Science of the Total Environment 612 (2018) 840–852

Contents lists available at ScienceDirect

Science of the Total Environment

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

Using spatial-stream-network models and long-term data to understand and predict dynamics of faecal contamination in a mixed land-use catchment

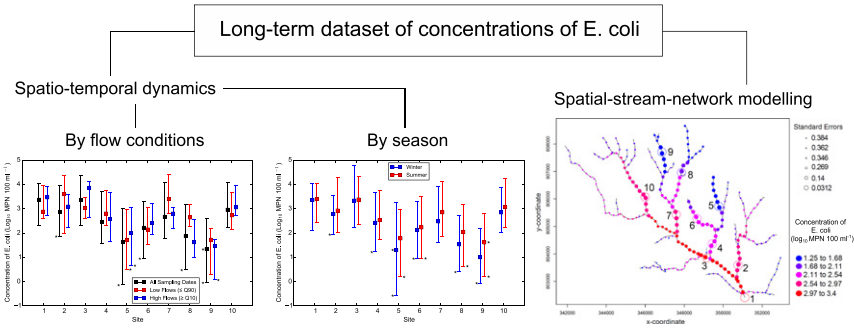
Aaron James Neill [a,b,*], Doerthe Tetzlaff [a,c,d], Norval James Colin Strachan [e], Rupert Lloyd Hough [b], Lisa Marie Avery [b], Helen Watson [b], Chris Soulsby [a]

[a] Northern Rivers Institute, School of Geosciences, St Mary's Building, Elphinstone Road, University of Aberdeen, Aberdeen AB24 3UF, Scotland, United Kingdom
[b] The James Hutton Institute, Craigiebuckler, Aberdeen AB15 8QH, Scotland, United Kingdom
[c] IGB Leibniz Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany
[d] Humboldt University Berlin, Berlin, Germany
[e] School of Biological Sciences, University of Aberdeen, Cruickshank Building, St Machar Drive, Aberdeen AB24 3UU, Scotland, United Kingdom

HIGHLIGHTS

- A key challenge is to understand and predict catchment-scale faecal contamination.
- We employ long-term *E. coli* data and novel use of spatial-stream-network models.
- Concentrations of *E. coli* not clearly associated with flow conditions or season
- A significant predictor of spatial patterns was an Anthropogenic Impact Index.
- Spatial-stream-network models helped predict potential "hot spots" of contamination.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 3 April 2017
Received in revised form 26 July 2017
Accepted 15 August 2017
Available online 25 September 2017

Editor: D. Barcelo

Keywords:
E. coli
Faecal indicator organism
Microbial pollution
Spatio-temporal dynamics
Surface water
Water quality

ABSTRACT

An 11 year dataset of concentrations of *E. coli* at 10 spatially-distributed sites in a mixed land-use catchment in NE Scotland (52 km^2) revealed that concentrations were not clearly associated with flow or season. The lack of a clear flow-concentration relationship may have been due to greater water fluxes from less-contaminated headwaters during high flows diluting downstream concentrations, the importance of persistent point sources of *E. coli* both anthropogenic and agricultural, and possibly the temporal resolution of the dataset. Point sources and year-round grazing of livestock probably obscured clear seasonality in concentrations. Multiple linear regression models identified potential for contamination by anthropogenic point sources as a significant predictor of long-term spatial patterns of low, average and high concentrations of *E. coli*. Neither arable nor pasture land was significant, even when accounting for hydrological connectivity with a topographic-index method. However, this may have reflected coarse-scale land-cover data inadequately representing "point sources" of agricultural contamination (e.g. direct defecation of livestock into the stream) and temporal changes in availability of *E. coli* from diffuse sources. Spatial-stream-network models (SSNMs) were applied in a novel context, and had value in making more robust catchment-scale predictions of concentrations of *E. coli* with estimates of uncertainty, and in enabling identification of potential "hot spots" of faecal contamination. Successfully managing faecal

* Corresponding author at: Northern Rivers Institute, School of Geosciences, St Mary's Building, Elphinstone Road, University of Aberdeen, Aberdeen AB24 3UF, Scotland, United Kingdom.
E-mail addresses: aaron.neill@abdn.ac.uk (A.J. Neill), d.tetzlaff@abdn.ac.uk (D. Tetzlaff), n.strachan@abdn.ac.uk (N.J.C. Strachan), rupert.hough@hutton.ac.uk (R.L. Hough), lisa.avery@hutton.ac.uk (L.M. Avery), helen.watson@hutton.ac.uk (H. Watson), c.soulsby@abdn.ac.uk (C. Soulsby).

http://dx.doi.org/10.1016/j.scitotenv.2017.08.151

contamination of surface waters is vital for safeguarding public health. Our finding that concentrations of *E. coli* could not clearly be associated with flow or season may suggest that management strategies should not necessarily target only high flow events or summer when faecal contamination risk is often assumed to be greatest. Furthermore, we identified SSNMs as valuable tools for identifying possible "hot spots" of contamination which could be targeted for management, and for highlighting areas where additional monitoring could help better constrain predictions relating to faecal contamination.

1. Introduction

When faecal material is transferred to surface waters, the delivery of faecal pathogens including *Escherichia coli* O157, *Campylobacter* and *Cryptosporidium parvum* may also occur (Oliver et al., 2005a). Such pathogens can lead to gastrointestinal illness in humans if exposure to contaminated water occurs through, for example, recreational uses of water or consumption of drinking water from poorly-treated private supplies (Fewtrell and Kay, 2015; Strachan et al., 2006). In the European context, legislation such as the Drinking Water Directive (Council Directive 98/83/EC) and revised Bathing Water Directive (Council Directive 2006/7/EC) stipulate acceptable concentrations of faecal indicator organisms (FIOs), used as a proxy for faecal contamination, that should be complied with for different uses of water in order to safeguard public health. Such legislation has prompted increased recognition of the need to better understand the dynamics and drivers of faecal contamination in surface waters, so that effective management strategies can be devised that permit microbiological water quality standards to be met (Kay et al., 2008a).

In rural areas, the potential for faecal contamination is often high due to potential for contributions from both point and diffuse sources. Sewage infrastructure is often more rudimentary in such areas, with septic tanks and combined sewer overflow waste water treatment works (WWTWs) being common, both of which represent important point sources of contamination (Kay et al., 2008b). Meanwhile, spread manure and faeces from grazing animals arising from intensive agriculture are examples of diffuse sources (Chadwick et al., 2008). The high potential for faecal contamination in rural areas can impact on a number of downstream water uses which, in turn, has implications for meeting legislative requirements and for public health. For example, exports of faecal contaminants from rural catchments have been suggested to account for large proportions of contamination observed in coastal bathing waters (Crowther et al., 2003). Furthermore, private water supplies are commonly relied upon to provide drinking water in rural areas, some of which may be drawn from surface waters. However, such supplies often employ only limited treatment mechanisms, meaning there is increased potential for human infection by faecal pathogens when the microbiological quality of the raw water of a private supply is poor (Kay et al., 2007). As such, there is a vital need to better manage faecal contamination in rurally-influenced catchments.

Compared with other types of water pollution, the evidence-base for understanding the behaviour and survival of faecal pathogens and FIOs in the environment has, historically, been more limited (Kay et al., 2008a). Significant knowledge gaps still persist in relation to understanding the spatio-temporal dynamics of faecal contamination, especially at the catchment scale where decisions regarding management of water quality need to be made (Oliver et al., 2016). In particular, understanding the response of concentrations of FIOs to hydrological conditions and season using datasets long enough to capture sufficient hydroclimatic variation, and developing models that can be used to infer potential sources of contamination from spatial patterns of FIOs and make robust predictions for unmeasured locations represent key challenges at this scale (Kay et al., 2010; Tetzlaff et al., 2012; Vitro et al., 2017).

Previous catchment-scale studies (e.g. Crowther et al., 2002, 2003; Kay et al., 2005, 2008b; McGrane et al., 2014; Tetzlaff et al., 2012) have offered important insights into the dynamics and controls of faecal contamination. In particular, high flow events and summer have often been identified as periods when concentrations of FIOs are likely to be elevated. In addition, multiple linear regression models (MLRMs) linking spatial patterns in concentrations to readily-available land-cover variables as proxies for different sources of contamination have generally identified intensive livestock farming and human sewage inputs as potentially important sources. Where more detailed datasets have been available, some studies have further identified physical, chemical and biological factors that can be significantly associated with spatial patterns of FIOs. For example, Dwivedi et al. (2013) found temperature, dissolved oxygen, phosphate, ammonia, suspended solids and chlorophyll to be important for estimating *E. coli* loads in Plum Creek, Texas. However, many past studies have generally been constrained by the availability of only short-duration (<1–2 year) datasets relating to concentrations of FIOs. Furthermore, many of the regression models based on land cover for FIOs are fairly simple in their implementation (Kay et al., 2010). For example, elevated concentrations of FIOs during high flow conditions are often attributed to increased hydrological connectivity between sources of contamination and the stream network, particularly via overland flow (Dwivedi et al., 2016; Kay et al., 2008b; Tyrrel and Quinton, 2003). However, conceptualisation of the connectivity potential of certain land covers within regression models is rare (an exception is Crowther et al., 2003, who showed that concentrations of FIOs during low flows were most influenced by land use within 1–2 km surface-flow distance of a sub-catchment outlet, whilst during high flows land use across the whole of a sub-catchment was important). Also rare is the recognition that concentrations of FIOs at flow-connected sampling sites along a stream network may not be independent of one another (although Vitro et al., 2017 successfully account for this with a spatial regression model). This may give rise to spatial autocorrelation between sampling sites, which, if not accounted for, may lead to significance being incorrectly assigned to the land-cover variables of the models (Isaak et al., 2014).

Whilst dataset length may be logistically constrained, representing hydrological connectivity in models is a possibility. A potential approach is the use of topographically-based indices, such as the Network Index (Lane et al., 2004). This is an extension of the topographic wetness index of Beven and Kirkby (1979) and accounts for the requirement that for a saturated area to be hydrologically connected to a stream via an overland flow path, the entire flow path must be saturated to prevent disconnection by processes such as re-infiltration (Lane et al., 2004, 2009). However, whilst this metric has potential in characterising the hydrological connectivity likelihood of diffuse sources of pollution, it has rarely been implemented in this context (Lane et al., 2009; an exception being SCIMAP outlined by Reaney et al., 2011).

