Aberystwyth University

Article

Generic Modelling of Faecal Indicator Organism Concentrations in the UK

John Crowther [1], Danyel I. Hampson [2],*, Ian J. Bateman [2], David Kay [3], Paulette E. Posen [2], Carl M. Stapleton [3] and Mark D. Wyer [3]

[1] Centre for Research into Environment and Health, University of Wales, Trinity Saint David, Lampeter, Ceredigion, SA48 7ED, UK; E-Mail: j.crowther@tsd.ac.uk

[2] School of Environmental Sciences, University of East Anglia, Norwich, Norfolk, NR4 7TJ, UK; E-Mails: I.Bateman@uea.ac.uk (I.J.B.); P.Posen@uea.ac.uk (P.E.P.)

[3] Catchment and Coastal Research Centre, River Basin Dynamics and Hydrology Research Group, IGES, Aberystwyth University, Aberystwyth, Ceredigion, SY23 3DB, UK; E-Mails: dave@crehkay.demon.co.uk (D.K.); carl@creh.karoo.co.uk (C.M.S.); mark@crehwyer.demon.co.uk (M.D.W.)

* Author to whom correspondence should be addressed; E-Mail: D.Hampson@uea.ac.uk; Tel.: +44-7708-857370; Fax: +44-1603-591327.

Abstract: To meet European Water Framework Directive requirements, data are needed on faecal indicator organism (FIO) concentrations in rivers to enable the more heavily polluted to be targeted for remedial action. Due to the paucity of FIO data for the UK, especially under high-flow hydrograph event conditions, there is an urgent need by the policy community for generic models that can accurately predict FIO concentrations, thus informing integrated catchment management programmes. This paper reports the development of regression models to predict base- and high-flow faecal coliform (FC) and enterococci (EN) concentrations for 153 monitoring points across 14 UK catchments, using land cover, population (human and livestock density) and other variables that may affect FIO source strength, transport and die-off. Statistically significant models were developed for both FC and EN, with greater explained variance achieved in the high-flow models. Both land cover and, in particular, population variables are significant predictors of FIO concentrations, with r^2 maxima for EN of 0.571 and 0.624, respectively. It is argued that

the resulting models can be applied, with confidence, to other UK catchments, both to predict FIO concentrations in unmonitored watercourses and evaluate the likely impact of different land use/stocking level and human population change scenarios.

1. Introduction

Under the Water Framework Directive (WFD) EU member states are legally required to design "*programmes of measures*" to manage point and diffuse sources of faecal indicator organisms (FIOs) that could cause non-compliance of bathing and shellfish-harvesting waters with microbial standards [1,2]. To meet this requirement, data are needed to define FIO concentrations and fluxes in individual rivers and streams so that the magnitude of the problem can be assessed and more heavily polluted waters identified for potential remedial action. Knowledge is also required of the effectiveness of specific measures to reduce FIO concentrations, especially those aimed at addressing diffuse pollution sources.

Previous studies have highlighted the importance of rainfall-induced hydrograph ("high-flow") conditions for the mobilisation and transport of FIOs within catchments, with FIO concentrations and discharge volumes both typically increasing by an order of magnitude compared with dry-weather ("base") flow, leading to a c.100-fold increase in export coefficients. Kay *et al.* [3], for example, report an increase in geometric mean (GM) faecal coliform (presumptive *Escherichia coli*) ("FC") export coefficient for 205 UK rivers and streams from 5.5×10^8 cfu km^{-2} h^{-1} at base flow to 3.6×10^{10} cfu km^{-2} h^{-1} at high flow, and a corresponding increase for enterococci ("EN") from 8.3×10^7 to 7.1×10^9 cfu km^{-2} h^{-1}. Unfortunately, FIO concentration data are lacking for many UK watercourses, and, even for those that are routinely monitored, there is often poor characterisation of high-flow conditions.

Given the legal requirements of the WFD, there is an urgent imperative from the research and policy communities for generic models that can accurately predict base- and high-flow FIO concentrations across the UK, thereby informing future integrated catchment management programmes.

While physically based watershed modelling (Hydrological Simulation Program Fortran (HSPF), Simulated Catchments (SIMCAT), Soil and Water Assessment Tool (SWAT), *etc.*) is now quite advanced in relation to nutrients and sediments, deployment of these model systems to FIOs is prevented by the absence of empirical data with which to parameterise and evaluate any of these models. For example, credible regionally specific in-channel deposition and re-entrainment coefficients, together with empirically based water column real-time T$_{90}$ coefficients, are simply not available. A similar absence of parameterisation data is evident for the terrestrial land phase of FIO flux, making application of process-based models, at this stage, difficult, if not impossible. The development of the Scotland and Northern Ireland Forum for Environmental Research (SNIFFER) "screening tool" (at a 1×1 km grid cell resolution) for identifying and characterising diffuse pollution in Scotland and Northern Ireland has provided valuable insights into the types and strengths of FIO

pollution sources and the factors affecting the risk of pollutant mobilisation and delivery to watercourses, thus enabling the potential FIO export coefficients for catchments to be determined [4,5]. However, SNIFFER does not provide a basis for characterising base- and high-flow FIO concentrations, and the veracity of the export coefficient calculations have yet to be fully evaluated against data from monitored catchments. Some of the most successful catchment-scale FIO modelling has been undertaken using linear regression techniques to model relationships between GM FIO concentrations recorded at monitored sites and land use within their catchments, using variables such as the proportions of grassland and built-up land as proxies for key sources of faecal pollution. Hitherto, this work has been primarily based on individual catchment studies [6,7].

The aim of the present study was to extend this latter approach by investigating: (i) whether improved models might be achieved by augmenting the predictor variables to include both direct measures of the key FIO sources (*i.e.*, human population and livestock density data) and factors that may affect source strength and the mobilisation, transport, die-off and sedimentation of FIOs within catchments (e.g., volume of runoff, soil hydrology and catchment size); and (ii) the extent to which models developed by combining data from discrete UK catchment studies, sampled at different times and under different antecedent weather conditions, are truly "generic" and transferable across the UK.

Previously, Kay *et al.* [3] have presented a synthesis of FIO concentration and export coefficient data for 205 sampling points in 15 study catchments, which are broadly representative of the diverse land use types and climatic regimes across the UK. Southern England and areas of chalk downland are the notable regions/landscapes that are not represented. In the present study, data from those points with catchments ≥ 5 km^2 (153 from 14 of the study catchments) have been used to model base- and high-flow FC and EN concentrations during the summer bathing season, using predictor variables that are readily available and have national coverage.

