (2024). 'Daily and sub-daily hydrometeorological and soil data (2013-2023) [COSMOS-UK]', NERC EDS Environmental Information Data Centre.

<https://doi.org/10.5285/399ed9b1-bf59-4d85-9832-ee4d29f49bfb>

Sokolova, E., Ivarsson, O., Lillieström, A., Speicher, N. K., Rydberg, H. and Bondelind, M. (2022) 'Data-driven models for predicting microbial water quality in the drinking water source using E. coli monitoring and hydrometeorological data', *Science of the Total Environment*, 802, pp. 149798. <https://doi.org/10.1016/j.scitotenv.2021.149798>

Sterk, A., de Man, H., Schijven, J. F., de Nijs, T. and de Roda Husman, A. M. (2016) 'Climate change impact on infection risks during bathing downstream of sewage emissions from CSOs or WWTPs', *Water Research*, 105, pp. 11–21.

<https://doi.org/10.1016/j.watres.2016.08.053>

Stream (2024) *Stream Blueprint: Open Data Framework for the Water Sector*, [Online]:

Stream Water Data. Available at:

<https://www.arcgis.com/sharing/rest/content/items/946544babdb144b5847b6cc8d0f284d8/data>.

Sturdee, A. P., Chalmers, R. M. and Bull, S. A. (1999) 'Detection of *Cryptosporidium* oocysts in wild mammals of mainland Britain', *Veterinary Parasitology*, 80 (4), 273–280.

[https://doi.org/10.1016/S0304-4017\(98\)00226-X](https://doi.org/10.1016/S0304-4017(98)00226-X)

Sturdee, A., Foster, I., Bodley-Tickell, A. T. and Archer, A. (2007) 'Water quality and *Cryptosporidium* distribution in an upland water supply catchment, Cumbria, UK', *Hydrological Processes*, 21(7), pp. 873–885. <https://doi.org/10.1002/hyp.6278>

Suradhaniwar, S., Kar, S., Durbha, S. S. and Jagarlapudi, A. (2021) 'Time series forecasting of univariate agrometeorological data: A comparative performance evaluation via one-step and multi-step ahead forecasting strategies', *Sensors* 21 (7), 2430.

<https://doi.org/10.3390/s21072430>

Suslovaite, V., Pickett, H., Speight, V. and Shucksmith, J. D. (2024) 'Forecasting acute rainfall driven *E. coli* impacts in inland rivers based on sewer monitoring and field runoff', *Water Research*, 248, pp. 120838. <https://doi.org/10.1016/j.watres.2023.120838>

Swaffer, B., Abbott, H., King, B., van der Linden, L. and Monis, P. (2018) 'Understanding human infectious *Cryptosporidium* risk in drinking water supply catchments', *Water Research*, 138, pp. 282–292. <https://doi.org/10.1016/j.watres.2018.03.063>

Swaffer, B. A., Vial, H. M., King, B. J., Daly, R., Frizenschaf, J. and Monis, P. T. (2014) 'Investigating source water *Cryptosporidium* concentration, species and infectivity rates during rainfall-runoff in a multi-use catchment', *Water Research*, 67, pp. 310–320.

<https://doi.org/10.1016/j.watres.2014.08.055>

- Takefuji, Y. (2025) 'Beyond XGBoost and SHAP: Unveiling true feature importance', *Journal of Hazardous Materials*, 488, 137382. <https://doi.org/10.1016/j.jhazmat.2025.137382>
- Tang, J., McDonald, S., Peng, X., Samadder, S. R., Murphy, T. M. and Holden, N. M. (2011) 'Modelling *Cryptosporidium* oocysts transport in small ungauged agricultural catchments', *Water Research*, 45 (12), 3665-3680.
<http://dx.doi.org/10.1016/j.watres.2011.04.013>
- Thames Water (2021) *System Operations Technical Briefing: TB 375 - Escalation of Crypto Trigger levels on Thames intakes*: Thames Water [Unpublished].
- Thames Water (2023) *TMS23 Enhancement Case: Long Term Water Quality Strategy Cryptosporidium*: Thames Water. Available at: <https://www.thameswater.co.uk/media-library/b41ffox5/water-quality-cryptosporidium.pdf>
- Thames Water (2025) *Crypto Log Reduction Calculator*. Thames Water [Unpublished].
- The Rivers Trust (2025) *Unpacking the 2024 Annual Sewage Spill Data*: The Rivers Trust. Available at: <https://theriverstrust.org/about-us/news/2024-sewage-spill-cso-data>
- Tran, Q. Q., De Niel, J. and Willems, P. (2018) 'Spatially Distributed Conceptual Hydrological Model Building: A Generic Top-Down Approach Starting From Lumped Models', *Water Resources Research*, 54 (10), 8064-8085. <https://doi.org/10.1029/2018WR023566>
- UK Government (2007) *Explanatory Memorandum to The Water Supply (Water Quality) Regulations 2000 (Amendment) Regulations 2007, S.I. 2007/2734*: legislation.gov.uk. Available at:
https://www.legislation.gov.uk/uksi/2007/2734/pdfs/uksiem_20072734_en.pdf
- UKWIR (2020) *Cryptosporidium: Enhancing the Water Industry's Capability to Respond*: UK Water Industry Research.
- UNESCO (2020) *United Nations World Water Development Report 2020: Water and Climate Change*, Paris: UNESCO. Available at:
<https://unesdoc.unesco.org/ark:/48223/pf0000372985.locale=en>
- UNESCO (2025) *Applications of artificial intelligence for water management*, Paris: UNESCO. Available at: <https://unesdoc.unesco.org/ark:/48223/pf0000393243>
- USEPA (2025) 'Drinking Water Treatment Technology Unit Cost Models'.

- Vásquez, G. A., Busschaert, P., Haberbeck, L. U., Uyttendaele, M. and Geeraerd, A. H. (2014) 'An educationally inspired illustration of two-dimensional Quantitative Microbiological Risk Assessment (QMRA) and sensitivity analysis', *International Journal of Food Microbiology*, 190, pp. 31–43. <https://doi.org/10.1016/j.ijfoodmicro.2014.07.034>
- Vermeulen, L. C., van Hengel, M., Kroeze, C., Medema, G., Spanier, J. E., van Vliet, M. T. H. and Hofstra, N. (2019) 'Cryptosporidium concentrations in rivers worldwide', *Water Research*, 149, pp. 202–214. <https://doi.org/10.1016/j.watres.2018.10.069>
- von Sperling, M., Verbyla, M. E. and Oliveira, S. M. A. C. (2020) *Assessment of Treatment Plant Performance and Water Quality Data: A Guide for Students, Researchers and Practitioners*. IWA Publishing.
- Walker Jr, F. R. W. and Stedinger, J. R. (1999) 'Fate and transport model of Cryptosporidium', *Journal of Environmental Engineering*, 125 (4), 325-333. [https://doi.org/10.1061/\(ASCE\)0733-9372\(1999\)125:4\(325\)](https://doi.org/10.1061/(ASCE)0733-9372(1999)125:4(325))
- Wallingford HydroSolutions [2016] *Time-of-Transport Model (Version 3b)*. [Property of Thames Water. Unpublished.]
- Wang, X., Wang, X. and Cao, J. (2023) 'Environmental Factors Associated with Cryptosporidium and Giardia', *Pathogens*, 12(3), pp. 420. <https://doi.org/10.3390/pathogens12030420>
- Wells, B., Shaw, H., Hotchkiss, E., Gilray, J., Ayton, R., Green, J., Katzer, F., Wells, A. and Innes, E. (2015) 'Prevalence, species identification and genotyping Cryptosporidium from livestock and deer in a catchment in the Cairngorms with a history of a contaminated public water supply', *Parasites & Vectors*, 8(1), pp. 1–13. <https://doi.org/10.1186/s13071-015-0684-x>
- WHO (2016) *Quantitative Microbial Risk Assessment: Application for Water Safety Management*. Geneva: World Health Organization. Available at: <https://www.who.int/publications/i/item/9789241565370> (Accessed: 5 February 2025).
- Widerström, M., Schönning, C., Lilja, M., Lebbad, M., Ljung, T., Allestam, G., Ferm, M., Björkholm, B., Hansen, A., Hiltula, J., Långmark, J., Löfdahl, M., Omberg, M., Reuterwall, C., Samuelsson, E., Widgren, K., Wallensten, A. and Lindh, J. (2014) 'Large outbreak of

Cryptosporidium hominis infection transmitted through the public water supply, Sweden', *Emerging Infectious Diseases*, 20(4), pp. 581–589.

