Science of the Total Environment 991 (2025) 179794

Contents lists available at ScienceDirect

Science of the Total Environment

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

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

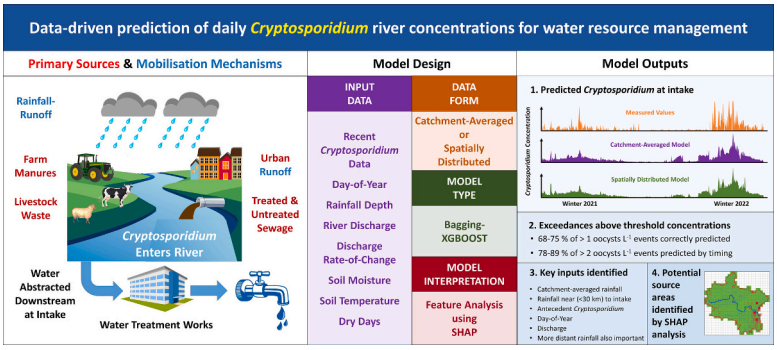
Alan L. Smalley [a,*], Isabel Douterelo [a], Michael Chipps [b], James D. Shucksmith [a]

[a] University of Sheffield, Sheffield S1 3JD, UK
[b] Thames Water Research, Development and Innovation, Kempton Park AWTW, Hanworth TW13 6XH, UK

HIGHLIGHTS

- *Cryptosporidium* river concentrations were modelled in a large, complex catchment.
- A range of potential inputs were trialled to explore effects on model performance.
- Final models predicted 69–75 % of >1 oocysts L^{-1} exceedances.
- Explainable AI methods revealed importance of both animal and human/urban sources.
- Models can inform abstraction strategy to reduce *Cryptosporidium* loadings to WTWs.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Ouyang Wei

Keywords:
Cryptosporidium
Water quality
Catchment modelling
Machine learning
Surface water
Abstraction management
Public health

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 *Cryptosporidium* 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

* Corresponding author.
E-mail addresses: alan.smalley@sheffield.ac.uk (A.L. Smalley), i.douterelo@sheffield.ac.uk (I. Douterelo), michael.chipps@thameswater.co.uk (M. Chipps), j.shucksmith@sheffield.ac.uk (J.D. Shucksmith).

https://doi.org/10.1016/j.scitotenv.2025.179794
Received 10 February 2025; Received in revised form 27 May 2025; Accepted 27 May 2025
Available online 20 June 2025

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

1. 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 4-6 μm 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. 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 (≈ 300 km^2) test catchments.

Previous attempts to explicitly model *Cryptosporidium* river concentrations are summarised in Fig. 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 km^2) homogenous catchments (Coffey et al., 2010; Tang et al., 2011) and a third (Liu et al., 2019) looking at a larger (≈ 4000 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 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; Ferguson et al., 2005) or the global scale (Hofstra and Vermeulen, 2016; Vermeulen et al., 2019) predict loadings to water courses averaged over the monthly, yearly or 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 km^2) 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 h) 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

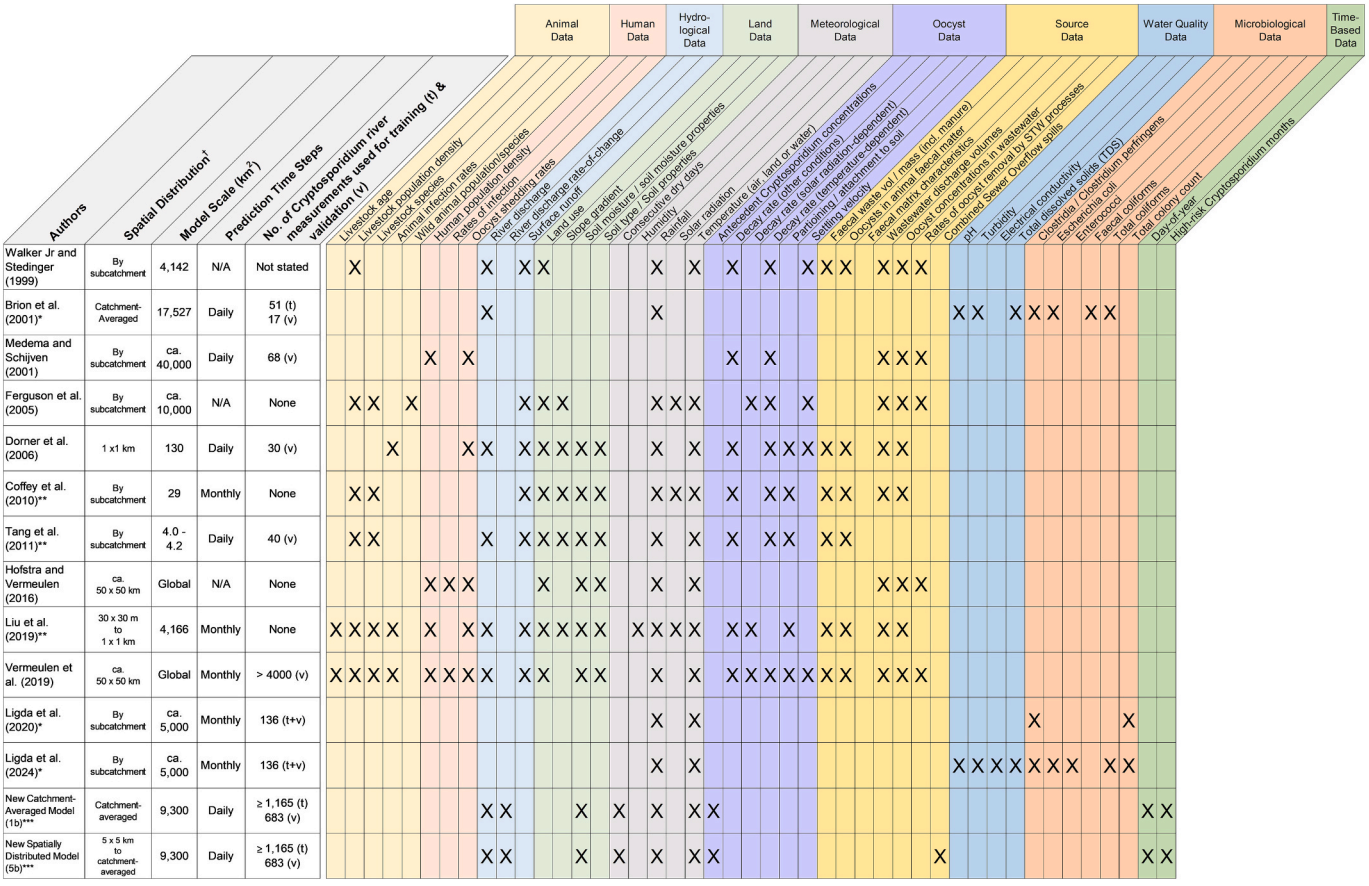


Fig. 1. Summary of key characteristics and input parameters used in previous *Cryptosporidium* models. (Notes: †Spatial distribution often varies with input type, hence sometimes two figures are provided; *ML models (all other models are deterministic); **SWAT models; ***New models (presented in this paper). For deterministic models, the number of *Cryptosporidium* river measurements was used for validation (v) only, whereas for ML models, the number of both training (t) and validation (v) data points is provided.)