Spatial-stream-network models (SSNMs) represent an advancement in geostatistical methods that mean it is also now possible to account for spatial autocorrelation between observations along stream networks (see Ver Hoef and Peterson, 2010 and Ver Hoef et al., 2006 for full details). Central to SSNMs is that, unlike traditional geostatistical methods, autocorrelation between observed locations is based on stream distance as opposed to Euclidian distance. Stream distance is the shortest distance between two points when following the stream network. Autocovariance functions based on stream distance are based on moving average constructions, and may be defined for sites

that are flow connected and unconnected. In this way, SSNMs are uniquely placed to account for spatial autocorrelation that may arise in stream networks due to both passive (e.g. downstream transport of bacteria) and active (e.g. upstream migration of fish) interactions with flow (Peterson and Ver Hoef, 2010; Ver Hoef and Peterson, 2010). Where SSNMs have been developed for water quality variables (e.g. temperature), accounting for spatial autocorrelation has helped to prevent significance being incorrectly assigned to dependent variables and also improved the accuracy of predictions (e.g. Isaak et al., 2014). However, to the authors' knowledge, there is only one instance where a SSNM-like approach has been applied to FIO data (Money et al., 2009).

Here, we aim to use a long-term dataset of spatially-distributed concentrations of *E. coli* and novel modelling approaches to understand and predict the spatio-temporal dynamics of faecal contamination in a mixed land-use catchment in NE Scotland. Our specific objectives were to:

1. Understand the response of concentrations of *E. coli* to hydrological conditions and season based on long-term data;
2. Investigate whether long-term spatial patterns of low, average and high concentrations of *E. coli* can be linked to land-cover proxies for different sources of contamination, and how accounting for hydrological connectivity potential might affect this;
3. Assess the value of SSNMs as tools to understand and predict long-term spatial patterns of concentrations of *E.coli* in water quality studies.

2. Study site

The Tarland Burn (71 km^2) is located in NE Scotland, and is a sub-catchment of the River Dee. Earlier assessments of diffuse pollution within the Dee identified that increasing agricultural land use in lowland tributaries adversely affected water quality (Langan et al., 1997). The Tarland is the first upstream tributary draining significant areas of intensive agriculture. This became possible following extensive field drainage and canalisation of the Tarland Burn in the 1800s for agricultural improvement. The Tarland catchment is a focus for research assessing point- and diffuse-source pollution and evaluating best management practices for mitigation (Bergfur et al., 2012). As such, long-term water quality monitoring has been taking place at 10 nested sampling sites (ranging from <1 to 52 km^2) in the upper catchment (Fig. 1a; Bergfur et al., 2012); the area of focus for this study.

Elevation in the catchment ranges from 136 m to 618 m (Fig. 1a). Freely-drained brown earth and humus iron podzols are the main soils, but poorly-drained peaty gley podzols and non-calcareous gleys are also present with the former dominating the upper catchment (Fig. 1b). Land cover is mixed (Fig. 1c). Based on the CORINE Land Cover 2012 dataset (Cole et al., 2015), pasture land for cattle and sheep grazing is the main land cover (39.8%), whilst forestry (24.9%) and arable land (18.6%) are other major land covers. The main settlement in the catchment is Tarland village (population 650; Bergfur et al., 2012), however, there are also a number of dispersed dwellings and farms across the catchment. There are 58 private water supplies serving a number of properties in the Tarland, 13 of which are sourced from surface waters (DWQR for Scotland, Personal Communication). Some properties in the catchment are connected to a WWTW ~3 km upstream of the catchment outlet (Fig. 1c), however, many of the dispersed dwellings and farms are served by septic tanks. The septic tanks are often of an older style and discharge into ditches or the stream itself. Treated effluent from the WWTW normally discharges into a wetland or occasionally to the stream under licenced consent; however, as the WWTW is a combined sewer overflow, effluent with limited treatment can be discharged directly to the stream during periods of heavy rainfall (Stutter et al., 2010).

3. Material and methods

3.1. Study period and hydrometric data

The study spans 11 water years (October 2004–September 2015). Hourly discharge was calculated at the catchment outlet, Site 1 (Coull Bridge; Lat: 57.111 | Lon: −2.810) using measured stage recorded by a sonic range sensor (SR50, Campbell Scientific, Loughborough, UK) and a rating curve based on velocity-area profiling (Stutter et al., 2008a, 2008b). Hourly air temperature was also recorded at Coull Bridge. Precipitation was measured at a meteorological station at Aboyne (Lat: 57.077 | Lon: −2.836, ~4 km SW of the catchment). Hydrometric data was amalgamated to daily time steps, by averaging discharge and temperature, and summing precipitation.

3.2. Water sampling and microbiological analysis

Samples were collected from 10 spatially-distributed sites on a monthly to three-monthly basis, with all sites sampled on the same day (Bergfur et al., 2012). This gave ~80 samples per site. Samples were collected in sterile bottles and analysed within 6 h of collection using the Colilert-18 most probable number (MPN) method (IDEXX Laboratories, Westbrook, Maine, USA) to determine concentrations of *E. coli* (in MPN 100 ml^{-1}). Where high concentrations of *E. coli* were likely present, appropriate dilutions were made using sterile Ringers solution. The detection limit for an undiluted sample is <1 MPN 100 ml^{-1}. These data form part of the dataset of Langan et al. (n.d., DOI pending).

3.3. Statistical summary of spatio-temporal patterns of concentrations of *E. coli*

Summary statistics of long-term concentrations of *E. coli* at each site were generated based on all sampling dates and in relation to flow and season to investigate spatio-temporal patterns. To understand the data in a hydrological context, samples were separated into those that were taken during high and low flows (corresponding to flows ≥ Q_{10} and ≤ Q_{90} at the Coull Bridge gauging station, respectively). Despite the long-term data, the coarse sample frequency meant that high and low flows only coincided with 6 and 8 sampling dates, respectively. For season, samples were separated into two periods, 'summer' and 'winter'. Summer (April–September) is the most biologically active period for this region (Dawson et al., 2008). Forty four sampling dates fell within the summer period and 37 in the winter. Given the range of the concentrations of *E. coli*, the log_{10}-transformed 5th, 50th and 95th percentiles were used as summaries, though at some sites this was not possible due to samples having concentrations of *E. coli* below detection limits. In these cases, a log-normal distribution was fitted to the data, accounting for samples below the various detection limits, using a maximum likelihood estimator, to estimate the required percentiles (Helsel, 1990).

3.4. Multiple linear regression modelling based on land cover

Spatial information required for fitting both the MLRMs and SSNMs (Section 3.5) was generated using the STARS package (Peterson and Ver Hoef, 2014) within ESRI ArcMap 10.2.1. The stream network and catchment areas of each sampling site were defined from a 5 m resolution LandMap digital terrain model (DTM). For each site, pasture and arable land cover (as % of catchment area) was derived from CORINE 2012 (Fig. 1c; Cole et al., 2015). These catchment characteristics represented potential diffuse sources of faecal contamination from manure application to arable land and defecation of livestock on pasture, and have been identified as significant predictors of spatial patterns of concentrations of FIOs in past studies (e.g. Crowther et al., 2002; Kay et al., 2008b; Tetzlaff et al., 2012). We also defined an Anthropogenic Impact Index (A.I.I.) as a lumped indicator of potential for contamination from

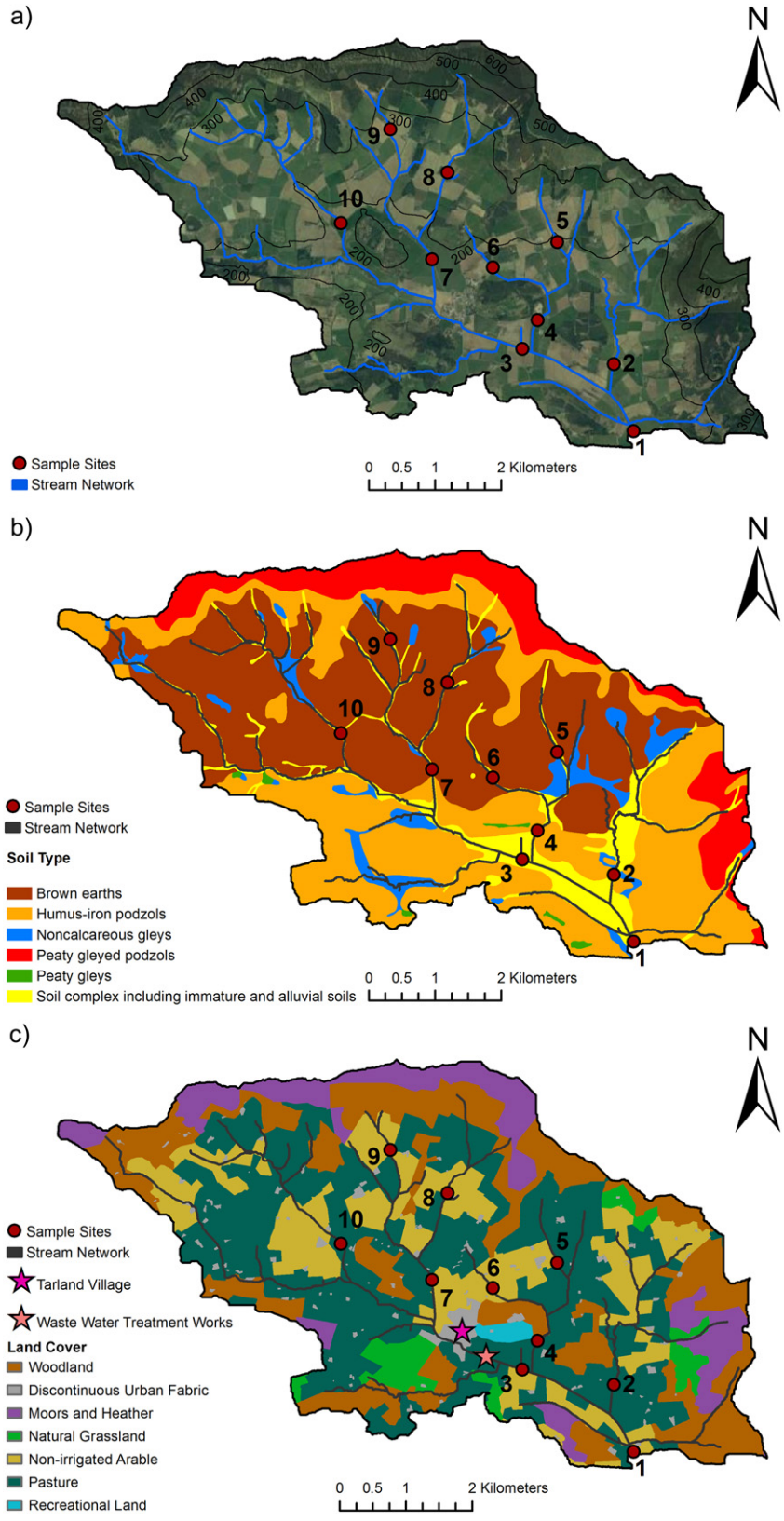


Fig. 1. Tarland Burn catchment showing a) Topography and location of long-term sampling sites; b) Distribution of soil classes according to Digitised Soil Map of Scotland, scale 1:25,000 (Soil Survey of Scotland Staff, 2014); c) Land cover based on CORINE 2012 dataset (Cole et al., 2015) and additional mapping of dispersed dwellings.

human point sources (e.g. leaking sewage pipes, septic tanks and open farmyards). One point was added to the A.I.I. of each site for every dwelling (either served by a septic tank or connected to the sewer) and every farmyard in its catchment. If a dwelling fell within a farmyard complex only a single point was given. Dwellings and farmyards were mapped using OS MasterMap data and septic tank locations were obtained from the DWQR for Scotland (Personal Communication). The A.I.I. only represents potential for contamination as data relating to

contributions from each source it encompasses in space and time were not available; indeed such data are very rare or otherwise impossible to obtain (e.g. Crowther et al., 2002; Richards et al., 2016). The catchment characteristics of each site are summarised in Table 1. Whilst we acknowledge that additional factors may be relevant for explaining spatial patterns of concentrations of *E. coli* (e.g. those identified by Dwivedi et al., 2013), we chose to focus on variables relating to land cover in our models as such data is readily available for most catchments.