The transferable models presented here provide a tool for characterising FIO concentrations in rivers and streams across the UK, though they clearly need to be applied with some degree of caution in Southern England. In addition, they can provide valuable insights into the key sources and factors affecting transfer and survival of FIOs at a catchment scale, thereby informing the development of policies and prioritisation of investment to reduce microbial pollution as an optimal mix of cost-effective regional and site specific policy remediation strategies will be required to achieve the highest reductions.

Importantly, too, these models can form a basis for quantifying the likely impact of different land use/stocking levels that might result from the implementation of measures designed to reduce FIO loadings, or reforms in agricultural policy/funding e.g. by linking to land use change models, such as those developed by Fezzi *et al.* [8] and Jones and Trantor [9].

The models presented here, developed to inform the Rural Economy and Land Use funded Catchment, Hydrology, Resources, Economics and Management programme [10], have been used to generate predictions of FIO concentrations at base- and high-flow, within the Humber river basin district (RBD), and generate predictions of FIO concentration reductions within that RBD following the implementation of seven land use management and policy instruments; fiscal constraint, production constraint, cost intervention, area intervention, demand-side constraint, input constraint, and micro-level land use management. The results of these analyses are detailed in Hampson *et al.* [11].

2. Materials and Methods

2.1. Study Catchments, Subcatchments, Sampling Periods, Field Methods and Laboratory Analysis

The present data set is derived from monitoring undertaken in 14 catchment-based studies during the period 1995–2005. These are identified in Table 1 and their locations shown in Figure 1.

Figure 1. Location of the 14 study catchments (1–14, as detailed in Table 1) used in the modelling data set and of the Haverigg catchment (labelled 15) used in independent model evaluation.
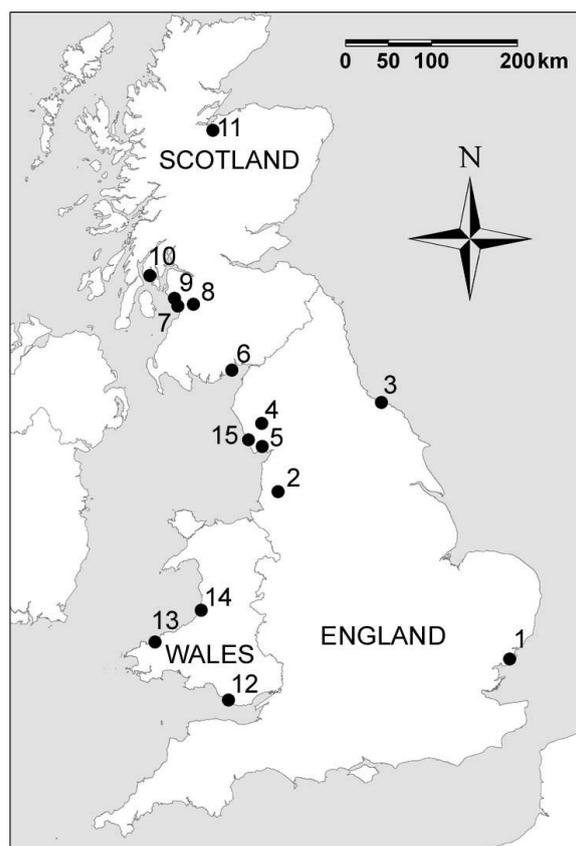
Typically, each study extended over a six- to eight-week period during the summer bathing season (*i.e.*, 15 May–30 September in England and Wales, and 1 June–15 September in Scotland), as the studies were aimed at improving understanding of bathing water compliance. To increase the robustness of the modelling, only sites that meet the following criteria have been included: (1) catchment (termed 'subcatchment' for individual sample points) area $\geq$ 5 km^2 (because of the relatively low resolution of the livestock census and soil hydrology data–see below); (2) < 50% of land within the subcatchment is located upstream of lake and/or reservoir outlets (see Section 2.2); (3) FIO data available for $\geq$ 5 samples taken under the flow conditions being modelled (*i.e.*, base or high flow); (4) river discharge records available; and (5) land within the subcatchments had not been subject to programmes of measures (e.g., riparian fencing and buffer strips) aimed at reducing FIO loadings.

Table 1 shows the distribution of the resulting 153 subcatchments included in the present analysis. The majority are < 100 km^2 (maximum, 1013 km^2). It should be noted that there are combined sewage overflows (CSOs) and wastewater treatment works (WwTWs) within many of the subcatchments. The FIO concentrations recorded therefore reflect inputs from a combination of point and diffuse sources. Details of the discharge monitoring, base-/high-flow separation, water sampling and microbial analysis are presented in Kay *et al.* [12]. Presumptive FC and EN concentrations were measured using standard UK methods based on membrane filtration, which have not changed substantially over the study period [13,14]. It should be noted that with membrane filtration, as with all standard methods of analysis (e.g., MPN), particle-attached FIOs, which are more likely present in turbid waters under high-flow conditions, may be underestimated using this procedure. Normal laboratory practice in the UK, involving sample agitation immediately prior to filtration, seeks to minimise this effect. FIO enumerations are expressed as colony-forming units (cfu) 100 mL^{-1}. The geometric mean (GM, calculated as: GM $= 10^x$, where $x =$ the mean of log$_{10}$-transformed values) concentration is used to characterise microbial water quality under base- and high-flow conditions for each sampling point, as this generally has greater validity as a measure of central tendency for FIO data.

Table 1. Catchments reported in present study.

Catchment		Study year	Subcatchments [a]
England			
1	Holland Brook	1998	10
2	River Ribble	2002	37
3	Staithes Beck	1995	6 (2)
4	Lake Windermere inputs	1999	3
5	River Leven/Crake	2005	16
Scotland			
6	Sandyhills	2004	4
7	Troon coastal inputs	2000	1
8	Killoch Burn[b]	2004	1
9	River Irvine/Garnock	1998	23
10	Ettrick Bay inputs	2004	1
11	River Nairn	2004	8
Wales			
12	Afon[c] Ogwr	1997	14
13	Afon Nyfer	1996	17 (2)
14	Afon Rheidol/Ystwyth	1999	12
	Total		153 (134)

[a] Number of subcatchments used in modelling. The Staithes Beck and Afon Nyfer study periods were relatively dry, and in each case ≥ 5 high-flow samples were only obtained for 2 subcatchments, giving a total of 134 high-flow sites (indicted in parentheses), *cf.* 153 base-flow sites;

[b] Killoch Burn is located within the headwaters of the River Irvine/Garnock catchment;

[c] "Afon" (Welsh) = "River".