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. 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 which (if any) produces the best performing model; 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.

3. Methods

3.1. 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). The catchment covers an area of approximately 9300 km^2 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,

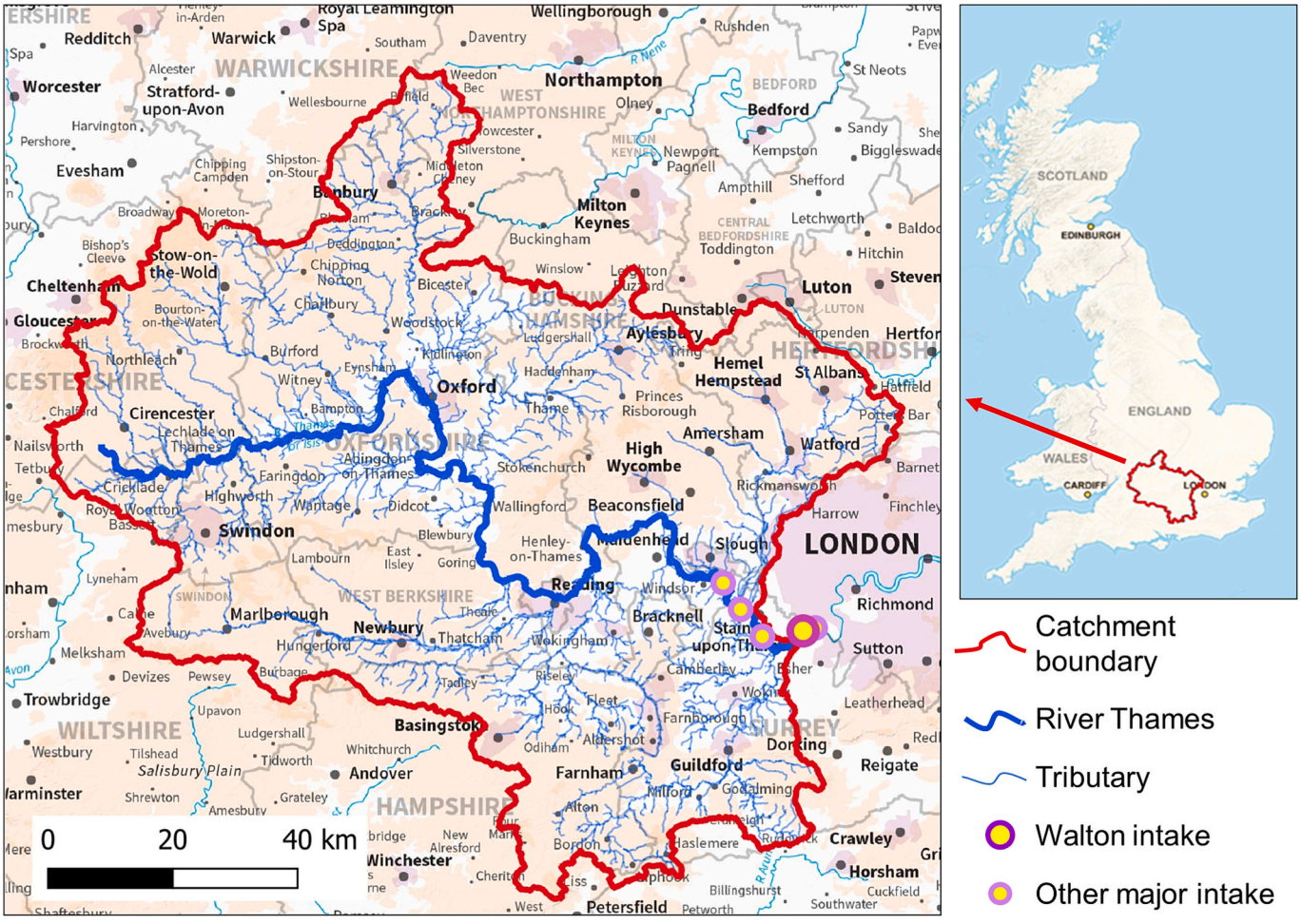


Fig. 2. Maps showing the catchment location on the mainland and the towns, cities and river network within the catchment boundary, along with the five major abstraction points on the River Thames which provide raw water for approximately 70 % of London's drinking water. Base map: OS MiniScale (Ordnance Survey, 2025).

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 5×5 km cells, corresponding to the cells for which rainfall radar data is available.

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 Blue Book method described in The Microbiology of Drinking Water, Part 14 (Environment Agency, 2010).

3.2. 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 decisions 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 and Ho, 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 3.5). XGBoost-based models were found to give the best overall predictive performance (based on the metrics outlined in Section 3.7), 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).

3.3. 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. 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.

3.3.1. 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 to 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.

3.3.2. 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 5x5 km cells at 15-min 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 5x5 km 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 3.4.2).

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 Supplementary material 1 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 m^3 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).

3.3.3. 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 (127) 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 (note this data also includes discharges from storm tanks, but for simplicity all releases of untreated sewage will be referred to henceforth as CSO spills). 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-min 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 (or *CSO_Sum*), was also calculated by summing all of the spilling CSOs for each day.

3.4. Travel times and time lags

3.4.1. 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 et al., 1998). The simplest (catchment-averaged) version of the model (see Table 1) was trialled using a minimum lag time of 1 day and maximum lag (*MaxLag*) times 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 5×5 km cells of the catchment under a Q50 flow regime (see Fig. 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).

