Centre for Ecology & Hydrology
NATURAL ENVIRONMENT RESEARCH COUNCIL

River Invertebrate Classification Tool Database and Delivery System

Project Report

Author(s)

Kral Filip, Robertson Oliver, Fry Matt, Laizé Cédric

Client Ref: R15056PUR

Issue Number 1

Date 11/10/201611/10/2016

Title	River Invertebrate Classification Tool Database and Delivery System
Client	Scottish Environment Protection Agency
Client reference	R15056PUR
CEH reference	NEC05827 SEPA RICT Database / 1
CEH contact details	Dr Cédric Laizé

Executive summary

This project was commissioned by the Scottish Environment Protection Agency (SEPA) acting on behalf of a consortium of funders including SEPA, the Environment Agency (EA), and Natural Resources Wales (NRW), with additional technical support from the Northern Ireland Environment Agency (NIEA). It aims to scope and to produce a database of the model input variables for the current River Invertebrate Classification Tool (RICT), and a demonstration delivery tool allowing users to get these variables for any location in the UK.

The River Invertebrate Prediction and Classification System (RIVPACS) is a predictive model dating back to 1979. The main feature of RIVPACS is that it can predict the invertebrate species composition or value of invertebrate biotic indices at any site on any watercourse in UK. Wright (2000) describes history and development until RIVPACS III. RIVPACS IV has been integrated into the RICT application, which provides a web-based platform for users to run the model free of charge.

From a small set of input variables, RICT predicts invertebrate communities at reference conditions. Some of the input variables of the original model are themselves influenced by environmental conditions, causing issues when assessing certain pressure influences, so a new RIVPACS model was developed using pressure-independent variables. These new variables were derived for the model calibration sites only, but are not currently available at a national scale in the RICT software.

This project aimed to develop a database of the input variables required by the latest version of RICT and to propose a solution for delivery of these variables to RICT users. RIVPACS for Great Britain (GB) and for Northern Ireland (NI) are two different models but this project aimed to generate data for GB as well as for NI as far as possible.

The key output of this project is the set of variables calculated along UK rivers at 50m grid interval in the following units:

- **Logarithm of upstream catchment area (LOGAREA, dimensionless)**
- **Logarithm of upstream catchment mean altitude (LOGALTBAR, dimensionless)**
- **Proportion of time upstream catchment soils are wet (PROPWET, as a number between 0 and 1)**
- **Upstream catchment cover of key geological types (as a number between 0 and 1 indicating proportion of catchment area)**
- **Distance from source (m)**
- **Altitude (m A.S.L.)**
- **Slope (m/km)**
- **Discharge category (integer from 1 to 10 as defined in the project specification).**

The project was organised around a number of Work Packages (WP), grouped in two main topics:

- RICT variable database

  WP A Scoping
  WP B Licensing
  WP C Generating datasets

- Demo delivery tool

  WP D Assessing data delivery options
  WP E Constructing demonstration data delivery system

WP A, B, and C were interconnected. Rather than a dedicated scoping section for WP A, the report covers scoping considerations in their relevant sections and sub-sections. For WP B, licensing and Intellectual Property Rights (IPR) issues are covered in Section 6. For WP C, data requirements and variable derivation methods are covered in Sections 2 (data common to all variables) and 3 (data specific to variable, and derivation). For WP D, possible database options and the final database specifications are covered in Section 4. Finally, WP E specifications of the demo delivery tool are given in Section 5. In addition to this report, the RICT input variable datasets and the code for the demonstration delivery tool are provided as separate deliverables (see Section 4.2).

Contents

1. Introduction 7
1.1 RICT background 7
1.2 Objectives 8
1.3 Report structure and project deliverables 8
2. General data sources 9
3. Variables 10
3.1 Required variables 11
3.1.1 LOGAREA 11
3.1.2 LOGALTBAR 11
3.1.3 PROPWET 14
3.1.4 Upstream catchment cover of key geological types 15
3.2 Additional variables 19
3.2.1 Distance from source 19
3.2.2 Altitude 23
3.2.3 Slope 25
3.2.4 Discharge category 30
4. Database 38
4.1 Database formats 38
4.1.1 Comma Separated Value (CSV) 38
4.1.2 Esri Shapefile (Shapefile) 38
4.1.3 Esri File Geodatabase (FGDB) 39
4.1.4 GeoPackage and Spatialite 39
4.1.5 ASCII Grid 39
4.1.6 Tagged Image File Format (TIFF) 40
4.1.7 Erdas Imagine (IMG) 40
4.1.8 Selected option 40
4.2 Database structure and delivery format 40
5. Demonstration delivery system 43
5.1 Components and configuration 43
5.2 Future development 45
5.2.1 Group policies and protecting specific variables 45
5.2.2 Display river names and improve snapping verification 45
5.2.3 Batch mode 45
6. Intellectual Property Rights 45
6.1 Licensing Approach for RICT 46
6.2 Additional IPR Considerations 50
7. Conclusions 52
8. Acknowledgements 52
Appendix 1 Types of bedrock geology for River Invertebrate Classification Tool 53

Appendix 2 Detailed comparison of proportion of key geological types between calibration data and results of this project 57
Appendix 3 Slope calculated using different elevation data 62
Appendix 4 Adjustments to the drainage grid 67
Appendix 5 Computational regions used for parallel processing 71
9. References 72

1. Introduction

This project was commissioned by the Scottish Environment Protection Agency (SEPA) acting on behalf of a consortium of funders including SEPA, the Environment Agency (EA), and Natural Resources Wales (NRW), with additional technical support from the Northern Ireland Environment Agency (NIEA). It aims to scope and to produce a database of the model input variables for the current River Invertebrate Classification Tool (RICT), and a demonstration delivery tool allowing users to get these variables for any location in the UK.

1.1 RICT background

The River Invertebrate Prediction and Classification System (RIVPACS) is a predictive model dating back to 1979. The main feature of RIVPACS is that it can predict the invertebrate species composition or value of invertebrate biotic indices at any site on any watercourse in UK. Wright (2000) describes history and development until RIVPACS III. RIVPACS IV has been integrated into the RICT application, which provides a web-based platform for users to run the model free of charge (Davy-Bowker et al. 2008, CEH 2016, SEPA 2016).

The RICT user community is estimated at a few hundred that can be split into two main categories: (i) agency staff (about 50% of model runs), eg members of EA, SEPA, FBA, NRW, policy makers, policy controllers (WFD), assessors; (ii) third party users (about 50% of model runs), eg contractors, universities, consultants. There are an average of 300 RICT runs every two weeks. The user community can be also divided into infrequent (casual) and frequent (heavy) users. Infrequent users are typically investigating specific sites failing WFD requirements anywhere on the network (eg contractors, universities, water companies, river trusts). Frequent users are typically from statutory agencies doing scheduled runs for hundreds (eg SEPA) or thousands (eg EA) of sites every year (sites are not always the same but are subsets of a larger site network). Agencies use RICT for classification purposes and also to aid with WFD investigations. The RICT classification has legal standing.

From a small set of input variables, RICT predicts invertebrate communities at reference conditions (“expected”). This is typically combined with invertebrate data at observed conditions to derive scores used in WFD assessments (eg observed/expected ratios). Some of the input variables of the original model are themselves influenced by environmental conditions (‘pressure-influenced variables’), eg substrate, width, depth, or alkalinity. This caused serious issues when assessing certain pressure influences so a new RIVPACS model was developed with input variables that are not influenced by conditions, as part of project WFD119 (Clarke et al., 2011).

These new time-invariant variables (‘replacement variables’) were derived for the model calibration sites only. However, these replacement variables are not currently available at a national scale in the RICT software for two reasons: (i) they are not easy to generate (GIS software and skills, computing time); (ii) they are based on datasets that may have licensing restrictions.

1.2 Objectives

This project aimed to develop a database of the input variables required by the latest version of RICT and to propose a solution for delivery of these variables to RICT users. The database should include four replacement variables (related to catchment area, altitude, wetness, and geology), and, budget permitting, a set of four of the existing variables generated by this novel method (Distance from source, Altitude, Slope, Discharge category). RIVPACS for Great Britain (GB) and for Northern Ireland (NI) are two different models but this project aimed to generate data for GB as well as for NI as far as possible.

The project objectives were:

- Collate, compile, and evaluate data sources needed for calculation of RICT input variables
- Resolve IPR so that the input variables are entirely open data if possible, but at the least are open data for internal users (ie they can see the raw input variables), and so that the model is open to all users (ie they can run the model but not necessarily see the raw input variables)
- Develop and implement methods for calculating the RICT input variables and evaluate results against data available at RICT calibration sites
- Create a database of these input variables across the GB and NI rivers networks
- Develop a demonstration delivery system showing how RICT variables can be accessed by users, both for one site at a time and for batches of sites.

1.3 Report structure and project deliverables

The project was organised around a number of Work Packages (WP), grouped in two main topics:

- RICT variable database
    - WP A Scoping
    - WP B Licensing
    - WP C Generating datasets
- Demo delivery tool
    - WP D Assessing data delivery options
    - WP E Constructing demonstration data delivery system

WP A, B, and C were interconnected. Rather than a dedicated scoping section for WP A, the report covers scoping considerations in their relevant sections and sub-sections. For WP B, licensing and Intellectual Property Rights (IPR) issues are covered in Section 6. For WP C, data requirements and variable derivation methods are covered in Sections 2 (data common to all variables) and 3 (data specific to variable, and derivation). For WP D, possible database options and the final database specifications are covered in Section 4. Finally, WP E specifications of the demo delivery tool are given in Section 5.

In addition to this report, the RICT input variable datasets and the code for the demonstration delivery tool are provided as separate deliverables (see Section 4.2).

2. General data sources

This section describes the datasets used for calculation of several variables. Datasets that were used for a specific variable only are described in its related section below.

A key dataset underpinning calculation of all variables in this project is a flow direction model consistent with a river network and elevation dataset. The starting point for deriving a flow direction model is generally an elevation grid that has to be conditioned for hydrological analysis (filling in depressions, cutting in river lines, etc). While most of the steps can be automated, manual input and extensive checking is always needed to ensure the resulting flow direction is consistent with reality. Conditioning an elevation grid for hydrological analysis was beyond the scope of this project and it was decided that an existing flow direction model would be used.

The CEH Integrated Hydrological Terrain Model (IHDTM; Morris and Flavin, 1990) is a suitable data source. The IHDTM is a set of gridded datasets with 50m cell size, originally derived from 1:50K maps. The outflow drainage direction grid (OUTF) and Cumulative Catchment Area grid (CCAR) were used as input for most of the calculated variables. In addition to the gridded data, river lines from CEH 1:50K Digitised River Network (DRN), from CEH Intelligent River Network (IRN) GIS application, were used to identify river sources.

The elevation data that IHDTM was derived from is available on the Ordnance Survey Open Data website (OS OpenData). However, OS has additional elevation and river datasets, which we considered if there was a need to create a new flow direction model:

- OS Terrain 50; 50m elevation grid for Great Britain (GB), available as open data. (https://www.ordnancesurvey.co.uk/business-and-government/products/terrain-50.html)
- OS Terrain 5; 5m elevation grid or contour lines for GB, not available as open data; this is the most accurate source for GB of elevation data (https://www.ordnancesurvey.co.uk/business-and-government/products/os-terrain-5.html)
- OS Open Rivers; vector river lines available as open data for GB; this dataset contains selected major watercourses and is suitable for cartographic representations and high-level views rather than detailed hydrological modelling; OS Open Rivers is not as detailed as the IRN rivers (https://www.ordnancesurvey.co.uk/business-and-government/products/os-open-rivers.html)
- OS MasterMap Water Network Layer; vector river lines for GB; released in 2016 but not available as open data (https://www.ordnancesurvey.co.uk/business-and-government/products/os-mastermap-water-network.html)

We evaluated parts of these datasets and concluded that constructing and validating a flow direction model from them was beyond the scope of this project. Indeed, there were significant issues with OS Terrain 5 (differences in height between individual tiles) and the OS MasterMap Water Network Layer had incomplete information required to resolve bifurcations. Deriving a flow direction model from OS Terrain 50 and OS Open Rivers would have required more resources than available, while not improving on the existing flow direction model. We therefore decided to use the IHDTM for GB. Note: the project stakeholders identified several locations in Scotland where the IHDTM drainage direction should be adjusted; we have corrected the flow where possible (Appendix 3); flow directions in flat areas such as East Anglia should be treated with caution.

For Northern Ireland (NI), Ordnance Survey Northern Ireland (OSNI) provides 10m and 50m elevation grids as OpenData. However, both of these datasets cover only NI and do not include areas that flow into NI from the Republic of Ireland. Ordnance Survey Ireland (OSI) has a Height Data Product but it is not free (http://www.osi.ie/products/professional-mapping/height-data/). OSI also have OpenData datasets but they do not include any elevation data (http://www.osi.ie/about/open-data/). The most complete and least restricted elevation dataset we found for Ireland was the 25m Digital Elevation Model over Europe (EU-DEM; http://www.eea.europa.eu/data-and-maps/data/eu-dem). EU-DEM was used as an alternative source of elevation for several variables in this project. Before processing the EU-DEM, it was clipped, projected, and resampled to 50m cell size to fit the resolution and alignment of the drainage direction grid.

3. Variables

This project derived a database of four required replacement variables, and four additional existing variables covering GB and NI:

- Replacement variables
    - Logarithm of upstream catchment area (LOGAREA)
    - Upstream catchment mean altitude (LOGALTBAR)
    - Proportion of time upstream catchment soils are wet (PROPWET)
    - Upstream catchment cover of key geological types
- Existing variables
    - Distance from source
    - Altitude
    - Slope
    - Discharge category

These variables were calculated at 50 m grid resolution across the UK river network and results were compared against the data for the 722 calibration sites originally produced for the WFD119 project by using scatter plots and/or calculating differences. As a reminder, most of the WFD119 variables were derived using the IRN or extracted from the CEH Flood Estimation Handbook (FEH) catchment descriptor database. The following sections describe in detail the method selected for derivation of the new variables, and their comparison to the calibration site data.

Discrepancies between this project and the WFD 119 calibration site values are generally due to:

- Differences in underlying datasets used (e.g. geology)
- Differences in method where a mixture of methods (manual or automated) was used previously (e.g. slope)
- Snapping issues (site locations are slightly different).

In general, the new derivation methods and data sources work well. Given the non-trivial licensing constraints for this project, we believe we achieved the best possible variables that: (i) provide an accurate representation of the physical world; (ii) can be produced consistently across the entire UK; (iii) are spatially consistent with each other; (iv) are available for the wider community of users (see licensing in Section 6).

3.1 Replacement variables

3.1.1 LOGAREA

LOGAREA is the decimal logarithm of the upstream catchment area. The IHDTM already contained a Cumulative Catchment Area (CCAR) grid based on the number of cells upstream from each cell as defined by the IHDTM drainage direction grid. LOGAREA was derived as the decimal logarithm of CCAR. A perfect match with calibration site data was achieved (Figure 1).

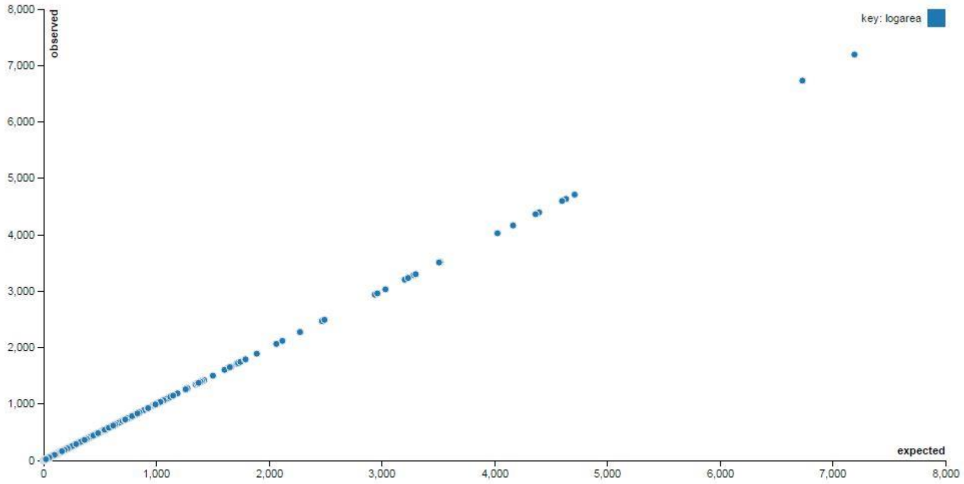


Figure 1 Logarithm of upstream catchment area (LOGAREA) at calibration sites (horizontal axis) and as extracted from FEH CCAR grid (vertical axis) in Great Britain.

3.1.2 LOGALTBAR

LOGALTBAR is the decimal logarithm of upstream catchment mean altitude. In WFD119, this was derived as the decimal logarithm of the FEH ALTBAR descriptor. Doing the same for this project would give a 100% match with the calibration sites (Figure 2), but FEH ALTBAR is not freely available. Alternative ways of calculating LOGALTBAR and alternative elevation data sources were therefore explored:

- IHDTM elevation grid (heights; HGHT); the same licensing limitations as per FEH ALTBAR apply (however, this dataset was used to quality-control the calculation methods and code)
- OS Landform-PANORAMA grid; only available for GB; while the IHDTM used elevation data from OS Landform-PANORAMA, a different interpolation method makes PANORAMA values different from the IHDTM therefore not subject to the same licensing restrictions
- OS Terrain 50 grid; only available for GB; this grid turned out to be significantly different (around 10 m differences were not unusual) from OS Landform-PANORAMA grid, but we were not able to find why
- EU-DEM grid; this dataset was used because it includes Ireland as a whole, thus covers all the areas needed for NI

The calculation always used the IHDTM OUTF drainage direction grid and an elevation grid. Sum of elevation values upstream from each cell was derived using the Flow Accumulation

Tool in ArcGIS Spatial Analyst. The sum was then divided by the total number of upstream cells to obtain mean altitude, the decimal logarithm of which gave LOGALTBAR.

Calculations based on IHDTM HGHT matched the calibration sites well but revealed several outliers which then appeared in results based on any of the remaining datasets. The best match for GB, based on scatter plots (Figures Figure 3 to Figure 6) and differences from calibration data (Table 1), was obtained with OS Landform-PANORAMA, which we selected to derive the final RICT variable.

In all cases, the same outliers were present (calibration sites 4003, 4309, HI10, 4701, SEPA_N47, 4703, 4705). Calibration values were lower than the new results. The two sites with the largest differences were 4309 (calibrated 385 m, OS Landform-PANORAMA 410 m) and 4003 (calibrated 540 m, OS Landform-PANORAMA 572 m). Results based on EU-DEM showed differences at additional sites (especially SEPA_N01, SEPA_N04, SEPA_N08, SEPA_N10), which are most likely caused by differences between EU-DEM and IHDTM HGHT.

For NI, we recommend using elevations from the EU-DEM since it is the only elevation dataset covering all required areas and fulfilling licensing requirements.

Table 1 Difference in upstream catchment mean altitude in metres between calibration data and different elevation data sources in Great Britain.

	IHDTM	EU-DEM	OS Landform-PANORAMA	OS Terrain 50
count	722	722	722	722
mean	0.15	0.91	-0.21	0.16
std	1.91	2.93	1.93	2.22
min	-2.61	-21.00	-5.68	-8.06
25%	-0.27	-0.11	-0.64	-0.43
50%	0.00	0.88	-0.35	0.04
75%	0.29	1.95	-0.05	0.51
max	32.92	33.53	32.48	34.15

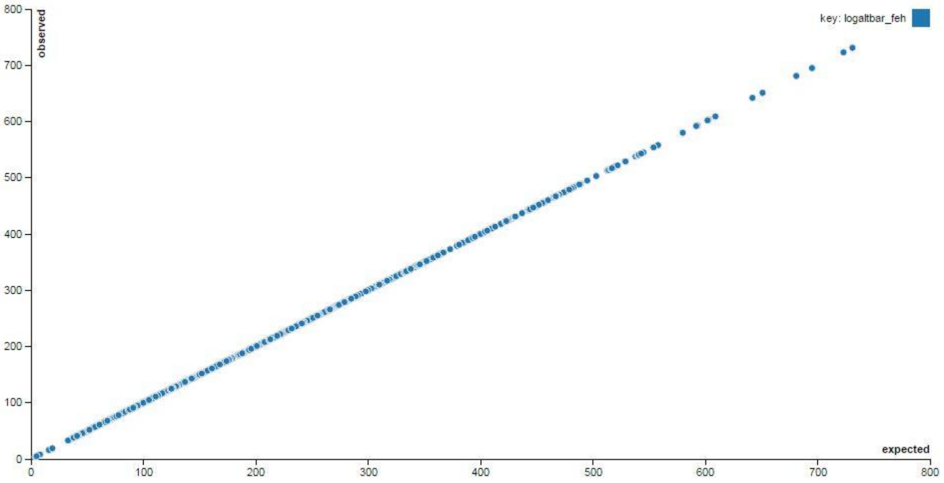


Figure 2 Upstream catchment mean altitude at calibration sites (horizontal axis) and as extracted from FEH ALTBAR grid (vertical axis) in Great Britain.