<https://doi.org/10.3201/eid2004.121415>

Wilkes, G., Edge, T. A., Gannon, V. P. J., Jokinen, C., Lyautey, E., Neumann, N. F., Ruecker, N., Scott, A., Sunohara, M., Topp, E. and Lapen, D. R. (2011) 'Associations among pathogenic bacteria, parasites, and environmental and land use factors in multiple mixed-use watersheds', *Water Research*, 45(18), pp. 5807–5825.

<https://doi.org/10.1016/j.watres.2011.06.021>

Wilkes, G., Edge, T., Gannon, V., Jokinen, C., Lyautey, E., Medeiros, D., Neumann, N., Ruecker, N., Topp, E. and Lapen, D. R. (2009) 'Seasonal relationships among indicator bacteria, pathogenic bacteria, Cryptosporidium oocysts, Giardia cysts, and hydrological indices for surface waters within an agricultural landscape', *Water Research*, 43(8), pp. 2209–2223. <https://doi.org/10.1016/j.watres.2009.01.033>

Wilkes, G., Ruecker Norma, J., Neumann Norman, F., Gannon Victor, P. J., Jokinen, C., Sunohara, M., Topp, E., Pintar Katarina, D. M., Edge Thomas, A. and Lapen David, R. (2013) 'Spatiotemporal Analysis of Cryptosporidium Species/Genotypes and Relationships with Other Zoonotic Pathogens in Surface Water from Mixed-Use Watersheds', *Applied and Environmental Microbiology*, 79(2), pp. 434–448.

<https://doi.org/10.1128/aem.01924-12>

Wu, C., Li, B., Xiong, N. (2021a) 'An Effective Machine Learning Scheme to Analyze and Predict the Concentration of Persistent Pollutants in the Great Lakes', 9. IEEE Access, pp. 52252–52265. <https://doi.org/10.1109/ACCESS.2021.3069990>

Wu, J., Song, C., Dubinsky, E. A. and Stewart, J. R. (2021b) 'Tracking Major Sources of Water Contamination Using Machine Learning', *Frontiers in Microbiology*, 11, pp. 616692.

<https://doi.org/10.3389/fmicb.2020.616692>

Yuan, H., Liu, M., Kang, L., Miao, C. and Wu, Y. (2023) 'An empirical study of the effect of background data size on the stability of SHapley Additive exPlanations (SHAP) for deep learning models. arXiv 2022', *arXiv preprint arXiv:2204.11351*.

<https://doi.org/10.48550/arXiv.2204.11351>

- Zhang, J., Ma, X., Zhang, J., Sun, D., Zhou, X., Mi, C. and Wen, H. (2023) 'Insights into geospatial heterogeneity of landslide susceptibility based on the SHAP-XGBoost model', *Journal of Environmental Management*, 332, pp. 117357. <https://doi.org/10.1016/j.jenvman.2023.117357>
- Zhang, Z., Huang, J., Duan, S., Huang, Y., Cai, J. and Bian, J. (2022) 'Use of interpretable machine learning to identify the factors influencing the nonlinear linkage between land use and river water quality in the Chesapeake Bay watershed', *Ecological Indicators*, 140, pp. 108977. <https://doi.org/10.1016/j.ecolind.2022.108977>
- Zhao, F. and Zeng, N. (2014) 'Continued increase in atmospheric CO₂ seasonal amplitude in the 21st century projected by the CMIP5 Earth system models', *Earth System Dynamics*, 5(2), pp. 423–439. <https://doi.org/10.5194/esd-5-423-2014>
- Zhu, J.-J., Yang, M. and Ren, Z. J. (2023) 'Machine learning in environmental research: common pitfalls and best practices', *Environmental Science & Technology*, 57(46), pp. 17671–17689. <https://doi.org/10.1021/acs.est.3c00026>
- Zhu, M., Wang, J., Yang, X., Zhang, Y., Zhang, L., Ren, H., Wu, B. and Ye, L. (2022) 'A review of the application of machine learning in water quality evaluation', *Eco-Environment & Health*, 1(2), pp. 107–116. <https://doi.org/10.1016/j.eehl.2022.06.001>
- Zuur, A. F., Ieno, E. N. and Smith, G. M. (2007) *Analysing ecological data. Statistics for biology and health* New York, Springer. <https://doi.org/10.1007/978-0-387-45972-1>

Appendix A: Supplementary Material for Chapter 2

A.1 Soil Moisture and Soil Temperature Data

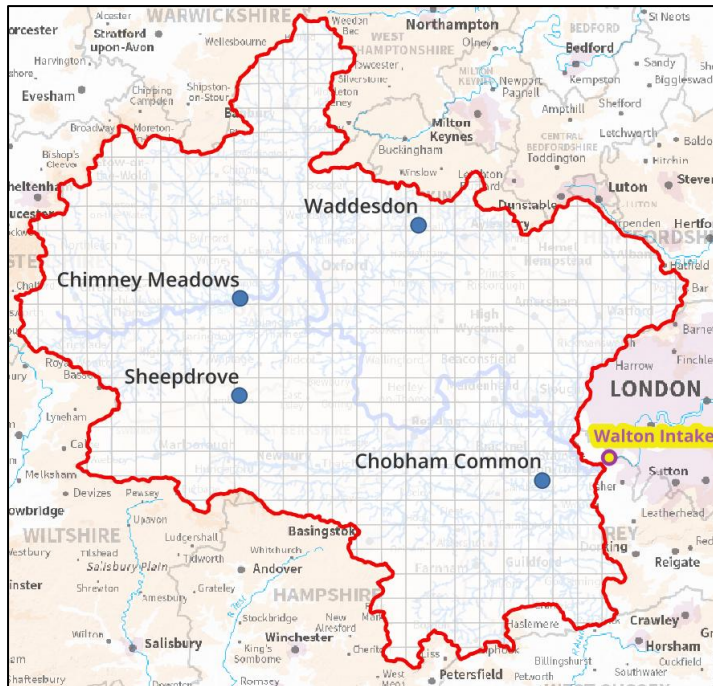


Fig. A.1 Locations of the 4 COSMOS monitoring stations in the Walton Catchment.