3.4.2. 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 5×5 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 datapoints obtained through a combination of tracer tests carried out by Environmental Tracing Systems Ltd. (unpublished results) and a spill time-of transport model developed by Wallingford HydroSolutions (unpublished results). Both sources provided data for limited areas of the catchment, including sections of the

Table 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									
				1a	1b	2a	2b	3a	3b	4a	4b	5a	5b
Core	Day of year	*Day_of_Year*	Day	■	■	■	■	■	■	■	■	■	■
	High Season Oct-Feb	*High_Season*	-	■	■	■	■	■	■	■	■	■	■
	Cryptosporidium 3-week mean	*Crypto_Rolling_Mean*	oocysts L^{-1}	■	■	■	■	■	■	■	■	■	■
Catchment-Averaged	Catchment-averaged rainfall	*Catchment_Rain*	mm		■		■		■		■		■
	Continuous antecedent dry days	*Dry_Days*	Days	■	■	■	■	■	■	■	■	■	■
	Discharge	*Discharge*	$m^3 s^{-1}$	■	■	■	■	■	■	■	■	■	■
	Discharge daily-rate-of-change	*Discharge_ROC*	$m^3 s^{-1}$	■	■	■	■	■	■	■	■	■	■
	Soil moisture	*Soil_Moisture*	% volume	■	■	■	■	■	■	■	■	■	■
	Soil temperature	*Soil_Temp*	ºC	■	■	■	■	■	■	■	■	■	■
	Catchment-wide CSO spills	*CSO_Sum*	Spills			■	■						
Sp. Dist.*	Cell-based rainfall	*Rainfall_Cell*	mm					■	■			■	■
	CSO spills (individual CSOs)	*CSO*	Spills							■	■	■	■

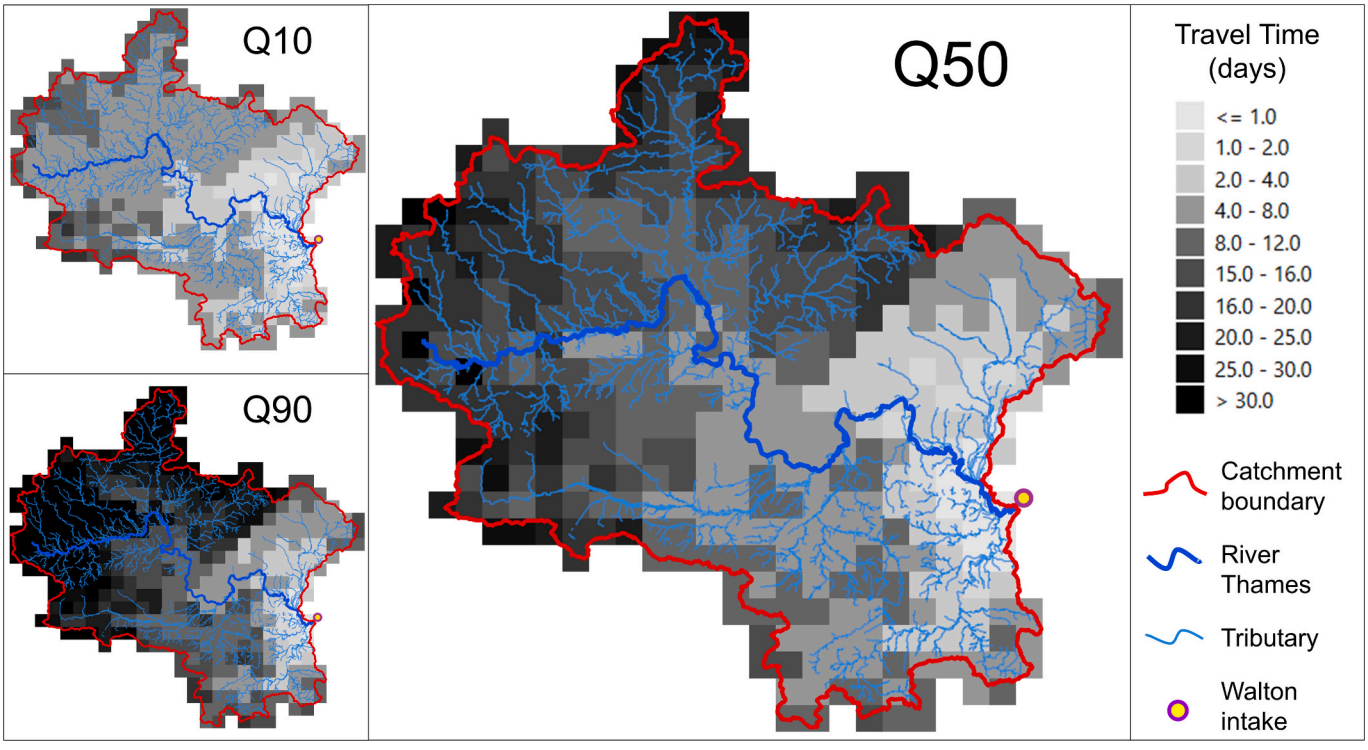


Fig. 3. Estimated travel times for 5×5 km cells under three different flow regimes in the study catchment.

Thames and a number of its tributaries. Curve-fitting was applied separately to the tributaries, to the upper Thames (with catchment areas of <1600 km^2), and to the lower Thames (with catchment areas of >1600 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 5×5 km catchment cell, following which 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. 3. Further information concerning curve-fitted equations and error analysis is provided in Supplementary material 1.

3.5. Input selection for models

Table 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. (A full list of inputs for each model is provided in Supplementary material 2.)

3.6. 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 *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.

3.7. 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. 4). This is as 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 et al., 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). 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

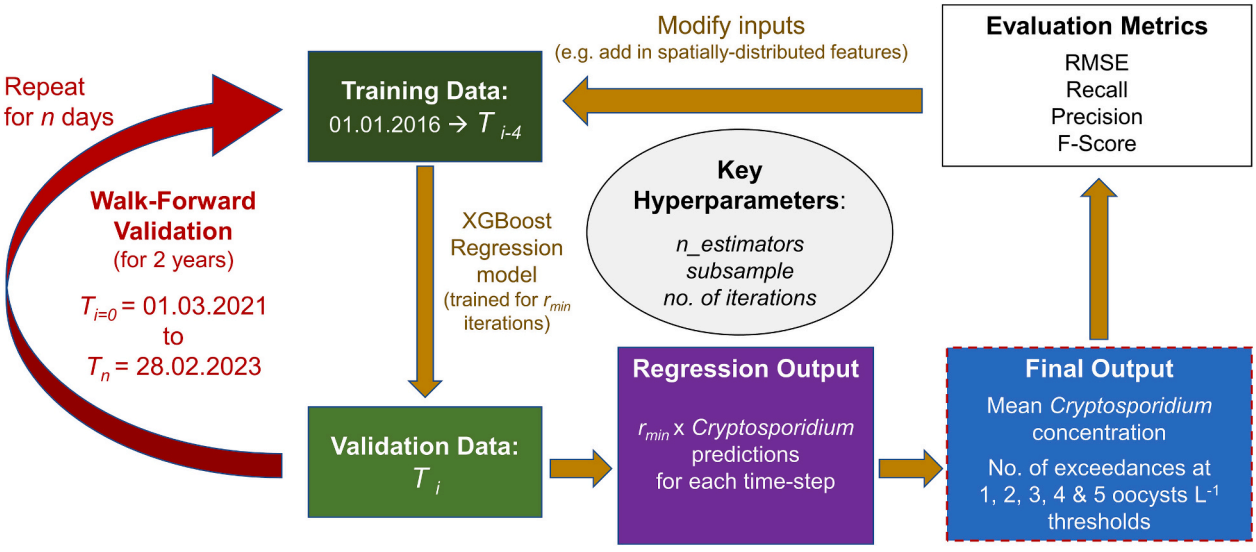


Fig. 4. Diagram of model processes and model outputs, showing dates (T) over which training and validation data is split for walk forward validation, alongside key hyperparameters, outputs and evaluation metrics.

method described in Section 3.2.