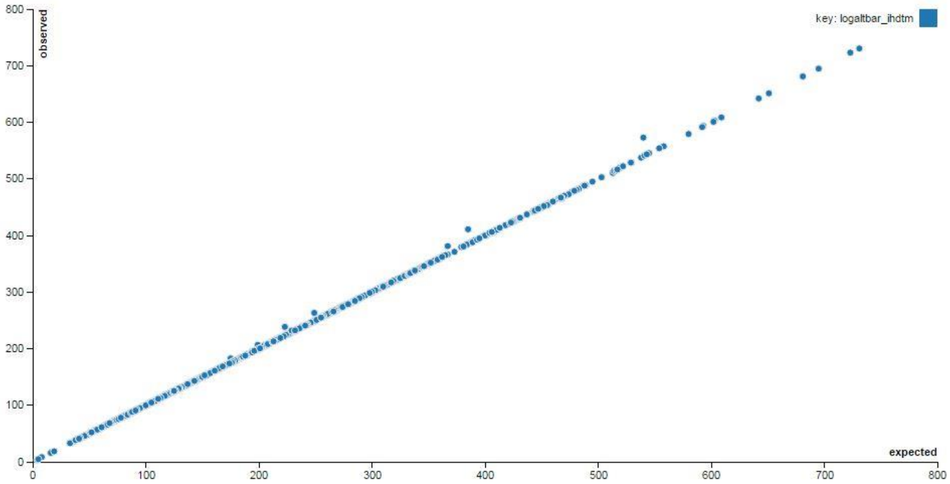


Figure 3 Upstream catchment mean altitude at calibration sites (horizontal axis) and as calculated based on accumulation of the IHDTM HGHT grid (vertical axis) in Great Britain.

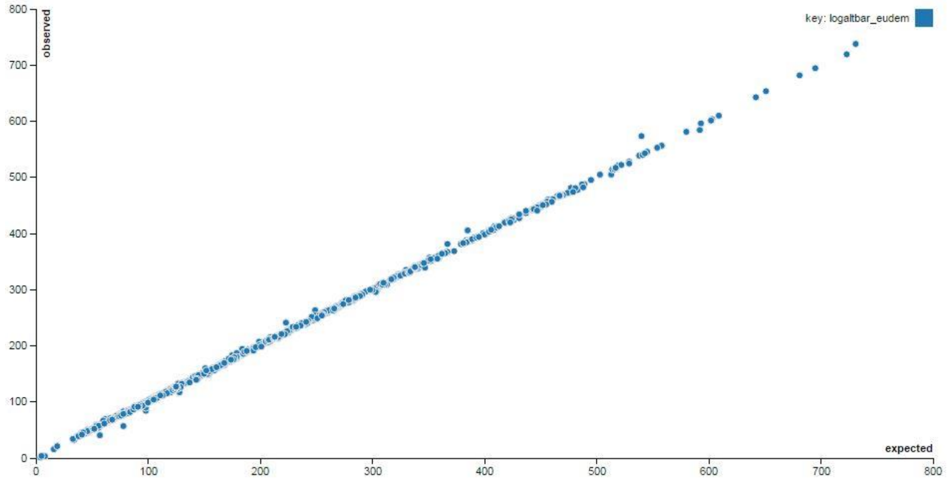


Figure 4 Upstream catchment mean altitude at calibration sites (horizontal axis) and as calculated based on accumulation of the EU-DEM grid (vertical axis) in Great Britain.

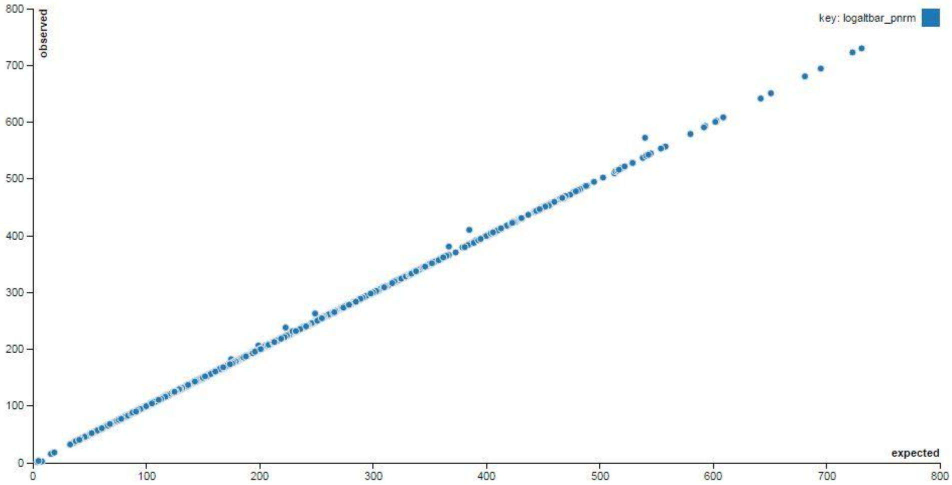


Figure 5 Upstream catchment mean altitude at calibration sites (horizontal axis) and as calculated based on accumulation of the OS Landform-PANORAMA grid (vertical axis) in Great Britain.

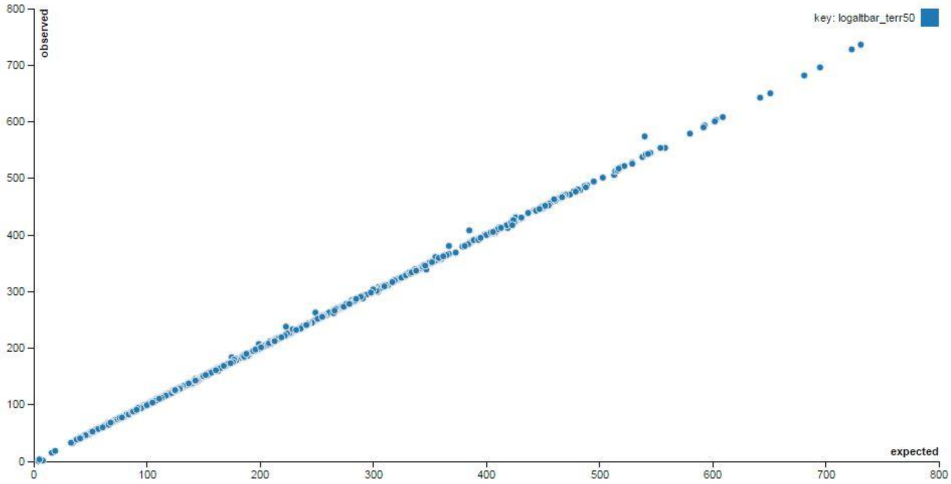


Figure 6 Upstream catchment mean altitude at calibration sites (horizontal axis) and as calculated based on accumulation of the OS Terrain 50 grid (vertical axis) in Great Britain.

3.1.3 PROPWET

PROPWET stands for ‘proportion of time upstream catchment soils are wet’. It is a Flood Estimation Handbook (FEH; Institute of Hydrology, 1999) catchment wetness index ranging between 0 (drier soils) and 1 (more saturated soils). The RICT PROPWET is a straight copy of the FEH PROPWET dataset.

An important point is that the FEH PROPWET exists only for catchments larger than 0.5 km^2. On the IHDTM, c. 5,027,000 cells can be identified as downstream from a source (ie using the 1:50K DRN) in GB, which leaves about 37% (c. 1,872,000) without PROPWET. In NI, nearly 25% of cells (approx. 85,000 out of 350,000) identified as cells downstream from a source are without PROPWET. It is worth noting that the IHDTM is generally considered to model catchments above 0.5 km^2 reasonably well, but that any catchment below 0.5 should be treated with caution. Generally, the proportion of cells without PROPWET is higher in mountainous regions and lower in flat areas (Figure 7). All the calibration sites had catchment area larger than 0.5 km^2 so they all have PROPWET value. After discussion with the project board, it was concluded that this may not be a major problem as the RICT model was not designed for such small catchments.

FEH PROPWET cannot be made freely available to all users, so PROPWET values will have to be send to RICT “behind the scenes” (eg the RICT system will have to request PROPWET values using an authenticated HTTPS request and users will not be able to see PROPWET values). An interim solution was discussed when members of organizations who licence FEH would be allowed to see the PROPWET values; see sections on IPR and demonstration delivery system below.

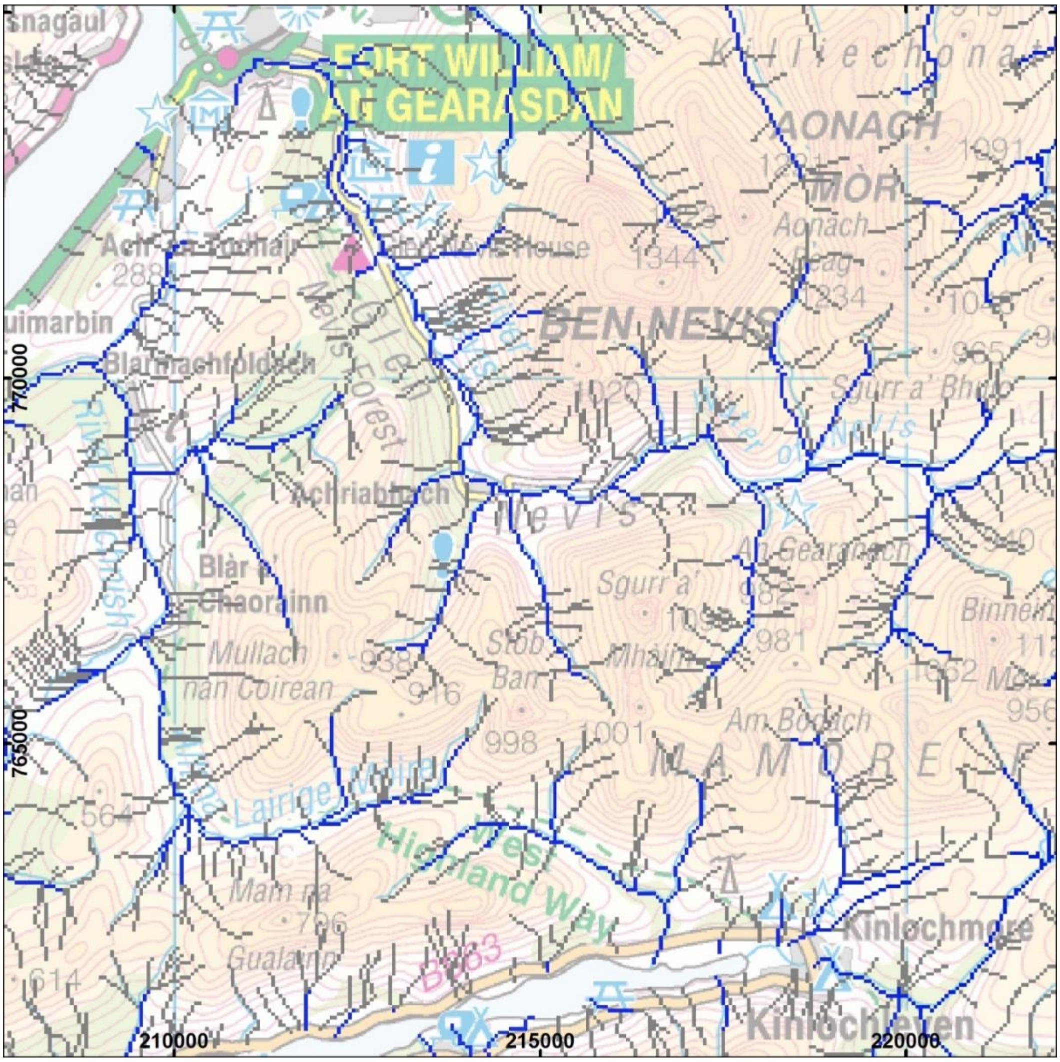


Figure 7 PROPWET is available only for catchments larger than 0.5 sq km. The proportion of cells where PROPWET is not available (grey) is higher in mountainous regions such as near Ben Nevis. Cells where PROPWET is available are in blue. Background is Ordnance Survey 1:250000 Raster.

3.1.4 Upstream catchment cover of key geological types

Upstream catchment cover of key geological types includes breakdowns of:

- bedrock geology: clay, chalk, limestone and hard rock bedrock (as defined in Clarke et al., 2011)
- superficial geology: peat.

The calibration data were derived from BGS 1:625K Geology Map version 4 (note: in version 4, bedrock geology was referred to as “Solid” and superficial geology was referred to as “Drift”). The latest version of the BGS 1:625K Geology Map is version 5. Clarke et al. (2011) define the RICT geology classes based on the MAP_CODE field, common to both versions except that values of version 5 MAP_CODE field are different from version 4. As a consequence, version 5 geology categories had to be re-assigned manually to the RICT classes. BGS developed a lookup table between MAP_CODE field versions 4 and 5. While this lookup table does not match all codes, we were able to find matching RICT geology class for each combination of ‘LEX’ and ‘RCS’ attributes in version 5 of the dataset. The lookup table from BGS also included a column with suggested RICT class, but in 45 cases the suggestion did not match the classification in WFD119 report. In such cases, the classification in the WFD119 report was preferred.

Visual comparison of the resulting maps revealed that the spatial distribution of RICT geological classes based on version 4 and version 5 agreed across most of the country but there were several notable differences (see Appendix 1):

- In version 4, most large lakes were classified as "no geology" but version 5 included geology “underneath” lakes.
- The new map showed outcrops of RICT sandstone in what was previously RICT clay (BGS map code v4 103) in Sussex and Kent and also on the fringes of some Chalk outcrops. Note that BGS suggested map code 103 to be classified as RICT clay, but the WFD119 report indicated sandstone.
- Hard rock outcrops in north of England and in Scotland were more common in version 4 than in version 5.
- What was unknown class near Isle of Wight in version 4 appeared as clay and sandstone in the new map.

BGS provides more detailed geological maps of Britain but none of these were open or free data.

BGS 1:625K Bedrock layer did not cover all the areas needed in Ireland so coverage of RICT geological classes for NI was compiled from multiple data sources:

- BGS 1:625K Bedrock Geology (version 5; used as a starting point since it already has RICT classes assigned based on version 4 as described above)
- GSI 1:500K Bedrock Geology (covers whole Ireland and closest to BGS 1:625K geology in terms of level of detail; http://www.dccae.gov.ie/natural-resources/en-ie/Geological-Survey-of-Ireland/Pages/Data-Downloads.aspx)
- GSI 1:100K Bedrock Geology (covers Republic of Ireland and parts of NI)
- GSNI 1:250K Bedrock Geology (covers NI only; https://www.opendatani.gov.uk/dataset/gsni-250k-geology).

Based on the inspection of the levels of detail and attributes of individual datasets, we decided to use overlap between GSI 1:500K Bedrock and BGS 1:625K Bedrock to manually transfer RICT classes from BGS 1:625K Bedrock onto the relevant GSI 1:500K Bedrock polygons, and use values from GSI 1:500K where BGS 1:625K was missing. Using the 1:250K and 1:100K layers would have required more effort to assign the right RICT class to individual polygons as there was no common attribute that would allow automatic classification. Lakes in GSI 1:500k Bedrock were manually filled-in based on surrounding geology (except for 'Lough Macnean Upper' where the boundaries of geological formations were too unclear).

To define peat coverage, BGS 1:625K Superficial Geology version 5 (where LEX='PEAT') was used for GB (polygons where drift MAPCODE=3 were used with version 4), while for NI, peat was selected from GSNI 1:250K Superficial Geology (where LEX='PEAT'). Peat in the parts of Republic of Ireland that flow into NI was taken from the ‘Soils Wet/Dry’ layer (where CATEGORY='Peat') published by the Irish Environmental Protection Agency (http://gis.epa.ie/GetData/Download). Note that using BGS 1:625K Superficial Geology in NI would leave a gap in western part of NI so the 1:250K map was preferred.

The key tool for calculating geology class catchment breakdowns was the Flow Accumulation Tool from ArcGIS Spatial Analyst. The tool can be used to count the number of cells upstream from any other cell in a specified flow direction grid. The tool has an optional parameter called weight raster. If the weight raster is specified, the result is not the number of cells upstream, but the sum of the values in the weight raster cells upstream. We converted individual types of geology into categorical raster grids where 1 indicated presence and 0 indicated absence of any given type. This categorical grid was then used as a weight raster with the Flow Accumulation Tool, and the resulting grid was divided by the normal flow accumulation grid.

We found several other packages with functionality similar to the Flow Accumulation Tool in ArcGIS Spatial Analyst. The GRASS GIS r.watershed module can accumulate weights values, but flow direction is always determined based on steepest descend of an elevation grid rather than based on a flow direction grid so we were not able to produce exactly the same results with GRASS GIS. Packages TauDEM and Python GeoProcessing should be able to accumulate weights based on a flow direction grid.

Comparison with calibration data was done in GB based on geology derived from version 4 and version 5 of BGS 1;625K geology datasets. There was generally a better match with version 4 than with version 5 (Figure 8 and Figure 9). Scatterplots of individual geology classes are in Appendix 2.

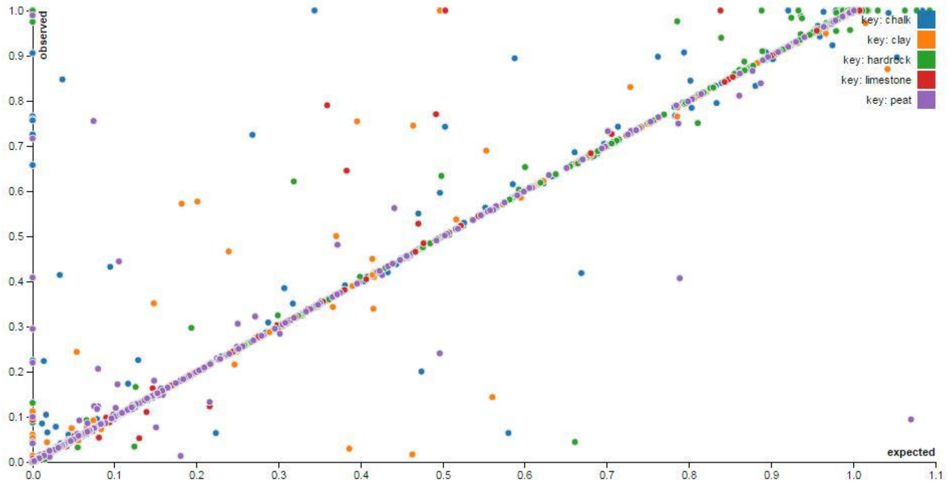


Figure 8 Proportion of key geological types based on version 4 of British Geological Survey 1:625000 Maps. Horizontal axis shows values used for calibration of RICT, vertical axis shows results calculated in this project. Data with calibration values outside the range are not shown.

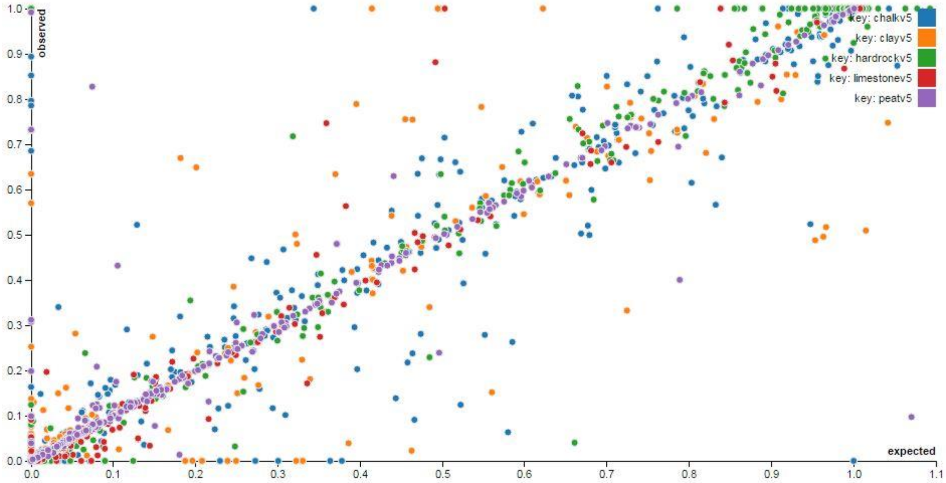


Figure 9 Proportion of key geological types based on version 5 of British Geological Survey 1:625000 Maps. Horizontal axis shows values used for calibration of RICT, vertical axis shows results calculated in this project. Data with calibration values outside the range are not shown.

Based on the scatter plots, there are only minor differences in proportion of peat between versions 4 and 5 (Table 2). Note that sites 2509 and SEPA_W16 have proportion of peat in the calibration dataset above 1.0. There are 22 sites where the difference between proportion of peat used for calibration and based on version 4 is greater than 0.05.