To quantify the hydrological connectivity potential of pasture and arable land associated with diffuse sources of faecal contamination, the Network Index (NI) of the Tarland catchment was generated from the 5 m DTM (Lane et al., 2004). From this, the Relative Network Index (RNI) was obtained (Fig. 2); this rescales the NI so that the smallest 5% of values are given a value of 0, the largest 5% a value of 1, and the remainder scale linearly between 0 and 1 (Lane et al., 2009). The NI and RNI were generated using the SCIMAP Risk Maps module (Reaney and Milledge, 2013) in SAGA 2.0 (Conrad, 2006). Two ranges of the RNI were then chosen to differentiate between levels of hydrological connectivity potential: $0.5 \leq RNI \leq 1$ and $RNI = 1$. Areas with an RNI in the former range are assumed to exhibit non-negligible durations of hydrological connectivity during storm events, whilst areas with $RNI = 1$ are assumed to be always connected (Lane et al., 2009; Reaney et al., 2011). The amount of arable and pasture land with $0.5 \leq RNI \leq 1$ and $RNI = 1$ as a percentage of catchment area with $0.5 \leq RNI \leq 1$ and $RNI = 1$, respectively, were then calculated as site characteristics (Table 1).

To investigate possible controls on spatial patterns of concentrations of *E. coli*, MLRMs were developed for the $\log_{10}$-transformed 5th, 50th and 95th percentile concentrations based on all the sampling dates. These summary statistics were used as dependent variables to represent average concentrations of *E. coli* at each site (50th percentile) as well as low and high concentrations (5th and 95th percentiles, respectively) without making assumptions about associations with flow and season. Stata 14 (StataCorp, 2015) was used to fit the MLRMs based on maximum likelihood and a backwards-stepwise procedure with $p \leq 0.05$ as the criteria for variable removal. Three models were fitted to each percentile of *E. coli* concentrations, meaning that nine models were fitted in total. Each model started with $\log_{10}$ A.I.I. and either amount of arable and pasture land based on (a) total site catchment area, (b) area with $0.5 \leq RNI \leq 1$, or (c) area with $RNI = 1$ as possible predictor variables. Predictors were tested for multi-collinearity, but no significant correlations were found. The MLRMs were assessed based on the coefficient of determination (R^2) for fit and leave-one-out cross-validation root-mean-squared prediction error (LOOCV RMSPE) as a measure of predictive capability.

3.5. Spatial-stream-network modelling

Spatial-stream-network models employ a linear mixed-modelling approach to explain the variance in observations, which takes the following general form:

$$y = X\beta + z + \varepsilon \quad (1)$$

where y is a vector of observations, X is a design matrix for fixed effects (variables that are measured and explain the general spatial patterns in the observations), β is a vector of coefficients for the fixed effects, z contains spatially-autocorrelated random effects (random variables with a spatial autocovariance structure modelled on the residuals of the observations after accounting for the fixed effects, that can represent both unmeasured and unknown factors influencing observations) and ε is a vector of independent random errors (Peterson and Ver Hoef, 2010; Peterson et al., 2013). We developed SSNMs using land cover for the fixed effects and a "tail-up" spatial autocovariance structure for the random effects. A tail-up autocovariance structure accounts for autocorrelation between flow-connected sites (Ver Hoef and Peterson, 2010), and was deemed sufficient for our purpose because: (a) *E. coli* is transported passively in the flow; (b) it was desirable to limit the degrees of freedom of the models due to the small number of sites used in their development (c.f. Isaak et al., 2014).

The STARS package was used to generate the network required to build a SSNM describing the structure of the Tarland stream network and how sampling sites are connected (Peterson and Ver Hoef, 2014). It was also used to calculate the spatial weightings between sites required when using tail-up autocovariance structures, which here were based on an additive function defined using catchment area (Fig. 3; see Peterson and Ver Hoef, 2010 Appendix A). SSNMs were then developed using the SSN package (Ver Hoef et al., 2014) in the R statistical software package (R Core Team, 2016). SSNMs were developed for the $\log_{10}$-transformed 5th, 50th and 95th percentile concentrations of *E. coli*. Three models were constructed for each percentile, each of which had the predictor variables from the best MLRMs for fixed effects, and random effects with either an exponential, spherical or linear-with-sill tail-up spatial autocovariance structure (see Ver Hoef et al., 2006). To reduce bias in the estimation of the parameters of the spatial autocovariance functions, we used restricted maximum likelihood as opposed to maximum likelihood when fitting the models (Cressie, 1993). An autocovariance function has three parameters: the nugget effect (variability between locations separated by distances close to zero), the partial-sill (variability of the random effects), and the range (distance beyond which locations are no longer autocorrelated). Here, the maximum value of the range parameter

Table 1
Summary of catchment characteristics for the 10 sampling sites.

Site	Total catchment			$0.5 \leq RNI^a \leq 1$			$RNI^a = 1$			Anthropogenic Impact Index[b]
	Area (km^2)	% Arable	% Pasture	Area (km^2)	% Arable	% Pasture	Area (km^2)	% Arable	% Pasture	
1	51.98	18.62	39.77	20.65	21.19	42.93	2.09	24.91	46.09	529
2	7.70	24.87	26.77	2.89	31.37	33.24	0.31	39.95	36.19	21
3	31.41	18.05	41.32	12.37	20.63	44.77	1.25	24.08	49.01	414
4	6.48	17.05	54.72	2.52	18.81	58.62	0.24	24.77	64.57	69
5	1.62	13.34	52.75	0.54	15.59	59.67	0.05	13.87	75.30	5
6	0.93	36.27	36.61	0.36	42.58	33.45	0.05	48.57	39.41	23
7	10.40	24.24	29.48	4.46	27.79	35.56	0.46	35.40	41.34	43
8	3.96	6.46	24.27	1.44	7.77	36.45	0.12	14.13	57.62	2
9	1.77	17.30	8.08	0.73	18.04	15.06	0.06	31.95	20.55	1
10	6.92	20.99	44.19	2.32	24.38	51.71	0.24	27.79	54.78	15

[a] Relative Network Index. Areas with $0.5 \leq RNI \leq 1$ are assumed to have non-negligible durations of hydrological connectivity during storm events, whilst areas with $RNI = 1$ are assumed to always be connected (Lane et al., 2009; Reaney et al., 2011).

[b] Anthropogenic Impact Index is assigned points for farmyards, and for dwellings attached to the sewer system or septic tanks. A single point is given if a dwelling falls within a farmyard complex.

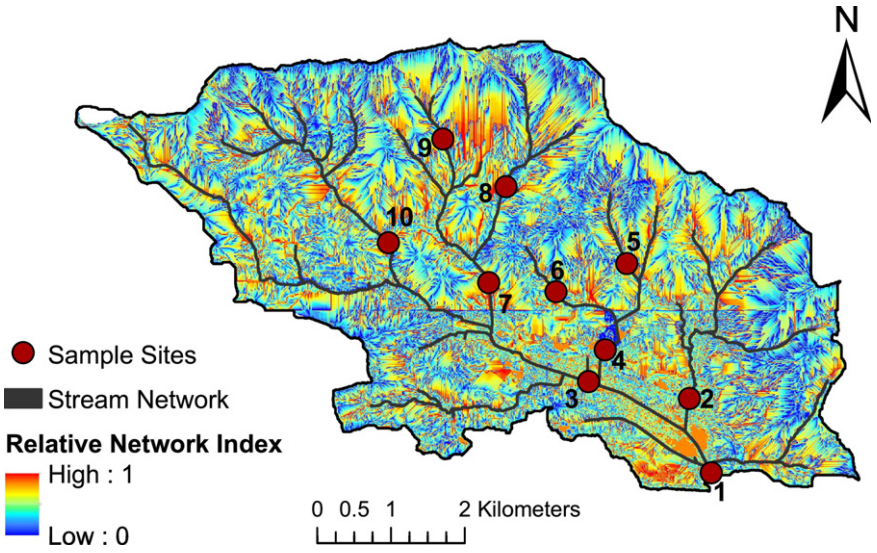


Fig. 2. The Relative Network Index for the Tarland catchment. (The line across the centre of the dataset is a result of tile joining in the Landmap digital terrain model from which the Relative Network Index was constructed).

was limited to the maximum stream distance between any of the sampling sites (8436.4 m). The performance of the SSNMs was assessed by calculating an R^2 to describe the amount of variability in the observations explained by the fixed and random effects, and using the LOOCV RMSPE as a measure of predictive capability. To allow comparison between the SSNMs and MLRMs, the best-performing MLRMs for each concentration percentile were refitted using restricted maximum likelihood.

3.6. Predicting long-term spatial patterns of concentrations of E. coli

The best models for the $\log_{10}$-transformed 5th, 50th and 95th percentile concentrations of *E. coli* were used to make predictions for sites evenly distributed at 200 m intervals along the stream network. Predictions were made using the SSN Package (Ver Hoef et al., 2014). The position of the prediction sites within the stream network and their connectedness to each other and observed sites was defined using the STARS package (Peterson and Ver Hoef, 2014).

4. Results

4.1. Hydroclimatic conditions

Fig. 4 shows total daily precipitation and mean daily discharge and temperature for the study period. Annual precipitation was either close to or above the long-term average (~1030 mm for 1961–1990; ~950 mm for 1971–2000) in NE Scotland. At Aboyne, the driest water years were 2005, 2006, 2009 and 2013, with annual precipitation ranging between 581.6 mm and 681.4 mm. The wettest years were 2007, 2010 and 2014, with annual precipitation ranging from 932.4 mm to 999.8 mm. Precipitation tended to be fairly evenly distributed throughout the year (Fig. 4a).