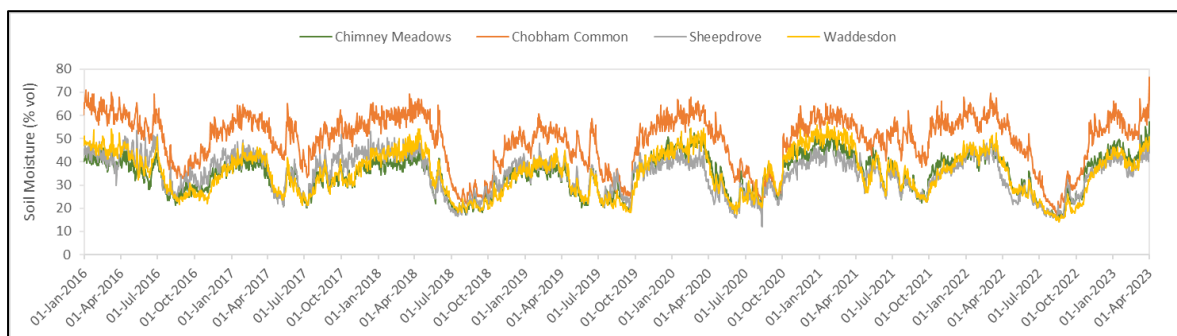


Fig. A.2 Mean daily soil moisture (as % volumetric water content at 5cm depth) for the 4 COSMOS monitoring sites in the Walton catchment.

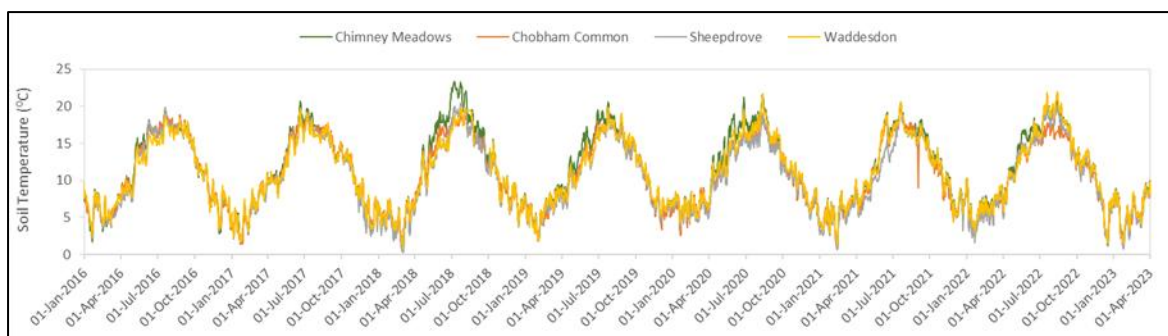


Fig. A.3 Mean daily soil temperature (in °C at 5cm depth) for the 4 COSMOS monitoring sites in the Walton catchment.

Table A.1 Matrix showing Pearson correlation coefficients for the soil moisture data at the 4 COSMOS monitoring sites in the Walton catchment, and the mean of the 4 sites (for 01.01.2016 – 01.04.2023 data).

	Chimney Meadows	Chobham Common	Sheepdrove	Waddesdon	Mean
Chimney Meadows	1.00	0.85	0.85	0.93	0.95
Chobham Common	0.85	1.00	0.85	0.87	0.95
Sheepdrove	0.85	0.85	1.00	0.86	0.93
Waddesdon	0.93	0.87	0.86	1.00	0.96
Mean	0.95	0.95	0.93	0.96	1.00

Table A.2 Matrix showing Pearson correlation coefficients for the soil temperature data at the 4 COSMOS monitoring sites in the Walton catchment, and the mean of the 4 sites (for 01.01.2016 – 01.04.2023 data).

	Chimney Meadows	Chobham Common	Sheepdrove	Waddesdon	Mean
Chimney Meadows	1.00	0.98	0.99	0.98	0.99
Chobham Common	0.98	1.00	0.99	0.98	0.99
Sheepdrove	0.99	0.99	1.00	0.99	1.00
Waddesdon	0.98	0.98	0.99	1.00	0.99
Mean	0.99	0.99	1.00	0.99	1.00

A.2 Estimation of travel times (spatially distributed model only)

A.2.1 Jobson travel time calculation

Jobson (1997) provides the following empirical equation for estimating the velocity, V_p (m s^{-1}) of an inert contaminant/tracer peak in a river system, calibrated based on data from 939 tracer test measurements taken from approximately 90 rivers in the USA.

$$V_p = a + (b \times (D'_a)^c \times (Q'_a)^{-d} \times S^e \times \frac{Q}{D_a}) \quad (\text{A.1})$$

where Q is discharge ($\text{m}^3 \text{s}^{-1}$), S is the mean slope (m m^{-1}) and D_a is the average drainage area (m^2) (based on the mean of upstream and downstream areas) and

$$D'_a = \frac{D_a^{1.25} \times \sqrt{g}}{Q_a} = \text{Dimensionless drainage area} \quad (\text{A.2})$$

$$Q'_a = \frac{Q}{Q_a} = \text{Dimensionless discharge} \quad (\text{A.3})$$

a , b , c , d and e are empirical coefficients (given as 0.094, 0.0143, 0.919, -0.469 and 0.1590, respectively), g is acceleration due to gravity (9.81 m s^{-2}) and Q_a is the average annual discharge ($\text{m}^3 \text{s}^{-1}$).

To estimate V_p (and hence travel time) under different flow regimes for each cell within the catchment, this method was recalibrated based on equivalent travel time data for the river Thames and its tributaries (see below). Travel time was then calculated for each cell, based on distance to the abstraction point. QGIS (v. 3.32.3) combined with OS Open Rivers Data and OS 1:50,000 terrain data (Ordnance Survey, 2024) was used to obtain the catchment area and river distances for the most downstream point in each 5x5 km catchment cell. Times were calculated separately for the tributaries and the River Thames and subsequently combined. Estimated travel time ranges were calculated under different flow regimes, from

a high discharge of Q10 through to a low discharge at Q90 (at intervals of Q10), with all times rounded to the nearest day. The quantile of each day's discharge at Kingston gauging station was calculated and rounded to the nearest Q10 value to give a quantile interval that could then be used to look up appropriate travel times for each cell.

A.2.2 Model Calibration

The constants in the original Jobson (1997) equation were found by curve-fitting. The present study applied this method using 110 Thames catchment travel time data points obtained through a combination of tracer tests carried out by Environmental Tracing Systems Ltd (1998) and a spill time-of-transport model developed by Wallingford HydroSolutions (2016) (only available for a limited area of the catchment). Combined, both data sources covered transport in the River Thames and a number of its tributaries. Curve-fitting was applied separately to the tributaries, to the upper Thames (with catchment areas of $\leq 1600 \text{ km}^2$), and to the lower Thames (with catchment areas of $> 1600 \text{ km}^2$). Table A.3 provides the revised coefficients, and Fig. A.4 presents the fit of the model against the calibration data.

Table A.3 Curve-fitted constants for Equation A.1 applied to three sections of the study catchment.

Catchment Section	Curve-fitted constants				
	a	b	c	d	e
Tributaries	0.000	0.0142	0.919	-0.469	0.201
Upper Thames	0.000	0.0143	0.881	-0.190	0.100
Lower Thames	0.000	0.0143	0.919	-0.491	0.272