2.2. Outputs from Lakes/Reservoirs

Because of die-off and sedimentation of FIOs within reservoirs and lakes, waters issuing from such waterbodies typically have very low FIO concentrations which may poorly reflect land use, stocking levels, *etc.* within the contributing catchment [15]. For this reason, the few subcatchments in which the majority ($\geq$ 50%) of land is located upstream of reservoir/lake outlets have been excluded from the modelling data set. In cases where any waterbody outlets are fed by areas occupying < 50% of the subcatchment, it has been assumed that: (i) the volume of flow derived from the waterbody is proportional to the area its catchment occupies within the subcatchment; and (ii) GM FC and EN concentrations in output waters from such waterbodies are as reported in Table 2.

Table 2. Geometric mean faecal coliform and enterococci concentrations (cfu 100 mL^{-1}) in waters issuing from lakes and reservoirs, based on data from Nant-y-Moch and Cwm Rheidol Reservoirs in Afon Rheidol/Ystywyth study, Lake Windermere in the Windermere study, and Fewston and Thruscross Reservoirs in Yorkshire—the latter from Kay [16].

	Base flow	High flow
Faecal coliforms	26	83
Enterococci	5	16

Using these assumptions, the GM FIO concentrations recorded at the subcatchment monitoring point have been separated into two components: that derived from the waterbody and that from the rest of the subcatchment. Data for the non-waterbody part of the subcatchment have then been used in the modelling. This procedure is considered preferable to excluding all subcatchments containing lakes and/or reservoirs.

2.3. Land Cover Data

The land cover data have been synthesised from the Centre for Ecology and Hydrology (CEH) Land Cover Map (LCM) 2000 [17] and the Ordnance Survey (OS) Meridian 2 digital boundary data [18] using GIS techniques [19]. The synthesis method and 'proportional' interpolation methods used subsequently to allocate land cover data (and also human population, livestock and soil hydrology data) to subcatchments are detailed in Posen *et al.* [20].

2.4. Human Population Data

Population data were derived from the Office for National Statistics (ONS) decennial census data for England, Scotland and Wales for 2001 [21,22], which lies roughly in the middle of the period that the catchment studies were undertaken (Table 1). Human population data at the "Output Area" (OA) level were used as this is the most detailed geographic unit for which 2001 census data are available. Each OA has approximately 300 residents and, more importantly, OA boundaries enclose the most spatially compact area possible, thereby minimising interpolation errors.

2.5. Livestock Data

Stocking densities have been derived from the Agcensus data set compiled from the June Agricultural Survey [23]. Livestock categories included in the Agcensus data sets can vary from year to year, and data sets are not necessarily available for every year. Additionally, there are occasional variations between the census forms used for England, Wales and Scotland, which can lead to minor inconsistencies in the data. In cases where census data are unavailable for the year an individual catchment was studied, or the data have been too coarsely aggregated to preserve the confidentiality of individual farmers (e.g., in the England 2000 and Wales 1999 censuses), the closest census year was used. Data for the smallest available grid-square resolution, 2×2 km, were used.

The livestock categories recorded were: dairy herd, beef herd, bulls, other cattle ≥ 1-yr old, other cattle < 1-yr old, sheep (which also include goats, deer and horses), total pigs, indoor pigs, outdoor pigs, total fowl, indoor fowl and free-range fowl. It should be noted, as detailed in Table 3, that some of these categories were aggregated for modelling purposes, and composite variables combining the different livestock types, and also human population with livestock, were derived using data on *E. coli* production per animal [24], thereby providing a FIO-related weighting for different animal types.

2.6. Soil Hydrological Data

Data from the HOST (Hydrology of Soil Type) database [25], at a grid-square resolution of 1×1 km, were used to calculate the mean standard percentage runoff (SPR) for each subcatchment.

2.7. Statistical Analysis

Statistical analysis was undertaken using SPSS v15.0 for Windows [26]. Multiple regression techniques, using a stepwise selection procedure, were used to model the relationships between GM FIO concentrations at base and high flow (the dependent variables, *y*) and the various independent variables (*x*) listed in Table 3. Log_{10} transformations were applied to those independent variables for which skewness exceeded 1.00. In the regression analysis, relationships of the following form were generated: $y = a + b_1x_1 + b_2x_2 + \ldots + b_ix_i + e$ where *a* is the intercept (*y* at $x = 0$), *b* is the slope (change in *y* per unit change in *x*) and *e* is a random error term. Independent variables with a variance inflation factor > 5 (*i.e.*, tolerance, 0.200) were excluded to minimise multicollinearity [27]; probability of *F* for a variable to enter was set at 0.05; the level of explained variance was assessed using the coefficient of determination (r^2), adjusted for degrees of freedom; and the normal probability plot of standardised residuals was examined to confirm the validity of each model. All statistical tests were assessed at $\alpha = 0.05$ (*i.e.*, 95% confidence level).

Regression models were developed for the following three sets of independent predictor variables: (1) all variables (as detailed in Table 3), (2) land cover variables alone and (3) population (human/livestock) variables alone. In the last two cases only those variables that represent likely sources of enhanced FIOs were included. Such a specification allows the most parsimonious models to be identified, thereby facilitating linkage to allied research, such as land-use change models [8,9].

Table 3. Independent variables [a] used in regression models to predict variations in $\log_{10}$ geometric mean faecal coliform and enterococci concentrations under base- and high-flow conditions, and summary statistics for the 153 subcatchments used in the modelling.

Variable (unit of measurement)	Code [b]	Mean	Minimum	Maximum	Std Dev
Population and population-related variables					
Human population (no. km^{-2})	HUMAN *	206.0	2.1	2,178.3	398.4
Dairy cattle (no. km^{-2})	DAIRY *	22.9	0	84.1	21.0
Other, *i.e.*, non-dairy, cattle (no. km^{-2})	OTHCATTLE *	44.7	0.2	109.3	29.0
All cattle (no. km^{-2})	ALLCATTLE *	67.6	0.2	189.5	47.5
Sheep (no. km^{-2})	SHEEP	292.8	6.7	953.5	206.1
E. coli input: all livestock (cfu km^{-2} h^{-1})	LSTOCKEC	2.4×10^{11}	5.1×10^{9}	7.4×10^{11}	1.5×10^{11}
E. coli input: humans + livestock (cfu km^{-2} h^{-1})	TOTALEC	2.6×10^{11}	1.0×10^{10}	7.4×10^{11}	1.5×10^{11}
Land use variables					
Temporary/permanent grassland (%)	GRASSLAND	45.66	0.45	88.33	23.92
Rough grazing (%)	RGRAZING *	18.69	0	85.33	19.50
Arable, including set-aside (%)	ARABLE *	12.39	0	81.14	19.72
Woodland (%)	WOODLAND *	11.84	0.38	65.26	12.85
Urban/suburban (%)	URBAN *	8.74	0	56.88	10.67
Other variables					
Subcatchment area (km^{2})	AREA *	71.19	5.01	1,013.18	137.49
Base-flow runoff (m^{3} km^{-2} h^{-1})	LFRUNOFF	49.47	2.43	196.33	38.99
High-flow runoff (m^{3} km^{-2} h^{-1})	HFRUNOFF	216.56	7.90	845.51	170.91
Total runoff (m^{3} km^{-2} h^{-1})	TOTRUNOFF	83.61	3.94	211.88	60.85
Standard percentage runoff (SPR) [c]: all land (%)	TOTSPR	40.28	22.47	59.41	40.28
Standard percentage runoff (SPR): grassland (%)	GRSPR	38.39	18.38	58.44	6.64