Discharge at Coull Bridge showed a strong, seasonally varying baseflow component most likely dominated by groundwater (Fig. 4b). This probably reflects the large areas of flat lowland topography in the catchment, dominance of freely-drained brown earth and humus iron podzols and the influence of underlying drift deposits. Nevertheless, there is also a flashy, non-linear response of discharge to precipitation

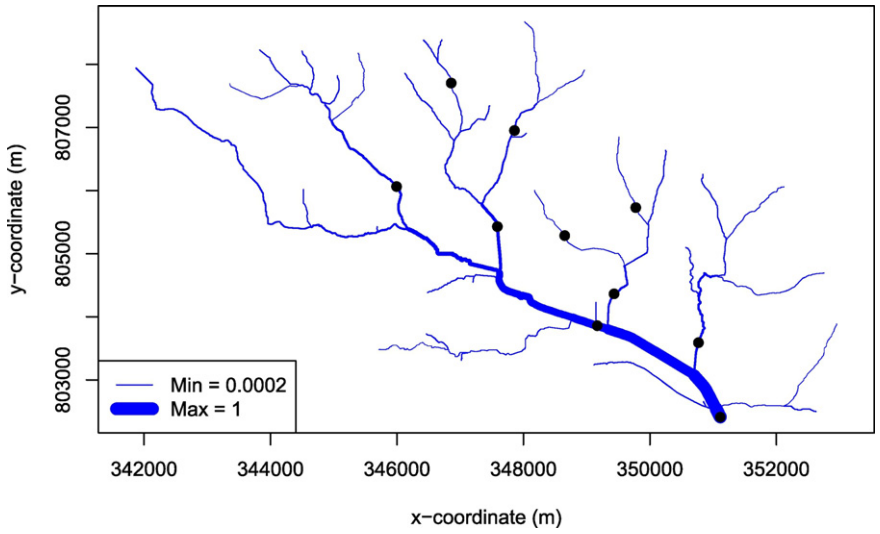


Fig. 3. The network created for the Tarland catchment. Black dots represent sampling sites whilst the blue line represents the stream network. The width of the blue line reflects the value of the additive function (dimensionless). For a given reach, this is defined by multiplying the proportional influences (based on reach catchment area) of each reach downstream to the catchment outlet (see Peterson and Ver Hoef, 2010 Appendix A).

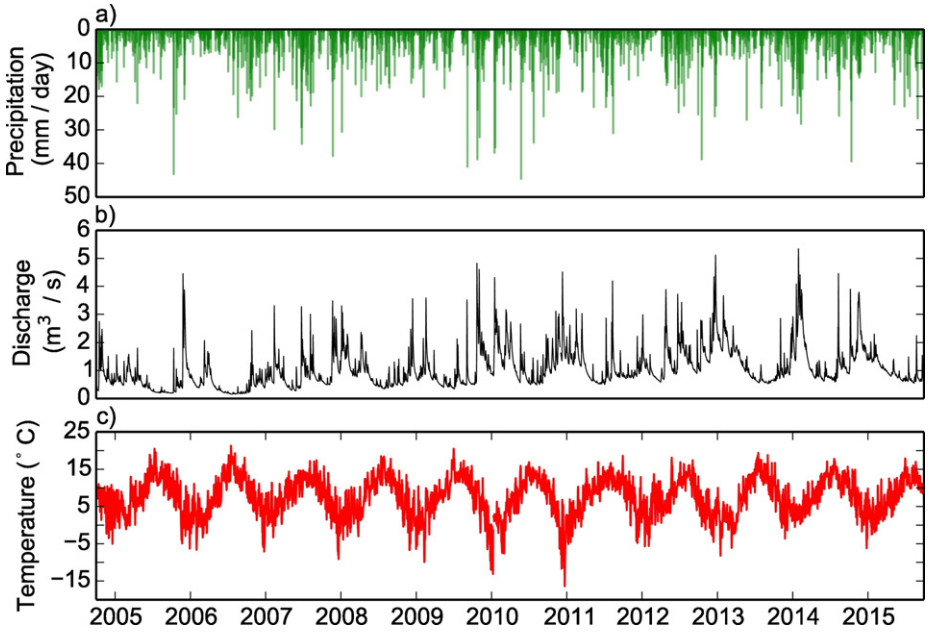


Fig. 4. Long-term time series of a) Total daily precipitation recorded at Aboyne meteorological station; b) Mean daily discharge at Coull Bridge; c) Mean daily air temperature at Coull Bridge.

events reflecting the hydrologically-responsive peaty gley soils in the headwaters and the rapid transport of water through agricultural field drains. Additionally, during larger events, the flat lowland areas are prone to saturation, which can activate overland flow pathways (Fig. 2). Highest mean daily flows occurred between December and February and the lowest between May and September (Fig. 4b). The Q_{90} flow for the study period was 0.41 $m^3\ s^{-1}$ and the Q_{10} flow was 1.94 $m^3\ s^{-1}$.

Air temperatures showed clear seasonality (Fig. 4c). The coldest mean monthly temperatures ranged between 1.84 °C and 2.2 °C for the months December to February, whilst July and August had the warmest mean temperatures (13.87 °C and 13.02 °C, respectively).

4.2. Long-term spatio-temporal patterns of concentrations of *E. coli*

Fig. 5 shows the percentile concentrations of *E. coli* at the 10 sites for all sampling dates and also for flows ≥ Q_{10} and ≤ Q_{90}. There was no clear relationship between concentrations and flow conditions. At Sites 1, 3, 5, 6 and 10, 50th percentile concentrations under high flow conditions were greatest, whilst for the remaining sites they were higher during low flows. In addition, the generally high intra-site variability (at least an order of magnitude) in concentrations during both high and low flows meant that there was usually overlap in the ranges between the 5th and 95th percentile concentrations observed under high and low flows.

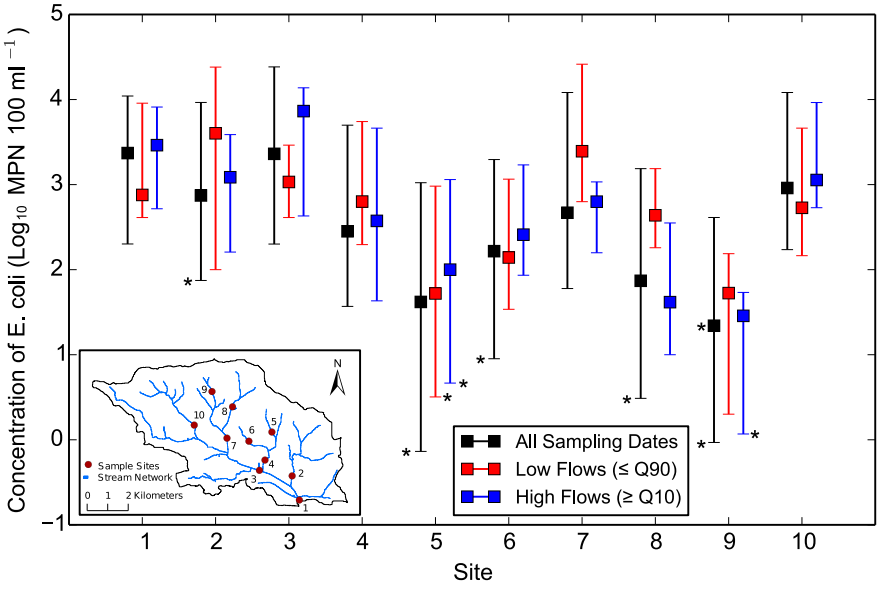


Fig. 5. The 5th, 50th and 95th percentile $\log_{10}$ most probable number (MPN) concentrations of *E. coli* for each sampling site for all sampling dates and by flow (low flows ≤ Q_{90} and high flows ≥ Q_{10}). The 50th percentile is shown by the square markers, whilst the 5th and 95th percentiles are given by the line caps. Stars indicate where percentiles were estimated using a log-normal distribution.

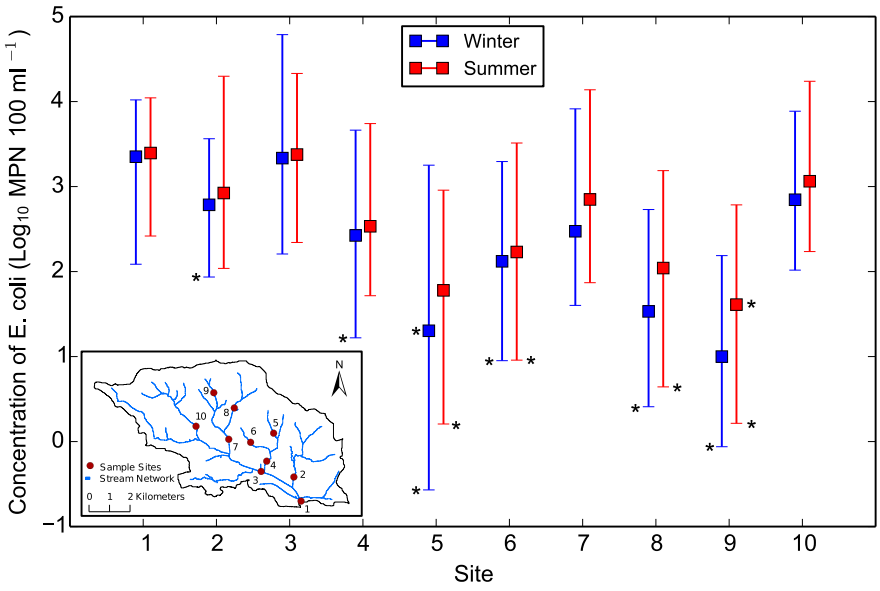


Fig. 6. The 5th, 50th and 95th percentile $\log_{10}$ most probable number (MPN) concentrations of *E. coli* for each sampling site for the summer and winter periods. The 50th percentile is shown by the square markers, whilst the 5th and 95th percentiles are given by the line caps. Stars indicate where percentiles were estimated using a log-normal distribution.

Fig. 6 shows the concentrations of *E. coli* for 'summer' and 'winter' sampling dates. For all sites, the 50th percentile concentrations were greater in summer than winter, but tended to be of the same order of magnitude. Additionally, winter 5th percentile concentrations were lower than those for summer, whilst summer 95th percentile concentrations tended to be higher than winter (exception is Site 3). There was a high degree of overlap between the summer and winter concentrations, owing to large intra-site variability.

Inter-site variability of concentrations was high when considered across all sampling dates, by flow and by season, with each percentile spanning at least two orders of magnitude across all sites (Figs. 5 and 6). Spatial patterns during high flows, summer and winter were all similar to those that emerged when considering all sampling dates. The highest 50th percentile concentrations were for Sites 1 and 3 in the lower catchment, whilst the lowest were for Sites 5, 8 and 9. A marked change during high flows, however, was that the 50th percentile concentration at Site 3 (~3.86 $\log_{10}$ MPN 100 ml^{-1}) was more than double that at Site 1 (~3.46 $\log_{10}$ MPN 100 ml^{-1}). Across all sampling dates, the 50th percentile concentrations at these two sites were very similar (~3.36 $\log_{10}$ MPN 100 ml^{-1}). Under low flows, the spatial pattern changed. Here, Sites 2, 3 and 7 had the highest 50th percentile concentrations, with Site 2 having the highest at ~3.60 $\log_{10}$ MPN 100 ml^{-1}. Sites 5 and 9 still had the lowest 50th percentile concentrations (~1.70 $\log_{10}$ MPN 100 ml^{-1} for both sites), however, Site 8 had a higher 50th percentile concentration at ~2.64 $\log_{10}$ MPN 100 ml^{-1}.