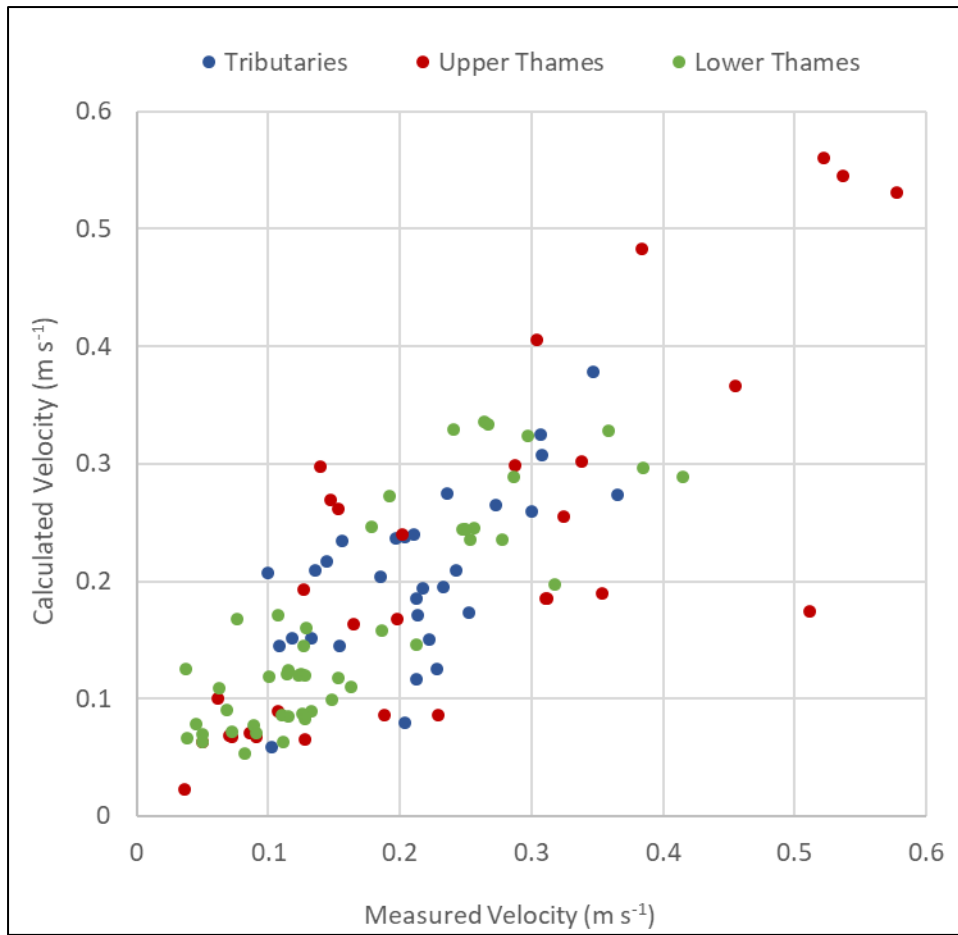


Fig. A.4 Measured and calculated (using Equation A.1) velocities for three sections of the study catchment.

A.2.3 Error Analysis

Fig. A.5 plots differences in measured velocities and calculated velocities for the calibration datasets as a cumulative distribution function (ECDF). To consider typical uncertainties in travel time estimates velocity error values at 1 standard deviation (SD) were then applied to the travel time calculations for each cell.

Example travel time errors for three cells at different distances from the abstraction point are presented in Table A.4, showing that error in travel time estimates is likely to exceed ± 1 day (i.e. model resolution) when the distance between cell and abstraction point exceeds 150 km at high (Q10) flow, 60 km at medium (Q50) flow, and 20 km at low (Q90) flow. The high error at low flow is not regarded as a critical limitation to the model because 70% of exceedances occur at flows of Q50 and above.

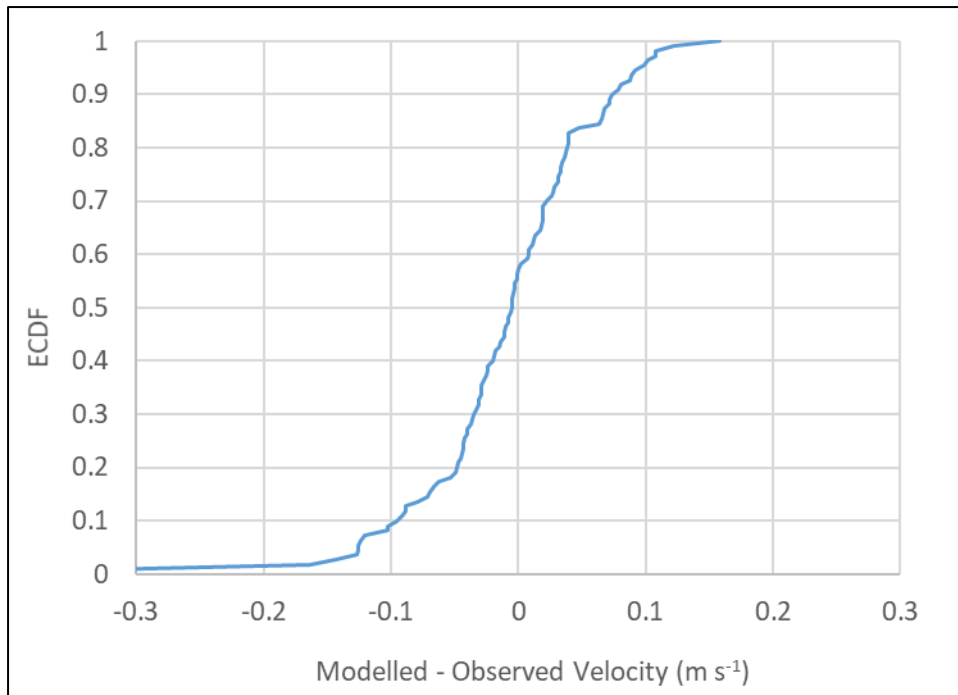


Fig. A.5 Empirical cumulative distribution error plots for velocity calculations.

Table A.4 Estimated error in the predicted travel times for cells at three different river distances from the abstraction point when ± 1 SD in the calculated velocities was applied.

Cell		499	383	345
River Distance (km)		17.10	60.59	153.00
Flow Regime	Velocity Error	Travel Time Error (days)		
High (Q10)	-1SD	+ 0.08	+ 0.42	+ 1.01
	+1SD	- 0.05	- 0.24	- 0.59
Medium (Q50)	-1SD	+ 0.30	+ 1.23	+ 4.01
	+1SD	- 0.15	- 0.57	- 1.74
Low (Q90)	-1SD	+ 0.93	+ 3.36	+ 13.16
	+1SD	- 0.32	- 1.15	- 3.79