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 most important for assessing model performance.

3.8. 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).

4. Results

4.1. 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. 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 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 Supplementary material 1 for a more detailed presentation of the hyperparameter tuning outputs.)

Table 2
Optimised maximum lag time and hyperparameter settings following initial trials.

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

4.2. Overall performance and comparison of models

Performance metric results for the different models are presented in Table 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 twice the value of the R_2 scores in every case). A maximum of two higher exceedances (>3 oocysts L^{-1}) were 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 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. 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 3). Measured *Cryptosporidium* peaks were higher and more frequent in Oct 2022-Feb 2023 than they 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

Table 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 the highest (green), intermediate (yellow) and lowest (red) performing models for each 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	
Metric Type	**Threshold** (oocysts L^{-1})	**Metric**	**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 Exceed-ances	1	64 (9.4%)										
	2	*23* (3.4%)										
	3	12 (1.8%)										
	4	6 (0.9%)										
	5	4 (0.6%)										
Total Measurements		683										

more and less severe *Cryptosporidium* seasons. Fig. 5b–c shows 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.

4.3. SHAP analysis

Fig. 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. 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. 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* ≥

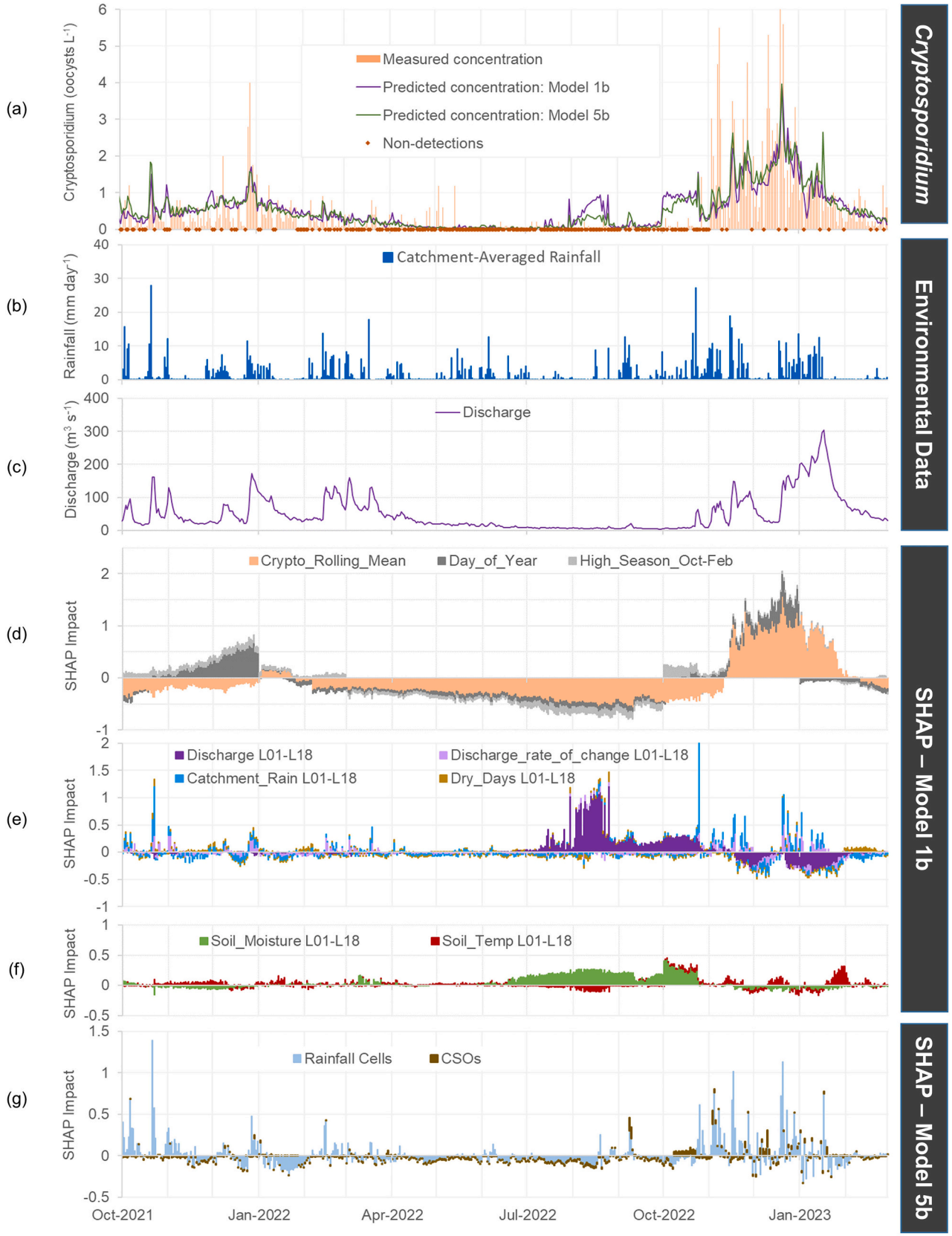


Fig. 5. (a) Measured and predicted concentrations of *Cryptosporidium* (generated by Models 1b and 5b), (b) catchment-averaged rainfall and (c) discharge. SHAP values showing the impact on Model 1b predictions of (d) core inputs, (e) hydrological inputs and (f) soil meteorological inputs (presented as 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.)