4.3. Multiple linear regression modelling based on land cover

The MLRMs developed for the $\log_{10}$-transformed 5th, 50th and 95th percentile concentrations of *E. coli* all had the $\log_{10}$ A.I.I. as the only significant predictor variable ($p < 0.01$; Table 2). The extents of arable and pasture land as percentages of the total area of each sampling sites' catchment were not significant predictors. This did not change when hydrological connectivity potential was accounted for. The model for 50th percentile concentrations performed best, with an R^2 of 0.78 and a LOOCV RMSPE of 0.371 $\log_{10}$ MPN 100 ml^{-1}. Whilst the model for the 5th percentile concentrations achieved an R^2 of 0.68, its high LOOCV RMSPE resulted in this being the poorest performing model. Despite the high R^2 achieved by all models, the LOOCV RMSPEs were always high, particularly for the 5th percentile model.

4.4. Spatial-stream-network modelling

The statistics for SSNMs developed for the various $\log_{10}$-transformed percentile concentrations of *E. coli* are given in Table 3. Accounting for spatial autocorrelation did not cause the $\log_{10}$ A.I.I. to be dropped as a significant fixed effect for any of the models. With respect to the parameters of the autocovariance structures, the nugget effect was consistently very close to 0, whilst the partial sill assumed a similar value to the nugget in the non-spatial MLRMs. This indicates that much of the unexplained variance in the observations after accounting for the fixed effects could be attributed to the spatially-autocorrelated random effects. This is also reflected in the R^2 (fixed + random effects) being close to 1. The range parameter for most SSNMs was at or close to the maximum imposed range (8436.4 m), suggesting that all sites were autocorrelated to some degree. The exceptions were the SSNMs for 95th percentile concentrations with spherical and linear-with-sill spatial autocovariance structures, and the 5th percentile SSNM with linear-with-sill spatial autocovariance structure.

In terms of the predictive capabilities of the SSNMs, for the 5th percentile concentrations, the LOOCV RMSPEs of the SSNMs were slightly poorer than the best MLRM (Table 3). However, increases in LOOCV RMSPEs were only ~1%. For 50th percentile concentrations, SSNMs showed markedly improved predictive capability. The best was for the SSNM that used a linear-with-sill autocovariance structure for the random effects, with the LOOCV RMSPE being reduced by 17% compared to the MLRM. For the 95th percentile concentrations, only the SSNM that used a linear-with-sill autocovariance structure showed a slight improvement in LOOCV RMSPE. Meanwhile, the SSNMs employing

Table 2
Summary statistics for multiple linear regression models for the 5th, 50th and 95th percentile concentrations of *E. coli* ($\log_{10}$ MPN 100 ml^{-1}).

Response variable	Model parameters		Model performance	
	$\log_{10}$ Anthropogenic Impact Index	Intercept	R^2	LOOCV RMSPE ($\log_{10}$ MPN 100 ml^{-1})[a]
$\log_{10}$ 5th	0.889*	0.1089	0.68	0.600
$\log_{10}$ 50th	0.7059*	1.5012*	0.78	0.371
$\log_{10}$ 95th	0.5337*	2.9012*	0.66	0.390

* Significant with at least $p < 0.01$.
[a] Leave-one-out cross-validation, root-mean-squared prediction error.

Table 3
Summary statistics for spatial-stream-network models for the 5th, 50th and 95th percentile concentrations of *E. coli* (log_{10} MPN 100 ml^{-1}).

		Fixed effects parameters		Autocovariance structure parameters			Model performance		
Response variable	Tail-up spatial autocovariance structure	log_{10} Anthropogenic Impact Index	Intercept	Nugget	Partial Sill	Range (m)	R^2 (fixed effects)	R^2 (fixed + random effects)	LOOCV RMSPE (log_{10} MPN 100 ml^{-1})[a]
log_{10} 5th	None	0.889*	0.1089	0.330	–	–	0.68	–	0.600
	Exponential	0.92205*	0.09893	0.000	0.349	8436.39	0.68	~1.0	0.617
	Spherical	0.96476*	0.04582	0.000	0.368	8436.37	0.70	~1.0	0.621
	Linear-with-sill	0.99095*	0.01552	0.000	0.384	7361.53	0.71	~1.0	0.616
log_{10} 50th	None	0.7059*	1.5012*	0.126	–	–	0.78	–	0.371
	Exponential	0.6944*	1.5459*	0.000	0.115	8436.40	0.79	~1.0	0.334
	Spherical	0.702*	1.5444*	0.000	0.113	8436.40	0.80	~1.0	0.320
	Linear-with-sill	0.7057*	1.546*	0.000	0.114	8436.40	0.81	~1.0	0.307
log_{10} 95th	None	0.5337*	2.9012*	0.125	–	–	0.66	–	0.390
	Exponential	0.558*	2.8854*	0.000	0.134	8116.88	0.67	~1.0	0.394
	Spherical	0.582*	2.8525*	0.000	0.140	7713.01	0.69	~1.0	0.390
	Linear-with-sill	0.5884*	2.843*	0.000	0.142	5926.96	0.69	~1.0	0.386

* Significant with at least $p < 0.01$.
[a] Leave-one-out cross validation, root-mean-squared prediction error.

exponential and spherical structures showed a slight increase and no change in LOOCV RMSPE, respectively. Therefore, the SSNMs did not show consistent improvement in predictive capability over the MLRMs, at least when quantified using the LOOCV RMSPE.

4.5. Predicting long-term spatial patterns of concentrations of E. coli

To investigate further the dominant controls on concentrations of *E. coli* and to identify "hot spots" of contamination, SSNMs that used linear-with-sill spatial autocovariance structures for the random effects were used to predict catchment-wide concentrations. The predicted concentrations are shown in Fig. 7. The SSNMs provided good predictions of concentrations near the sampling sites. Standard errors were smallest for predictions made at prediction sites near to sampling sites to which they were flow-connected, and greatest for predictions made along tributaries which did not contain a sampling site.

Predicted 5th percentiles ranged between ~ −0.65 to 2.45 log_{10} MPN 100 ml^{-1} (Fig. 7a). Lowest concentrations were predicted for around and upstream of Sites 5 and 9, whilst highest concentrations were predicted from around Sites 2, 7 and 10 to the catchment outlet. Predictions of the 50th percentile concentrations ranged from ~1.25 to 3.4 log_{10} MPN 100 ml^{-1} (Fig. 7b). Compared with the 5th percentile predictions, it was more common for headwater tributaries to have lower predictions than more lowland parts of the network. Highest concentrations were predicted from the lower confluence on the tributary of Site 10 to the catchment outlet. Predictions of the 95th percentile concentrations ranged from ~2.5 to 4.5 log_{10} MPN 100 ml^{-1} (Fig. 7c). The lowest predictions were made for the tributary of Site 9, reflecting the very low observed 95th percentile concentration of *E. coli* at this site (Fig. 5). Highest predictions were generally made downstream of Site 7 and of the lower confluence on the tributary of Site 10.

The inclusion of spatially-autocorrelated random effects in the SSNMs caused some interesting features to emerge in the predicted concentrations. Firstly, it was possible to identify parts of the stream network where higher concentrations had been predicted than would be expected based on the A.I.I. alone. Key examples of this are the tributaries of Sites 2 and 10. The 5th, 50th and 95th percentile concentrations at both sites were all strongly under-predicted by just the fixed effects of the SSNMs (Table 4). Consequently, the random effects of the models increased the final predictions for prediction sites most strongly autocorrelated with each site, thus elevating predicted concentrations along their tributaries and helping identify these areas as "hot spots" of contamination. Areas where predicted concentrations were lower than expected based on just the A.I.I. could also be identified, such as at locations autocorrelated with Sites 4, 5 and 6 where fixed effects over-predicted observed concentrations (Table 4).

Secondly, whilst 95th percentile concentrations generally increased towards the catchment outlet, there was an exception just above the confluence of the tributary of Site 2 with the main stem, where predicted concentrations decreased before increasing downstream of the confluence (Fig. 7c). The decrease likely resulted from the random effects reducing the predicted concentrations for prediction sites increasingly autocorrelated with Site 1 to reflect a) Site 1 having an observed 95th percentile concentration that was much lower than Site 3's (Fig. 5); b) Site 1 having a large negative residual after accounting for the fixed effects whilst Site 3's residual was close to 0 (Table 4). Below the confluence, however, the increase in log_{10} A.I.I. arising from the additional contributing area, and possibly the influence of the random effects increasing the predictions made for prediction sites autocorrelated with Site 2 competed with this, thus, causing predicted concentrations to increase.

5. Discussion

5.1. How do concentrations of E. coli vary in response to hydrological conditions and season?

Our first objective was to understand how concentrations of *E. coli* responded to hydrological conditions and season using the long-term dataset. We found no clear association between flow and concentrations, with half the sites showing higher 50th percentile concentrations at high flows ($\geq Q_{10}$) and half at low flows ($\leq Q_{90}$). High intra-site variability dictated no clear distinction of concentrations under each flow type. This is in marked contrast to previous studies which have usually identified significant increases in concentrations of FIOs during high flows (Crowther et al., 2002, 2003; Kay et al., 2005, 2008b; Pachepsky and Shelton, 2011). Furthermore, we found that whilst 50th percentile concentrations were highest in summer at all sites, large overlap in summer and winter concentrations at individual sites prevented clear seasonal differentiation. Other studies have generally found concentrations to be higher in summer than winter; however, the increase has not always been significant (McGrane et al., 2014; Tetzlaff et al., 2012).