Table 2 Five sites with the largest difference between proportion of peat based on version 4 and proportion of peat based on Version 5 of British Geological Survey 1:625000 Superficial Geology Map. Sites where proportion of peat in calibration dataset was above 1.0 were not considered.

Rict ID	Peat for calibration	Peat based on version 4	Peat based on version 5
381	0.146	0.142	0.176
1603	0.074	0.755	0.827
2709	0.787	0.750	0.695
2903	0.083	0.084	0.117
4885	0.000	0.408	0.307
9205	0.441	0.562	0.630

Proportion of all bedrock geology classes changed considerably between version 4 and 5. The extra differences between calibration data and version 5 can be attributed to changes in distribution of RICT geology classes between the two versions. There are 89 sites where the difference between the proportion of at least one of bedrock geology categories based on version 4 and the corresponding calibration value is greater than 0.05. At 68 of these 89 sites, the calibration data indicated that catchment delineation was used directly from the FEH IHDTM, while the remaining 21 were derived manually, using an alternative catchment, or not available at all. From the 89 sites, 17 have also difference in proportion of peat larger than 0.05.

3.2 Existing variables

3.2.1 Distance from source

Distance from source (DFROMSRC) is for the purpose of this project the distance between the selected location and the source that is furthest upstream. Different ways of calculating distance from source were explored using vector (along river line geometries) and raster data (along FEH drainage direction grid). Calculating distance from source along river line geometries has been partially implemented but substantial effort would be needed to account for special cases near bifurcations and for situations when the river network is incomplete or incorrect.

Calculating distance from the furthest source along the flow direction grid was performed is the following steps. First, source points were established by selecting those start nodes of the CEH DRN layer which did not coincide with any end node. All cells downstream from any source were selected and converted to a raster containing 1 for cells downstream from any source and 0 otherwise. This raster was then used as a weight raster in the Flow Length Tool in ArcGIS Spatial Analyst (Esri, 2016) to calculate DFROMSRC for each cell. This method produces results consistent with the drainage direction grid but it works correctly only for river reaches that are represented in the raster data model. For example, some stretches of braided rivers end up with underestimated distance from source. Some inaccuracies also result from the simplification of the river network due to the flow direction grid cell size; based on the comparison with calibration data the differences caused by this simplification were considered acceptable (within around 1 km where DFROMSRC <50 km and within around 5 km where DFROMSRC >50 km). Reservoirs and other impoundments were treated as part of the watercourse as defined in the drainage direction grid. However, distance measured through lakes followed the path defined by the drainage direction grid, which is not necessarily the shortest (straight line) distance within the water body. Again, inaccuracies resulting from this simplification appear acceptable based on the calibration data. Comparison with calibration data in GB revealed several types of significant differences (Figure 10).

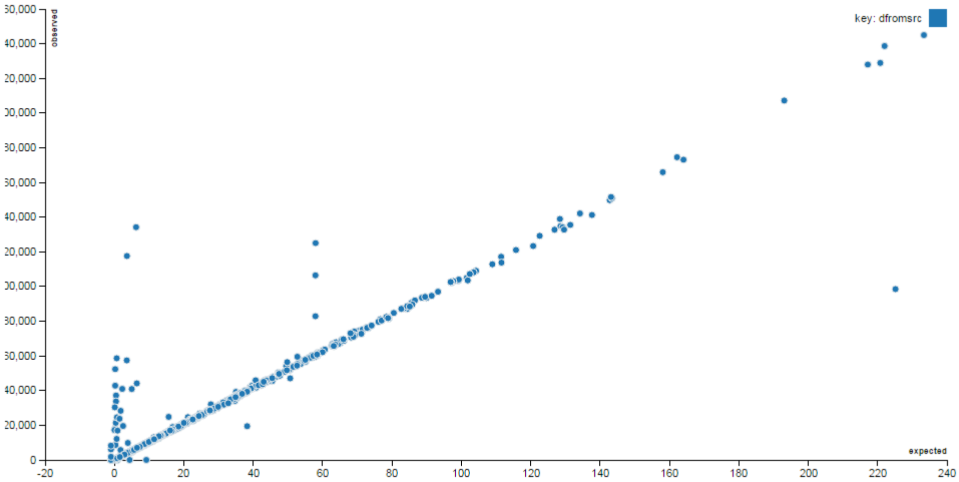


Figure 10 Distance from furthest source at calibration sites (horizontal axis, kilometres) and as calculated based on flow path length downstream from any source (vertical axis, metres) in Great Britain.

Sites 2509, 6261, and AN03 had no distance from source in the calibration dataset but according to the new results they were 6.1, 8.2, and 1.8 km downstream from the furthest source, respectively. Sites 5852, 6242, 6381, and 6844 had no distance from source either, but the new results gave 0 km. Sites 6111 (River Great Ouse) and AN05 (Forty Foot or Vermuden's Drain) both had calibration values much higher than what the new results indicated. These differences occur because the sites are downstream from bifurcations that cannot be fully represented in the flow direction grid (Figure 11). The same problem occurred at 5203 located on River Axe (Figure 12) and at 4885 (Figure 13). These differences point out a significant limitation of calculations based on the 8-directional drainage direction grid. The distance from source parameter will be affected by this limitation in every cell downstream from a bifurcation until flow path joins another dominant flow path.

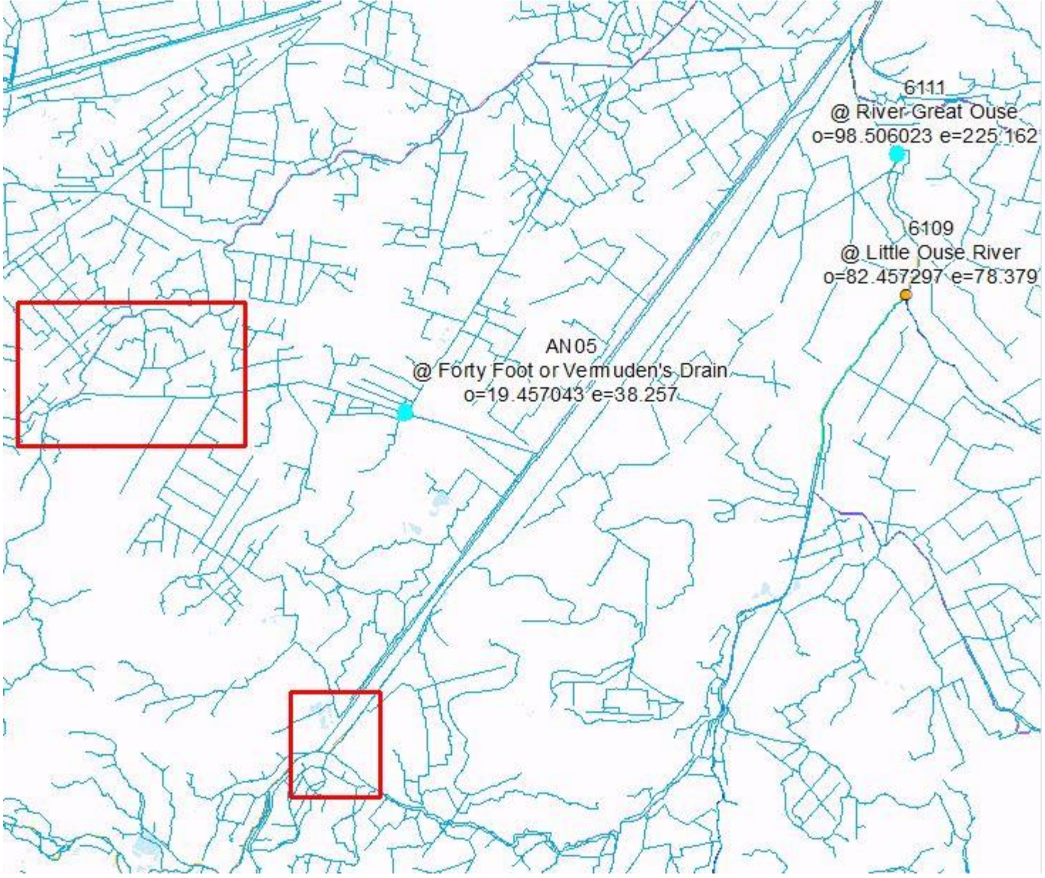


Figure 11 Two sites (turquoise) where distance from source in the calibration dataset was much higher than the new results. These differences occur because bifurcations in the red rectangles cannot be fully represented in the flow direction grid. Values in the labels are 'o' for the new results and 'e' for calibration data.

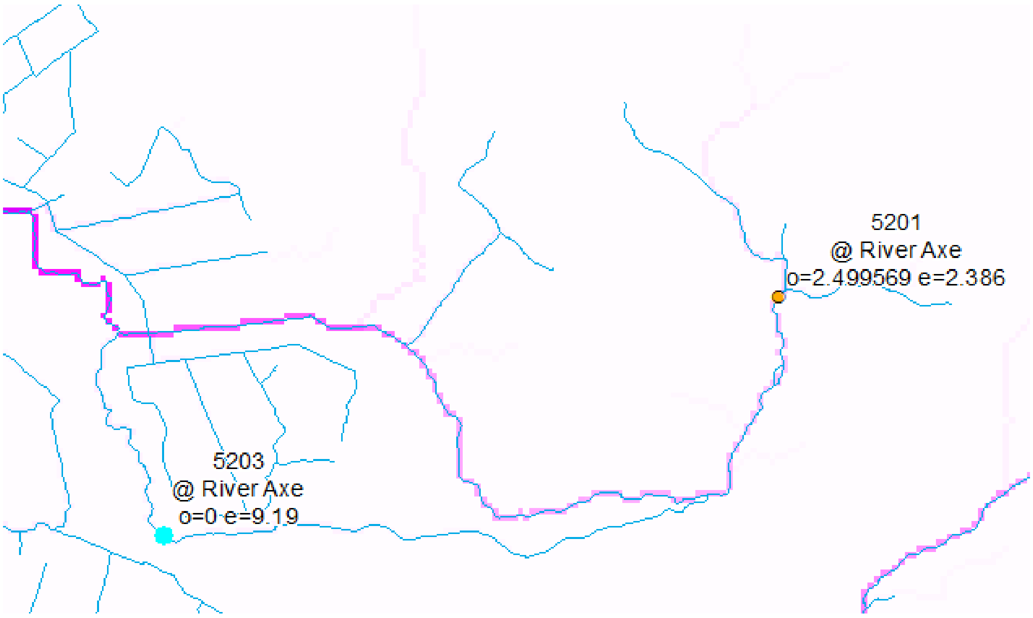


Figure 12 A site (turquoise) where distance from source in the calibration dataset was much higher than the new result. The difference occurs because a bifurcation upstream of the site cannot be fully represented in the flow direction grid. Values in the labels are 'o' for the new results and 'e' for calibration data.

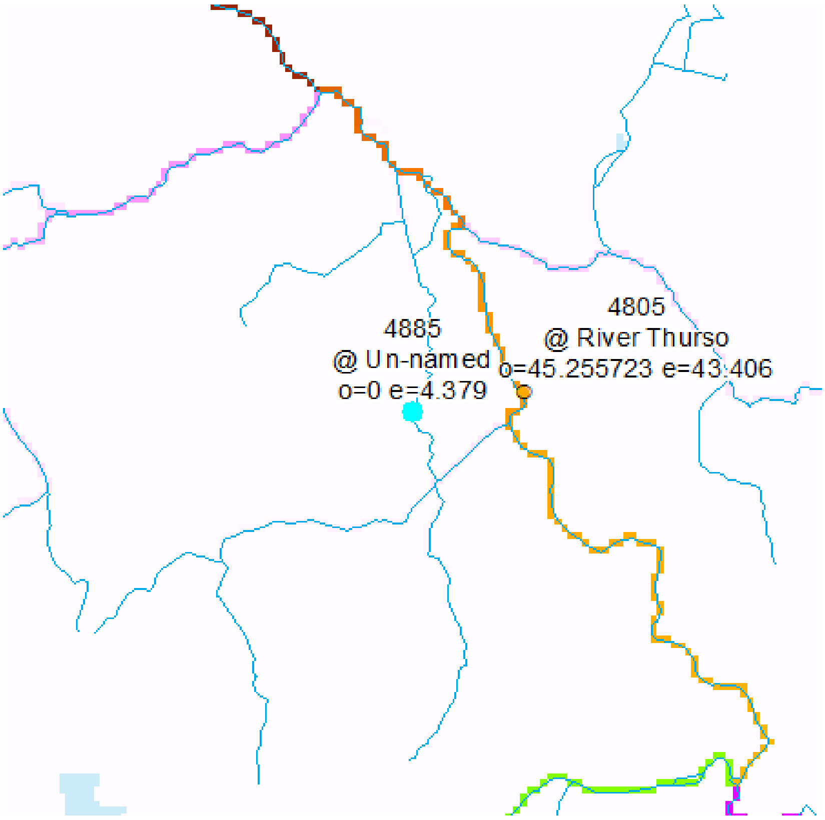


Figure 13 A site (turquoise) where distance from source in the calibration dataset was much higher than the new result. These differences occur because a bifurcation upstream of the site cannot be fully represented in the flow direction grid. Values in the labels are 'o' for the new results and 'e' for calibration data.

At 50 more sites, distance from source in the calibration dataset was more than 10% larger than the new results indicated. Many of these differences are also caused by disparity between vector data model (calibration distances derived from IRN) and raster data model (8-directional drainage direction grid). Examples are sites 0607 and 5107 (Figure 14 andFigure 15).

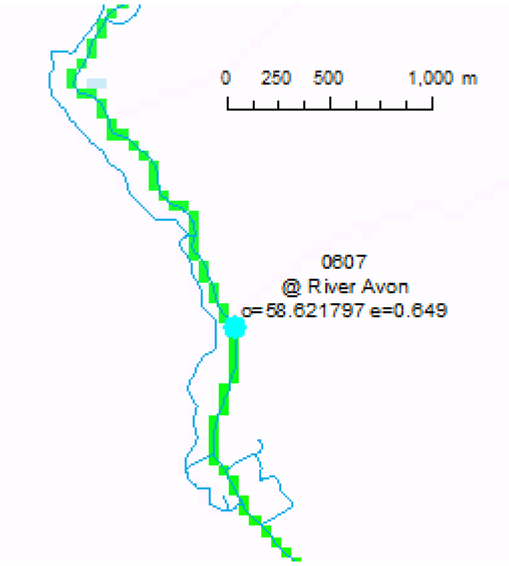


Figure 14 A site where distance from source based on the vector river network (blue) used in the calibration dataset is much lower than based on drainage direction grid.

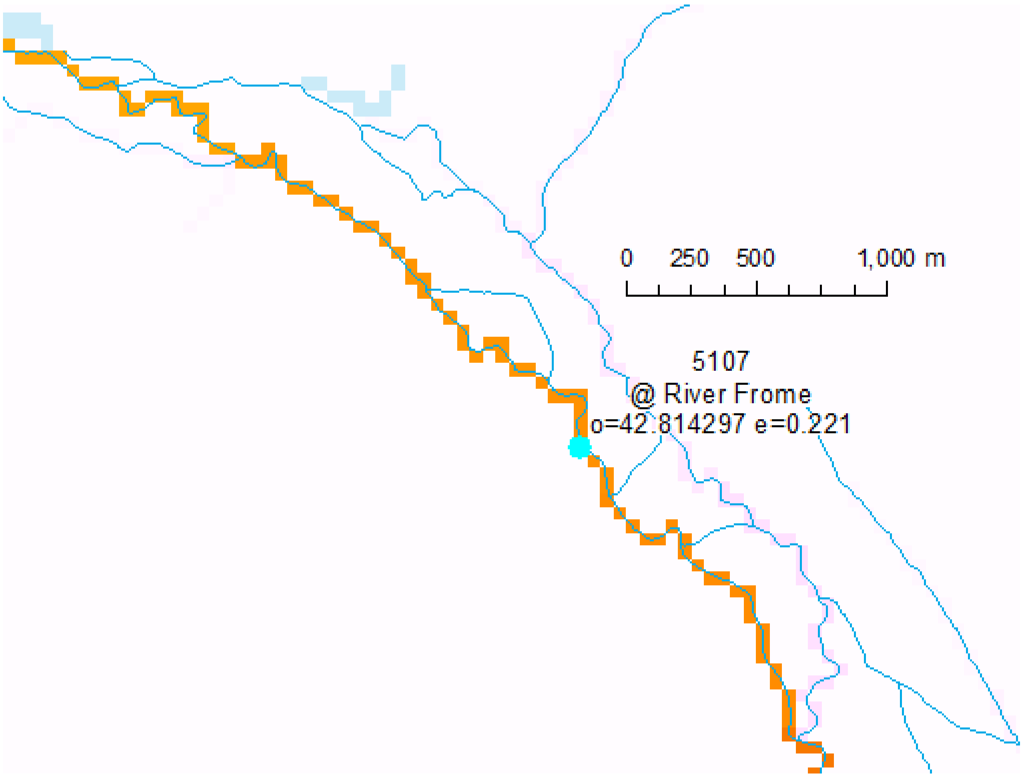


Figure 15 A site where distance from source based on the vector river network (blue) used in the calibration dataset is much lower than based on drainage direction grid.

Another common reason for WFD119 distances to be lower than the new results is that calibration sites were snapped to an IHDTM cell representing a different river (Figure 16).

Note that for all comparisons, we used the IHDTM coordinates included in the calibration dataset.

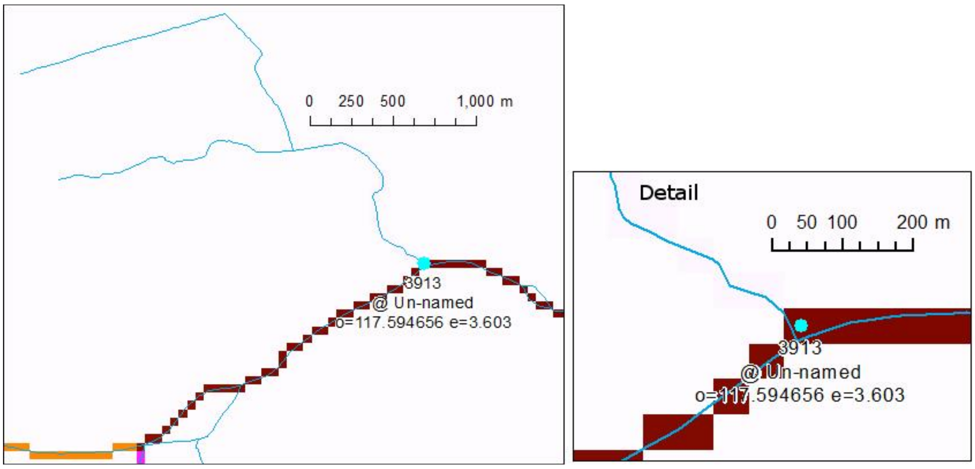


Figure 16 A site where calibration point was likely snapped to a DTM cell representing a different river which resulted in incorrect distance from source.

In the operational RICT data delivery tool, the cases where distance from source is lower than the new results should be filtered out or corrected during the snapping phase (eg manually or semi-automatically). Implications for the calibration of RICT are not clearly known.

3.2.2 Altitude

Altitudes for the WFD119 calibration data were derived with CEH IRN and based on the FEH HGHT grid. In order to make licensing of RICT variables as open as possible, we explored alternative sources of elevation (see Section 2). Altitude was delivered as a copy of this original data source. In GB, all elevation data sources matched well with calibration data. The variable that matched best, based on scatter plots (Figure 17 to Figure 20) and differences from calibration data (Table 3), was OS Landform-PANORAMA, which was retained for this project. With all elevation data sources, site HI04 did not match the calibration data. A manual check against the elevation data sources suggests that the value in the calibration dataset was incorrect. All remaining outliers are sites where calibration data had missing altitude. In NI, we recommend using elevations from the EU-DEM since that was the only elevation dataset covering all required areas that can be licensed to the project.

Table 3 Difference in altitude between calibration data and different elevation data sources in Great Britain.