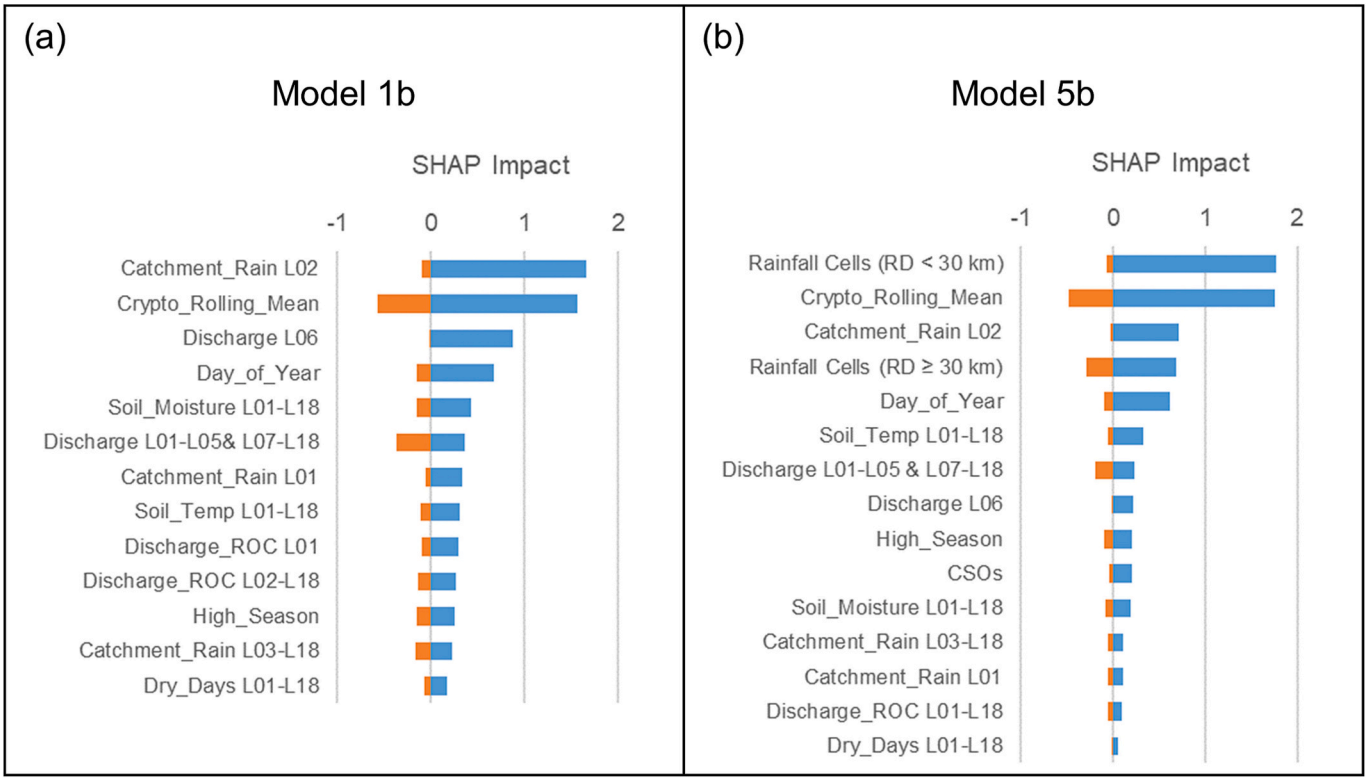


Fig. 6. Maximum positive and negative SHAP impact scores for key individual features and grouped features in (a) Model 1b and (b) Model 5b. (Note: catchment-averaged and spatially distributed features with an absolute maximum impact of <0.25 are grouped together by feature type; RD = River Distance (km) from Walton intake; Ln = lag time, expressed as n days.)

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.

4.4. Identifying possible source locations/areas

Fig. 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* sources areas/locations (see further discussion on this in Section 5.4). 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. 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 livestock distributions, although the higher-impact cells in the west do coincide with higher cattle densities (Fig. 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.

5. Discussion

5.1. 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 time scales 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.

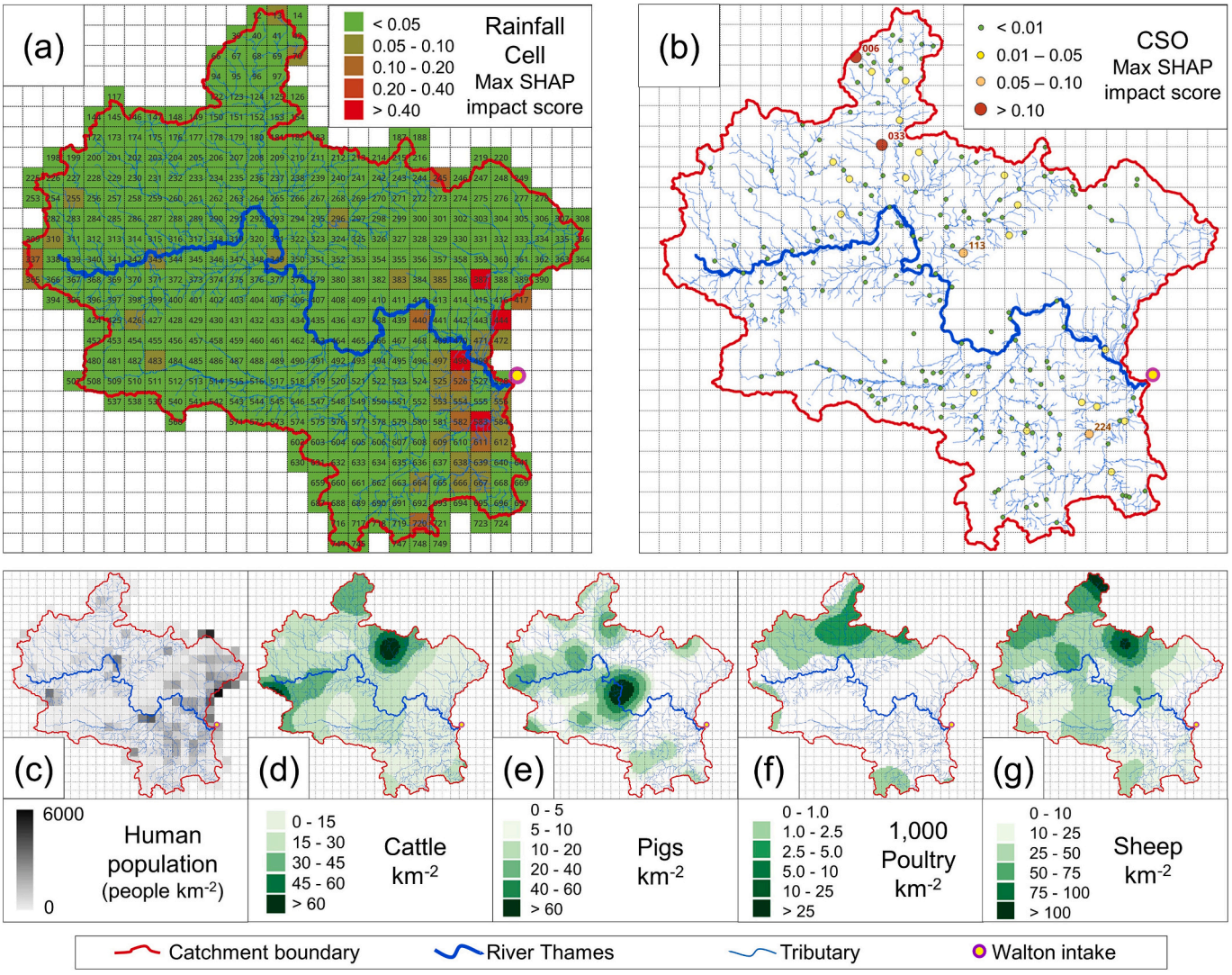


Fig. 7. Maximum SHAP impact scores for (a) 5×5 km rainfall cells and (b) CSO binary inputs, with (c) human population density from 2011 Census (Reis et al., 2017), alongside livestock densities for (d) cattle, (e) pigs (f) poultry and (g) sheep from APHA (2022).