Consideration of the hydrological context may help explain the lack of a clear flow-concentration relationship. Firstly, it may reflect the spatial organisation of hydrological source areas and heterogeneous distribution of concentrations of *E. coli* within the stream network. During high flows, the headwater sites (Sites 5, 8 and 9) had the lowest concentrations (Fig. 5). The flashy discharge response to precipitation (Fig. 4b) probably reflects the activation of hydrological source areas in the headwaters, where poorly-drained peaty soils (Fig. 1b) are highly responsive to rainfall. With greater fluxes of less-contaminated water from the headwaters, it is likely that concentrations of *E. coli* dilute downstream,

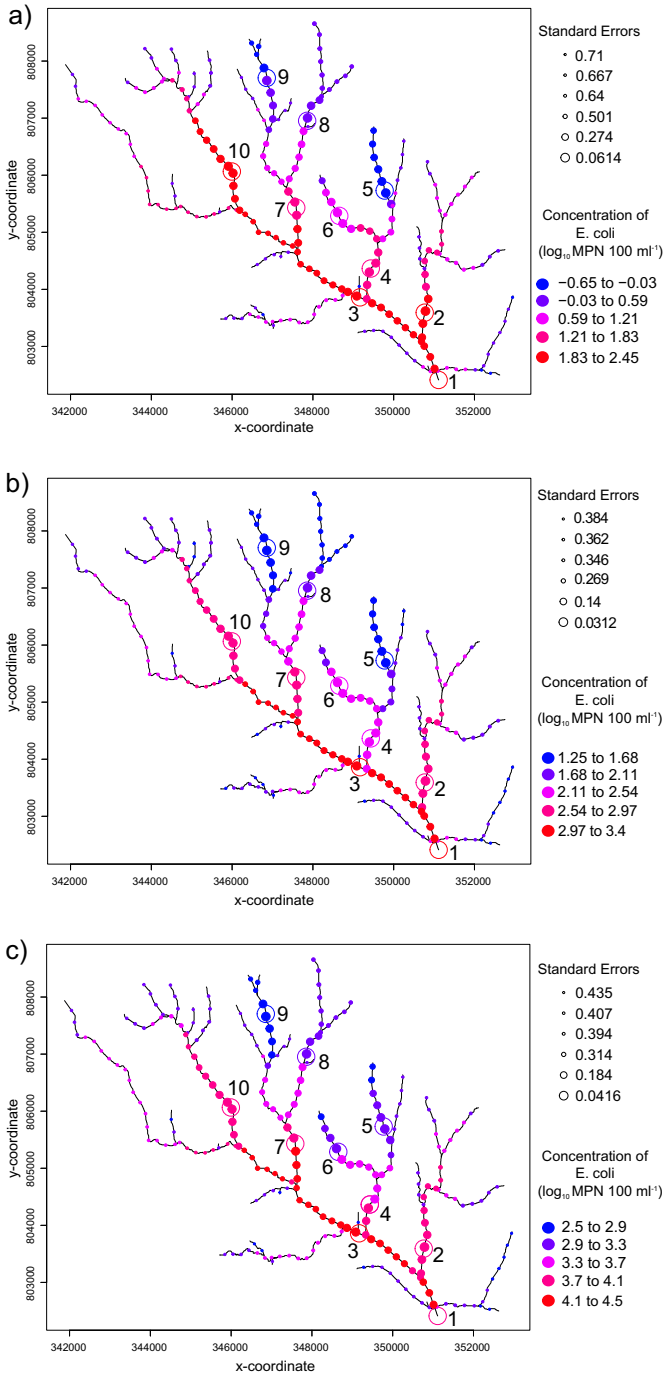


Fig. 7. Predictions of a) 5th percentile; b) 50th percentile; c) 95th percentile $\log_{10}$ most probable number (MPN) concentrations of *E. coli*. Filled circles are prediction sites, with colour representing concentration of *E. coli* and size the standard error of the prediction. Large open circles are observed sites.

Table 4
The residuals (observed – predicted) obtained based on the fixed effects alone of the spatial-stream-network models for the 5th, 50th and 95th percentile concentrations of *E. coli* ($\log_{10}$ MPN 100 ml^{-1}).

Site	Residuals ($\log_{10}$ MPN 100 ml^{-1})		
	5th percentile	50th percentile	95th percentile
1	−0.414	−0.098	−0.406
2	0.538	0.387	0.346
3	−0.309	−0.033	−0.003
4	−0.268	−0.394	−0.225
5	−0.837	−0.411	−0.228
6	−0.421	−0.291	−0.358
7	0.141	−0.032	0.273
8	0.218	0.141	0.194
9	−0.181	−0.301	−0.312
10	1.060	0.586	0.547

obscuring a clear flow-concentration relationship, even though pollutant inputs from the lower agricultural/urban areas may increase at the same time. Previous work in the Tarland has identified the headwaters as areas that are responsive to rainfall with the potential to effect downstream water quality (Stutter et al., 2008a, 2008b). Furthermore, McKergow and Davies-Colley (2010) identified a similar diluting effect on concentrations of *E. coli* due to activation of hydrological source areas in less contaminated parts of the larger mixed land-use Motueka catchment in New Zealand.

Contributions of *E. coli* from different sources may also have obscured a clear flow-concentration relationship. The most marked increases in concentrations during high flows were at Site 1 and, in particular, Site 3 (Fig. 5). However, this may predominantly reflect the influence of the WWTW upstream of Site 3 (Fig. 1c), which, as a combined sewer overflow, will likely discharge poorly-treated effluent directly to the stream during high flows (c.f. Stapleton et al., 2008). Higher concentrations at low flows could result from the main sources of contamination being from point sources which become diluted at high flows. Anthropogenic point sources may help explain elevated concentrations during low flows at sites with higher values of the A.I.I. (e.g. Sites 2, 4 and 7; Table 1). There may also be persistent agricultural "point sources" (highly localised sources without a fixed spatial location) such as chronic seepage of faecal pollutants from saturated areas of animal congregation and direct defecation of livestock into the stream (Davies-Colley et al., 2004; Tetzlaff et al., 2012) that are important. Field drains may further contribute as an agricultural point source. However, their ability to enhance or dilute concentrations of *E. coli* is likely to depend on soil type (Hunter et al., 1999; Oliver et al., 2005b). Changing spatial distribution of livestock or depletion of existing stores of *E. coli* in previous high flow events may also prevent the emergence of a clear flow-concentration relationship by causing temporal variability in the availability of *E. coli* from diffuse sources (McKergow and Davies-Colley, 2010).

A final possible explanation for the lack of a defined flow-concentration relationship might be the coarse monthly/three-monthly sampling resolution. Despite the dataset being temporally long, this sampling resolution caused both high and low flows to be under-sampled (each <10% of the total samples). As such, it is probable that concentrations at high flows were under-estimated (Tetzlaff et al., 2012). This may have been compounded by the timing of sample collection during high flows - concentrations of *E. coli* have been found to peak just prior to the peak in flow, which monthly sampling is more likely to miss (McKergow and Davies-Colley, 2010).

The general expectation that concentrations of *E. coli* are lower in winter is attributed to housing of cattle, reduced manure spreading and depletion of stores of *E. coli* during wetter weather (Kay et al., 2008b; Tetzlaff et al., 2012). However, where seasonality in concentrations has been illusive, anthropogenic point sources of contamination related to sewage, and year-round grazing of livestock (important in the Tarland) have been suggested to contribute to persistence of high concentrations in winter (McGrane et al., 2014). Furthermore, colder temperatures during wet winter conditions may help increase the length of time that *E. coli* persists in the environment for mobilisation by reducing die-off rates (Blaustein et al., 2013; Tyrrel and Quinton, 2003), so long as freeze-thaw cycles are limited (Natvig et al., 2002).

These findings in relation to how concentrations of *E. coli* respond to flow conditions and season may have important implications for the management of faecal contamination. A number of potential mitigation measures (e.g. riparian buffer strips and retention ponds) are predicated on the assumption that the greatest risk of faecal contamination is

during high flow events, and thus seek to reduce contamination by disrupting surface flow paths that connect pollutant sources to the stream (Oliver et al., 2007). However, where the influence of factors such as hydrologically-responsive, less-contaminated headwaters and persistent point sources of contamination prevent a clear relationship emerging between flow and concentrations of FIOs, as appears to be the case in the Tarland, such mitigation measures are unlikely to be entirely effective. Furthermore, whilst summer is often considered as a priority time for managing faecal contamination (e.g. Kay et al., 2008b), our findings suggest that point sources, livestock being left out over winter, and conditions potentially being favourable for the longer-term survival of *E. coli* in the environment may allow elevated levels of contamination to persist during the winter months as well. These factors should be considered when planning management strategies, and may require the implementation of mitigation measures that also target persistent point sources (e.g. sewage systems, farmyards, direct access of livestock to streams, seepages from saturated areas and field drains) and that can be effective year-round.

5.2. Can spatial patterns of long-term concentrations of E. coli be linked to land cover, and what is the effect of accounting for hydrological connectivity potential?

When developing MLRMs to link spatial patterns of low, average and high concentrations of *E. coli* to land cover, only the A.I.I. was identified as a significant predictor, explaining 66–78% of the variation in the observations (Table 2). This corroborates previous work that has associated metrics relating to human influence (e.g. population, percentage urban area) with spatial patterns of FIOs (Crowther et al., 2003; Kay et al., 2005, 2008b; McGrane et al., 2014; Tetzlaff et al., 2012).

Contrary to previous studies, however, neither the extent of pasture nor arable land (as % of catchment area) was identified as significant. Whilst this may seem counter-intuitive, it can be explained. Temporal dynamics of contamination in the Tarland suggested that important agricultural sources of contamination include direct defecation into the stream, seepage of contaminants from areas of high saturation potential where animals congregate, and potentially field drains. As these represent "point sources" as opposed to diffuse sources of contamination, they may not be adequately captured in coarse-scale land-cover data (Crowther et al., 2002; Kay et al., 2008b). Furthermore, such data cannot account for temporal variation in the availability of *E. coli* from diffuse sources. Whilst such smaller-scale details may average out in larger catchments (c.f. Kay et al., 2005, 2008b), most sampling sites in the Tarland had catchments ≤10 km^2, possibly precluding the identification of agricultural land cover as being significant in the MLRMs. For smaller catchments, more focused approaches such as microbial source tracking may be necessary to properly characterise the contributions to faecal contamination from human and animal sources (e.g. Flynn et al., 2016).

The importance of "point sources" of contamination from farmed land probably contributed to the insignificance of agricultural land-cover variables that accounted for hydrological connectivity in the MLRMs. The RNI represents hydrological connectivity (via overland flow) based on topographically-driven saturation potential (Lane et al., 2004). Whilst this can be a useful connectivity metric if agricultural pollution sources are diffuse and mobilised by surface runoff (Lane et al., 2009), this conceptualisation is likely to be less helpful where agricultural "point sources" are important. In addition, whilst the RNI helped to identify areas which may become hydrologically connected to the stream, the issue still remains that the availability of *E. coli* from diffuse sources in these areas for mobilisation and transfer will vary temporally. Where more detailed data are available, it may be possible to better parameterise in statistical models the potential processes that set spatial patterns of concentrations of FIOs, to permit a more process-based understanding of the drivers of faecal contamination (e.g. Dwivedi et al., 2013).

5.3. What is gained from using spatial-stream-network models in understanding and predicting long-term spatial patterns of concentrations of E. coli?