	IHDTM HGHT	EU-DEM	OS Landform-PANORMA	OS Terrain 50
count	718	718	718	718
mean	-0.46	4.68	0.28	0.98
std	1.98	5.30	2.99	3.89
min	-44.40	-42.77	-49.00	-44.00
25%	-0.55	1.48	-0.90	-0.70
50%	-0.50	3.73	0.00	0.55
75%	-0.35	6.97	1.35	2.30
max	9.90	35.03	9.65	21.65

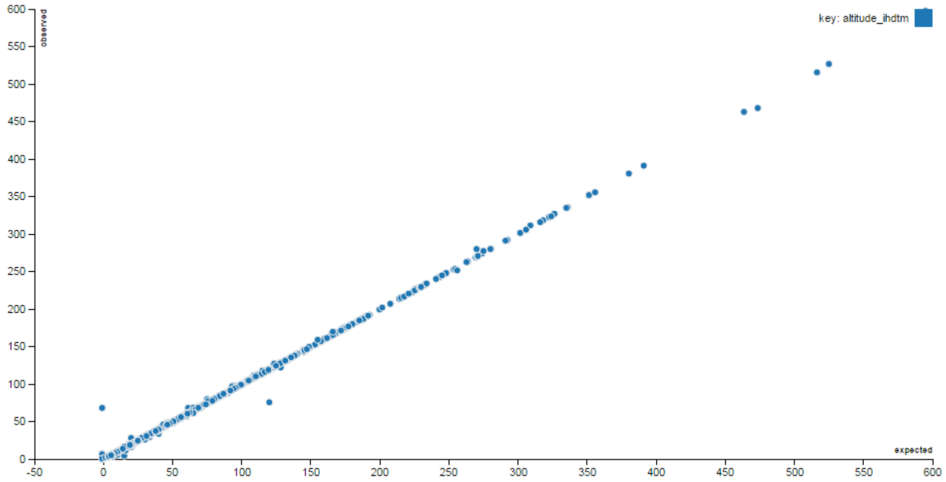


Figure 17 Altitude at calibration sites (horizontal axis) and altitude calculated from IHDTM HGHT grid (vertical axis) in Great Britain.

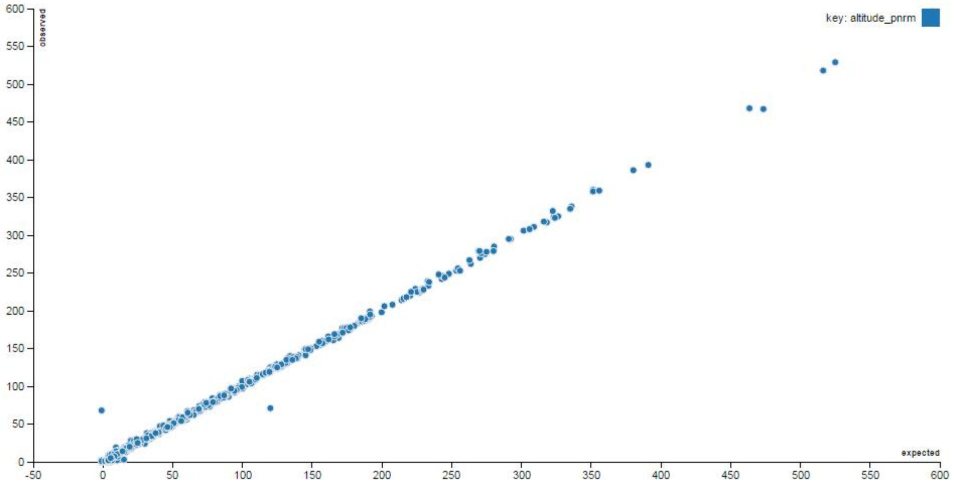


Figure 18 Altitude at calibration sites (horizontal axis) and altitude calculated from OS Landform-PANORAMA grid vertical axis) in Great Britain.

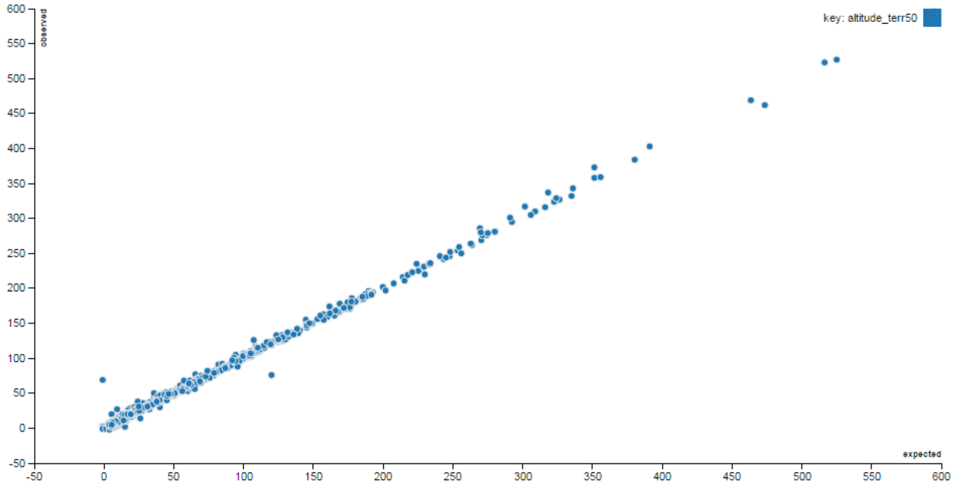


Figure 19 Altitude at calibration sites (horizontal axis) and altitude calculated from OS Terrain 50 grid vertical axis) in Great Britain.

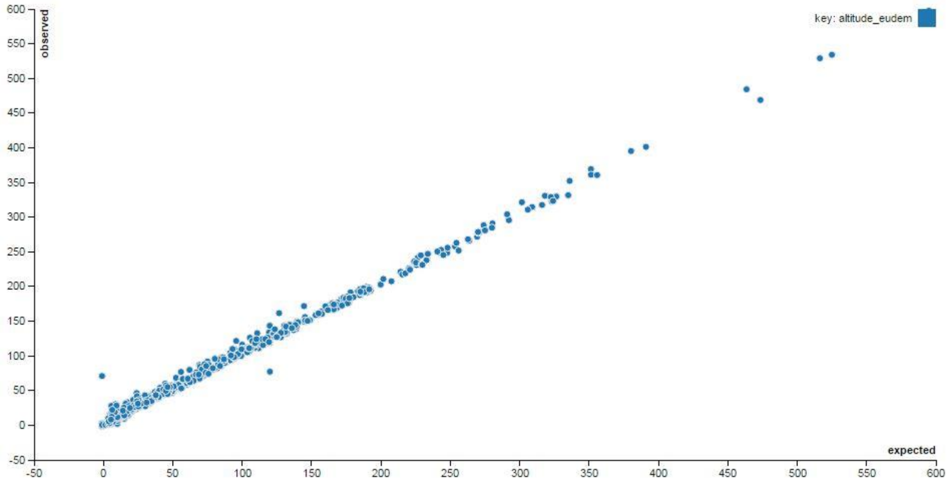
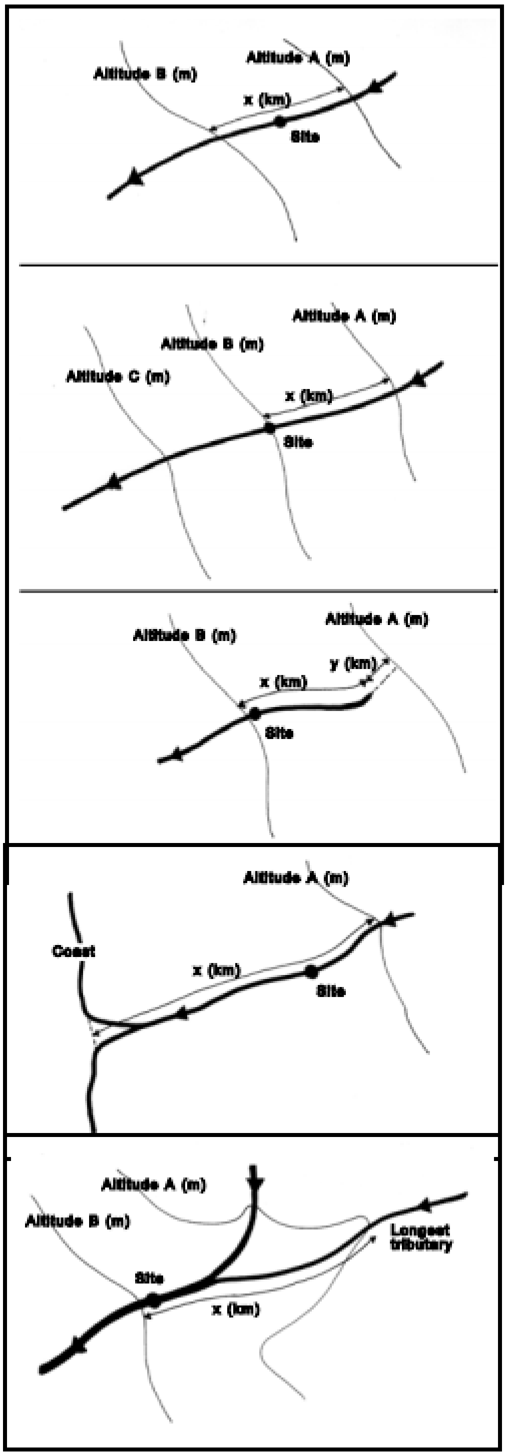


Figure 20 Altitude at calibration sites (horizontal axis) and altitude calculated from EU-DEM grid vertical axis) in Great Britain.

3.2.3 Slope

Slope at a site (m.km^{-1}) was originally manually derived from 1:50K OS maps. Slope was calculated as the height difference between the closest upstream and downstream contours divided by the horizontal distance between the two contours measured along the river. The RIVPACS Macro-invertebrate Sampling Protocol (http://eu-star.at/pdf/RivpacsMacroinvertebrateSamplingProtocol.pdf) defines how special cases should be resolved (Figure 21). In the WFD119 project, slope was derived using the IRN built-in method: height difference between points 500 m upstream and 500 m downstream divided by distance along the river; points within 500 m of lakes or sea have null slopes.



Site is between contours

$$\text{Slope} = \frac{A - B}{x}$$

More than one site may lie between the same contour lines. They would both have the same slope.

Site situated on a contour

$$\text{Slope} = \frac{A - B}{x}$$

The distance between contours is measured between the contour intersected and the next contour upstream. The slope upstream from a site is more likely to affect it than the slope downstream.

Upstream limit is the source

$$\text{Slope} = \frac{A - B}{x + y}$$

x is the distance between the site and the source
y is the shortest distance between the source and the next highest contour.

Downstream limit is the coast

$$\text{Slope} = \frac{A}{x}$$

The altitude at the coast is zero. Distance x is measured from contour A to the theoretical line that extends the natural line of the coast across the estuary.

Site is downstream from a tributary

$$\text{Slope} = \frac{A - B}{x}$$

x is measured along the longest tributary marked on the 1 : 50 000 scale map, even if that tributary has a different name or a smaller discharge.

Figure 21 Computing slope under variety of circumstances. This figure is an exact copy of the figure on page 42 of the RIVPACS Macro-invertebrate Sampling Protocol. Originally adopted from Furse et al. (1986).

With any method, information about lakes is important so that slope upstream from lakes can be calculated according to the rules in Figure 21. The IHDTM contains gridded representation of lakes but this is part of a layer which is not licensed to any of the partner organizations.

Alternative data sources were therefore used to represent lakes. In GB, lake shorelines were extracted from OS Landform-PANORAMA and converted to polygons. These were reviewed and polygons that represented broad river sections and that obviously did not match the IHDTM representation were removed. In NI, data were compiled from two data sources. In parts of the Republic of Ireland that drain into NI, polygons published by Environmental Protection Agency Ireland were used (Soils_IE_WetDry.CATEGORY='Water'). In NI, the Northern Ireland Lake Water Bodies dataset available under UK Open Government License published by Northern Ireland Environment Agency was used (https://www.opendatani.gov.uk/dataset/https-www-daera-ni-gov-uk-sites-default-files-publications-doe-lakewaterbodygml-zip). This dataset contains only lakes of size 50 hectares and above, so the resulting lake layer contains much fewer polygons than the original IHTDM.

Lake polygons were converted to a raster using the maximum of combined area rule. This rule produced the closest match to the representation of lakes in the IHDTM, but some cells near the lake shores did not match (Figure 22). The minimum slope that RICT accepts is 0.1 m.km-1, so this value was used to replace lower values and negative values.

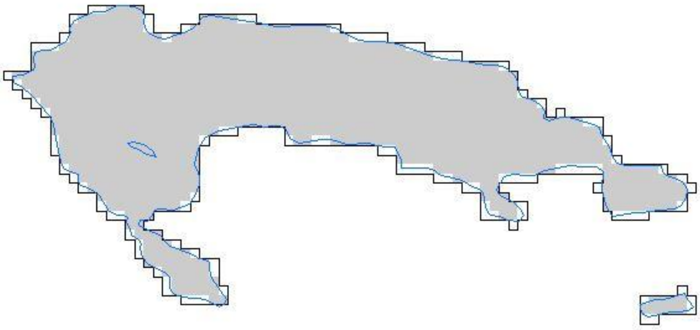


Figure 22 Illustration of differences in representation of lakes in IHDTM (black line), vector polygon from available open datasets (blue line) and rasterized representation of the vector polygon using the maximum combined area rule (grey area). This figure shows Loch Ard in Scotland.

We considered several methods for calculating slope, described in the following sub-sections.

3.2.3.1 Slope from contours

This was an attempt to mimic the original method based on OS maps. The main idea is to intersect river lines with contours, extract elevation at the start and end point of each segment, and divide the difference by the length of the segment. We partially implemented this method with IRN river lines, OS Landform-PANORAMA contours, and IHDTM HGHT grid. Trial results revealed several issues that stopped us from using this method:

- In many cases the river line followed and crossed a single contour several times. This was a problem especially in areas of lower slope and it was the main reason why this method was not developed further.
- River lines are often split into two features at a point between two contours which may result calculation of slope for very short segments.
- Incorporating lakes and coastlines into the calculation would be required.

- Incorporating the requirement that slope is calculated along the longest tributary (i.e. following the lines with highest distance from source) would be difficult to implement.
- Transfer from river lines to river channels derived by drainage direction would be problematic

3.2.3.2 Slope from river lines as average slope between contours

Add Surface Information Tool in ArcGIS 3D analyst toolbox can calculate average slope of a surface over a line feature. We split rivers at points where they intersected with OS Landform-PANORAMA contours and applied the Add Surface Information Tool. The results were very different from the calibration data. Furthermore, all the issues listed under 'Slope from contours' method apply also to this method.

3.2.3.3 Slope from river lines at vertices

In our GIS vector data model, rivers are approximated as line geometries. Each line geometry consists of a start node, an end node, and zero or more points in between – so called vertices. This method would calculate slope at every vertex of the river geometries. First, elevation of each river vertex would be interpolated from an elevation grid (e.g. using the Interpolate Shape Tool in ArcGIS Spatial Analyst toolbox). Then, distance from source for each vertex would have to be established. Finally, slope could be calculated at each vertex based on elevations at a vertex downstream and a vertex upstream, and the distance between them. The distances could either be set to fixed value (e.g. 500 m), or it could be established as the first distance where the difference in elevation is higher than a threshold (e.g. 10 m). The distance from source at each vertex has to be known so that points upstream can be selected at the longest tributary. Translating this verbal formulation of the method into computer code proved challenging and naïve implementations would not process the whole network in acceptable time. In addition to developing a computationally efficient implementation, the following issues would have to be resolved to make the method fully operational:

- Handling of bifurcations (our naïve implementation was too slow as it had to process each path from source to mouth 2^n times where n is the number of bifurcations on the path).
- Incorporating lakes and coastlines into the calculations.
- Handling of conflicts between flow direction and the direction of digitization and fixing any other errors in the river network.

3.2.3.4 Slope from shifting rasters

This method has the potential to calculate slope along flow direction for each cell (in our case each 50m by 50m cell). It relies on the drainage (out)flow direction, drainage inflow grid (encoding which neighbours flow into each cell), an elevation grid, and grid of distance from the furthest source. For each cell, the drainage (out)flow direction grid is used to obtain elevation at a next downstream cell. The inflow direction grid in combination with the distance from furthest source grid is used to obtain elevation at the upstream cell on the longest tributary. Distances are counted as cell size times $\sqrt{2}$ for diagonal moves and as cell size otherwise. By applying this process multiple times, slope can be calculated over ever longer distances. The distances will be different for different cells as it depends on the specific configuration of the input grids around each cell. This method was computationally very intensive and would require incorporating lakes and coastlines to make it fully operational.

3.2.3.5 Slope from flow segments

This method is similar to 'Slope from shifting rasters' in that it calculates slope at cell centres rather than at river line vertices, and that distance is based on direct distance between cells rather than distance between vertices. The difference is that this method uses much smaller set of cells. It focuses only on the cells that are downstream from river sources. This allowed an alternative implementation, which is much faster than processing all cells in the grid. The implementation relies on the NetworkX Python package (https://networkx.github.io/) for construction of a network of flow segments representing flow from one cell to another. Other Python packages such as arcpy and pandas used for other variables are also required. First, flow segments downstream from any source are converted to lines (see also Section **Error! Reference source not found.**) and a network of these lines is constructed using NetworkX. Elevation and distance from furthest source is established for each cell. For each cell, elevation of the upstream and downstream cells are established (taking into account lakes and the sea). Slope is calculated from the first encountered pair of cells where the difference in elevation is above a predefined threshold. Another predefined threshold determines maximum number of moves allowed in both upstream and downstream directions. If there is no pair with difference in elevation larger than the threshold, the pair of cells that have the longest distance between them and that are not a lake or sea is used.

3.2.3.6 Performance of retained method

The method 'slope from flow segments' performed best in terms of results and computationally. This method was fully implemented to derive slope at site, using several different elevation data sets: IHDTM HGHT, OS Landform-PANORAMA (GB only), and EU-DEM. The maximum number of moves was set to 10 so that points approximately 500 m downstream and 500 m upstream were used. The maximum difference in elevation was set to 50 m. The value of 50 m was selected because it resulted in the best match with the calibration data when different thresholds were used (10, 30, 50, 75, and 100 m). Plots comparing the calibration data to results obtained with all the different elevation datasets are in Appendix 3. From the freely available elevation data, slope calculated from OS-Landform PANORMA had the best match with calibration data and, from the available results, it is the best to use as RICT input for GB sites (Figure 23 and Figure 24). In NI, we recommend using the slope based on EU-DEM since that was the only elevation dataset covering all required areas that could be licensed.

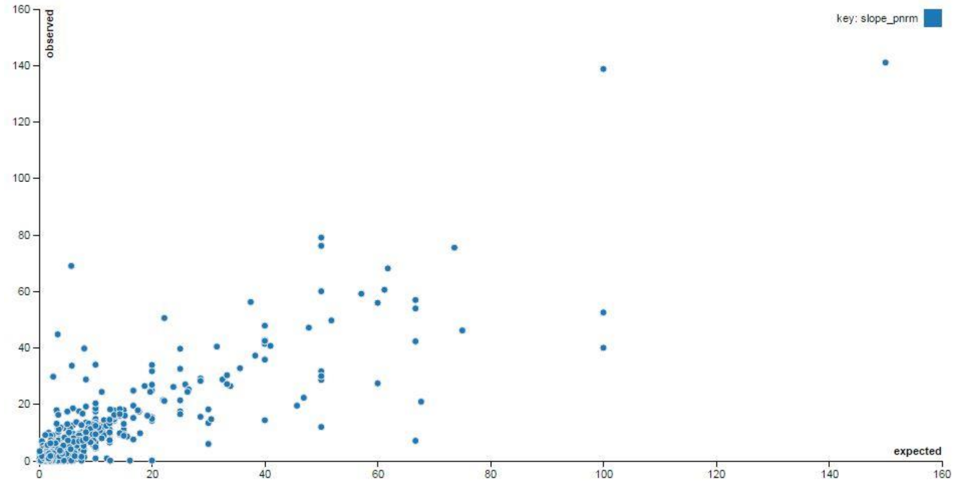


Figure 23 Slope at calibration sites (horizontal axis) and slope calculated along drainage direction grid segments (vertical axis) based on OS Landform-PANORAMA in Great Britain.

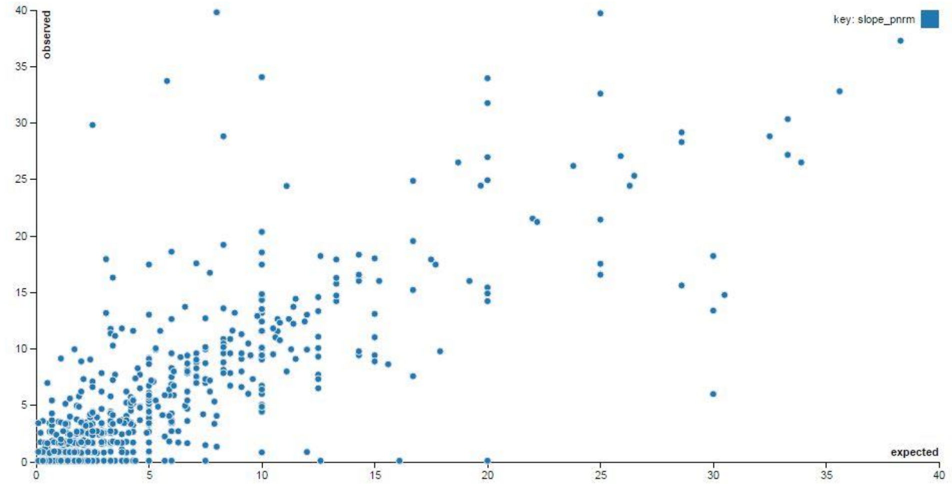


Figure 24 Slope at calibration sites (horizontal axis) and slope calculated along drainage direction grid segments (vertical axis) based on OS Landform-PANORAMA in Great Britain. This plot shows only sites where either slope was less than 40 m per km.