A.3 Results from Hyperparameter Tuning

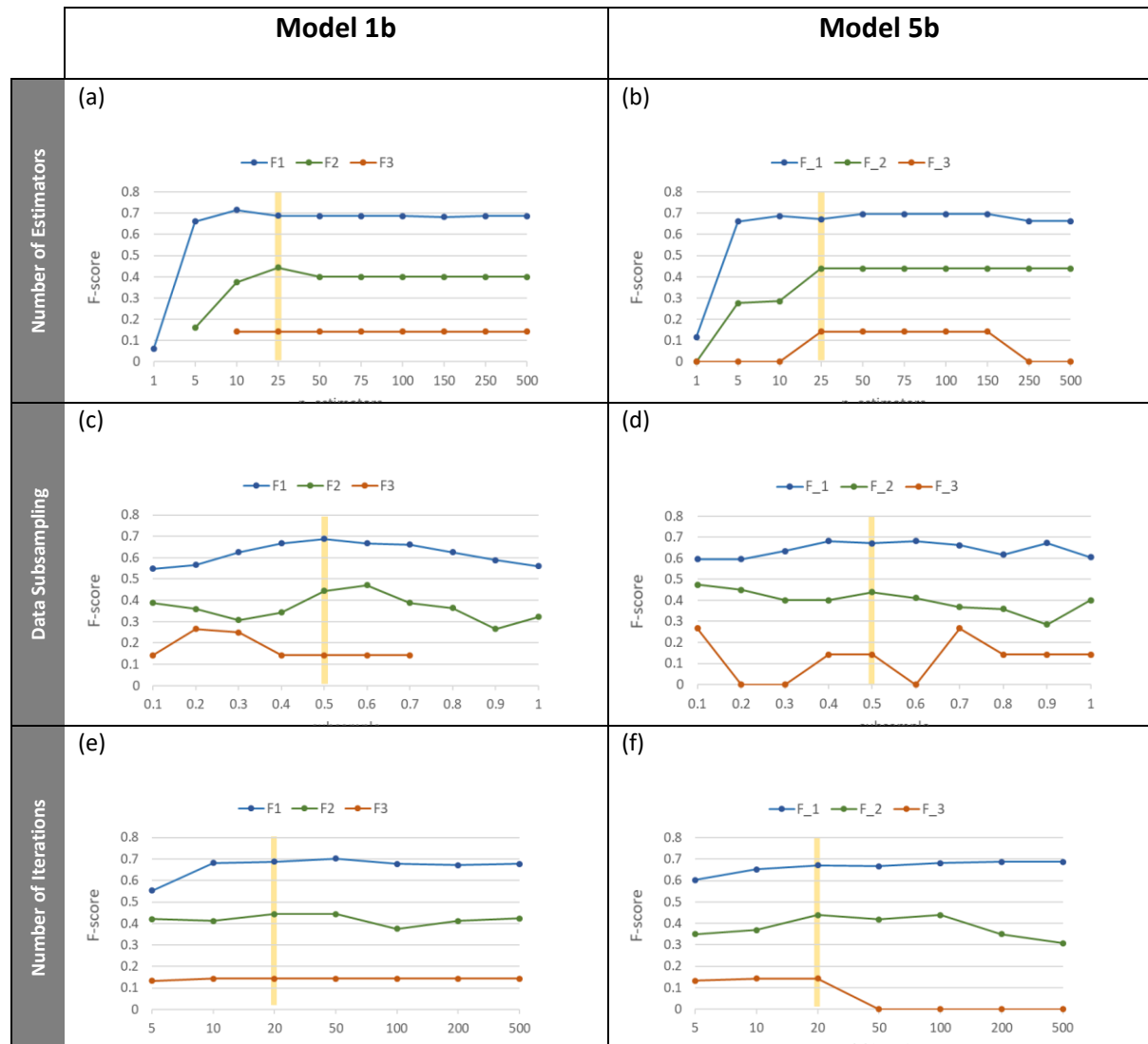


Fig. A.6 F-score outputs from hyperparameter tuning using Model 1b (with the second fewest features) and model 5b (with the largest number of features). Vertical yellow lines indicate “optimum” value based on an assessment of combined performance at the three thresholds. Because of the greater number of low exceedances, high F1-score and F2-scores were regarded as more important than high F3-scores. (Note that in each trial the optimum values for the other three elements were used.)

Page intentionally left blank

Appendix B: Supplementary Material for Chapter 3

B.1 Parametric Modelling to Fix Lower Limits

To determine appropriate lower limits (C_{LL}) for the subsequent concentration modelling of censored values by random sampling from log-uniform distributions, the entirety of each intake, WTW inlet and WTW outlet data was modelled using the censored Maximum Likelihood Estimation (MLE) method. The best match was found by selecting the fit with the lowest Akaike Information Criterion (AIC) score (Helsel, 2012). Figs. B.1-B.3 show the fitted models alongside the censored and uncensored concentration data points, together with the Q_{50} concentration given by the “best” fitted model. For each dataset, the lower limit (in oocysts L^{-1}) was given as $C_{LL} = 0.1 \times Q_{50}$.

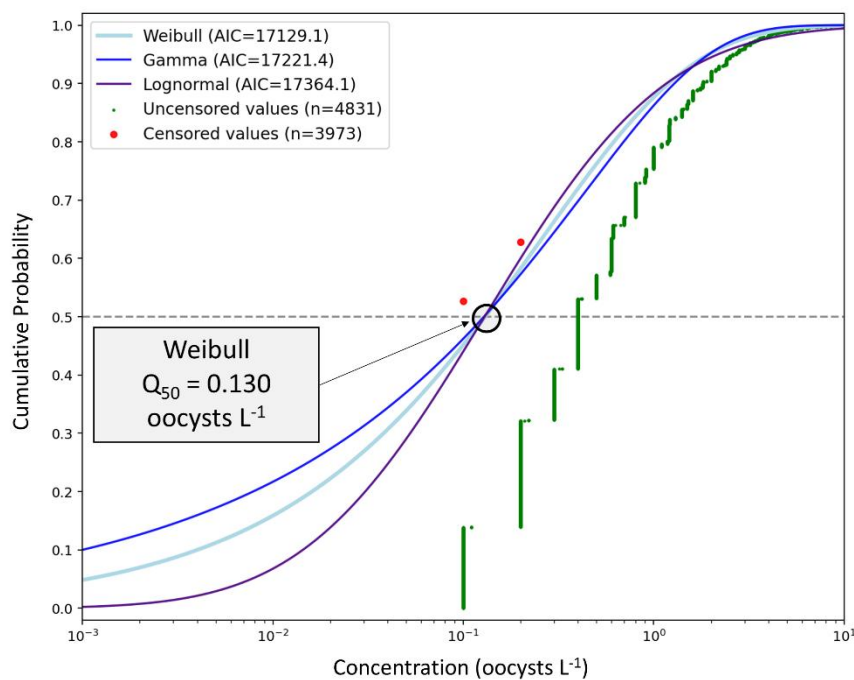


Fig. B.1 MLE-fitted models for *Cryptosporidium* concentration data from all intakes. The “best” model (based on AIC outputs) is given in the shaded box, along with the Q_{50} value for that model.

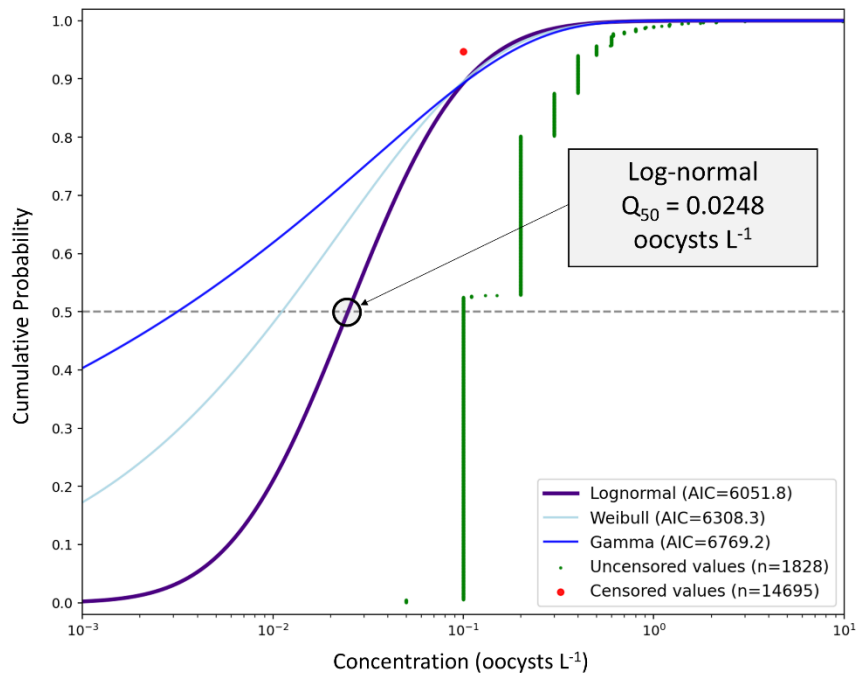


Fig. B.2 MLE-fitted models for *Cryptosporidium* concentration data from all WTW inlets. The “best” model (based on AIC outputs) is given in the shaded box, along with the Q50 value for that model.