To the authors' knowledge, the application of SSNMs to understand and predict spatial patterns of concentrations of *E. coli* has been very rare, with only Money et al. (2009) employing a similar approach to predict concentrations for the Raritan River basin. Here, we found the best SSNMs for the 5th, 50th and 95th percentile concentrations to be those that included a linear-with-sill tail-up autocovariance structure, with the $\log_{10}$ A.I.I. remaining a significant fixed effect variable in all models (Table 3). The inclusion of spatially-autocorrelated random effects in the SSNMs meant that nearly all the variance in the observed data could be explained, thus improving on the MLRMs. The LOOCV RMSPEs of the SSNMs did not show consistent improvement over those of the MLRMs, and were relatively high for all models. However, this may reflect the relatively small number of sampling sites at which long-term *E. coli* data was available. The LOOCV method suffers from high variance and pessimistic bias, which in turn may limit its utility as a metric of model performance in such instances (Beleites et al., 2005). The LOOCV RMSPEs of the 5th percentile models may have been further impacted by these concentrations being more uncertain, since half the sites had these estimated from a log-normal distribution (Fig. 5). Overall, predictions made by all three SSNMs in the vicinity of the sampling sites reflected the observed concentrations of *E. coli* quite well (Fig. 7).

Arguably, the key value in the SSNMs developed in this study was making semi-continuous catchment-wide predictions of concentrations of *E. coli* with associated estimates of uncertainty, and this is likely to have transferability value to other sites. By accounting for additional variance in the data not explained by the A.I.I. with the spatially-autocorrelated random effects, it was possible to more robustly predict how concentrations of *E. coli* varied across the stream network at the catchment scale. This helped to identify "hot spots" of contamination (e.g. around Sites 2 and 10) and areas where contamination was reduced (e.g. around Sites 4–6) that would have been missed if predictions had been based on the A.I.I. alone. By better identifying "hot spots" of contamination, one could use SSNM predictions to determine areas for smaller-scale study to further explore drivers of contamination (e.g. with microbial source tracking methods), which in turn would aid identification of suitable mitigation strategies (Money et al., 2009; Oliver et al., 2016). It is also possible to use the uncertainty associated with each prediction so that additional monitoring efforts can be targeted to better constrain predicted concentrations of *E. coli* in areas where uncertainty is greatest (Peterson and Urquhart, 2006; Peterson and Ver Hoef, 2010). Similar value of SSNMs in catchment-wide assessment of other water quality variables such as temperature has been recently shown (Isaak et al., 2014; Peterson and Urquhart, 2006), and further application to microbiological parameters appears to have promising potential.

There is also value in the SSNMs in allowing stream network interactions to be inferred, due to the inclusion of a spatial autocovariance function (c.f. Ver Hoef et al., 2006). For example, 95th percentile concentrations of *E. coli* in the Tarland were predicted to decrease downstream of Site 3 until the confluence with the tributary of Site 2 (Fig. 7c). A physically-meaningful interpretation of this could be that elevated concentrations at Site 3, likely reflecting the WWTW influence, are increasingly diluted by cleaner downstream inputs of water, perhaps in particular from the tributary joining the main stem just downstream of Site 3. This would suggest that inputs of *E. coli* from the WWTW are moderated within the stream network, thus limiting the concentrations at the catchment outlet.

The SSNMs, although based on an unusually rich FIO dataset, were based on just 10 sampling locations, which may give rise to some limitations in comparison to applications to other water quality variables that can be sampled at higher spatial resolution (Garreta et al., 2009).

Firstly, the 10 locations gave 17 pairs of flow-connected sites for modelling the spatial autocovariance structures; this is lower than recommended for less uncertain characterisation of spatial autocorrelation and the parameters of spatial autocovariance structures (e.g. Cressie, 1993). Uncertainty in the range parameter may have also been increased by, at most, only one sampling site being present between network confluences (Garreta et al., 2009). In this case, the range parameter may be poorly defined and tend towards ∞, with the spatial weightings required by tail-up autocovariance structures controlling most of the changes in autocorrelation along the stream network (Garreta et al., 2009; Ver Hoef and Peterson, 2010). There is some evidence for this having been an issue with the SSNMs, with some models in Table 3 having ranges that were at the maximum site separation distance to which the range parameter was limited. Secondly, the small number of sampling sites will have impacted on the SSNM predictions and their certainty. For prediction sites that were distant from a flow-connected sampling site or along a tributary with no sampling site and a low spatial weighting (Fig. 3), uncertainty was increased (Fig. 7), and predictions will have increasingly reflected the fixed effects component of the model, losing the benefit of the spatially-autocorrelated random effects (Garreta et al., 2009; Peterson and Urquhart, 2006). As stated earlier, however, the estimates of uncertainty provided by SSNMs could be used to help highlight where additional monitoring could usefully take place, which, in turn, could help to better constrain both the predictions and parameters of the models (Peterson and Urquhart, 2006).

Whilst it is important to recognise these limitations in the SSNMs, they have been valuable in understanding and predicting faecal contamination at the catchment scale. As previously stated, such applications of SSNMs to FIO data have been very rare, likely reflecting the paucity of long-term, spatially-extensive datasets of FIOs (Kay et al., 2008a). However, the potential of such models to identify "hot spots" of faecal contamination and offer insights into how the influence of these is moderated by the stream network, as well as to help inform future monitoring efforts, should not be ignored. Our use of SSNMs with a unique long-term dataset provides a "proof of concept" in this regard, for the more reliable characterisation of catchment-scale spatial patterns of faecal contamination.

6. Conclusions

Understanding what drives the spatio-temporal dynamics of faecal contamination and being able to accurately predict "hot spots" of contamination are vital precursors to improving the microbiological quality of surface waters. Using a long-term, spatially-distributed dataset of concentrations of *E. coli* in a mixed land-use catchment, we found that, contrary to other catchment-scale studies, concentrations could not be clearly associated with flow conditions. This potentially reflects a combination of factors, including greater water fluxes from the less-contaminated headwaters of the catchment diluting downstream concentrations during high flows, the importance of contributions of *E. coli* from persistent point sources both anthropogenic and agricultural, and possibly the temporal resolution of the dataset. Also, a clear seasonality in concentrations was not evident, however 50th percentile concentrations were elevated at all sites in summer. These findings may have implications for the development of management strategies to improve surface water quality by suggesting that efforts should not solely be focused on high flow conditions or summer when concentrations of FIOs might normally be assumed to be greatest.

Spatial patterns of concentrations of *E. coli* could be significantly linked to an Anthropogenic Impact Index, a lumped metric used to indicate potential for contamination by anthropogenic point sources. The lack of an association with pasture or arable land may have reflected the inability of coarse-scale land-cover data to characterise "point sources" of agricultural contamination and the temporal variability in the spatial distribution of diffuse sources. SSNMs were found to have value in making more robust predictions of catchment-scale concentrations of *E. coli* with estimates of uncertainty, and in better characterising potential "hot spots" of contamination. Whilst the models may have been limited by the relatively sparse (in comparison to other water quality variables) spatial nature of the *E. coli* dataset used to define their autocovariance structures, the identification of potential "hot spots" of faecal contamination at the catchment scale can inform possible locations for management or smaller-scale, process-based studies that would seek to confirm and then understand why certain areas are "hot spots". In addition, the estimates of uncertainty provided for predictions made by SSNMs could be used to help design more spatially-intensive monitoring programmes, which, in turn, would allow for the parameters and predictions of SSNMs to be better constrained. The method, therefore, has the potential for wider application in microbiological water quality studies. Future work in the Tarland will focus on the "hot spot" area of the Site 10 sub-catchment, and will employ microbial source tracking and environmental tracers to further explore dominant sources of contamination and the hydrological flow paths which facilitate their connection to the stream.

Acknowledgements

Thanks to the Scottish Government's Hydro Nation Scholars Programme for funding AJN to do this research. Thanks to the Drinking Water Quality Regulator (DWQR) for Scotland for providing information about private water supplies and septic tanks in the Tarland.

References

Beleites, C., Baumgartner, R., Bowman, C., Somorjai, R., Steiner, G., Salzer, R., Sowa, M.G., 2005. Variance reduction in estimating classification error using sparse datasets. Chemom. Intell. Lab. Syst. 79, 91–100.
Bergfur, J., Demars, B.O.L., Stutter, M.I., Langan, S.J., Friberg, N., 2012. The Tarland catchment initiative and its effect on stream water quality and macroinvertebrate indices. J. Environ. Qual. 41, 314–321.
Beven, K.J., Kirkby, M.J., 1979. A physically based, variable contributing area model of basin hydrology. Hydrol. Sci. J. 24 (1), 43–69.
Blaustein, R.A., Pachepsky, Y., Hill, R.L., Shelton, D.R., Whelan, G., 2013. *Escherichia coli* survival in waters: temperature dependence. Water Res. 47 (2), 569–578.
Chadwick, D., Fish, R., Oliver, D.M., Heathwaite, L., Hodgson, C., Winter, M., 2008. Management of livestock and their manure to reduce the risk of microbial transfers to water – the case for an interdisciplinary approach. Trends Food Sci. Technol. 19, 240–247.
Cole, B., King, S., Ogutu, B., Palmer, D., Smith, G., Balzter, H., 2015. Corine Land Cover 2012 for the UK. NERC Environmental Information Data Centre, Jersey and Guernsey.
(Computer software)Conrad, O., 2006. SAGA Version 2.0 (System for Automated Geoscientific Analysis). Geographisches Institut, Göttingen.
COUNCIL DIRECTIVE 2006/7/EC of 15 February, 2006. Concerning the Management of Bathing Water Quality and Repealing Directive 76/160/EEC.
COUNCIL DIRECTIVE 98/83/EC of 3 November, 1998. on the Quality of Water Intended for Human Consumption.
Cressie, N.A.C., 1993. Statistics for Spatial Data. John Wiley and Sons, New York.
Crowther, J., Kay, D., Wyer, M.D., 2002. Faecal-indicator concentrations in waters draining lowland pastoral catchments in the UK: relationships with land use and farming practices. Water Res. 36, 1725–1734.
Crowther, J., Wyer, M.D., Bradford, M., Kay, D., Francis, C.A., 2003. Modelling faecal indicator concentrations in large rural catchments using land use and topographic data. J. Appl. Microbiol. 94, 962–973.
Davies-Colley, R.J., Nagels, J.W., Smith, R.A., Young, R.G., Phillips, C.J., 2004. Water quality impact of a dairy cow herd crossing a stream. N. Z. J. Mar. Freshw. Res. 38 (4), 569–576.
Dawson, J.J.C., Soulsby, C., Tetzlaff, D., Hrachowitz, M., Dunn, S.M., Malcolm, I.A., 2008. Influence of hydrology and seasonality on DOC exports from three contrasting upland catchments. Biogeochemistry 90, 93–113.
Dwivedi, D., Mohanty, B.P., Lesikar, B.J., 2013. Estimating *Escherichia coli* loads in streams based on various physical, chemical and biological factors. Water Resour. Res. 49, 2896–2906.
Dwivedi, D., Mohanty, B.P., Lesikar, B.J., 2016. Impact of the linked surface water-soil water-groundwater system on transport of *E. coli* in the subsurface. Water Air Soil Pollut. 227, 351.
Fewtrell, L., Kay, D., 2015. Recreational water and infection: a review of recent findings. Current Environ. Health Rep. 2 (1), 85–94.
Flynn, R.M., Deakin, J., Archbold, M., Cushnan, H., Kilroy, K., O'Flaherty, V., Misstear, B.D., 2016. Using microbiological tracers to assess the impact of winter land use restrictions on the quality of stream headwaters in a small catchment. Sci. Total Environ. 541, 949–956.
Garreta, V., Monestiez, P., Ver Hoef, J.M., 2009. Spatial modelling and prediction on river networks: up model, down model or hybrid? Environmetrics 21 (5), 439–456.