The differences between calibration data and new variables may be caused by multiple factors. An obvious factor is the simplification of vector river lines to flow segments between cells. However, based on visual inspection and results from other variables, we believe this simplification is acceptable. Another factor is that, in the new method, slope can be calculated over different distances at different points depending on configuration of the terrain. In flat areas, longer distance (up to around 1.4 km) will be used while in areas of high gradient the vertical threshold can be reached over a distance of a few hundred metres. This adaptive nature of the new method should be seen as an advantage since it is better suited to pick up localized changes in elevation.

3.2.4 Discharge category

Discharge Category (QMEANCAT) is based on naturalized mean annual discharge (m^3s^{-1}; Table 4).

Table 4 Discharge categories for RIVPACS

Discharge Category	Mean Annual Discharge (m^3s^{-1})
1	<0.31
2	0.31 – 0.62
3	0.62 – 1.25
4	1.25 – 2.50
5	2.50 – 5.00
6	5.00 – 10.00
7	10.00 – 20.00
8	20.00 – 40.00
9	40.00 – 80.00
10	>80.00

The RIVPACS Macro-invertebrate Sampling Protocol (http://eu-star.at/pdf/RivpacsMacroinvertebrateSamplingProtocol.pdf) suggests that discharge category was originally calculated using the Micro Low Flows System (MLFS). Lewis (1994) provides more details about the method of estimating flow at ungauged sites with MLFS. The method relied on average annual rainfall in the standard period 1961-90, potential evapotranspiration, and an adjustment factor representing the effect of soil moisture deficit in limiting evaporation (while not providing details on how to define this factor at a given location). Several alternative options have been investigated and we decided to capitalise on naturalized flows produced by the CEH Grid-to-Grid (G2G) Hydrological Model (Bell et al. 2009, Bell et al. 2016). G2G is a distributed model operating at 1 km cell size. Naturalized monthly flows for each 1 km cell were obtained for 1961-90 from the G2G team (one grid per month), and averaged as a single grid. The G2G cumulative catchment area grid was also provided. Therefore, our approach was to transfer the G2G average flow values from 1 km cells to the 50 m cells used in the IHDTM flow direction model, and fill-in gaps where cell match could not be done. Two approaches were developed, which are presented in the next two sub-sections. These were combined in the finalised method (third sub-section). Note that in NI, the modelled discharge, and therefore also the derived discharge categories, should be regarded as purely indicative because the G2G model has not been formally validated for NI. More work would be needed to verify that modelled discharge agrees well with discharge observed at gauging stations in NI as noted in 5.2.4 Data improvements. However, visual inspection of our results suggested that the discharge categories in NI are broadly consistent with what we expected.

3.2.4.1 Downscaling modelled flows using regression

Given a G2G mean annual discharge 1-km grid ($QMEAN_{G2G}$) and a G2G cumulative catchment area 1-km grid ($CCAR_{G2G}$), we established the following linear relationship:

$$QMEAN_{G2G} = a \cdot CCAR_{G2G}$$

(no intercept)

This relationship was used to estimate mean annual discharge for every 50-m cell based on the 50-m IHDTM cumulative catchment area grid. To account for regional variations in rainfall, evapotranspiration, and other factors affecting runoff, the regression parameter *a* was fitted individually for each IHU Group defined by Kral et al. (2015).

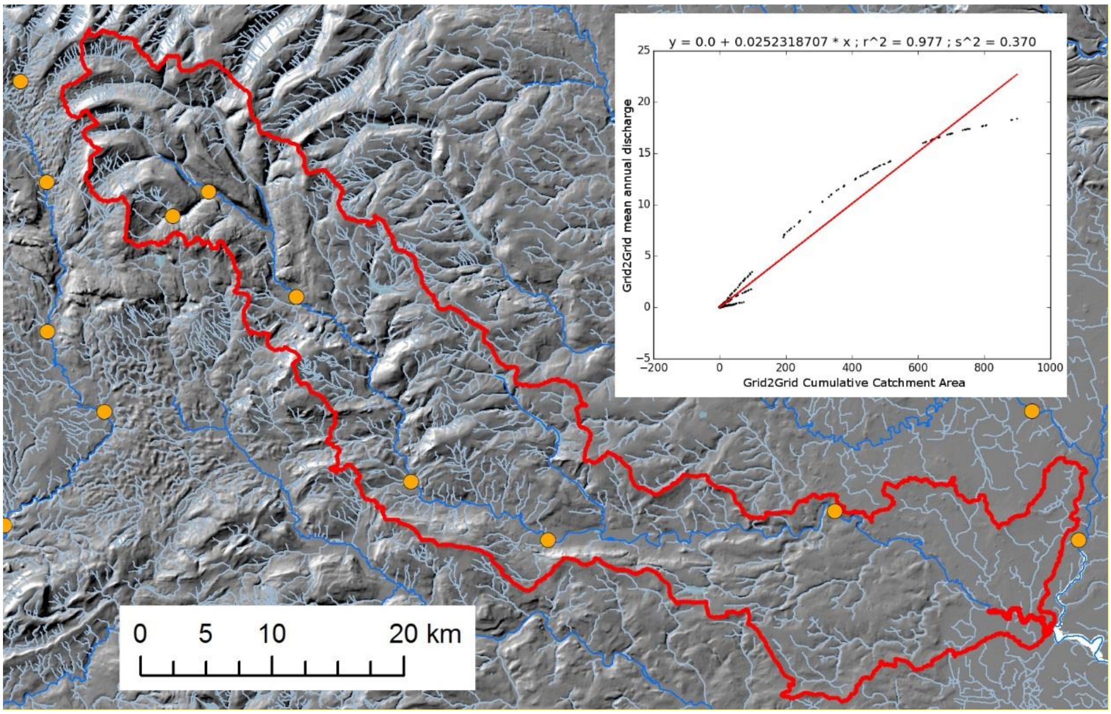

y = 0.0 + 0.0252318707 * x ; r^2 = 0.977 ; s^2 = 0.370
Grid2Grid mean annual discharge
Grid2Grid Cumulative Catchment Area
25
20
15
10
5
0
−5
−200
0
200
400
600
800
1000
0
5
10
20 km

Figure 25 Integrated Hydrological Units Group HA27G10 Wharfe (Source to Sea) and correlation between Grid2Grid Cumulative Catchment Area and Grid2Grid mean annual discharge within this group. Orange points show RICT calibration sites, rivers are CEH Intelligent River Network, and hillshade is based on Ordnance Survey Open Data.

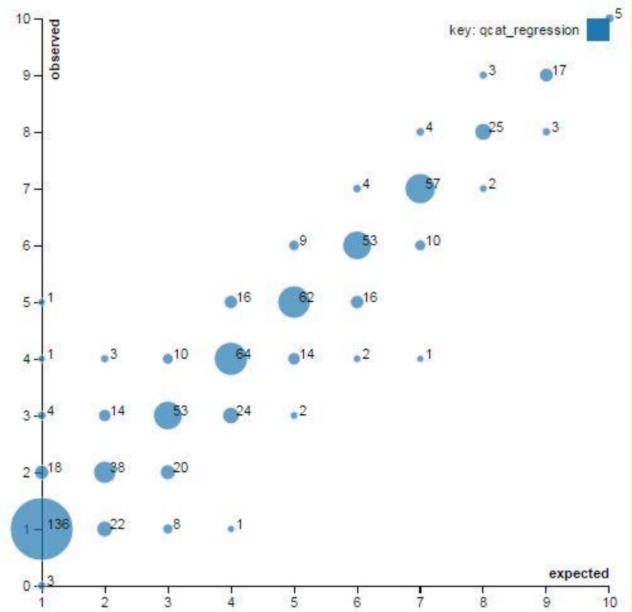


Figure 26 Discharge category at calibration sites (horizontal axis) and discharge category obtained using regression (vertical axis). It was regression of mean annual discharge with cumulative catchment area within Integrated Hydrological Units - Groups. Labels and sizes of points indicate number of calibration sites in each combination. For 3 sites discharge category using regression could not be established (observed=0).

3.2.4.2 Flow interpolation

Given a G2G mean annual discharge 1-km grid ($\mathrm{QMEAN_{G2G}}$) and a G2G cumulative catchment area 1-km grid ($\mathrm{CCAR_{G2G}}$), this method interpolates $\mathrm{QMEAN_{G2G}}$ along flow segments defined by the drainage flow direction grid downstream from any source.

The first step was to find the corresponding 50-m cell within each 1-km square. The corresponding 50-m cell is the cell with the largest CCAR where the difference from $\mathrm{CCAR_{G2G}}$ is small (specifically, cells with $[(CCAR_{G2G} - CCAR) \,/\, CCAR] < 0.05$ were accepted).

Then, continuous flow lines were established by climbing from each mouth upstream and following the path with the largest CCAR. When all mouths had been used, the same process was applied on the points that were not labelled, and so on until all points were labelled with a flow line identifier.

Finally, within each flow line with at least two $\mathrm{QMEAN_{G2G}}$ values, the QMEAN value at a given cell (‘*this*’) is calculated according to the following formula:

$$QMEAN^{this} = QMEAN^{US} + \left(\frac{QMEAN_{G2G}^{DS} - QMEAN_{G2G}^{US}}{CCAR_{G2G}^{DS} - CCAR_{G2G}^{US}}\right) \cdot (CCAR^{this} - CCAR^{US})$$

Where *this* is the current cell, *US* the upstream cell, *DS* the downstream one, and *G2G* is the QMEAN from G2.

At each source of a flow line, $\mathrm{QMEAN_{G2G}}$ is assumed to be zero. At each end of a flow line, QMEAN is assumed to follow the gradient based on the previous available pair of values. Note that simply performing linear interpolation along flow lines would mean that discharge in a small tributary would be affected by the flows of the river it flows into. This method is trying to mitigate that by first defining individual flow lines. However, within one flow line, flows above confluences may be influenced by the flows of the tributary downstream. For example, in Figure 27d, discharge at point Y will be affected by the river branch C because the branch contributes to the discharge at point A_5. One possible way to improve this would be to establish discharge at each confluence from all available input points, to take the largest discharge at each confluence, and only then to perform the interpolation. In practice, there were rarely more than one input point between any two confluences so it would not be possible to calculate the discharge at confluences before the interpolation. While we believe the interpolation technique is more accurate than the regression-based downscaling, the drawback of this approach is that, for many river branches, the gradient of discharge over cumulative catchment area cannot be established because there are fewer than two points on that branch (e.g. branches D, E, F, G in Figure 27b).

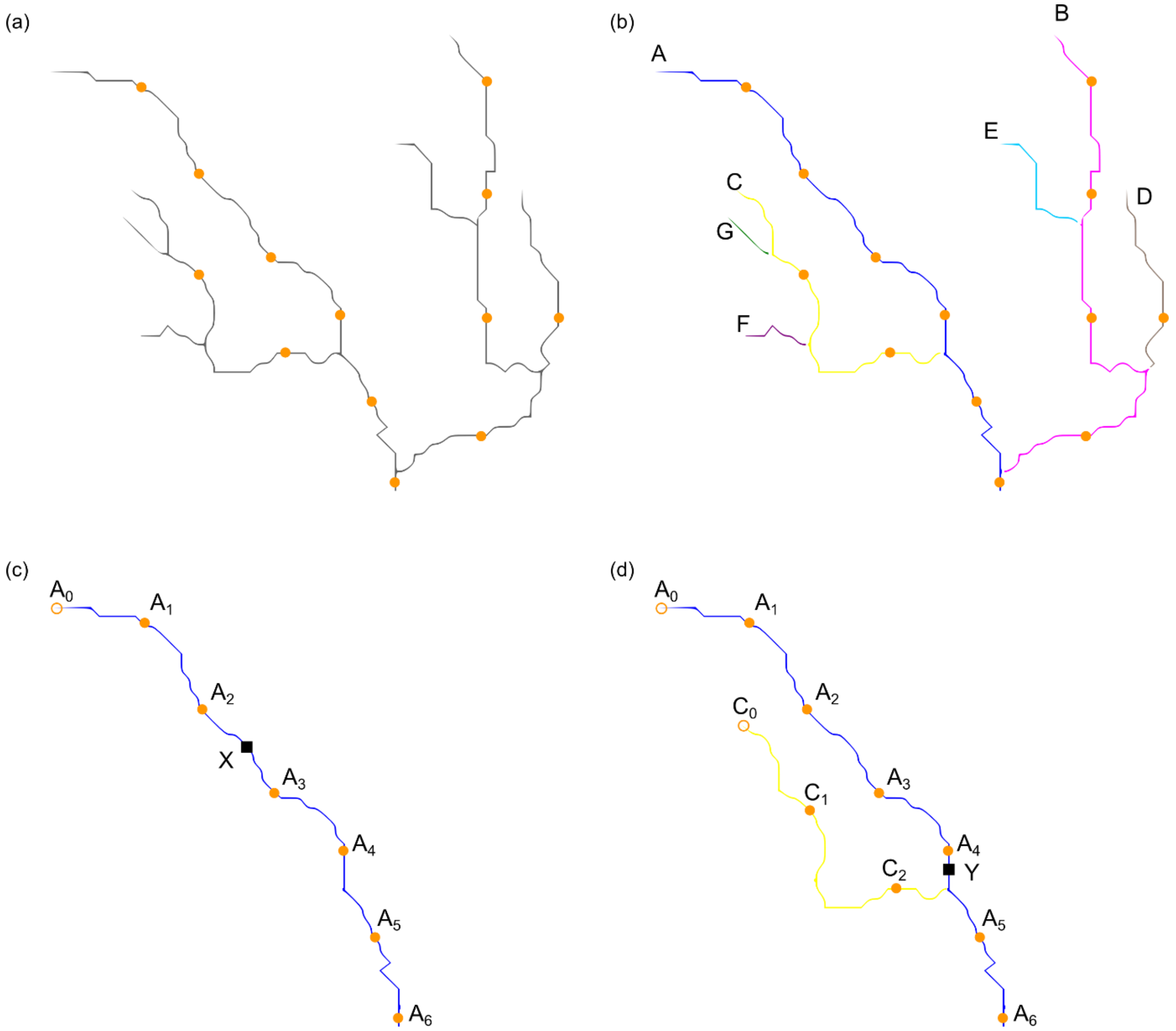


Figure 27 Conceptual illustration of the interpolation of discharge calculated by a model in cells with 1 km cell size along flow segments of 50 m length with known cumulative catchment area. (a) river network is represented by a set of links between flow cells downstream from any source and a set of points (orange) with input discharge is established based on similarity of cumulative catchment area; (b) dominant flow paths are selected, marked here with unique colours and labelled with capital letters; (c) taking each flow line in isolation (A in this case), discharge at any point is calculated based on the gradient of discharge between the closest upstream and the closes downstream input points over cumulative catchment area, point X is an example of a point where interpolation works well because the point is between two points where modelled discharge is available and it is not affected by any other flow line; (d) input discharge at sources is assumed to be zero and discharge at end point of every flow line is calculated based on the previous known gradient of discharge over cumulative catchment area, point Y is an example of a point where interpolated discharge may be overestimated because it is between a point where modelled discharge is available and a confluence with another flow line.

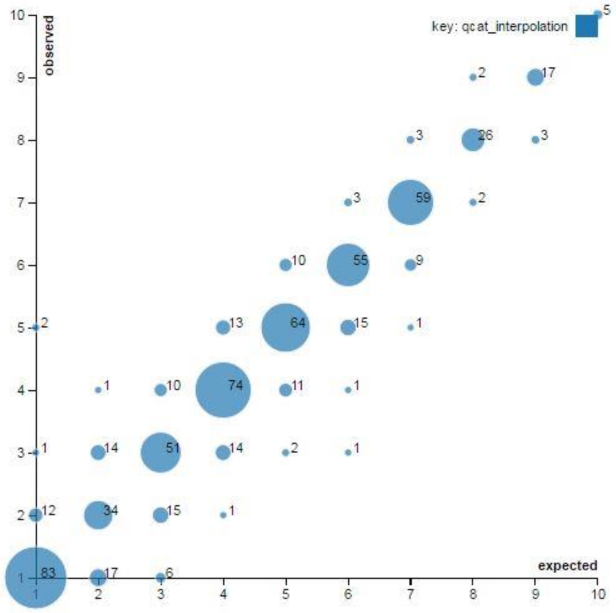


Figure 28 Discharge category at calibration sites (horizontal axis) and discharge category obtained using interpolation of mean annual discharge along flow paths (vertical axis). For 80 calibration sites discharge category using interpolation could not be established are they are not shown in the plot.

3.2.4.3 Discharge category from combined techniques

The comparison of the two previous sets of results revealed that most calibration sites were classified into the same category with either method. The largest mismatch was in classes 1 to 4. For 80 sites, it was not possible to estimate discharge category using interpolation due to the limitations mentioned above. There was no obvious indication whether the regression technique or the interpolation technique performed substantially better than the other. Discharge category established using the different techniques was generally within one category of the other (Figure 29).

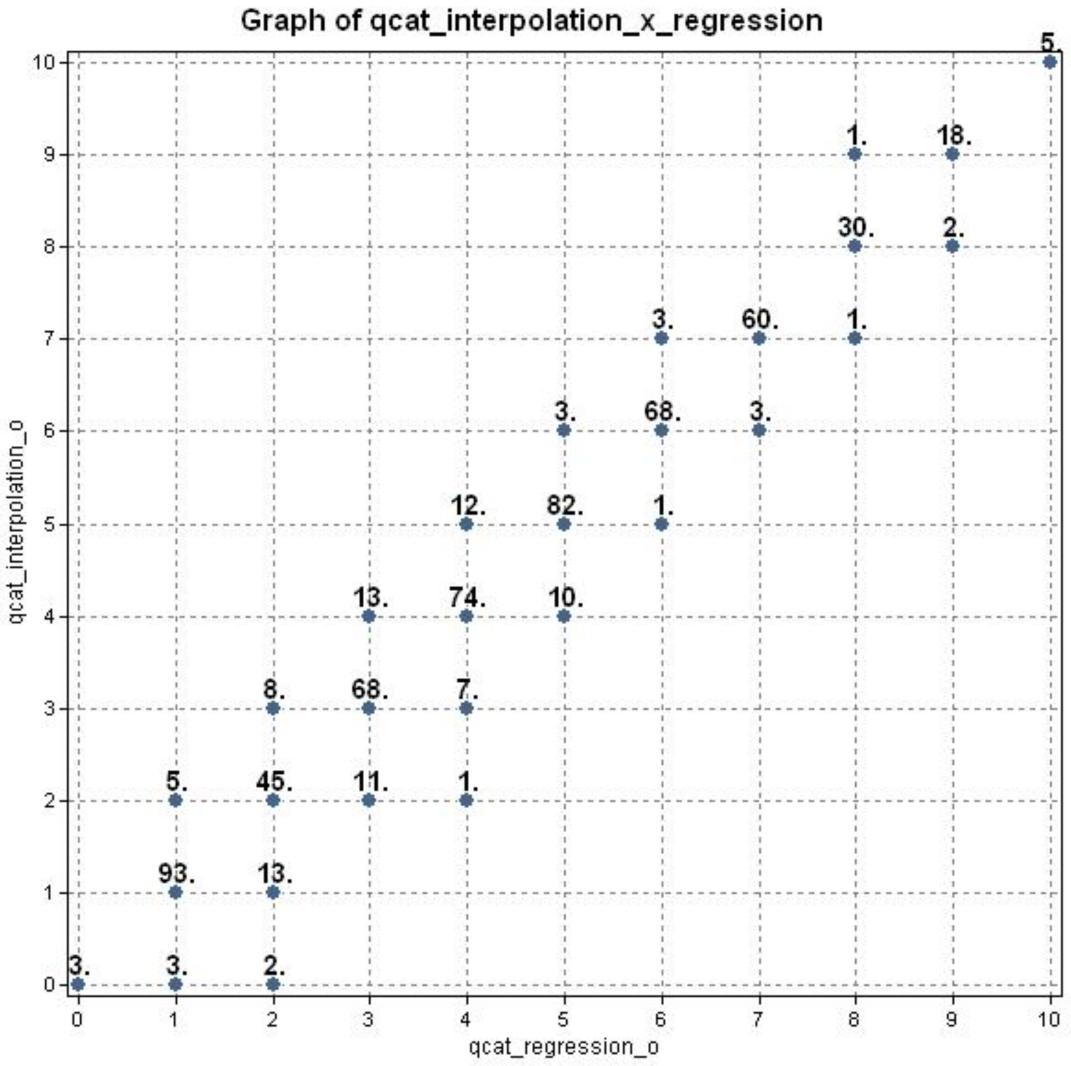


Figure 29 Comparison of categories at calibration sites produced using two different techniques. Regression of discharge with cumulative catchment area (horizontal axis) and interpolation of discharge along flow paths (vertical axis). Numerical labels indicate number of sites in each combination.