[a] Data sources/derivations (see further details in text): Human population = 2001 census; Livestock numbers = Agricultural census for year nearest to study period; *E. coli* input = number of each livestock type (km^{-2}) × mean *E. coli* output for each livestock type (based on Jones and White, 1984); Land use data = Centre for Ecology and Hydrology Land Cover Map 2000 in combination with OS Meridian 2; Runoff data = actual runoff during study period; and Standard percentage runoff (SPR) variables = Institute of Hydrology: Hydrology of Soil Types ("HOST"), with area-weighted mean SPR derived using procedure described by Boorman *et al.* (Table 4.17, p. 60) [25];

[b] * indicates $\log_{10}$ transformation applied in cases where this reduces skewness of data set;

[c] SPR values for individual soil types in the UK range from 2–60%.

2.8. Assessment of Inter-Study Transferability of Models

A programme of out-of-sample testing was undertaken to evaluate the extent to which the models are truly generic and transferable to other UK catchments [28,29]. In order to minimise the effects of unexplained variance in the models, attention focused on the model that provided the highest level of explained variance. This model was re-run seven times with data for one of the seven catchment studies with $\geq$ 5 subcatchments omitted in turn. The resulting models were then used to predict the GM concentration for subcatchments in the omitted catchments (termed "test catchments"), and the mean error (*i.e.*, predicted–actual concentration) and mean absolute error (*i.e.*, absolute difference between predicted and actual concentration) were calculated for each study catchment. The mean error provides a measure of whether a model is under- (+ve values) or over-estimating (−ve values) GM FIO concentrations within the test catchments. In cases where the mean error and mean absolute error have the same value, then the GM concentrations for all of the subcatchments in the test catchment are either over- or under-estimated. As a further independent check, this model was applied to three sampling points in a further catchment (the Haverigg catchment, Cumbria), which was monitored in 2008.

3. Results

3.1. All-Variable Models

The results, summarised in Table 4, show statistically significant ($p < 0.05$) base- and high-flow regression models for both FC and EN.

Table 4. Summary of results of stepwise multiple regression models of relationship between log_{10} geometric mean faecal coliform and enterococci concentrations at base and high flow and all the independent variables listed in Table 3.

	Step	Variable	Sign of *b* [a]	Adjusted r^2	Sig level (p)
Base-flow models (*n* = 153)					
Faecal coliforms					
	1	HUMAN	+	0.363	
	2	DAIRY	+	0.418	
	3	ALLCATTLE	−?	0.481	
	4	ARABLE	+?	0.505	
	5	AREA	+?	0.518	<0.001
Enterococci					
	1	URBAN	+	0.294	
	2	DAIRY	+	0.325	
	3	ALLCATTLE	−?	0.369	
	4	ARABLE	+?	0.394	<0.001
High-flow models (*n* =134)					
Faecal coliforms					
	1	DAIRY	+	0.439	
	2	HUMAN	+	0.595	
	3	TOTALEC	+	0.631	
	4	OTHCATTLE	−?	0.653	<0.001

Table 4. *Cont.*

Step	Variable	Sign of b [a]	Adjusted r^2	Sig level (p)
Enterococci				
1	DAIRY	+	0.388	
2	HUMAN	+	0.598	
3	TOTALEC	+	0.632	<0.001

[a]? indicates that the sign does not conform with prior expectation.

The levels of explained variance are higher in the two high-flow models (maximum r^2, 0.632 for FC) than the base-flow models (maximum r^2, 0.518 for FC). In each case at least three independent variables were entered. With the exception of AREA, which is entered at Step 5 in the base-flow FC model, all the variables entered are either population- or land cover-related variables (*i.e.*, neither runoff during the study period nor soil hydrology (SPR) are sufficiently significant to warrant inclusion in the models).

Overall, the models are dominated by the population variables, particularly HUMAN and DAIRY, though some land cover variables are also significant. DAIRY is entered first in the high-flow models, whereas HUMAN or URBAN are entered at Step 1 in the base-flow models. For all of the more significant predictors the sign of the slope (*b*) value is consistent with prior expectations. However, some of the less significant variables (as highlighted in Table 4) have unexpected effects.

3.2. Land Cover-Based Models

In each case statistically significant regression models have been generated, Table 5, with URBAN and GRASSLAND being the only two variables entered (*i.e.*, RGRAZING, the other potentially significant FIO source, proved insufficiently significant to be included).

While these models inevitably have lower levels of explained variance than those including all potential predictors, it is notable that all the models generated using land cover variables conform to prior expectations, with both URBAN and GRASSLAND land cover types being significant.

Table 5. Summary of results of stepwise multiple regression models of relationship between $\log_{10}$ geometric mean faecal coliform and enterococci concentrations at base and high flow and those land use variables that represent likely enhanced sources of faecal indicator organisms—*i.e.*, grassland, rough grazing and urban/suburban (Table 3).

Step	Variable	Sign of b	Adjusted r^2	Sig level (p)
Base-flow models (*n* = 153)				
Faecal coliforms				
1	URBAN	+	0.339	
2	GRASSLAND	+	0.388	<0.001
Enterococci				
1	URBAN	+	0.294	
2	GRASSLAND[a]	+	0.301	<0.001

<center>**Table 5.** *Cont.*</center>

	Step	Variable	Sign of b	Adjusted r^2	Sig level (p)
High-flow models ($n = 134$)					
Faecal coliforms					
	1	GRASSLAND	+	0.316	
	2	URBAN	+	0.540	<0.001
Enterococci					
	1	URBAN	+	0.331	
	2	GRASSLAND	+	0.571	<0.001

[a] Only URBAN is entered with PIN = 0.05. In order to include GRASSLAND (the key agricultural FIO-source variable), PIN was relaxed to 0.12.

3.3. Population- (Human/Livestock-) Based Models

In this case the composite LSTOCKEC and TOTEC variables were excluded in order to remove the inevitable overlap with the individual population variables. The results, shown in Table 6, highlight the importance of HUMAN and DAIRY, which are entered at Steps 1 and 2 in all four models.