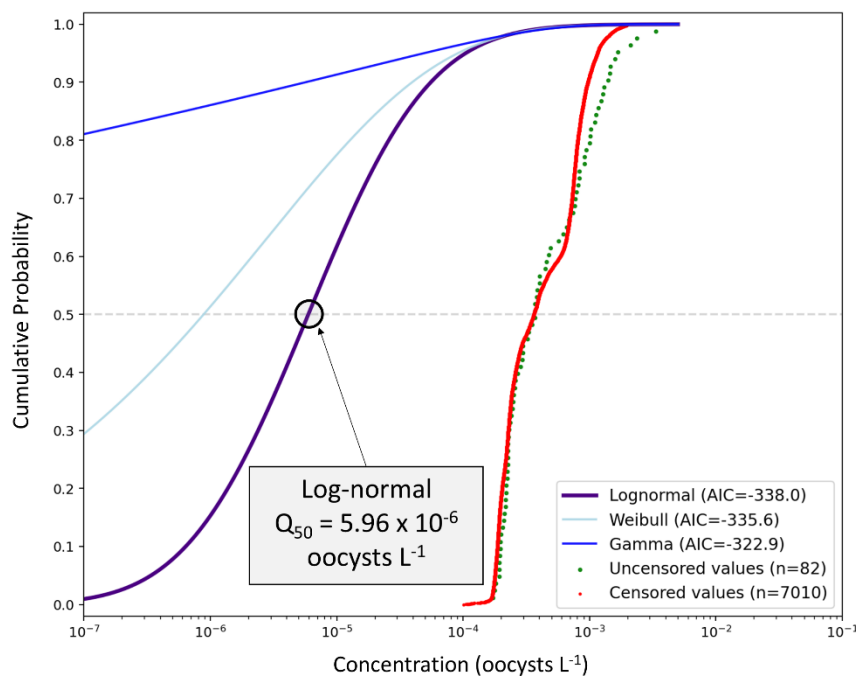


Fig. B.3 MLE-fitted models for *Cryptosporidium* concentration data from all WTW outlets. The “best” model (based on AIC outputs) is given in the shaded box, along with the Q50 value for that model.

B.2 Pooled Dose-Response Model

B.2.1 Overview

To evaluate the influence of between-strain variability in *Cryptosporidium* infectivity on QMRA outputs, a hierarchical Bayesian extension of the exponential dose-response model described by Messner et al. (2001) was implemented. The model combined human challenge data for the Iowa, UCP and TAMU isolates into a single probabilistic infectivity distribution (the pooled DR model) for use in the pooled DR QMRA.

B.2.2 Isolate-Specific Bootstrap Modelling

The pooled DR model used grouped infection data for three *C. parvum* isolates (Table B.1). For each dose level, the numbers of infected and uninfected volunteers were extracted and fitted to the exponential model (Equation 11).

Table B.1 Challenge test data used in pooled DR QMRA model. (from Messner et al., 2001).

Isolate	Dose (oocysts)	Volunteers infected	Volunteers uninfected	Total volunteers
Iowa	30	2	3	5
	100	4	4	8
	300	2	1	3
	500	5	1	6
	1000	2	0	2
	10,000	3	0	3
	100,000	1	0	1
	1,000,000	1	0	1
UCP	500	3	2	5
	1000	2	1	3
	5000	2	3	5
	10,000	4	0	4
TAMU	10	2	1	3
	30	2	1	3
	100	3	0	3
	500	5	0	5

Non-parametric bootstrap resampling was first applied independently to each isolate dataset. For each bootstrap iteration, volunteer outcomes were resampled with replacement. The exponential dose-response parameter k was refitted by maximum

likelihood estimation, and the fitted k -value was retained. A total of 50,000 bootstrap replicates were generated for each isolate, producing three empirical distributions (Fig. B.4).

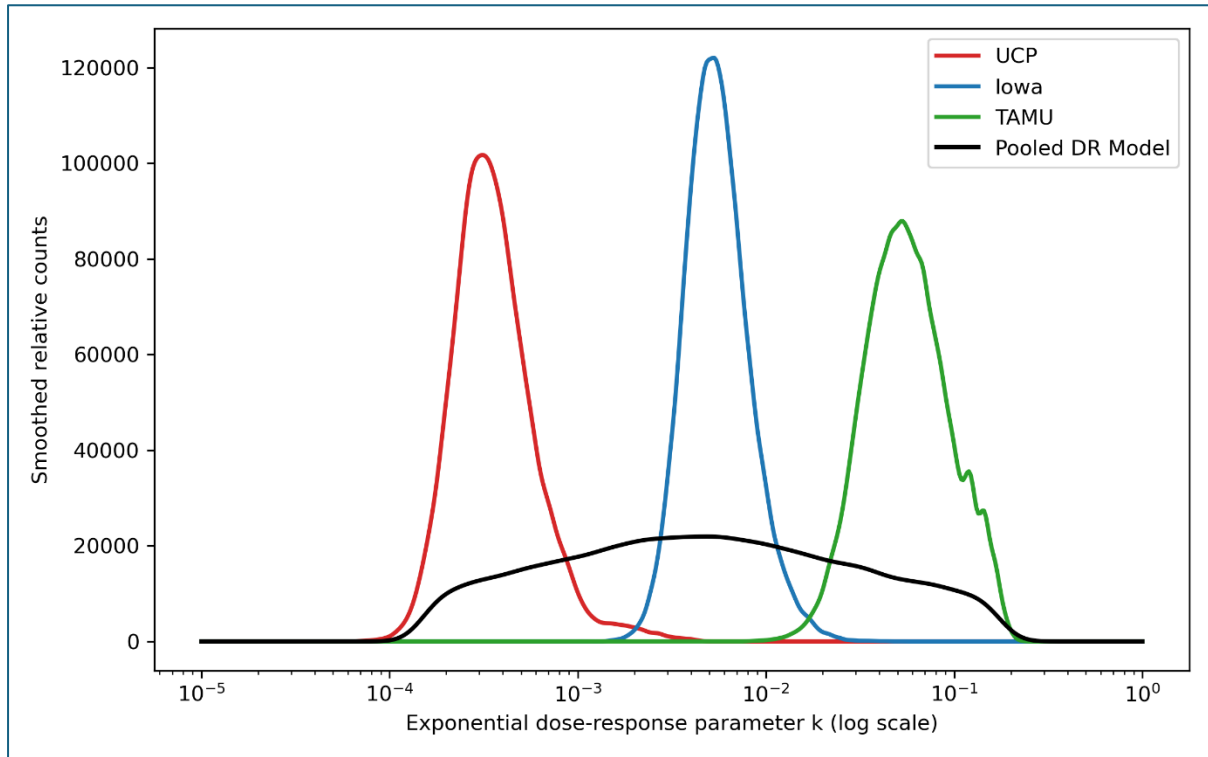


Fig. B.4 Bootstrapped dose-response (k) value distributions for three *C. parvum* isolates and combined distribution generated by hierarchical Bayesian model.

B.2.3 Hierarchical Bayesian Model

The isolate-specific infectivity parameters were assumed to arise from a wider population of human-infective *Cryptosporidium* isolates, as shown in Equation B.1:

$$\log(k_i) \sim \text{Normal}(\mu, \sigma) \quad (\text{B.1})$$

where (k_i) is the dose-response (infectivity) coefficient for isolate i , μ is the population mean log-infectivity, and σ is between-isolate variability. The posterior distribution of μ and σ was estimated using Markov Chain Monte Carlo (MCMC) simulation following the hierarchical approach described in Messner et al. (2001), using 100,000 iterations, 20,000 burn-in iterations and a thinning interval of 10. Weakly informative priors were used for μ and σ .

Posterior samples of μ and σ were used to generate a posterior predictive distribution of k -values representing a broader population of plausible *Cryptosporidium* dose-response coefficients. The generated distribution consisted of 50,000 k -values, which were sampled directly in the pooled DR QMRA simulations (Fig. B.4).