While the interpolation technique is theoretically more accurate, it cannot be used on its own because the interpolation could not be performed for many cells. Simply in-filling missing values with regression would be technically easy but could introduce situations where discharge category in a cell is higher than discharge category of its downstream cells. We decided to combine the two methods as follows:

- If discharge category can be established using the interpolation technique, use the value from the interpolation technique
- Else inspect the value from the regression technique; if the value from the regression technique is lower or equal to all downstream values of discharge category, use the value established using regression
- Else use the closest downstream discharge category, or if no downstream discharge category is available, use the value from regression technique.

Comparing the combined results with the calibration data revealed that most discharge category values were within 2 categories from calibration data (Figure 30). Investigation of 7 sites where discharge category could not be established revealed that:

- Sites 4885 and 5203 were located on branches of rivers not represented by the flow segments because it was not possible to identify these segments as downstream from a source (see also Figure 12 and Figure 13).
- Site 5845 was not properly snapped to the DTM but manual snapping confirmed discharge category 1 as expected based on the calibration data.
- Site 6381 was on a river not represented in CEH rivers (IHDTM indicated catchment area of 1.14 sq km).
- Sites 5852, 6242, and 6844 did not have valid DTM coordinates.

Investigation of the sites with the largest differences from calibration data revealed that the newly calculated values are plausible considering the site position on the river network and considering G2G model outputs:

- 379; expected 6 (at calibration site), interpolated 3, regression 4, accepted 3.
- 338; expected 4 (at calibration site), interpolated NA, regression 1, accepted 1.
- 5905; expected 2 (at calibration site), interpolated 4, regression 4, accepted 4.

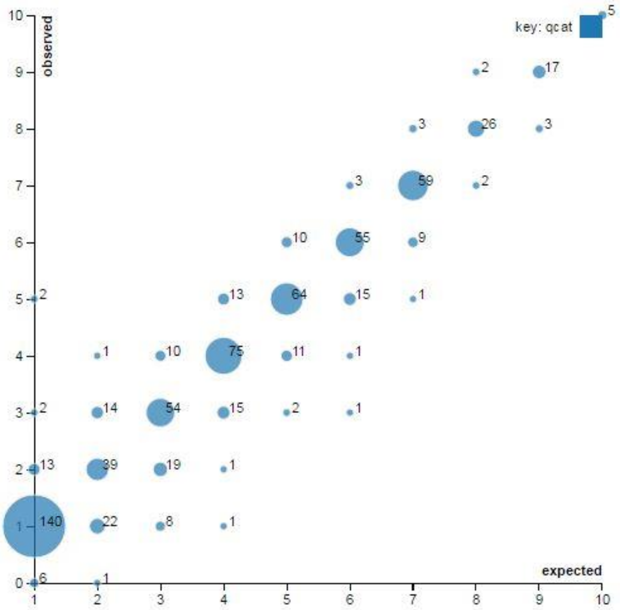


Figure 30 Discharge category at calibration sites (horizontal axis) and discharge category derived as a combination of regression and interpolation techniques (vertical axis). At 7 sites discharge category could not be established (observed=0).

4. Database

4.1 Database formats

Most data in this project have spatial reference so we mostly considered spatial data formats that can be loaded in conventional GIS packages. Data will be ultimately be delivered to users through a web-based system. Therefore, we considered how suitable different formats are for underpinning a web application based on *GeoServer* and *Django* (see Section 5). While suitability for deployment on the server is key, it is still desirable to be able to access, query, and visualize the data offline (eg during development and maintenance of the web tool).

There are two main data models for which different storage formats are suitable: vector model and raster model. With the vector model, objects are represented by rows with a set number of columns and geometry are stored in one or more columns of each row (e.g. river lines, river flow segments, etc). The raster model is generally suitable for continuous fields which are recorded in cells on a regular grid (e.g. elevation grid; Burrough and McDocnnell, 1998).

For vector data, the optimal storage solution on a production server is a PostGIS database (http://postgis.net/), which can achieve excellent performance thanks to many tuning options. For raster data, file-based raster formats are generally good for the types of queries required by RICT (retrieving dozens or a few hundreds of pixels) and are easier to manage.

4.1.1 Comma Separated Value (CSV)

CSV is a widespread set of conventions often used and implemented in different ways, rather than an entirely defined format. CSV is very popular mainly because users can open a CSV file in any text editor and see values directly. However, compared to binary formats, CSV requires more storage space. CSV also does not retain information about data types of individual columns. It is also not suitable for spatial data storage and it does not support indexing of any kind. Because of these drawbacks we did not use CSV for the core datasets delivered for this project, except for very small explanatory tables.

4.1.2 Esri Shapefile (Shapefile)

Shapefile has been widely used since Esri (1998) defined the format. It has been used in proprietary as well as in open source GIS packages and is one of the most popular vector data formats. However, it has several disadvantages because of which we did not select it as the delivery format. The properties especially relevant for this project were:

- Column heading can have at most 10 characters.
- There is no concept of missing data; missing data would have to be encoded as special values like -9999 or empty string which can lead to confusion and difficulties in future processing and visualization.
- Shapefile data types may not be interpreted correctly in all GIS packages (see https://geonet.esri.com/thread/159997).
- The component files of a Shapefile should not exceed 2 GB; some of the datasets produced in this project exceed this threshold (.dbf files) when exported to Esri Shapefile.

4.1.3 Esri File Geodatabase (FGDB)

Esri File Geodatabase is the native and recommended storage container for ArcGIS. We have seen consistently the best performance when compared to other file-based formats, not only in ArcGIS. The format is proprietary but a free C++ application programming interface released by Esri means that Esri File Geodatabase can be (and has been) used by other GIS (eg GDAL, QGIS). As most of the delivered datasets were created in ArcGIS, appropriate attribute and spatial indices have been created which allow fast access in QGIS too. Esri file geodabase does not have any of the drawbacks of a shapefile and can be used (or imported) in different systems.

In addition to vector data, Esri File Geodatabase can store rasters and other types of data but these are generally readable only by ArcGIS. Therefore, we used Esri File Geodatabase for vector data storage but we recommend storing rasters as stand-alone datasets in either TIFF or IMG as described below.

Note: GeoServer cannot read Esri File Geodatabase; data would have to be exported to another format, ideally into a PostGIS database. However, importing vector data from any of the file based formats listed here into a PostGIS database on the production server is highly recommended anyway.

Commands to import data from Esri File Geodatabase to PostGIS using ogr2ogr follow a pattern:

```
ogr2ogr -f "PostgreSQL" PG:"connectionString" "/path/to/file.gdb" "feature_class_name"
```

For example:

```
ogr2ogr -f "PostgreSQL" PG:"host=localhost port=54321 dbname=geoserver user=postgres
password=geoserver" "/data/rict.gdb" "flowsegments"
```

Note that spatial and attribute indices on the imported table need to be built afterwards using the CREATE INDEX command in SQL.

4.1.4 GeoPackage and Spatialite

GeoPackage is an Open Geospatial Consortium (OGC) standard which aspires to define a universal spatial data storage container. The GeoPackage standard has been implemented in several GIS software including ArcGIS, QGIS, and GDAL. To our best knowledge, the implementations rely on a Spatialite database which extends SQLite file-based database.

Spatialite database formatted according to the GeoPackage format, or just plain Spatialite database were interesting delivery options because they are open source resources. There are also GeoServer plugins that enable GeoServer to read these data formats directly. However, these plugins are community plugins not officially supported by the GeoServer community. For more information about GeoPackage and SQLite see http://www.geopackage.org/ and http://www.gaia-gis.it/gaia-sins/.

4.1.5 ASCII Grid

A text-based raster data format. It can be opened in any text editor and values will be directly visible. However, ASCII Grid requires more data storage than the binary raster formats listed below. Rendering speed is generally also lower and requires more computational resources than with binary raster formats. Furthermore, different GIS packages handle the file header information (such as coordinates of the anchor point of the raster) in different ways which may lead to inconsistencies.

4.1.6 *Tagged Image File Format (TIFF)*

One of the most widely used formats for raster data storage. Natively supported by most GIS packages. However, we have experienced inconsistencies in rendering between GIS packages (e.g. loss or change in colour ramps). TIFF has been widely used with GeoServer and it offers several tuning options that can improve performance in rendering on the web. More details at http://www.gdal.org/frmt_gtiff.html

4.1.7 *Erdas Imagine (IMG)*

Another popular format designed for raster data storage. IMG has similar capabilities to TIFF. We have experienced consistently better performance with IMG over TIFF, especially when it comes to reading and rendering so we recommend IMG and the raster data delivered for this project are in the IMG format. While the IMG format is proprietary, it has been widely used for decades and most GIS packages can natively read this format as they often rely on the GDAL library which supports IMG out-of-the box. However, GeoServer requires a (free) plugin to read this format. More details at http://www.gdal.org/frmt_hfa.html

4.1.8 *Selected option*

The preferred format selected is Esri File Geodatabase for vector data and Erdas Imagine for raster data. These files can be transferred to the server and, where vector data, imported into a PostGIS database. Raster data will be stored outside the database either in the IMG format or they can be converted to TIFF if necessary. Note that raster data stored in a TIFF format produced in ArcGIS often cannot be used by GeoServer directly so translation using the gdal_translate tool would be necessary anyway.

Use gdal_translate to convert an IMG file to a tiff optimal for GeoServer:

```
gdal_translate -of GTiff -co "TILED=YES" in.img out.tiff
```

It is also strongly recommended to build overlays for each tiff file:

```
gdaladdo -r average the.tif 2 4 8 16
```

More information about data optimization for GeoServer can be found at:

http://docs.geoserver.org/latest/en/user/production/data.html

4.2 Database structure and delivery format

Results of this project are delivered in a folder which contains the following items:

- rict_variables_gb.gdb – Esri File Geodatabase with flow_segments feature class for GB.
- rict_variables_ni.gdb – Esri File Geodatabase with flow_segments feature class for NI.
- rict_rasters_gb – Folder with several variables stored as gridded data in GB. These files are not required by the delivery tool but could be useful in future development.
- rict_rasters_ni – Folder with several variables stored as gridded data in NI. These files are not required by the delivery tool but could be useful in future development.

- rict_validation – Folder with workbook with all variables calculated in this project against calibration sites in GB from WFD119 (not available for NI).
- rictrepo.zip – code repository with the web application and details about configuration of the demonstration tool.
- lut_geology_bedrock_map_code.xlsx – classification of BGS 1:625K maps to RICT geology classes.

Input datasets such as the IHDTM grids, river network, geology layers, etc., and the code to process them are not part of the delivery database. The key output of this project is a feature class of flow segments with individual RICT variables stored in the attribute table (Table 5). Start nodes of the segments represent centres of 50m cells located downstream of river sources as defined by CEH 1:50K river network and the IHDTM drainage direction grid. End nodes of the flow segments are located at centres of cells that the segment flows into. Data is stored in British National Grid (http://epsg.io/27700) coordinate system for GB and in TM65 / Irish Grid also known as Irish National Grid (http://epsg.io/29902) for NI. The gridded files and validation results are provided in case any further validation should be performed. Naming conventions used for these are closer to the names used during computation as summarized in Table 6. Note that in the final results, geology without a version suffix refers to the latest version of geology while during the calculation in GB it referred to the geology based on version 4 of BGS data.

Table 5 Attributes included in the main output dataset of flow segments.

Name	Description	Notes
altitude	Altitude in metres above sea level	Based on OS Landform-PANORAMA in GB and on EU-DEM in NI
dfromsrc	Distance from the furthest source in metres	Calculated using ArcGIS Flow Length Tool
logaltbar	Base 10 logarithm of upstream catchment mean altitude	Based on OS Landform-PANORAMA in GB and on EU-DEM in NI
logarea	Base 10 logarithm of upstream catchment area	
chalk	Proportion of upstream catchment area covered by Chalk	Based on the latest geology maps if *_v4 or *_v5 not specified
clay	Proportion of upstream catchment area covered by Clay	Based on the latest geology maps if *_v4 or *_v5 not specified
hardrock	Proportion of upstream catchment area covered by Hard rock	Based on the latest geology maps if *_v4 or *_v5 not specified
limestone	Proportion of upstream catchment area covered by Limestone	Based on the latest geology maps if *_v4 or *_v5 not specified
peat	Proportion of upstream catchment area covered by Peat	Based on the latest geology maps if *_v4 or *_v5 not specified
propwet	Proportion of time upstream catchment soils are wet.	Available only for catchments greater than 0.5 km^2
qcat	Discharge category	
slope	Slope	Slope calculated along the flow segments based on OS Landform-

		PANORAMA in GB and EU-DEM in NI
oid	Internal ID.	Not required but can be useful
SX	Easting of the segment start	Not required but can be useful
SY	Northing of the segment start	Not required but can be useful
EX	Easting of the segment end	Not required but can be useful
EY	Northing of the segment end	Not required but can be useful
end	Internal ID of segment end	Not required but can be useful
lake	1 if start is in a lake, 0 otherwise	Not required but can be useful
length	Length of the segment in metres	Not required but can be useful
region	Internal computational region ID	Not required but can be useful. See Appendix 5.
source	Internal ID of the furthest source	Not required but can be useful
start	Internal ID of segment start	Not required but can be useful

Table 6 Summary of naming conventions used during calculations.

Name	**Description**
oid	Internal ID
SX	Easting of the segment start
SY	Northing of the segment start
EX	Easting of the segment end
EY	Northing of the segment end
altitude_*	Altitude based on a certain digital elevation model
ccar	Cumulative catchment area. Different units may be used in different context, one of m^2, km^2, or 0.0025*km^2
d_ds	Distance to the next point downstream
d_us	Distance to the previous point upstream with the largest cumulative catchment area
dfromsrc_x	Distance from the furthest source calculated by accumulating length of flow segments from each source segment. In metres.
dz	Difference in a quantity, usually difference in altitude between the upstream and downstream points selected for calculation of slope
end	Internal ID of the segment end point
lake or is_lake	1 if the start or a segment is in a lake, 0 otherwise
length	Generally indicates length of the segment in metres
rawslope	Raw result of calculation of slope before conversions and clean up.
region	Internal computational region ID
slope	Slope
slopedistance	Distance used for calculation of slope
source	Internal ID of the furthest source
start	Internal ID of the segment start point
z	Used for various variables during calculation but stored values generally refer to altitude.
altitude_*	Altitude based on a certain elevation grid
dfromsrc, dfromsrc_y	Distance from the furthest source calculated using ArcGIS

	Flow Length Tool. In metres.
***_eudem**	EU-DEM elevation grid
***_feh**	Dataset based on Flood Estimation Handbook
***_ihdtm**	IHDTM HGHT elevation grid
***_panorama or *_pnrm**	OS Landform-PANORAMA elevation grid
***_terrain50**	OS Terrain 50 elevation grid
logarea*	Logarithm of catchment area
proportion_*	Proportion of catchment area covered by category of certain type, e.g. proportion_chalk indicates proportion of catchment covered by Chalk.
propwet	Proportion of time upstream catchment soils are wet for RICT.
qcat	Discharge category derived using combination of regression and interpolation.
qcat_interpolation	Discharge category derived using interpolation of mean annual from Grid2Grid model.
qcat_regression	Discharge category derived using regression of mean annual discharge from Grid2Grid model with cumulative catchment area.
slope_from_segments_* _raster	Slope along flow direction segments derived using certain elevation grid.

5. Demonstration delivery system

A demonstration system was developed allowing users to visualise and retrieve RICT variables at any location. While the demonstration system has some limitations outlined below, the components and technologies used in the demonstration tool have the potential to resolve these limitations. All components are based on free open source software so that operational costs do not entail any licensing fees but mainly hosting fees and staff time (maintenance). This section describes how the demonstration tool is constructed, and how it could be developed further.

5.1 Components and configuration

The demonstration tool is hosted on a single Linux server (Ubuntu 14.04). The roles of individual system components are showed in the conceptual diagram (Figure 31).

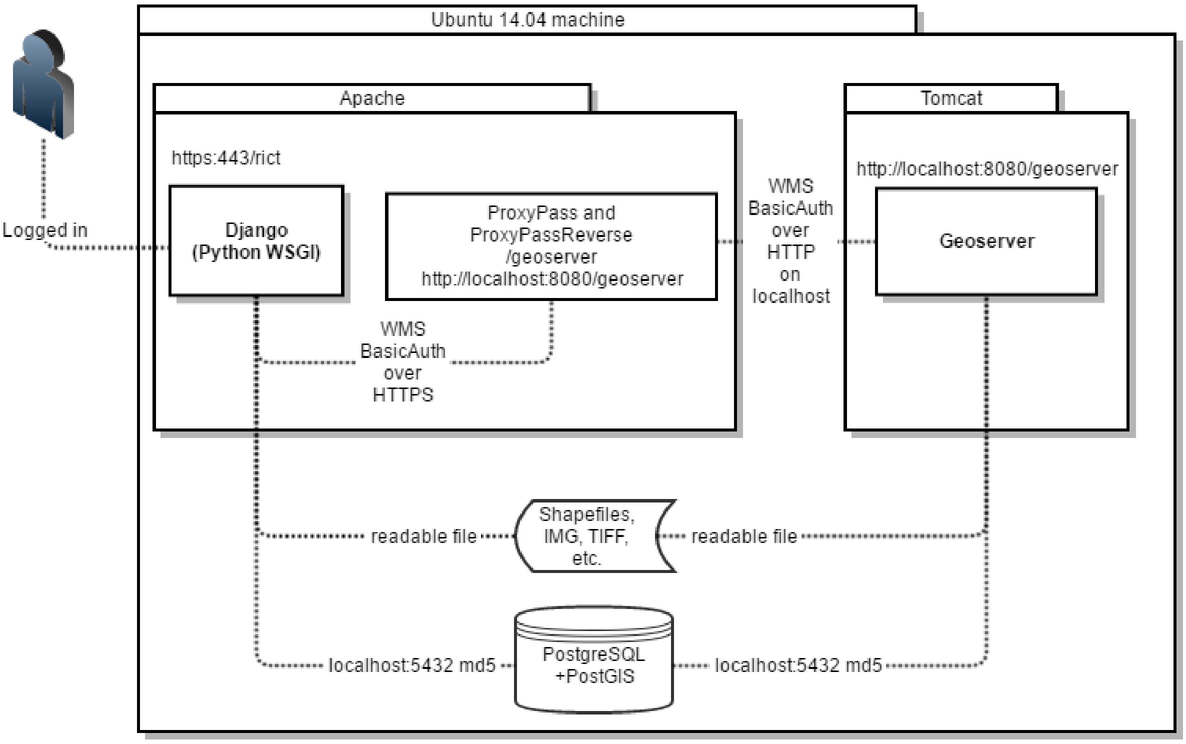


Figure 31 Conceptual diagram of the data delivery demonstration system.

The *Apache* web server (version 2.4.7; http://httpd.apache.org/) listens to HTTPS requests and passes them to the *Django* application (version 1.9.7; https://www.djangoproject.com/). HTTPS requests for web map services are passed onto *GeoServer*. The *Django* project contains applications written in the Python Django framework. It handles user authentication, requests for RICT variables (including snapping), and provides the interface visible to the user as a web page:

- *Django* RICT application homepage at https://fkvm.cloudapp.net/rict/rictdata/
- *Django* administration interface to manage users at https://fkvm.cloudapp.net/rict/admin

The 'geoproxy' application included in the project allows the display of the secured web map services from *GeoServer* only for users who are logged in. The *Git* repository with the *Django* project is provided in a password protected .zip archive. The README file in this repository includes further details. *Tomcat* (version 7.0.52; http://tomcat.apache.org/) is a Java Servlet container required to run *GeoServer*. *GeoServer* (version 2.8.5; http://geoserver.org/) is used to provide web map services so that RICT variables can be visualised in the *Django* application. *GeoServer* has powerful security options which were used to restrict access to the web map services. Only logged users are able to see the map services. The *Django* application requests maps from the web services using a dedicated user name and password. The web map services thus appear to be an integral part of the *Django* application and user authorisation can be handled by *Django*. Names of workspaces, services, and styles exposed by *GeoServer* must match names expected by *Django* (see code for details). Style files used for the demonstration system are included in the *Git* repository (geoserver_style_files folder). *PostgreSQL* with the spatial extension *PostGIS* are used to store the RICT variables at individual flow segments. Web map services provided by *GeoServer* read the data from the *PostgreSQL* database.