Table 6. Summary of results of stepwise multiple regression models of relationship between $\log_{10}$ geometric mean faecal coliform and enterococci concentrations at base and high flow and the population variables—*i.e.*, human, dairy, other cattle, all cattle, sheep (Table 3).

	Step	Variable	Sign of b	Adjusted r^2	Sig level (p)
Base-flow models ($n = 153$)					
Faecal coliforms					
	1	HUMAN	+	0.363	
	2	DAIRY	+	0.418	<0.001
Enterococci [a]					
	1	HUMAN	+	0.290	
	2	DAIRY	+	0.311	
	3	SHEEP	−?[a]	0.326	<0.001
High-flow models ($n = 134$)					
Faecal coliforms					
	1	DAIRY	+	0.439	
	2	HUMAN	+	0.595	
	3	SHEEP	+	0.622	<0.001
Enterococci					
	1	DAIRY	+	0.388	
	2	HUMAN	+	0.598	
	3	SHEEP	+	0.624	<0.001

[a] ? indicates that the sign does not conform with prior expectation.

HUMAN is entered first in the base-flow models, whereas DAIRY is the key variable at high flow. SHEEP is also entered at Step 3 in three of the models: with a +ve b value in the two high-flow models,

though with a counter-intuitive −ve value (for EN) at base flow. The levels of explained variance are notably higher in the high-flow models (maximum r^2, 0.624 for EN).

3.4. Inter-Study Transfer Errors

Transfer errors were investigated for the high-flow, population-based EN model, since this had the highest level of explained variance of the more parsimonious land cover- and population-based models.

Table 7. Inter-study transfer errors[a] in high-flow, population-based enterococci model.

Study catchment tested[b]		Mean error[c] ($\log_{10}$ cfu 100 mL^{-1})	Mean absolute error[d] ($\log_{10}$ cfu 100 mL^{-1})
1	Holland Brook	0.4975	0.4975
2	River Ribble	−0.1883	0.2985
5	River Leven/Crake	−0.0513	0.2215
9	River Irvine/Garnock	0.6227	0.6227
11	River Nairn	−0.2126	0.3059
12	Afon Ogwr	0.0417	0.2116
14	Afon Rheidol/Ystwyth	0.4609	0.4912
	Mean	0.1672	0.3784

[a] Determined by deriving a model with data for the tested study catchment omitted and using the resulting model to predict the geometric mean concentration for subcatchments in the omitted study;

[b] Only study catchments with ≥ 5 subcatchments with valid high-flow data were included, see Table 1;

[c] Mean of predicted–actual $\log_{10}$ EN concentrations for each of the subcatchments in the study catchment being tested;

[d] Mean of absolute difference between predicted and actual $\log_{10}$ EN concentrations for each of the subcatchments in the study catchment being tested.

The results in Table 7 reveal considerable inter-study variability that is not accounted for by the model. In fact, only the Leven/Crake and Ogwr studies have mean errors close to zero. In the case of Holland Brook, Irvine/Garnock and Rheidol/Ystwyth studies the models based on the other study catchments tend to overestimate the actual EN concentrations that were recorded (mean errors: 0.4975, 0.6227 and 0.4609 $\log_{10}$ cfu 100 mL^{-1}, respectively); whereas for the Ribble and Nairn studies the models tend to underestimate the actual concentrations (mean errors: −0.1883 and −0.2126 $\log_{10}$ cfu 100 mL^{-1}, respectively). The mean absolute error recorded is 0.3784 $\log_{10}$ cfu 100 mL^{-1}, with values ranging from 0.2116 (Ogwr) to 0.6227 (Irvine/Garnock) $\log_{10}$ cfu 100 mL^{-1}. The pattern in these results is closely reflected in the plot of predicted against actual high-flow EN concentrations based on the overall model, shown in Figure 2. Application of the model to the three sites in the Haverigg catchment produced a mean error of −0.1810 $\log_{10}$ cfu 100 mL^{-1} (range, −0.0513 to −0.2467 $\log_{10}$ cfu 100 mL^{-1}). It should be noted that inter-study transfer errors will tend to be greater where levels of explained variance in the models are lower, notably in the base-flow models.

Figure 2. Plot of actual high-flow $\log_{10}$ geometric mean enterococci concentration against predicted values using population-based model reported in Table 6, with values from those studies showing clear +ve or −ve anomalies from transferability testing (Table 7) identified.
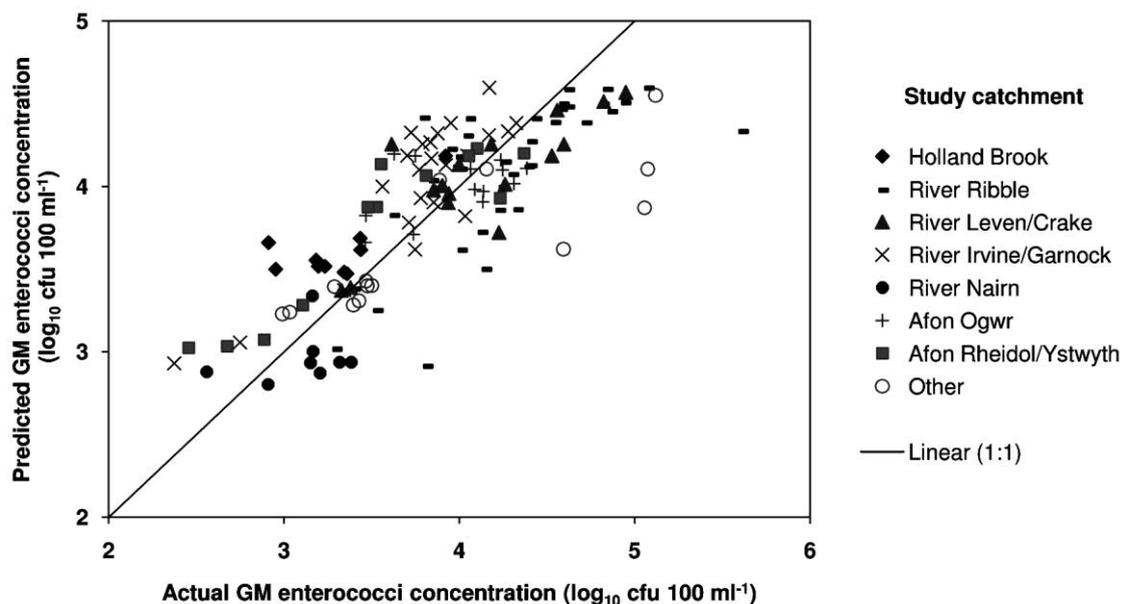
4. Discussion

4.1. Dominant Faecal Indicator Organism Sources within Catchments

The regression models clearly identify both humans and livestock as key FIO sources within catchments. It should be noted that some FIOs from both sources, especially particle-attached FIOs, may be deposited on the stream bed under base-flow conditions and re-suspended at times of high-flow. The FIO concentrations reported are therefore derived from both newly entrained and newly added organisms into the water column. Indeed, a significant proportion of the elevated concentration at high flow may well be from the stream bed.