5.2. 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. 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.

5.3. 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. 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).

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 period 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. 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. 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 most impactful features – *Crypto_Rolling_Mean* and *Day_of_Year* – serve as surrogates for other processes for which we lack the 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 Oct 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, 2019; 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 purpose of providing a scaling factor to inform the model of the severity of the current *Cryptosporidium* season.

5.4. 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). 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 (using sub-daily *Cryptosporidium* and environmental data) may reveal subtleties that are not visible in the current models. 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 the 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).

6. Conclusions

The present study represents 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 data for *Cryptosporidium* in raw water. 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 of 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 to support future modelling and advance understanding of *Cryptosporidium* sources, transport and temporal trends.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2025.179794.

CRediT authorship contribution statement

Alan L. Smalley: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Isabel Douterelo:** Writing – review & editing, Supervision, Methodology, Investigation. **Michael Chipps:** Writing – review & editing, Supervision, Resources, Methodology, Investigation. **James D. Shucksmith:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Investigation, Formal analysis, Conceptualization.

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.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Alan L. Smalley and Michael Chipps report financial support was provided by Thames Water Utilities Limited (via indirect studentship funding and by direct employment, respectively). The other authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors are extremely grateful for the assistance provided by staff at Thames Water, including Dean Khanna, Jo Clint, James Townsend, Lucy Parkes and many others.

Data availability

The authors do not have permission to share data.

References

Alomari, Y., Andó, M., 2024. SHAP-based insights for aerospace PHM: temporal feature importance, dependencies, robustness, and interaction analysis. Res. Eng. Des. 21, 101834. https://doi.org/10.1016/j.rineng.2024.101834.

Alpaydin, E., 2014. Introduction to Machine Learning. MIT Press, pp. 547–591. http://ieeexplore.ieee.org.sheffield.idm.oclc.org/document/6917149.

Alsahaf, A., Petkov, N., Shenoy, V., Azzopardi, G., 2022. A framework for feature selection through boosting. Expert Syst. Appl. 187, 115895. https://doi.org/10.1016/j.eswa.2021.115895.

APHA, 2022. Livestock Demographic Data Group. Available at:, Animal & Plant Health Agency. https://www.gov.uk/guidance/view-apha-surveillance-reports-publications-and-data#disease-surveillance-dashboards. (Accessed 8 February 2025).

APHA, 2024. The Cattle Disease Surveillance Dashboard. Animal & Plant Health Agency. Available at: https://public.tableau.com/app/profile/siu.apha/viz/CattleDashboard/Overview. (Accessed 8 February 2025).

Asfaw, A., Maher, K., Shucksmith, J.D., 2018. Modelling of metaldehyde concentrations in surface waters: a travel time based approach. J. Hydrol. 562, 397–410. https://doi.org/10.1016/j.jhydrol.2018.04.074.

Atherholt, T.B., LeChevallier, M.W., Norton, W.D., Rosen, J.S., 1998. Effect of rainfall on Giardia and crypto. J. Am. Water Works Assoc. 90 (9), 66–80. https://doi.org/10.1002/j.1551-8833.1998.tb08499.x.

Betancourt, W.Q., Rose, J.B., 2004. Drinking water treatment processes for removal of *Cryptosporidium* and *Giardia*. Vet. Parasitol. 126 (1), 219–234. https://doi.org/10.1016/j.vetpar.2004.09.002.

Bhattarai, R., Kalita, P., Trask, J., Kuhlenschmidt, M.S., 2011. Development of a physically-based model for transport of *Cryptosporidium parvum* in overland flow. Environ. Model Softw. 26 (11), 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. https://doi.org/10.1007/978-1-4842-4470-8.

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.

Bouzid, M., Hunter, P.R., Chalmers, R.M., Tyler, K.M., 2013. Cryptosporidium pathogenicity and virulence. Clin. Microbiol. Rev. 26 (1), 115–134. https://doi.org/10.1128/cmr.00076-12.

Brion, G.M., Neelakantan, T.R., Lingireddy, S., 2001. Using neural networks to predict peak *Cryptosporidium* concentrations. J. Am. Water Works Assoc. 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., Fricker, C.R., 1997. Occurrence of *Cryptosporidium* spp oocysts and *Giardia* spp cysts in sewage influents and effluents from treatment plants in England. Water Sci. Technol. 35 (11–12), 385–390. https://doi.org/10.1016/S0273-1223(97)00290-4.

Chan, J.Y.-L., Leow, S.M.H., Bea, K.T., Cheng, W.K., Phoong, S.W., Hong, Z.-W., 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.

Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting system. ArXiv.org. https://doi.org/10.48550/arXiv.1603.02754.

Chen, J., Chen, S., Fu, R., Li, D., Jiang, H., Wang, C., Peng, Y., Jia, K., 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.

Choubin, B., Darabi, H., Rahmati, O., Sajedi-Hosseini, F., Kløve, B., 2018. River suspended sediment modelling using the CART model: a comparative study of machine learning techniques. Sci. Total Environ. 615, 272–281. https://doi.org/10.1016/j.scitotenv.2017.09.293.

Coffey, R., Cummins, E., Flaherty, V.O., Cormican, M., 2010. Analysis of the soil and water assessment tool (SWAT) to model *Cryptosporidium* in surface water sources. Biosyst. Eng. 106 (3), 303–314. https://doi.org/10.1016/j.biosystemseng.2010.04.003.

Crockett, C.S., 2004. The Significance of Streambed Sediments as a Reservoir of *Cryptosporidium* oocysts. Drexel University (PhD thesis).

Daraei, H., Oliveri Conti, G., Sahlabadi, F., Thai, V.N., Gholipour, S., Turki, H., Fakhri, Y., Ferrante, M., Moradi, A., Mousavi Khaneghah, A., 2020. Prevalence of *Cryptosporidium* spp. in water: a global systematic review and meta-analysis. Environ. Sci. Pollut. Res. Int. 28 (8), 9498–9507. https://doi.org/10.1007/s11356-020-11261-6.

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., Chen, X., 2022. Bagging–XGBoost algorithm based extreme weather identification and short-term load forecasting model. Energy Rep. 8, 8661–8674. https://doi.org/10.1016/j.egyr.2022.06.072.

Dhal, P., Azad, C., 2022. A comprehensive survey on feature selection in the various fields of machine learning. Appl. Intell. 52 (4), 4543–4581. https://doi.org/10.1007/s10489-021-02550-9.

Dobson, B., Mijic, A., 2020. Protecting rivers by integrating supply-wastewater infrastructure planning and coordinating operational decisions. Environ. Res. Lett. 15 (11), 114025. https://doi.org/10.1088/1748-9326/abb050.

Dormann, C.F., Calabrese, J.M., Guillera-Arroita, 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., Hartig, F., 2018. Model averaging in ecology: a review of Bayesian, information-theoretic, and tactical approaches for predictive inference. Ecol. Monogr. 88 (4r), 485–504. https://doi.org/10.1002/ecm.1309.