5.2 Future development

The demonstration system provides basic access to RICT variables. We make the following recommendations to improve its usefulness and robustness.

5.2.1 Group policies and protecting specific variables

The demonstration system requires users to log in, which gives them access to all variables including PROPWET. The final system should allow access to PROPWET only to authorised users. This can be implemented by restricting access to PROPWET only to members of specific groups. The *Django* framework is well suited for this purpose. However, creating and managing groups (adding/removing users) was outside the scope of the demonstration tool and should be implemented during the development of the final system. This limitation implies that only users who already have access to the FEH PROPWET product are allowed to use the demonstration system.

5.2.2 Display river names and improve snapping verification

Currently the demonstration system has a limited number of river names visible on the backdrop maps. More river names could be added and mechanisms for verifying that the river name of a snapped location matches the expected river name of the input location (if known) could be implemented. The snapping currently finds the cell with the largest cumulative catchment area within a specified search radius. More advanced snapping algorithm providing a measure of confidence could be implemented.

5.2.3 Batch mode

The current demonstration system handles requests for each point individually. A basic user interface for batch mode has been implemented, where user can specify coordinates of multiple points to extract values from. In addition to the basic batch mode interface, the demonstration system can be queried for multiple points programmatically. The project *Git* repository includes a script that shows how to do that. The script is for internal project use only as general public use could overload the demonstration server. The capabilities for processing multiple points could be substantially improved in future development.

5.2.4 Data improvements

It is possible that user feedback will include suggestions for improvements of the data. This may include recommendations for specific sites but also some suggestions that can be applied over wider areas. Future improvements to the data may include validations of G2G modelled discharge in NI or production of a variable that can be used instead of PROPWET.

6. Intellectual Property Rights

Careful consideration of Intellectual Property Rights (IPR) and Licensing are essential to the successful completion of this project and future work streams. Balancing the objective of an Open, or “as Open as possible” output with the desire to use the best possible input data unavoidably causes some conflict between RICT objectives and the established licensing practices of IPR owners.

The conflict can most clearly be seen in the case of the PROPWET replacement variable. Here a replacement variable which is fundamental to RICT is also a fundamental component of CEH's FEH Web Service (https://fehweb.ceh.ac.uk/). It is essential for flood estimation in the UK that the FEH Web Service is maintained and developed to a high standard. Funding for this maintenance and development is derived in part from fees paid for access to PROPWET values. Any "free of charge" release of PROPWET values via RICT would undermine the future of a vital tool in UK flood estimation. As PROPWET values 1) are required to drive RICT; but, 2) cannot be given away for free, and 3) RICT cannot be a charged for service there is a clear conflict. Similar conflicts exist for other replacement variables and this work package will propose a resolution for each.

The IPR/Licensing resolutions will be mindful that the ambition of the RICT project is to produce an output which is, where possible, less restrictively available than the "point and click system" envisaged by the IPR recommendations in Clarke et al. (2011).

6.1 Licensing Approach for RICT

In selecting input data to derive the replacement variables CEH has chosen datasets whose licensing conditions provide varying levels of restriction on the use of derived data (NB. the replacement variables are all considered to be derived data).

All input datasets that are proposed as sources for the replacement variables will satisfy at least one of the following:

1) Input data are publicly available under an Open Data licence;
2) Input data are already held by the regulators under existing licensing agreements;
3) Input data are CEH owned datasets where 1) and 2) do not apply, but where use by CEH in production of replacement variables is incidental or will lead to a derived dataset that CEH has no objection to making openly available (e.g. see Discharge Category where input data are not open and are not currently licensed by the regulators, but where the derived data are sufficiently "removed" from the originating input data that they cannot act as a substitutes or otherwise compete with the originating input data).

The different licensing restrictions associated with the various input datasets will govern the type of use that replacement variables can be made available for. These different types of use can be defined in the context of RICT as '**Internal Use**,' '**Open Use,**' '**Use to Drive a Closed RICT'** and **'Evaluation Use**.' The four use types are defined below.

It should be noted that the replacement variables will be provided in two forms, 1) as **values for individual sites** obtained by a user clicking on the relevant part of a map in the web based demonstration delivery tool, and 2) as **standalone gridded datasets**. The types of use will vary in some cases for the two different forms of replacement variable delivery:

- **Open Use**: means the variable can be made available publicly.
- **Internal Use**: means the variable will be made available to the staff of regulators and their contractors only.

- **Use to Drive a Closed RICT**: means the variable can be used by all users within a closed RICT tool where individual values are not seen by, or otherwise available to, users.
- **Evaluation Use**: means the variable will be made available to the staff of regulators and their contractors only. Access will be granted for an initial 6 month period allowing evaluation of the data and demonstration tool within the day to day activities of EA, SEPA, NRW and Northern Ireland Environment Agency. Licensing negotiation may be required to extend the evaluation period beyond 6 months or convert use type to one of the other three categories.

The replacement variables are listed in Table 7 below along with the chosen input data and details of the types of permitted use.

Table 7 RICT Variable, Chosen Input Data and Permitted Use Type; CEH means Centre for Ecology & Hydrology; OS means Ordnance Survey; EU means European Union; LPS means Land and Property Systems (previously Ordnance Survey Northern Ireland); BGS means British Geological Survey; GSNI means Geological Survey Northern Ireland

RICT Variable	Input Data and owner	Permitted Use Type
Upstream catchment mean altitude (LIGALTBAR)	**GB:** • Landform-PANORAMA (OS) • IHDTM Flow Grid (CEH)	**GB:** • Open Use of values for individual sites • Internal Use of standalone gridded dataset
	NI: • Digital Elevation Model over Europe (EU-DEM) (EU) • IHDTM Flow Grid (CEH)	**NI:** • Open Use of values for individual sites • Evaluation Use of standalone gridded dataset
Proportion of time upstream catchment soils are wet (PROPWET)	**GB:** • FEH descriptor PROPWET (CEH)	**GB:** • Internal Use of standalone gridded dataset • Use to Drive a Closed RICT[1] of values for individual sites
	NI: • FEH descriptor PROPWET (CEH)	**NI:** • Evaluation Use of standalone gridded dataset • Use to Drive a Closed RICT of values for individual sites

[1] PROPWET data and the values derived from it will be directly available to EA, SEPA and NRW staff, but will initially only be available to DAERA staff during a 6 month evaluation period.

Upstream catchment cover of key geological types	**GB:** • DiGMapGB-625: Bedrock geology (BGS); • DiGMapGB-625: Superficial theme (BGS); • IHDTM Flow Grid (CEH)	**GB:** • Open Use of values for individual sites • Internal Use of standalone gridded dataset
	NI: • DiGMapGB-625: Bedrock geology (BGS); • NI 250k SUPERFICIAL Geology (GSNI) • GSI 500K Bedrock • IHDTM Flow Grid (CEH)	**NI:** • Open Use of values for individual sites • Evaluation Use of standalone gridded dataset
Logarithm of upstream catchment area (LOGAREA)	**GB:** • IHDTM Cumulative Catchment Area grid (CEH)	**GB:** • Open Use of values for individual sites • Internal Use of standalone gridded dataset
	NI: • IHDTM Cumulative Catchment Area grid (CEH)	**NI:** • Open Use of values for individual sites • Evaluation Use of standalone gridded dataset
Altitude	**GB:** • Landform-PANORAMA (OS) • IHDTM Flow Grid (CEH)	**GB:** • Open Use of values for individual sites • Internal Use of standalone gridded dataset
	NI: • Digital Elevation Model over Europe (EU-DEM) (EU) • IHDTM Flow Grid (CEH)	**NI:** • Open Use of values for individual sites • Evaluation Use of standalone gridded dataset
Distance from source	**GB:** • 1:50,000 Watercourses data (CEH) • IHDTM Flow Grid (CEH)	**GB:** • Open Use of values for individual sites • Internal Use of standalone gridded dataset

	NI: • 1:50,000 Watercourses data (CEH/LPS) • IHDTM Flow Grid (CEH)	**NI:** • Open Use of values for individual sites • Evaluation Use of standalone gridded dataset
Slope	**GB:** • Landform-PANORAMA (OS) • 1:50,000 Watercourses data (CEH) • IHDTM Flow Grid (CEH)	**GB:** • Open Use of values for individual sites • Internal Use of standalone gridded dataset
	NI: • Digital Elevation Model over Europe (EU-DEM) (EU) • 1:50,000 Watercourses data (CEH/LPS) • IHDTM Flow Grid (CEH)	**NI:** • Open Use of values for individual sites • Evaluation Use of standalone gridded dataset
Discharge Category	**GB:** • Grid2Grid mean annual discharge 1km grid (CEH) • Grid2Grid cumulative catchment area 1km grid (CEH) • IHDTM Cumulative Catchment Area grid (CEH) • IHDTM Flow Grid (CEH) • IHU Groups (CEH)	**GB:** Open Use of values for individual sites Internal Use of standalone gridded dataset
	NI: • Grid2Grid mean annual discharge 1km grid (CEH) • Grid2Grid cumulative catchment area 1km grid (CEH) • IHDTM Cumulative Catchment Area grid (CEH/LPS) • IHDTM Flow Grid (CEH) • IHU Groups (CEH)	**NI:** • Open Use of values for individual sites • Evaluation Use of standalone gridded dataset

6.2 Additional IPR Considerations

1. Where replacement variables are to be supplied for **Internal Use**, **Evaluation Use** or **Use to Drive a Closed RICT** a separate licence will be issued by CEH to EA, SEPA, NRW and DAERA ahead of data supply. It should be noted that the long term viability of the solution described above is subject to the continued existence of these licences and the regulator’s other existing licences with CEH described in Table 8 as ‘Internal Business Use licences’ and ‘memoranda of understanding regarding access to the FEH Web Service.’ The continued existence of these agreements are not currently under any doubt.
2. **Acknowledgement Requirements:** The use of input data to create replacement variables, even where the input is licensed under an open data licence, will usually come with an acknowledgment requirement. Any RICT tool that is publicly visible should acknowledge the use of all input datasets and identify the owners of each input dataset (Table 8).
3. **Data Security:** Where the RICT Data Delivery Tool, or in future the RICT itself, contains a copy of any datasets which are not available under an Open Data licence, adequate security mechanisms must be in place to prevent unauthorized use or access to the relevant dataset.
4. **On-line Terms of Use:** any public facing web tool supplying replacement variable values should include a Terms of Use section that will state that users are not permitted to bulk extraction of replacement variables or to use replacement variable values for commercial activity unrelated to the normal use of RICT.

Table 8 Input data and relevant licences

Input Data	**Licence**
Landform-PANORAMA (OS)	Licensed openly under Open Government Licence[2]
IHDTM Flow Grid (CEH)	Licensed to EA, SEPA and NRW under Internal Business Use licences.
Proportion of time upstream catchment soils are wet (PROPWET)	Licensed to EA, SEPA and NRW under memoranda of understanding regarding access to the FEH Web Service
DiGMapGB-625: Bedrock geology (BGS) DiGMapGB-625: Superficial theme (BGS)	Licensed openly under Open Government Licence[3]
NI 250k SUPERFICIAL Geology	Licensed openly under Open Government Licence[4]

[2] http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/
[3] http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/

(GSNI)	
GSI 500K Bedrock	Licensed openly under Creative Commons Attribution 4.0 Licence[5]
IHDTM Cumulative Catchment Area grid (CEH)	Used by CEH to generate RICT variables without licence being held by regulators
Digital Elevation Model over Europe (EU-DEM) (EU)	Provided openly subject to acknowledgement and non-endorsement conditions[6]
1:50,000 Watercourses data (CEH)	Licensed to EA, SEPA and NRW under Internal Business Use licences.
Grid2Grid mean annual discharge 1km grid (CEH) Grid2Grid cumulative catchment area 1km grid (CEH)	Used by CEH to generate RICT variables without licence being held by regulators

[4] http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/
[5] https://creativecommons.org/licenses/by/4.0/
[6] http://www.eea.europa.eu/data-and-maps/data/eu-dem#tab-metadata

7. Conclusions

In agreement with the project board, the project focused more resources on the derivation of the eight RICT input variables for both GB and NI generated than originally specified. As a consequence, the demonstration delivery system is relatively simple, but includes all the technologies needed to fully implement the features required before a public release, in particular, authorization to control access to PROPWET.

This report provides a detailed description of the data and methods selected to derive the eight RICT variables, and points out their limitations, so that users can make informed decisions about how to use the results appropriately. Many of the limitations are inherent to the 8-directional flow direction model used for catchment definition. Half of the variables required this model to be used, so it was used for the other variables as well to achieve consistency. However, the final database is the best technical compromise that can be produced across the entire UK while achieving spatial consistency, meeting licensing requirements, and being useable in a web-based delivery system.

The final database provides RICT input variables for both GB and NI at a 50-m grid cell resolution. The selected database formats are Esri File Geodatabase for vector data and Erdas Imagine for raster data.

The demonstration delivery tool was developed using free open-source software components so that operational costs do not entail any licensing fees but mainly hosting fees and staff time (maintenance). All software components are common, mainstream tools of the trade. Some functions would require further development, especially batch processing and the point-to-river snapping validation tool. It was agreed at the final project meeting that the demonstration tool would be available for 6 months after the formal end of the project.

8. Acknowledgements

We would like to thank Melinda Lewis (BGS) for her input on matching categories between version 4 and 5 of the 1:625K Bedrock geology map, and Robert Bailie (Department of the Environment Northern Ireland) and colleagues for providing valuable advice regarding data sources for Northern Ireland. We also want to thank the team working on the CEH Grid-to-Grid model for providing the modelled discharge as a result of their long term work. Without this dataset, it would not have been possible to produce the discharge category variable.

Appendix 1 Types of bedrock geology for River Invertebrate Classification Tool

The original WFD119 report used BGS 1:625K Bedrock Geology Map version 4 to define broad classes of geology. However, BGS have since released version 5 of this dataset. There is no direct mapping of attributes between versions 4 and 5 so a BGS expert assigned the broad classes required by RICT to each combination of LEX and RCS values in version 5 of the dataset. This appendix illustrates the differences in spatial distribution of RICT bedrock types based on the two different versions. In Northern Ireland, only version 5 was used. The BGS expert and several RICT project board members indicated that some classes might have been misclassified in the original WDF119 report and that a more detailed review may be needed (for example, areas around Wales classified as ‘chalk’). Given the specifications of this project, the objective was to produce variables consistent with the WFD119 calibration data, so a classification close to the original WFD119 report was used. Refer to the original WFD119 report for lookup table between version 4 and RICT classes (Clarke et al., 2011). A lookup table between version 5 and RICT classes is provided as an Excel spreadsheet lut_geology_bedrock_map_code.xlsx. The spreadsheet also indicates which categories may need review.

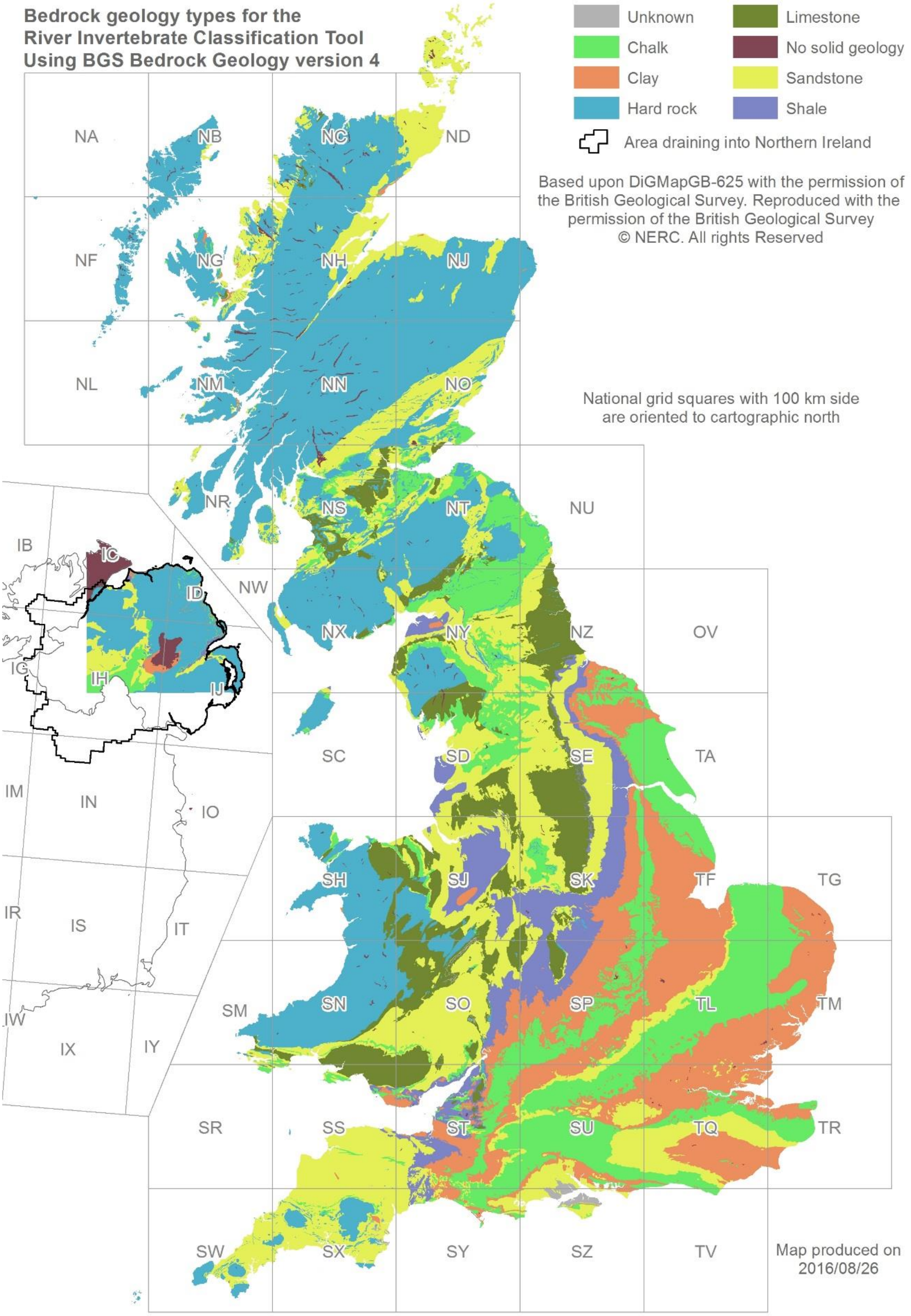
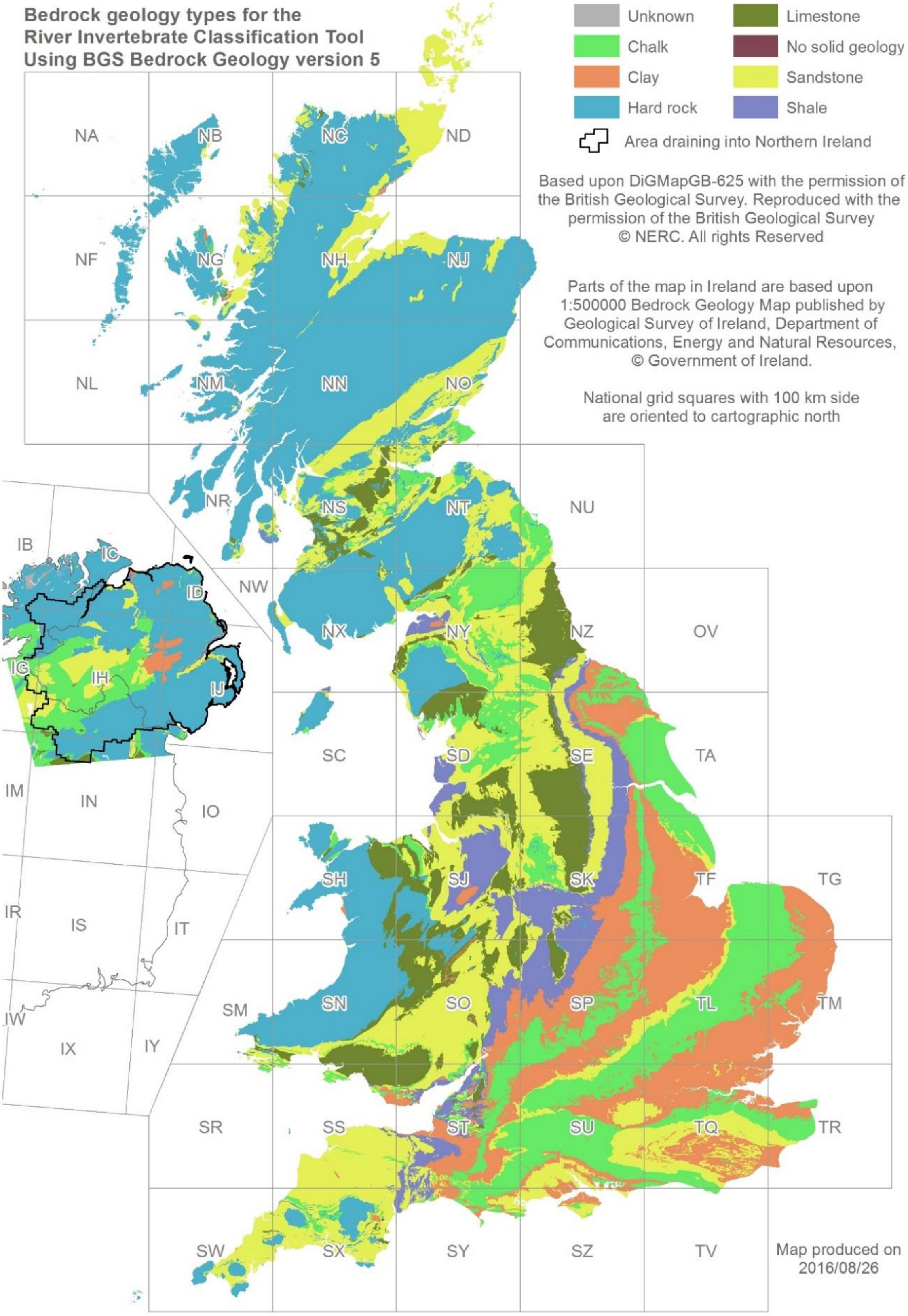

Bedrock geology types for the
River Invertebrate Classification Tool
Using BGS Bedrock Geology version 4
Unknown
Chalk
Clay
Hard rock
Limestone
No solid geology
Sandstone
Shale
Area draining into Northern Ireland
Based upon DiGMapGB-625 with the permission of the British Geological Survey. Reproduced with the permission of the British Geological Survey © NERC. All rights Reserved
National grid squares with 100 km side are oriented to cartographic north
Map produced on 2016/08/26
NA
NB
NC
ND
NF
NG
NH
NJ
NL
NM
NN
NO
NR
NS
NT
NU
IB
IC
ID
NW
NX
NY
NZ
OV
IG
IH
IJ
SC
SD
SE
TA
IM
IN
IO
SH
SJ
SK
TF
TG
IR
IS
IT
SM
SN
SO
SP
TL
TM
IW
IX
IY
SR
SS
ST
SU
TQ
TR
SW
SX
SY
SZ
TV

Figure 32 Spatial distribution of bedrock geology types based on version 4 of British Geological Survey 1:625000 Bedrock Geology Map.