Under base-flow conditions human sources (as reflected in the HUMAN and URBAN variables) are more important than livestock sources in accounting for the observed variance in FC and EN concentrations. Indeed, the DAIRY or GRASSLAND variables that are entered at Step 2 in the base-flow regression models provide only very limited additional explanation. This suggests that sewage-related sources are dominant at base flow, with relatively little FIO input from agricultural sources. The former will be largely treated effluents from WwTWs, which generally have much lower FIO concentrations under base-flow conditions than high flow [12]. The relatively low levels of explained variance in the base-flow models probably reflects the fact that in this 'black box' modelling, no account is taken of the nature of the effluent quality of individual WwTWs, which varies with the type of treatment [12]; and also that the URBAN and HUMAN data for individual subcatchments will poorly reflect the magnitude of sewage effluent inputs to the subcatchment watercourses in cases where WwTWs serving a significant proportion of the built-up area are located downstream of the monitoring point (*i.e.*, sewage is exported out of the subcatchment for treatment). It is also interesting to note that the HUMAN and URBAN variables provide very similar levels of explained variance—

suggesting that, for the purpose of catchment-scale modelling, built-up land is a good proxy for human population.

At high flow both human and livestock sources assume importance, with the latter generally being the more dominant. Under such conditions some untreated sewage from CSOs or overflows from WwTW storage tanks is likely to be discharged to watercourses, and the quality of treated effluents from many WwTWs will be reduced as a consequence of more rapid transmission through the plant [12]. The importance of human sources is evidenced by the inclusion of URBAN and HUMAN as key variables in the various high-flow models.

The general importance of livestock sources at high flow is reflected in the land cover-based models by the prominence of GRASSLAND, which is entered first for FC and makes a major contribution to the explained variance achieved for EN (Table 5). It should be noted that the GRASSLAND land use category comprises all temporary/permanent grassland, other than that which is mapped as rough grazing. As such it encompasses quite a wide range in terms of quality and productivity, extending from very fertile lowland pastures, which tend to be dominated by dairy farming, up to quite high altitudes in some subcatchments, where beef and sheep production systems tend to dominate. Because of this, GRASSLAND is simply a proxy variable for the more intensive areas of livestock production. Consequently, land cover data, as are traditionally used in FIO modelling, inevitably have limited explanatory power and potential for scenario modelling. By incorporating livestock density data, the present study provides insight into the relative significance of different production systems. Of the various livestock variables used in the modelling (Table 3), DAIRY emerges consistently as the key variable, with levels of explained variance that are consistently higher than GRASSLAND. In the case of the high-flow FC models, for example, the DAIRY has an r^2 value of 0.439, compared with 0.316 for GRASSLAND, which clearly highlights the importance dairy farming systems (*cf.* beef cattle and sheep) as a FIO source. This presumably reflects the high intensity of most dairy farming operations, which tend to be largely confined to the better land in the lowlands; the concentration of animals close to farm buildings for milking; and the storage and disposal to land of large quantities of waste (mostly in form of slurry) from yard areas and indoor winter housing–all of which pose potential pollution risks in terms of both diffuse sources (e.g., faeces voided directly in fields and slurry/manure applications to land) and point-source pollution (e.g., runoff from farmyards and milking parlours, slurry stores and manure heaps). By contrast, beef and sheep systems are not so confined to the better land, are often less intensive, and generate smaller amounts of waste for disposal. Sheep may, however, be present in quite large numbers in some catchments, both in areas of temporary/permanent grassland and rough grazing. They therefore represent a potentially significant FIO source, and this is reflected in SHEEP being entered at Step 3 with a +ve *b* value in both high-flow population-based models (Table 6). On the basis of these results, the design and implementation of measures to address FIO pollution from agricultural sources should be targeted initially on areas of dairy production.

4.2. Other Potential Factors Affecting Faecal Indicator Concentrations

Several "other" (*i.e.*, non-source) variables (Table 3) were included in the all-variable modelling. These relate to three catchment characteristics that may affect source strength and the mobilisation,

transport, die-off and sedimentation of FIOs within catchments, namely: runoff volume, soil hydrology and catchment size.

Volume of runoff during the study period may be an important factor since, during prolonged periods of wet weather, certain FIO sources (especially those associated with diffuse sources, such as animal faeces in fields and stream source contributory areas) will tend to become depleted. It might be anticipated, therefore, that a period of high flow will tend to be associated with higher FIO concentrations if preceded by a long spell of dry weather than if it followed a relatively wet period. Due to differences in weather conditions between the 6–8 weeks of each of the 14 catchment studies, there is very marked inter-subcatchment variability in runoff volumes (e.g., "TOTRUNOFF": range, 3.94–211.88 m^3 km^{-2} h^{-1}). In the case of soil hydrology, in subcatchments with more poorly drained soils (*i.e.*, with a higher mean SPR) there will likely be more surface runoff per unit rainfall and hence increased mobilisation and transport of FIOs from land to adjacent watercourses, which may well lead to increases in FIO concentrations. In the present study the SPR for both the subcatchments as a whole ("TOTSPR": range, 22.47–59.41%) and for the areas of permanent/temporary grassland ("GRSPR": range, 18.38–58.44%) were used as predictor variables. Catchment size may also be an important factor, since the opportunity for die-off of FIOs along watercourses as a result of exposure to UV light is increased within larger catchments as a result of the greater length of channel flow. This is particularly likely under base-flow conditions when flow velocities, water depth and turbidity are all at a minimum, thereby maximising UV exposure. The 153 subcatchments used in the modelling range in size ("AREA") from 5.01–1,013.18 km^2.

Despite the marked inter-subcatchment variability of runoff volume, soil hydrology and catchment area, only AREA was entered in any of the models, and that with a (counter-intuitive) +ve *b* value in the base-flow FC model (Table 4). Clearly, controlled experimental studies are needed to assess more fully the effects of these factors. On present evidence, however, it would seem that their role in affecting FIO concentrations in watercourses at the regional and national scales is minor compared with differences in human population density, stocking levels and associated land use types (URBAN and GRASSLAND)–*i.e.*, the factors that relate directly or indirectly to the key FIO sources.