Dorner, S.M., Anderson, W.B., Slawson, R.M., Kouwen, N., Huck, P.M., 2006. Hydrologic modeling of pathogen fate and transport. Environ. Sci. Technol. 40 (15), 4746–4753. https://doi.org/10.1021/es060426z.

Drobnič, F., Kos, A., Pustišek, M., 2020. On the interpretability of machine learning models and experimental feature selection in case of multicollinear data. Electronics (Basel) 9 (5), 761. https://doi.org/10.3390/electronics9050761.

Drummond, J.D., Boano, F., Atwill, E., Li, X., Harter, T., Packman, A., 2018. Cryptosporidium oocyst persistence in agricultural streams–a mobile-immobile model framework assessment. Sci. Rep. 8 (1), 4603.

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).

Efstratiou, A., Ongerth, J.E., Karanis, P., 2017a. Waterborne transmission of protozoan parasites: review of worldwide outbreaks - an update 2011–2016. Water Res. 114, 14–22. https://doi.org/10.1016/j.watres.2017.01.036.

Efstratiou, A., Ongerth, J., Karanis, P., 2017b. Evolution of monitoring for Giardia and Cryptosporidium in water. Water Res. 123, 96–112. https://doi.org/10.1016/j.watres.2017.06.042.

Environment Agency, 2010. The Microbiology of Drinking Water (2010) - Part 14 - Methods for the Isolation, Identification and Enumeration of Cryptosporidium Oocysts and Giardia Cysts. Environment Agency, Bristol. Available at: https://assets.publishing.service.gov.uk/media/5be9956aed915d6a0d6f6fe8/Part_14-oct20-234.pdf. (Accessed 5 February 2025).

Environmental Tracing Systems Ltd, unpublished results. River Thames Catchment Time-of-Travel Studies.

Ferguson, C.M., Croke, B., Ashbolt, N.J., Deere, D.A., 2005. A deterministic model to Rquantify pathogen loads in drinking water catchments: pathogen budget for the Wingecarribee. Water Sci. Technol. 52 (8), 191–197. https://doi.org/10.2166/wst.2005.0262.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29 (5), 1189–1232. https://doi.org/10.1214/aos/1013203451.

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

Giakoumis, T., Voulvoulis, N., 2023. Combined sewer overflows: relating event duration monitoring data to wastewater systems' capacity in England. Environ. Sci.: Water Res. Technol. 9 (3), 707–722. https://doi.org/10.1039/D2EW00637E.

Gibson III, C.J., Stadterman, K.L., States, S., Sykora, J., 1998. Combined sewer overflows: a source of *Cryptosporidium* and *Giardia*? Water Sci. Technol. 38 (12), 67–72. https://doi.org/10.1016/S0273-1223(98)00802-6.

Golomazou, E., Mamedova, S., Eslahi, A.V., Karanis, P., 2024. Cryptosporidium and agriculture: a review. Sci. Total Environ. 916, 170057. https://doi.org/10.1016/j.scitotenv.2024.170057.

Hassan, E.M., Örmeci, B., Derosa, M.C., Dixon, B.R., Sattar, S.A., Iqbal, A., 2021. A review of Cryptosporidium spp. and their detection in water. Water Sci. Technol. 83 (1), 1–25. https://doi.org/10.2166/wst.2020.515.

Hastie, T., Friedman, J.H., Tibshirani, R., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edition. Springer, New York: New York. https://doi.org/10.1007/978-0-387-84858-7.

Hofstra, N., Vermeulen, L.C., 2016. Impacts of population growth, urbanisation and sanitation changes on global human *Cryptosporidium* emissions to surface water. Int. J. Hyg. Environ. 219 (7), 599–605. https://doi.org/10.1016/j.ijheh.2016.06.005.

Jobson, H.E., 1997. Predicting travel time and dispersion in rivers and streams. J. Hydraul. Eng. 123 (11), 971–978. https://doi.org/10.1061/(ASCE)0733-9429(1997)123:11(971).

King, B.J., Keegan, A.R., Monis, P.T., Saint, C.P., 2005. Environmental temperature controls Cryptosporidium oocyst metabolic rate and associated retention of infectivity. Appl. Environ. Microbiol. 71 (7), 3848–3857. https://doi.org/10.1128/aem.71.7.3848-3857.2005.

Knapp, J.L.A., Li, L., Musolff, A., 2022. Hydrologic connectivity and source heterogeneity control concentration–discharge relationships. Hydrol. Process. 36 (9), e14683. https://doi.org/10.1002/hyp.14683.

Ligda, P., Claerebout, E., Kostopoulou, D., Zdragas, A., Casaert, S., Robertson, L.J., Sotiraki, S., 2020. Cryptosporidium and Gi*ardia* in surface water and drinking water: animal sources and towards the use of a machine-learning approach as a tool for predicting contamination. Environ. Pollut. 264, 114766. https://doi.org/10.1016/j.envpol.2020.114766.

Ligda, P., Mittas, N., Kyzas, G.Z., Claerebout, E., Sotiraki, S., 2024. Machine learning and explainable artificial intelligence for the prevention of waterborne cryptosporidiosis and giardiasis. Water Res. 262, 122110. https://doi.org/10.1016/j.watres.2024.122110.

Liu, W., An, W., Jeppesen, E., Ma, J., Yang, M., 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. J. Environ. Manage. 232, 462–474. https://doi.org/10.1016/j.jenvman.2018.10.064.

Lundberg, S., Lee, S.-I., 2017. A unified approach to interpreting model predictions. arXiv.org. https://doi.org/10.48550/arXiv.1705.07874.

Mamun, A.A., Shams, S., Nuruzzaman, M., 2020. Review on uncertainty of the first-flush phenomenon in diffuse pollution control. Appl Water Sci 10 (1), 53. https://doi.org/10.1007/s13201-019-1127-1.

Masood, A., Niazkar, M., Zakwan, M., 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., Schijven, J.F., 2001. Modelling the sewage discharge and dispersion of Cryptosporidium and Giardia in surface water. Water Res. 35 (18), 4307–4316. https://doi.org/10.1016/S0043-1354(01)00161-0.

Met Office, 2003. 5 km resolution UK composite rainfall data from the Met Office Nimrod system. Available at:, NCAS British Atmospheric Data Centre. https://catalogue.ceda.ac.uk/uuid/f91b2c5399c5bf689e29bb15ab45da8a. (Accessed 5 February 2025).

Met Office, Hollis, D., McCarthy, M., Kendon, M., Legg, T., Simpson, I., 2021. HadUK-Grid Gridded Climate Observations on a 5km Grid Over the UK, v1.0.3.0 (1862–2020). NERC EDS Centre for Environmental Data Analysis. https://doi.org/10.5285/f2da35c56afb4fa6aebf44094b65dff3 (08 September 2021).