To avoid biologically implausible values arising from the tails of the hierarchical distribution, predictive samples were constrained using the minimum and maximum values derived from the combined bootstrap distributions for the three isolates.

The resulting pooled DR distribution of k -values incorporates both uncertainty in individual challenge studies and variability in infectivity between *Cryptosporidium* isolates.

Consequently, the pooled DR QMRA represents a broader uncertainty analysis than the primary Iowa DR model and should not be interpreted as a precise prediction of annual infection risk.

B.2.4 Pooled DR QMRA Outputs

TW-QMRA risk outputs based on the pooled DR model (Fig. B.5a) differ from those based on the Iowa DR model (Fig. 3.5a) in that the distributions for all seasons overlap the USEPA guideline value of 10^{-4} infections person⁻¹ yr⁻¹. Median values are all below the guideline, but in each case Q95 values are approximately an order of magnitude above it at $\approx 1 \times 10^{-3}$ infections person⁻¹ yr⁻¹. There is only a very weak pattern of seasonality, with winter risk distributions slightly higher than summer distributions. The range within each distribution (based on Q5-Q95 values) is approximately 3 log units, between 1×10^{-6} and 2×10^{-3} infections person⁻¹ yr⁻¹ (i.e. 0.0001% to 0.2%).

Pooled DR RW-QMRA outputs without selective abstraction (Fig. B.5b) show the same strong seasonality seen in the Iowa DR QMRA outputs (Fig. 3.5b), with higher risk distributions for all Winter periods compared with Summer periods. In three periods, the Winter median falls above or almost on the USEPA guideline, and for all Summer and Winter periods the Q95 values exceed the guideline. The spread (Q5-Q95) of risk across all periods is greater than for the treated outputs, ranging across almost 4 log units from 4×10^{-7} to 3×10^{-3} infections person⁻¹ yr⁻¹ (0.00004% to 0.3%). Grouped separately, the Q5-Q95 range is ≈ 3 log units for the Winter periods and ≈ 3 log units for the Summer. Summer periods are extremely consistent across the 6-year study period.

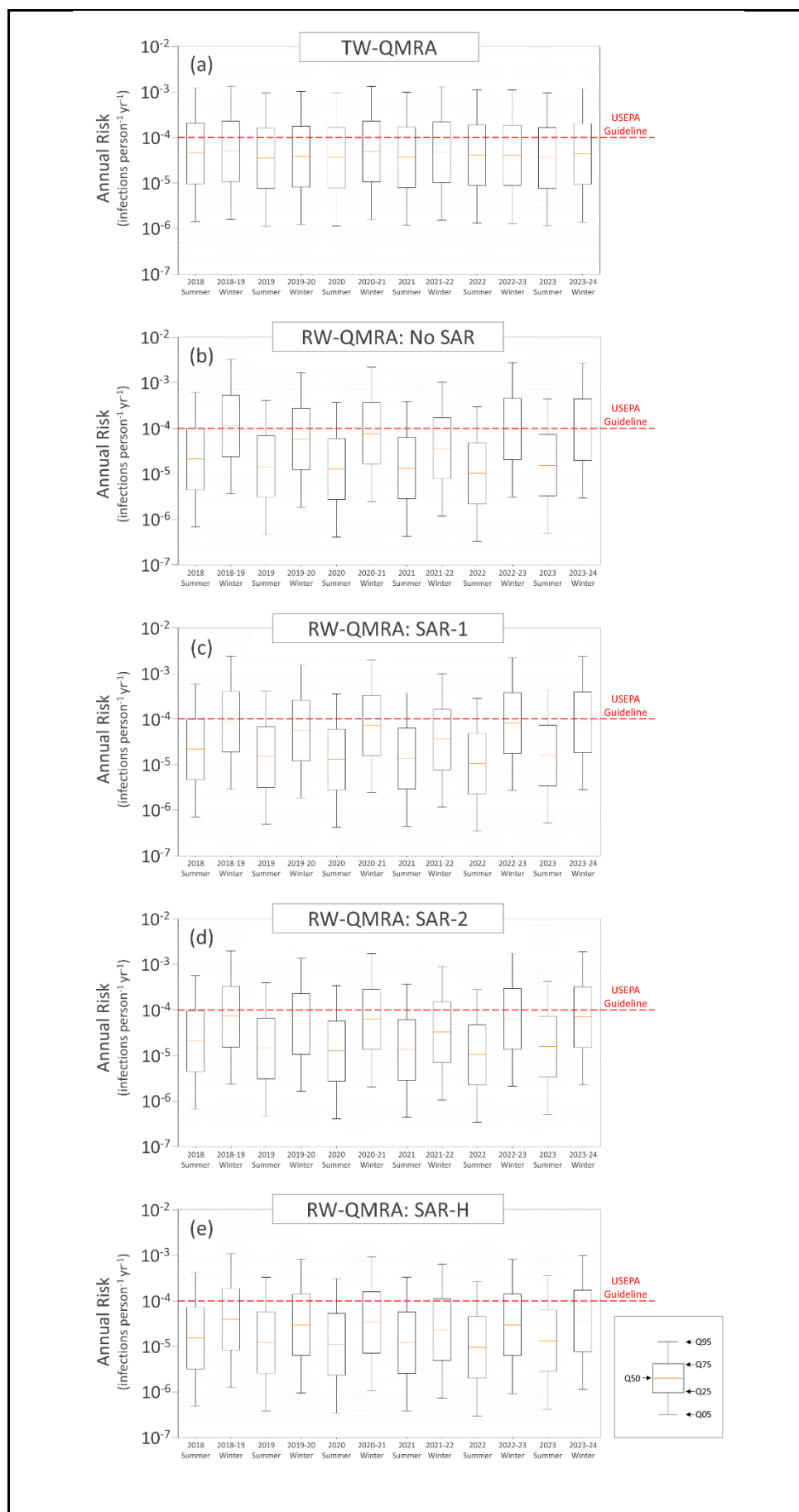


Fig. B.5 Annual infection risk distribution outputs from the pooled DR QMRA models for biannual periods, based on *Cryptosporidium* concentrations in (a) treated water, and (b-e) river water with (b) no selective abstraction, (c) SAR-1, (d) SAR-2 and (e) SAR-H.