4.3. Inter-Catchment Transferability of the Models

The out-of-sample testing reveals some degree of inter-study variability in the model evaluated, and this will inevitably tend to be greater in models with lower levels of explained variance, notably the base-flow models. This is not unexpected [29]; and is likely to be attributable to a combination of both inter-catchment and temporal factors. The former reflect systematic differences between the catchments affecting the sources, survival and mobility of FIOs that are not accounted for by the variables in the final regression models (*i.e.*, the unexplained variance). For example, there may be inter-catchment variations in livestock farming facilities and management practices that limit the extent to which key predictor variables such as GRASSLAND and DAIRY provide a measure of FIO sources. Also, soil hydrology (as outlined above) seems likely to account for some degree of inter-catchment variability, but its influence is not sufficiently strong to be included in the all-variable models; and other factors that were not included as potential predictor variables (e.g., temperature and topography) are likely to have a similar effect. The temporal factors, on the other hand, reflect the fact

that the individual studies were undertaken over 6–8 week monitoring periods with markedly contrasting weather conditions, both before and during the studies; and at different times during the bathing season, which could, for example, affect FIO source strength in grazed fields as a result of the progressive accumulation over the summer months of faeces from dairy cattle (which are housed over winter). Volume of runoff during the individual study periods, which was considered most likely to be the key temporal factor, was included in the predictor variable set, but (*cf.* hydrology) was not sufficiently significant to be entered in the all-variable models.

The strength of the present models lies in the fact that they are based on a FIO database that has such extensive geographical coverage (land use, climate, topography, soils, *etc.*) and encompasses a wide range of weather conditions during the individual monitoring periods. Some of the inter-study transfer errors are inevitably quite high, and these are partly attributable to temporal factors. Clearly, by combining the data from all 14 catchment studies the effects of the temporal factors are minimised and the inter-catchment errors reduced. The resulting land cover- and population-based models developed in the present study can therefore be applied with some confidence for predicting base- and high-flow GM FC and EN concentrations during the summer bathing season in UK watercourses with catchments areas between 5 and approximately 1,000 km^2. While the lower size threshold is determined by the level of resolution of the available agricultural census data, the upper limit simply reflects the size of the larger catchments used in the present modelling. Further investigations are needed to establish the validity of applying these models to much larger catchments. By combining these GM FIO concentrations with discharge data, then the contribution that the individual rivers/streams make to overall FIO loadings to coastal waters can be estimated.

The models can also be used to evaluate the likely impact of different land use/stocking level and human population change scenarios, as might result from the implementation of measures designed to reduce FIO loadings, or reforms in agricultural policy/funding, as reported in Hampson *et al.* [11].

5. Conclusions

In order to meet European WFD requirements there is an urgent need for generic (*i.e.*, transferable) models that can accurately predict base- and high-flow GM FIO concentrations in UK watercourses. Previous studies of individual catchments have successfully developed regression models based on relationships between GM FIO concentrations recorded at monitored sites and land use within their subcatchments. The present study extended this approach by augmenting the predictor variables to include direct measures of key FIO sources (*i.e.*, human population and livestock density data) and various factors (catchment size, runoff and soil hydrology) that may affect source strength and the mobilisation, transport and die-off of FIOs; and exploring the development of generic models by combining data from 14 different catchment studies across the UK.

Statistically significant base- and high-flow regression models have been developed for both FC and EN, with levels of explained variance consistently higher in latter models. Population variables (notably HUMAN and DAIRY) generally provide higher levels of explained variance than the land cover variables. Under base-flow conditions human, sewage-related, sources are dominant, whereas livestock sources tend to assume greater significance at high flow, with dairy farming systems (*cf.* beef cattle and sheep) being particularly important sources. Neither runoff, soil hydrology or catchment size

were significant predictor variables. In more parsimonious land cover- and population-based models, developed for ease of transferability to other UK catchments, relatively high levels of explained variance were achieved for all the high-flow models, with adjusted r^2 values ranging from 0.540 (land use model for FC) to 0.624 (population model for EN).

A programme of out-of-sampling testing on the high-flow EN model indicated some degree of inter-study variability, which is likely attributable to a combination of: (i) inter-catchment factors, which reflect systematic differences between the catchments that affect the sources, survival and mobility of FIOs that are not accounted for by the variables in the models; and (ii) temporal factors, which reflect the fact that the FIO monitoring was undertaken under different weather conditions and at different times during the summer bathing season. However, it is argued that by combining data from all 14 studies, which have a wide geographical distribution across UK and encompass a wide range of weather conditions, the effects of the temporal factors are minimised and the inter-catchment errors reduced.

Our research contributes to the emerging international debate on the use of farm best management practices and policy instruments to reduce FIOs and agricultural diffuse pollution (e.g., Bateman *et al.* [10], Chadwick *et al.* [30], Monaghan *et al.* [31], Helming and Reinhard [32], Hutchins *et al.* [33], Maringanti *et al.* [34] and Oliver *et al.* [35]) and the resulting land cover- and population-based models can be employed, with some confidence, in UK catchments both to predict base- and high-flow FC and EN concentrations in unmonitored watercourses and to evaluate the likely impact of different land use/stocking level and human population change scenarios.

Acknowledgments

Survey/EDINA supplied service. Copyright of the Land Cover Map 2000 is owned by the Natural Environment Research Council (NERC). All rights reserved. HOST soil data are based on digital spatial data licensed from the Centre for Ecology and Hydrology © NERC, © Cranfield University and © Macaulay Land Use Research Institute.