Helsel, D.R., 1990. Less than obvious – statistical treatment of data below the detection limit. Environ. Sci. Technol. 24 (12), 1766–1774.
Hunter, C., Perkins, J., Tranter, J., Gunn, J., 1999. Agricultural land-use effects on the indicator bacterial quality of an upland stream in the Derbyshire peak district in the UK. Water Res. 33, 3577–3586.
Isaak, D.J., Peterson, E.E., Ver Hoef, J.M., Wenger, S.J., Falke, J.A., Torgersen, C.E., Sowder, C., Steel, E.A., Fortin, M., Jordan, C.E., Ruesch, A.S., Som, N., Monestiez, P., 2014. Applications of spatial statistical network models to stream data. Wiley Interdiscip. Rev. 1 (3), 277–294.
Kay, D., Wyer, M., Crowther, J., Stapleton, C., Bradford, M., McDonald, A., Greaves, J., Francis, C., Watkins, J., 2005. Predicting faecal indicator fluxes using digital land use data in the UK's sentinel Water Framework Directive catchment: the Ribble study. Water Res. 39, 3967–3981.
Kay, D., Watkins, J., Francis, C.A., Wyn-Jones, A.P., Stapleton, C.M., Fewtrell, L., Wyer, M., Drury, D., 2007. The microbiological quality of seven large commercial private water supplies in the United Kingdom. J. Water Health 05 (4), 523–538.
Kay, D., Crowther, J., Fewtrell, L., Francis, C.A., Hopkins, M., Kay, C., McDonald, A.T., Stapleton, C.M., Watkins, J., Wilkinson, J., Wyer, M.D., 2008a. Quantification and control of microbial pollution from agriculture: a new policy challenge? Environ. Sci. Pol. 11, 171–184.
Kay, D., Crowther, J., Stapleton, C.M., Wyer, M.D., Fewtrell, L., Anthony, S., Bradford, M., Edwards, A., Francis, C.A., Hopkins, M., Kay, C., McDonald, A.T., Watkins, J., Wilkinson, J., 2008b. Faecal indicator organism concentrations and catchment export coefficients in the UK. Water Res. 42, 2649–2661.
Kay, D., Anthony, S., Crowther, J., Chambers, B.J., Nicholson, F.A., Chadwick, D., Stapleton, C.M., Wyer, M.D., 2010. Microbial water pollution: a screening tool for initial catchment-scale assessment and source apportionment. Sci. Total Environ. 408, 5649–5656.
Lane, S.N., Brookes, C.J., Kirkby, M.J., Holden, J., 2004. A network-index-based version of TOPMODEL for use with high-resolution digital topographic data. Hydrol. Process. 18, 191–201.
Lane, S.N., Reaney, S.M., Heathwaite, A.L., 2009. Representation of landscape hydrological connectivity using a topographically driven surface flow index. Water Resour. Res. 45, W08423.
Langan, S.J., Wade, A.J., Smart, R.P., Edwards, A.C., Soulsby, C., Billett, M.F., Jarvie, H.P., Cresser, M.S., Owen, R., Ferrier, R.C., 1997. The prediction and management of water quality in a relatively unpolluted major Scottish catchment: current issues and experimental approaches. Sci. Total Environ. 194 (195), 419–435.
Langan, S., Watson, H., Johnston, L., Cook, Y., Cooper, R., Taylor, C., Masson, L., Stutter, M.I., Riach, D.J., McIntyre, S., Thomson, C., Curran, C., Sturgeon, F., McKeen, M., (DOI pending). Water Quality Data for the Tarland Catchment, NE Scotland n.d.
McGrane, S.J., Tetzlaff, D., Soulsby, C., 2014. Application of a linear regression model to assess the influence of urbanised areas and grazing pastures on the microbiological quality of rural streams. Environ. Monit. Assess. 186, 7141–7155.
McKergow, L.A., Davies-Colley, R.J., 2010. Stormflow dynamics and loads of *Escherichia coli* in a large mixed land use catchment. Hydrol. Process. 24, 276–289.
Money, E.S., Carter, G.P., Serre, M.L., 2009. Modern space/time geostatistics using river distances: data integration of turbidity and *E. coli* measurements to assess fecal contamination along the Raritan River in New Jersey. Environ. Sci. Technol. 43, 3736–3742.
Natvig, E.E., Ingham, S.C., Ingham, B.H., Cooperband, L.R., Roper, T.R., 2002. Salmonella enterica serovar *Typhimurium* and *Escherichia coli* contamination of root and leaf vegetables grown in soils with incorporated bovine manure. Appl. Environ. Microbiol. 68 (6), 2737–2744.
Oliver, D.M., Clegg, C.D., Haygarth, P.M., Heathwaite, A.L., 2005a. Assessing the potential for pathogen transfer from grassland soils to surface waters. Adv. Agron. 85, 125–180.
Oliver, D.M., Heathwaite, L., Haygarth, P.M., Clegg, C.D., 2005b. Transfer of *Escherichia coli* to water from drained and undrained grassland after grazing. J. Environ. Qual. 34, 918–925.
Oliver, D.M., Heathwaite, A.L., Hodgson, C.J., Chadwick, D.R., 2007. Mitigation and current management attempts to limit pathogen survival and movement within farmed grassland. Adv. Agron. 93, 95–152.
Oliver, D.M., Porter, K.D.H., Pachepsky, Y.A., Muirhead, R.W., Reaney, S.M., Coeffey, R., Kay, D., Milledge, D.G., Hong, E., Anthony, S.G., Page, T., Bloodworth, J.W., Mellander, P., Carbonneau, P., McGrane, S.J., Quilliam, R.S., 2016. Predicting microbial water quality with models: over-arching questions for managing risk in agricultural catchments. Sci. Total Environ. 544, 39–47.
Pachepsky, Y.A., Shelton, D.R., 2011. *Escherichia coli* and fecal coliforms in freshwater and estuarine sediments. Crit. Rev. Environ. Sci. Technol. 41, 1067–1110.
Peterson, E.E., Urquhart, N.S., 2006. Predicting water quality impaired stream segments using landscape-scale data and a regional geostatistical model: a case study in Maryland. Environ. Monit. Assess. 121, 615–638.
Peterson, E.E., Ver Hoef, J.M., 2010. A mixed-model moving-average approach to geostatistical modeling in stream networks. Ecology 91 (3), 644–651.
Peterson, E.E., Ver Hoef, J.M., 2014. STARS: an ArcGIS toolset used to calculate the spatial information needed to fit spatial statistical models to stream network data. J. Stat. Softw. 56 (2), 1–17.
Peterson, E.E., Ver Hoef, J.M., Isaak, D.J., Falke, J.A., Fortin, M., Jordan, C.E., McNyset, K., Monestiez, P., Ruesch, A.S., Sengupta, A., Som, N., Steel, E.A., Theobald, D.M., Torgersen, C.E., Wenger, S.J., 2013. Modelling dendritic ecological networks in space: an integrated network perspective. Ecol. Lett. 16, 707–719.
(Computer software)R Core Team, 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria http://www.R-project.org/.
(Computer software)Reaney, S.M., Milledge, D., 2013. SCIMAP Risk Maps. Retrieved from:. http://www.scimap.org.uk/2013/03/march-2013-version-of-scimap-for-saga-gis/.
Reaney, S.M., Lane, S.N., Heathwaite, A.L., Dugdale, L.J., 2011. Risk-based modelling of diffuse land use impacts from rural landscapes upon salmonid fry abundance. Ecol. Model. 222 (4), 1016–1029.
Richards, S., Withers, P.J.A., Paterson, E., McRoberts, C.W., Stutter, M., 2016. Temporal variability in domestic point source discharges and their associated impact on receiving waters. Sci. Total Environ. 571, 1275–1283.
Soil Survey of Scotland Staff, 2014. Digitized Soil Map of Scotland, Scale 1:25 000. James Hutton Institute.
Stapleton, C.M., Wyer, M.D., Crowther, J., McDonald, A.T., Kay, D., Greaves, J., Wither, A., Watkins, J., Francis, C., Humphrey, N., Bradford, M., 2008. Quantitative catchment profiling to apportion faecal indicator organism budgets for the Ribble system, the UK's sentinel draining basing for Water Framework Directive research. J. Environ. Manag. 87, 535–550.
(Computer software)StataCorp, 2015. Stata Statistical Software: Release 14. StataCorp LP, College Station, TX.
Strachan, N.J., Dunn, G.M., Locking, M.E., Reid, T.M., Ogden, I.D., 2006. *Escherichia coli* O157: burger bug or environmental pathogen? Int. J. Food Microbiol. 112 (2), 129–137.
Stutter, M.I., Langan, S.J., Cooper, R.J., 2008a. Spatial and temporal dynamics of stream water particulate and dissolved N, P and C forms a long a catchment transect, NE Scotland. J. Hydrol. 350, 187–202.
Stutter, M.I., Langan, S.J., Cooper, R.J., 2008b. Spatial contributions of diffuse inputs within-channel processes to the form of stream water phosphorus over storm events. J. Hydrol. 350, 203–214.
Stutter, M.I., Demars, B.O.L., Langan, S.J., 2010. River phosphorus cycling: separating biotic and abiotic uptake during short-term changes in sewage effluent loading. Water Res. 44, 4425–4436.
Tetzlaff, D., Capell, R., Soulsby, C., 2012. Land use and hydroclimatic influences on faecal indicator organisms in two large Scottish catchments: towards land use-based models as screening tools. Sci. Total Environ. 434, 110–122.
Tyrrel, S.F., Quinton, J.N., 2003. Overland flow transport of pathogens from agricultural land receiving faecal wastes. J. Appl. Microbiol. 94, 87S–93S.
Ver Hoef, J.M., Peterson, E.E., 2010. A moving average approach for spatial statistical models of stream networks. J. Am. Stat. Assoc. 105 (489), 6–18.
Ver Hoef, J.M., Peterson, E., Theobald, D., 2006. Spatial statistical models that use flow and stream distance. Environ. Ecol. Stat. 13, 449–464.
Ver Hoef, J.M., Peterson, E.E., Clifford, D., Shah, R., 2014. SSN: an R package for spatial statistical modeling on stream networks. J. Stat. Softw. 56 (3), 1–45.
Vitro, K.A., BenDor, T.K., Jordanova, T.V., Miles, B., 2017. A geospatial analysis of land use and stormwater management on fecal conliform contamination in North Carolina streams. Sci. Total Environ. 603-604:709–727. http://dx.doi.org/10.1016/j.scitotenv.2017.02.093.