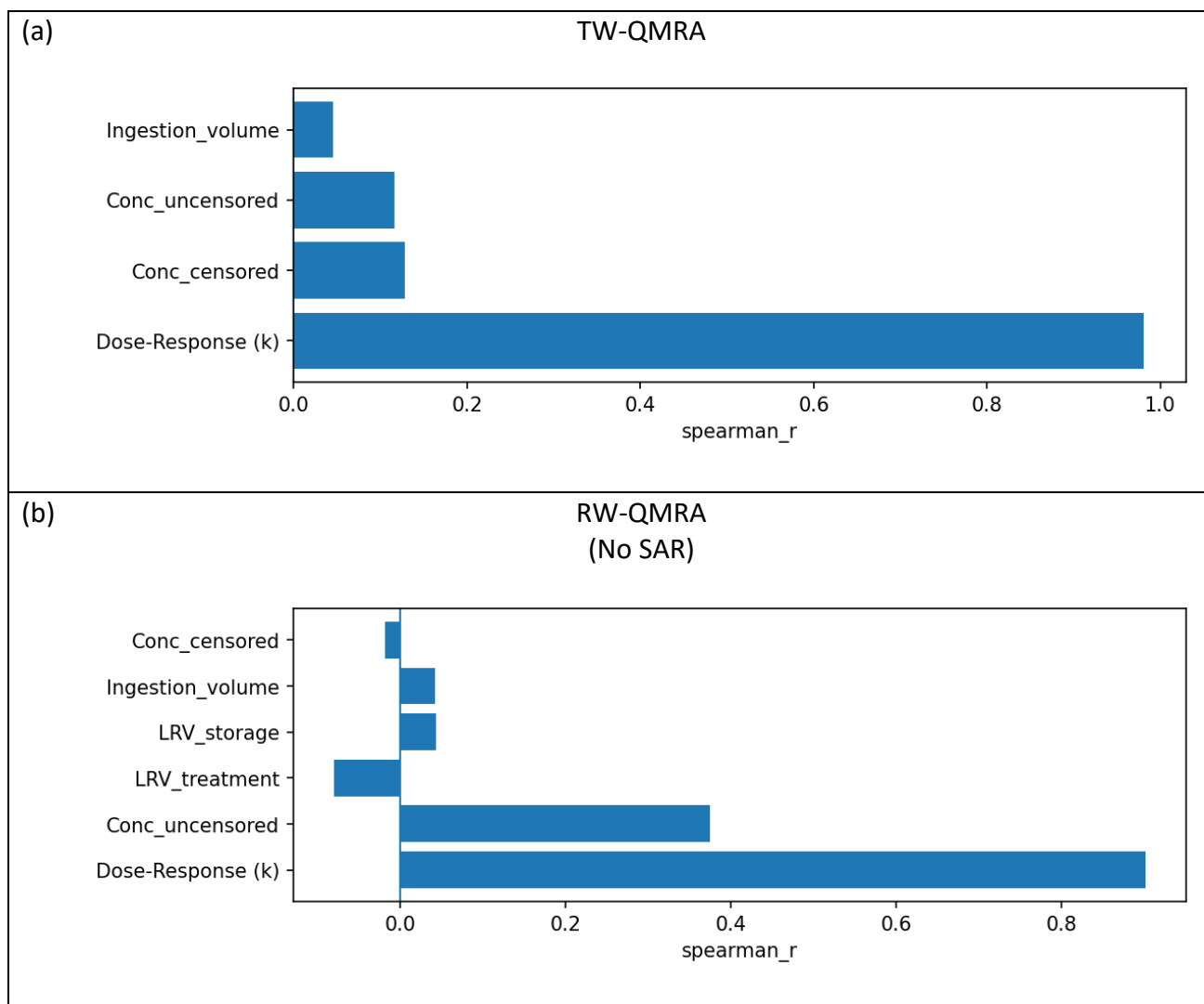


Fig. B.6 Sensitivity analysis based on Spearman correlation coefficient (r), for (a) TW-QMRA (b) RW-QMRA (without SAR), using the pooled DR model.

Annual risk statistics across the full six-year study period are also provided in Chapter 3 (Table 3.4), showing that the medians for both TW-QMRA and RW-QMRA fall well below the USEPA guideline, whereas Q95 for RW-QMRA (No SAR) falls very slightly above the guideline. This illustrates the innately conservative nature of biannual modelling (i.e. inclined to generate higher risk outputs for some periods) when compared to models which include data from one or more complete years. Applying dose-response k -values from the pooled DR model resulted in similar median values across all models, but raised Q95 values by at least an order of magnitude.

Sensitivity analysis (Fig. B.6) shows that dose-response (k), has a dominant effect on model outputs. This is particularly true for the TW-QMRA, where it is almost entirely dominant

(Spearman's $r \approx 1$). For RW-QMRA, dose-response is the primary variable ($r \approx 0.9$), but uncensored concentration is still an important secondary variable ($r \approx 0.37$). The pooled DR QMRA therefore captures broader dose-response uncertainty and may be appropriate where a conservative upper-percentile (Q95) risk estimate is required.

Appendix C: Supplementary Material for Chapter 4

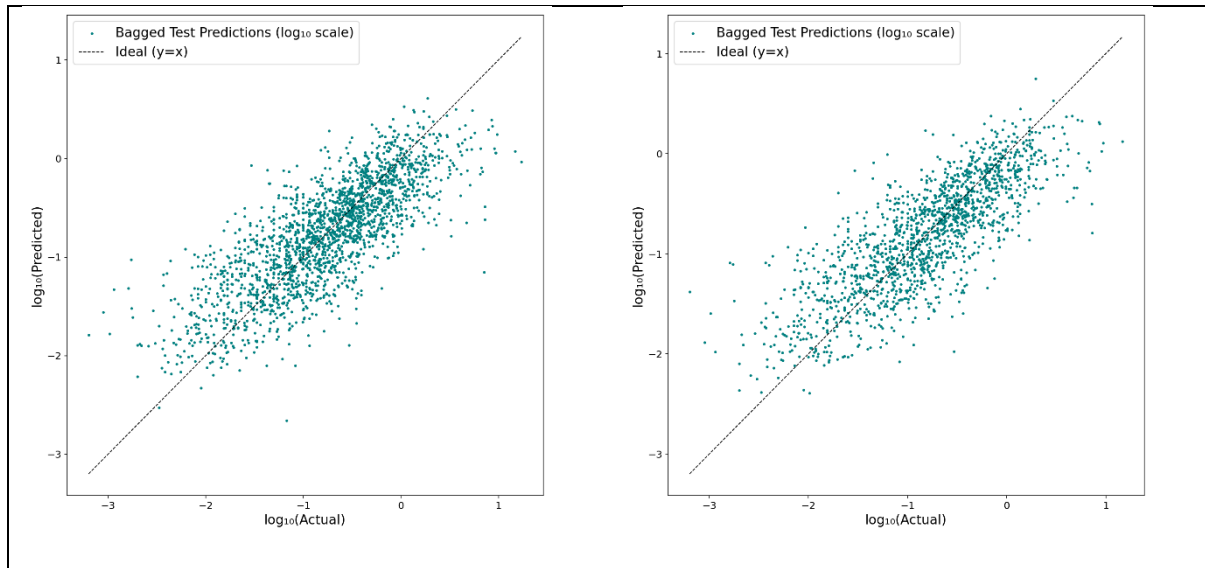


Fig. C.1 Out-of-fold (OoF) predictions from (a) Model 2 and (b) Model 3.

Table C.1 Top 20 SHAP interaction contributions for humidity features with land use and source-related features.

Rank	Humidity Feature	Interacting Feature	Mean Absolute Interaction Contribution ($\times 10^{-3}$)
1	Humidity_(25km)	Human_Pop_(25km)	9.44
2	Humidity_(fc)	Human_Pop_(25km)	7.54
3	Humidity_(05km)	Human_Pop_(25km)	7.49
4	Humidity_(fc)	Land_Urban_(25km)	7.33
5	Humidity_(fc)	Land_Grassland_(25km)	5.15
6	Humidity_(25km)	Land_Grassland_(fc)	4.90
7	Humidity_(fc)	Land_Heath_&_Bog_(25km)	4.64
8	Humidity_(fc)	Land_Grassland_(fc)	3.76
9	Humidity_(25km)	Land_Urban_(25km)	3.69
10	Humidity_(05km)	Land_Urban_(25km)	3.69
11	Humidity_(25km)	Land_Grassland_(25km)	3.56
12	Humidity_(fc)	Land_Arable_(25km)	3.41
13	Humidity_(fc)	Land_Urban_(05km)	3.39
14	Humidity_(05km)	Land_Grassland_(25km)	3.36
15	Humidity_(fc)	Poultry_Pop_(05km)	3.36
16	Humidity_(05km)	Land_Grassland_(fc)	3.32
17	Humidity_(fc)	Land_Forested_(fc)	3.29
18	Humidity_(fc)	Human_Pop_(05km)	3.15
19	Humidity_(fc)	Land_Heath_&_Bog_(fc)	3.15
20	Humidity_(25km)	Poultry_Pop_(05km)	3.12