Müller, A., Österlund, H., Marsalek, J., Viklander, M., 2020. The pollution conveyed by urban runoff: a review of sources. Sci. Total Environ. 709, 136125. https://doi.org/10.1016/j.scitotenv.2019.136125.

Nasser, A.M., 2016. Removal of *Cryptosporidium* by wastewater treatment processes: a review. J. Water Health 14 (1), 1–13. https://doi.org/10.2166/wh.2015.131.

Niazkar, M., Menapace, A., Brentan, B., Piraei, R., Jimenez, D., Dhawan, P., Righetti, M., 2024. Applications of XGBoost in water resources engineering: a systematic literature review (Dec 2018–May 2023). Environ. Model. Softw. 174. https://doi.org/10.1016/j.envsoft.2024.105971.

O'Brien, R.M., 2017. Dropping highly collinear variables from a model: why it typically is not a good idea. Soc. Sci. Q. 98 (1), 360–375. https://doi.org/10.1111/ssqu.12273.

Ordnance Survey, 2025. OS data download and API products. Available at: https://www.ordnancesurvey.co.uk/products. (Accessed 8 February 2025).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830. Available at: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf. (Accessed 8 February 2025).

Pham, X.T.T., Ho, T.H., 2021. Using boosting algorithms to predict bank failure: an untold story. Int. Rev. Econ. Financ. 76, 40–54. https://doi.org/10.1016/j.iref.2021.05.005.

Piraei, R., Afzali, S.H., Niazkar, M., 2023. Assessment of XGBoost to estimate total sediment loads in rivers. Water Resour. Manag. 37 (13), 5289–5306. https://doi.org/10.1007/s11269-023-03606-w.

Public Health England, 2019. Research and analysis: Cryptosporidium data 2008 to 2017. Available at: https://www.gov.uk/government/publications/cryptosporidium-national-laboratory-data/cryptosporidium-data-2008-to-2017. (Accessed 8 February 2025).

Rawat, P., Bajaj, M., Sharma, V., Vats, S., March 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.

Reis, S., Liska, T., Steinle, S., Carnell, E., Leaver, D., Roberts, E., Vieno, M., Beck, R., Dragosits, U., 2017. UK Gridded Population 2011 Based on Census 2011 and Land Cover Map 2015. NERC Environmental Information Data Centre. https://doi.org/10.5285/0995e94d-6d42-40c1-8ed4-5090d82471e1 (Accessed: 5 February 2025).

Robertson, L.J., Campbell, A.T., Smith, H.V., 1992. Survival of Cryptosporidium parvum oocysts under various environmental pressures. Appl. Environ. Microbiol. 58 (11), 3494–3500. https://doi.org/10.1128/aem.58.11.3494-3500.1992.

Rose, J.B., Huffman, D.E., Gennaccaro, A., 2002. Risk and control of waterborne cryptosporidiosis. FEMS Microbiol. Rev. 26 (2), 113–123. https://doi.org/10.1016/S0168-6445(02)00090-6.

Singh, N.K., Emanuel, R.E., McGlynn, B.L., Miniat, C.F., 2021. Soil moisture responses to rainfall: implications for runoff generation. Water Resour. Res. 57 (9), e2020WR028827. https://doi.org/10.1029/2020WR028827.

Sinha, S., Hammond, A., 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. Hydrol. Process. 36 (3), e14544. https://doi.org/10.1002/hyp.14544.

Smeets, P.W.M.H., van Dijk, J.C., Stanfield, G., Rietveld, L.C., Medema, G.J., 2007. How can the UK statutory Cryptosporidium monitoring be used for quantitative risk assessment of cryptosporidium in drinking water? J. Water Health 5 (1), 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.

Sturdee, A.P., Chalmers, R.M., Bull, S.A., 1999. Detection of Cryptosporidium oocysts in wild mammals of mainland Britain. Vet. Parasitol. 80 (4), 273–280. https://doi.org/10.1016/S0304-4017(98)00226-X.

Sturdee, A., Foster, I., Bodley-Tickell, A.T., Archer, A., 2007. Water quality and *Cryptosporidium* distribution in an upland water supply catchment, Cumbria, UK. Hydrol. Process. 21 (7), 873–885. https://doi.org/10.1002/hyp.6278.

Suradhaniwar, S., Kar, S., Durbha, S.S., 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., Shucksmith, J.D., 2024. Forecasting acute rainfall driven E. coli impacts in inland rivers based on sewer monitoring and field runoff. Water Res. 248, 120838. https://doi.org/10.1016/j.watres.2023.120838.

Swaffer, B.A., Vial, H.M., King, B.J., Daly, R., Frizenschaf, J., Monis, P.T., 2014. Investigating source water Cryptosporidium concentration, species and infectivity rates during rainfall-runoff in a multi-use catchment. Water Res. 67, 310–320. https://doi.org/10.1016/j.watres.2014.08.055.

Takefuji, Y., 2025. Beyond XGBoost and SHAP: unveiling true feature importance. J. Hazard. Mater. 488, 137382. https://doi.org/10.1016/j.jhazmat.2025.137382.

Tang, J., McDonald, S., Peng, X., Samadder, S.R., Murphy, T.M., Holden, N.M., 2011. Modelling *Cryptosporidium* oocysts transport in small ungauged agricultural catchments. Water Res. 45 (12), 3665–3680. https://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.).

Tran, Q.Q., De Niel, J., Willems, P., 2018. Spatially distributed conceptual hydrological model building: a generic top-down approach starting from lumped models. Water Resour. Res. 54 (10), 8064–8085. https://doi.org/10.1029/2018WR023566.

Vermeulen, L.C., van Hengel, M., Kroeze, C., Medema, G., Spanier, J.E., van Vliet, M.T.H., Hofstra, N., 2019. *Cryptosporidium* concentrations in rivers worldwide. Water Res. 149, 202–214. https://doi.org/10.1016/j.watres.2018.10.069.

Walker Jr., F.R.W., Stedinger, J.R., 1999. Fate and transport model of *Cryptosporidium*. J. Environ. Eng. 125 (4), 325–333. Available at: https://api.semanticscholar.org/CorpusID:98242550. (Accessed 8 February 2025).

Wallingford HydroSolutions, Unpublished results. Time-of-Transport Model (Version 3b) [Excel spreadsheet] (Not publicly available. Property of Thames Water).

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., Stewart, J.R., 2021b. Tracking Major Sources of Water Contamination Using Machine Learning. Front. Microbiol. 11, 616692. https://doi.org/10.3389/fmicb.2020.616692.

Zhu, J.-J., Yang, M., Ren, Z.J., 2023. Machine learning in environmental research: common pitfalls and best practices. Environ. Sci. Technol. 57 (46), 17671–17689. https://doi.org/10.1021/acs.est.3c00026.