References

1. Council of the European Communities (CEC). Council directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for community action in the field of water policy. *Off. J. Eur. Comm.* **2000**, *L327*, 1-72.
2. Kay, D.; McDonald, A.T.; Stapleton, C.M.; Wyer, M.D.; Fewtrell, L. The challenges of the water framework directive: Proceedings of the institution of civil engineers. *Water Manage.* **2006**, *159*, 58-64.
3. Kay, D.; Crowther, J.; Stapleton, C.M.; Wyer, M.D.; Fewtrell, L.; Anthony, S.; Bradford, M.; Edwards, A.; Francis, C.A.; Hopkins, M.; *et al*. Faecal indicator organism concentrations and catchment export coefficients in the UK. *Water Res*. **2008**, *42*, 2649-2661.
4. Scottish and Northern Ireland Forum for Environmental Research. Provision of a screening tool to identify and characterise diffuse pollution pressures: Phase II, Project FWD19, 2006. Available online: http://www.sniffer.org.uk/ (accessed on 20 May 2009).
5. Scottish and Northern Ireland Forum for Environmental Research. Diffuse pollution screening tool: Stage 3, Project FWD77, 2006. Available online: http://www.sniffer.org.uk/ (accessed on 20 May 2009).
6. Crowther, J.; Wyer, M.D.; Bradford, M.; Kay, D.; Francis, C.A. Modelling faecal indicator concentrations in large rural catchments using land use and topographic data. *J. Appl. Microbiol.* **2003**, *94*, 962-973.
7. Kay, D.; Wyer, M.D.; Crowther, J.; Stapleton, C.; Bradford, M.; McDonald, A.T.; Greaves, J.; Francis, C.; Watkins, J. Predicting faecal indicator fluxes using digital land use data in the UK's sentinel water framework directive catchment: The Ribble study. *Water Res.* **2005**, *39*, 655-667.
8. Fezzi, C.; Rigby, D.; Bateman, I.J.; Hadley, D.; Posen, P. Estimating the range of economic impacts on farms of nutrient leaching reduction policies. *Agric. Econ.* **2008**, *39*, 197-205.
9. Jones, P.J.; Tranter, R.B. *Implications of a Nutrition Driven Food Policy for Land Use and the Rural Environment: Modelling the Impact of Different Policy Scenarios on farm Business Management*, *Land Use and Rural Employment*; Work Package No. 5.2; Interim Report to the Rural Economy and Land Use Programme; University of Reading: Reading, UK, August 2008.
10. Bateman, I.J.; Brouwer, R.; Davies, H.; Day, B.H.; Deflandre, A.; Falco, S.D.; Georgiou, S.; Hadley, D.; Hutchins, M.; Jones, A.P.; *et al*. Analysing the agricultural costs and non-market benefits of implementing the water framework directive. *J. Agric. Econ.* **2006**, *57*, 221-237.
11. Hampson, D.; Crowther, J.; Bateman, I.; Kay, D.; Posen, P.; Stapleton, C.; Wyer, M.; Fezzi, C.; Jones, P.; Tzanopoulos, J. Predicting microbial pollution concentrations in response to land use change. *Water Res.* **2010**, *44*, 4748-4759.

12. Kay, D.; Crowther, J.; Stapleton, C.M.; Wyer, M.D.; Fewtrell, L.; Edwards, A.; Francis, C.A.; McDonald, A.T.; Watkins, J.; Wilkinson, J. Faecal indicator organism concentrations in sewage and treated effluents. *Water Res.* **2008**, *42*, 442-454.

13. HMSO. *The Microbiology of Water. Part I: Drinking Water. Reports in Public Health and Medical Subjects*, *No. 71*; Methods for the Examination of Waters and Associated Materials; Her Majesty's Stationery Office: London, UK, 1994.

14. Environment Agency. *The Microbiology of Recreational and Environmental Waters* 2000; Methods for the Enumeration of Waters and Associated Material; Standing Committee of Analysts, The Environment Agency: Bristol, UK, 2000.

15. Kay, D.; McDonald, A. Reduction of coliform bacteria in two upland water supply reservoirs: The significance of distance decay relationships. *Water Res.* **1980**, *14*, 305-318.

16. Kay, D. *Factors Affecting Water Quality within Two Pennine Reservoirs under a System of Multiple-Use Catchment Management*; PhD Thesis; University of Leeds: Leeds, UK, 1979.

17. Natural Environment Research Council. 2000 Land cover map, 2008. Available online: http://www.ceh.ac.uk/sci_programmes/BioGeoChem/LandCoverMap2000.html (accessed on 5 October 2009).

18. Ordnance Survey. OS Meridian 2, 2008. Available online: http://edina.ac.uk/digimap/description/products/meridian.shtml (accessed on 5 October 2009).

19. ESRI Inc. ArcMap 9.1, 2005. Available online: http://www.esri.com/ (accessed on 17 October 2009).

20. Posen, P.E.; Hutchins, M.G.; Lovett, A.A.; Davies, H.N. Identifying the catchment size at which robust estimations of agricultural land use can be made, and implications for diffuse pollution modelling. *Appl. Geog.* **2011**, *31*, 919-929.

21. Office for National Statistics. 2001 Census: Digitised Boundary Data (England, Scotland and Wales). Available online: http://edina.ac.uk (accessed on 4 October 2009).

22. Office for National Statistics. 2001 Census: Standard Area Statistics (England, Scotland and Wales). Available online: http://edina.ac.uk (accessed on 28 September 2009).

23. EDINA. Agricultural census: Grid square agricultural census data for England, Scotland and Wales. Available online: http://edina.ac.uk (accessed on 9 July 2009).

24. Jones, F.; White, R.W. Health and amenity aspects of surface waters. *Water Pollut. Control* **1984**, *83*, 215-225.

25. Boorman, D.B.; Hollis, J.M.; Lilly, A. *Hydrology of Soil Types: A Hydrologically Based Classification of the Soils of the United Kingdom*; Report No. 126; Institute of Hydrology: Wallingford, CT, USA, 1995.

26. SPSS Inc. SPSS for Windows, Version 15.0, 2006. Available online: http://www.spss.com (accessed on 14 June 2007).

27. Rogerson, P.A. *Statistical Methods for Geography*; Sage: London, UK, 2001.

28. Tashman, L.J. Out-of-sample tests of forecasting accuracy: An analysis and review. *Int. J. Forecast.* **2000**, *16*, 437-450.

29. Bateman, I.J.; Brouwer, R.; Hime, S.; Provins, A. *Development of Guidelines for Use of Benefits Transfer in Policy and Project Appraisal*; Report to the Department for Environment, Food and Rural Affairs; Eftec: London, UK, 2009.

30. Chadwick, D.; Fish, R.; Oliver, D.M.; Heathwaite, L.; Hodgson, C.; Winter, M. 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.* **2008**, *19*, 240-247.

31. Monaghan, R.M.; de Klein, C.A.M.; Muirhead, R.W. Prioritisation of farm scale remediation efforts for reducing losses of nutrients and faecal indicator organisms to waterways: A case study of New Zealand dairy farming. *J. Environ. Manage*. **2008**, *87*, 609-622.

32. Helming, J.; Reinhard, S. Modelling the economic consequences of the EU water framework directive for dutch agriculture. *J. Environ. Manage*. **2009**, *91*, 114-123.

33. Hutchins, M.; Fezzi, C.; Bateman, I.; Posen, P.; Deflandre-Vlandas, A. Cost-effective mitigation of diffuse pollution: Setting criteria for river basin management at multiple locations. *Environ. Manage*. **2009**, *44*, 256-267.

34. Maringanti, C.; Chaubey, I.; Popp, J. Development of a multiobjective optimization tool for the selection and placement of best management practices for nonpoint source pollution control. *Water Resour. Res*. **2009**, *45*, 1-15.

35. Oliver, D.M.; Heathwaite, A.L.; Fish, R.D.; Chadwick, D.R.; Hodgson, C.J.; Winter, M.; Butler, A.J. Scale appropriate modelling of diffuse microbial pollution from agriculture. *Prog. Phys. Geog*. **2009**, *33*, 358-364.