Bedrock geology types for the River Invertebrate Classification Tool Using BGS Bedrock Geology version 5

Based upon DiGMapGB-625 with the permission of the British Geological Survey. Reproduced with the permission of the British Geological Survey © NERC. All rights Reserved

Parts of the map in Ireland are based upon 1:500000 Bedrock Geology Map published by Geological Survey of Ireland, Department of Communications, Energy and Natural Resources, © Government of Ireland.

National grid squares with 100 km side are oriented to cartographic north

Map produced on 2016/08/26

Figure 33 Spatial distribution of bedrock geology types based on version 5 of British Geological Survey 1:625000 Bedrock Geology Map and on data from Geological Survey Ireland in parts of the island of Ireland.

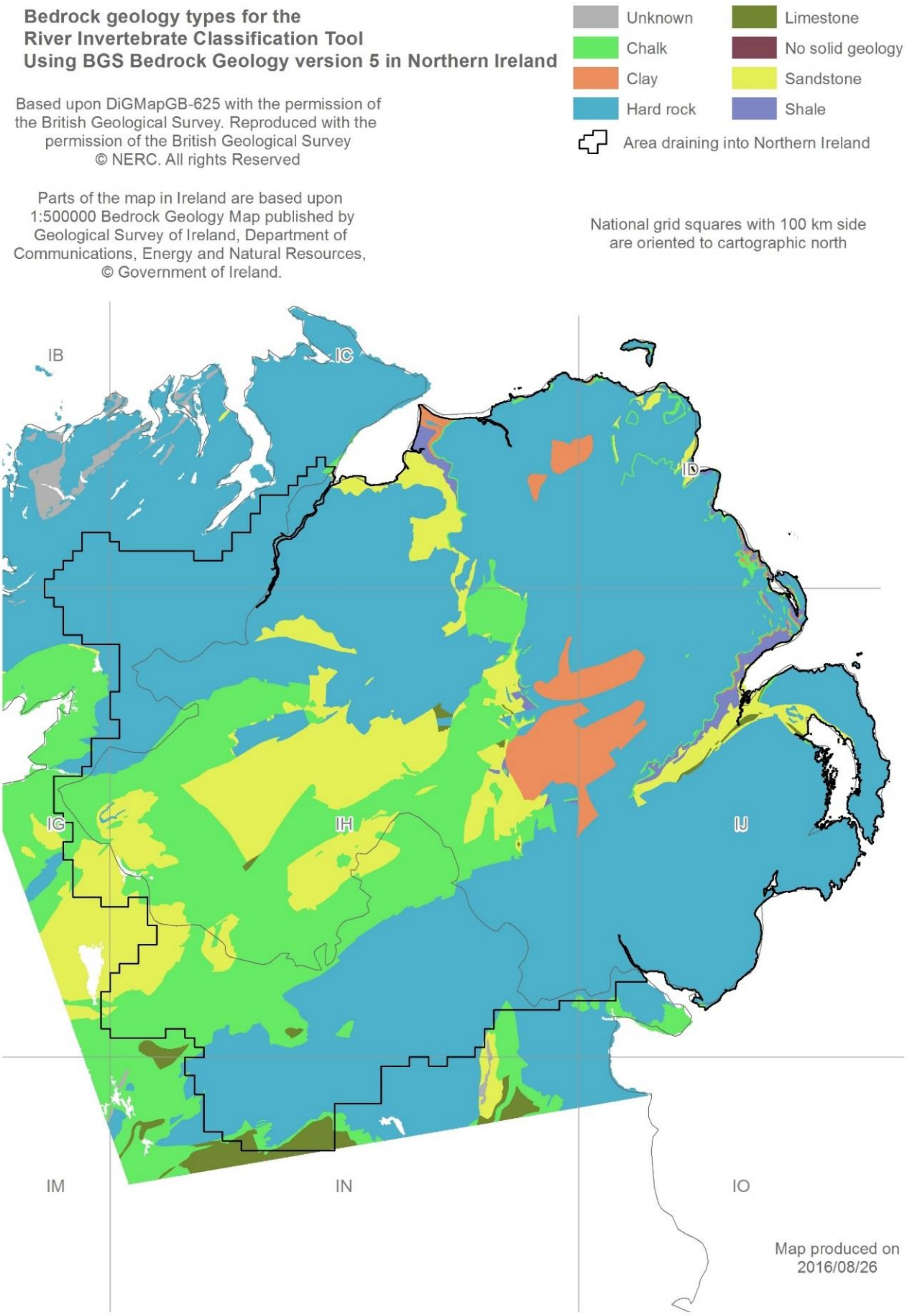

Bedrock geology types for the
River Invertebrate Classification Tool
Using BGS Bedrock Geology version 5 in Northern Ireland
Based upon DiGMapGB-625 with the permission of the British Geological Survey. Reproduced with the permission of the British Geological Survey © NERC. All rights Reserved
Parts of the map in Ireland are based upon 1:500000 Bedrock Geology Map published by Geological Survey of Ireland, Department of Communications, Energy and Natural Resources, © Government of Ireland.
Unknown
Chalk
Clay
Hard rock
Limestone
No solid geology
Sandstone
Shale
Area draining into Northern Ireland
National grid squares with 100 km side are oriented to cartographic north
IB
IC
ID
IG
IH
IJ
IM
IN
IO
Map produced on 2016/08/26

Figure 34 Spatial distribution of bedrock geology types in Northern Ireland based on version 5 of British Geological Survey 1:625000 Bedrock Geology Map and on data from Geological Survey Ireland.

Appendix 2 Detailed comparison of proportion of key geological types between calibration data and results of this project

Scatter plots of proportion of individual geological types based on BGS 1:625K Bedrock Geology and BGS 1:625K Superficial Geology Map compared to WFD119 calibration data.

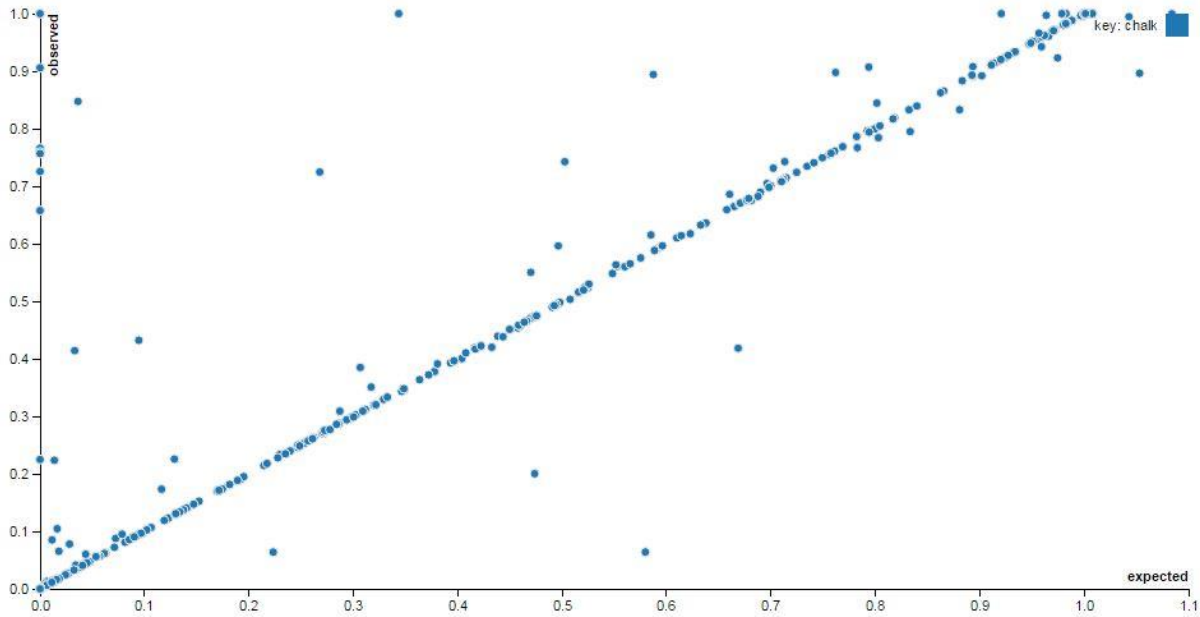


Figure 35 Proportion of chalk RICT category in a catchment at calibration sites (horizontal axis) and as calculated based on accumulation of the BGS Bedrock Geology Map v4 (vertical axis) in Great Britain. Values out of range not shown.

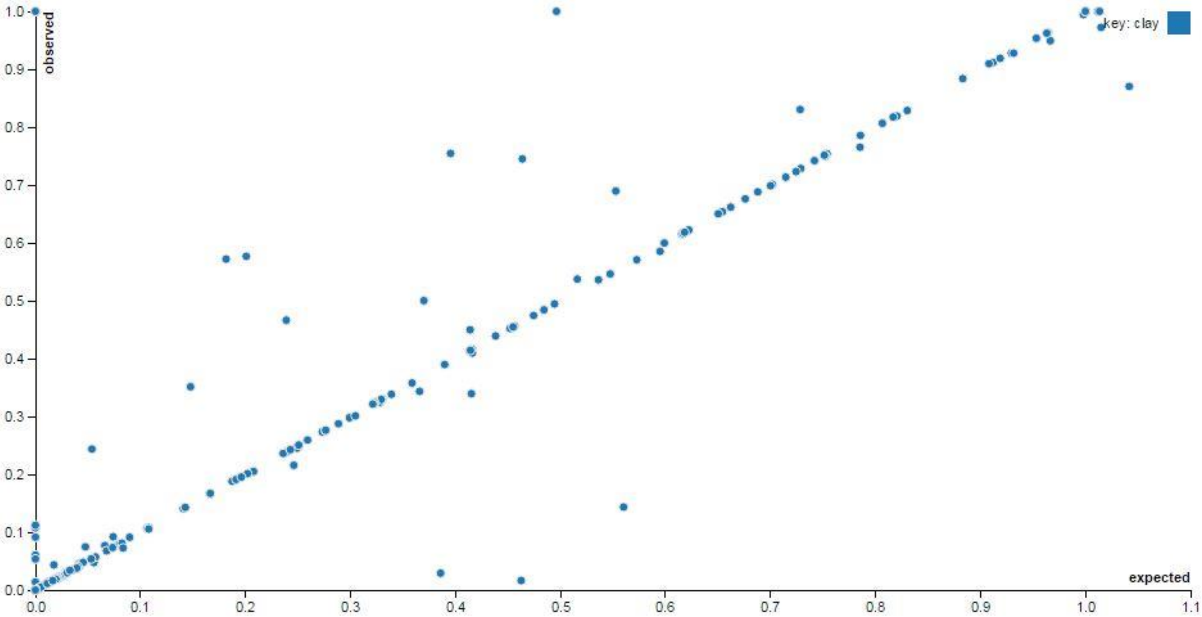


Figure 36 Proportion of clay RICT category in a catchment at calibration sites (horizontal axis) and as calculated based on accumulation of the BGS Bedrock Geology Map v4 (vertical axis) in Great Britain. Values out of range not shown.

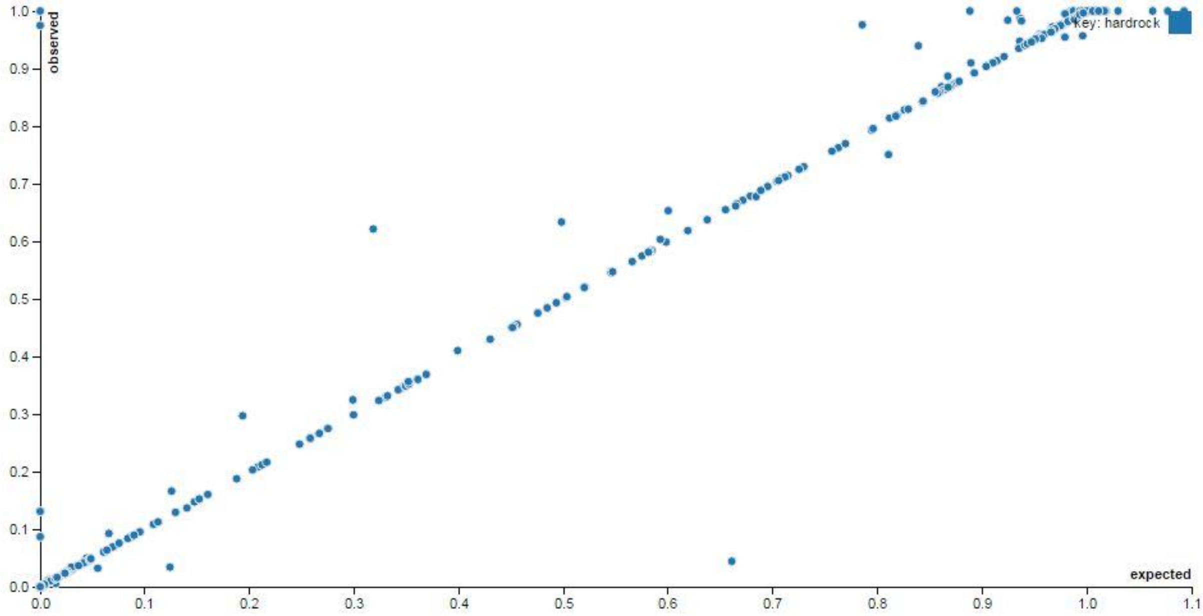


Figure 37 Proportion of hard rock RICT category in a catchment at calibration sites (horizontal axis) and as calculated based on accumulation of the BGS Bedrock Geology Map v4 (vertical axis) in Great Britain. Values out of range not shown.

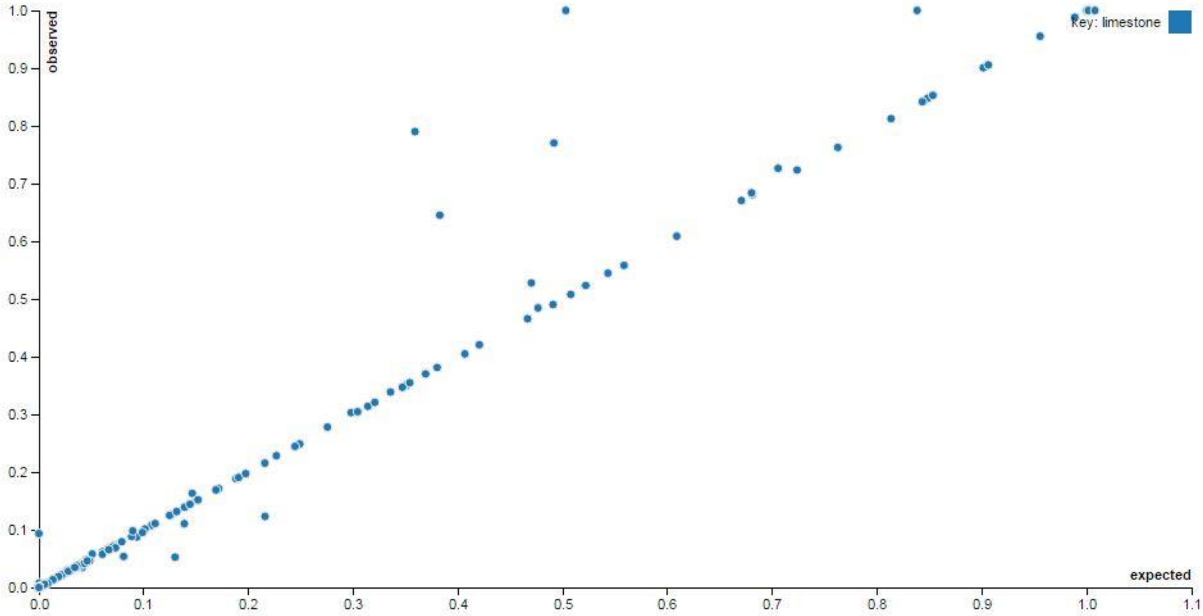


Figure 38 Proportion of limestone RICT category in a catchment at calibration sites (horizontal axis) and as calculated based on accumulation of the BGS Bedrock Geology Map v4 (vertical axis) in Great Britain. Values out of range not shown.

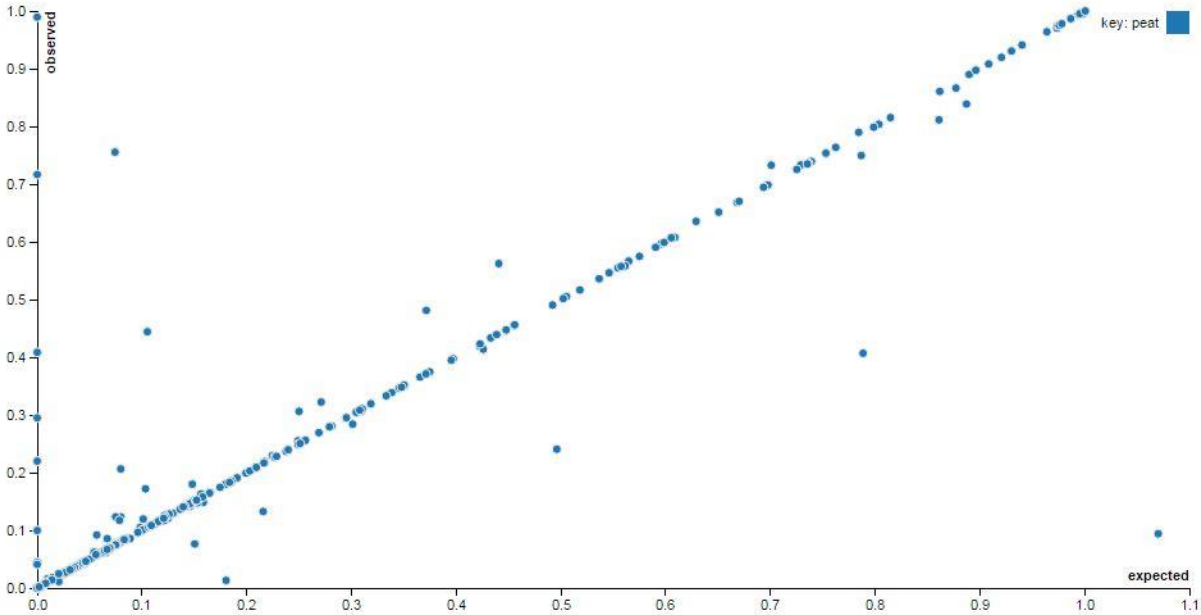


Figure 39 Proportion of peat RICT category in a catchment at calibration sites (horizontal axis) and as calculated based on accumulation of the BGS Bedrock Geology Map v4 (vertical axis) in Great Britain. Values out of range not shown.

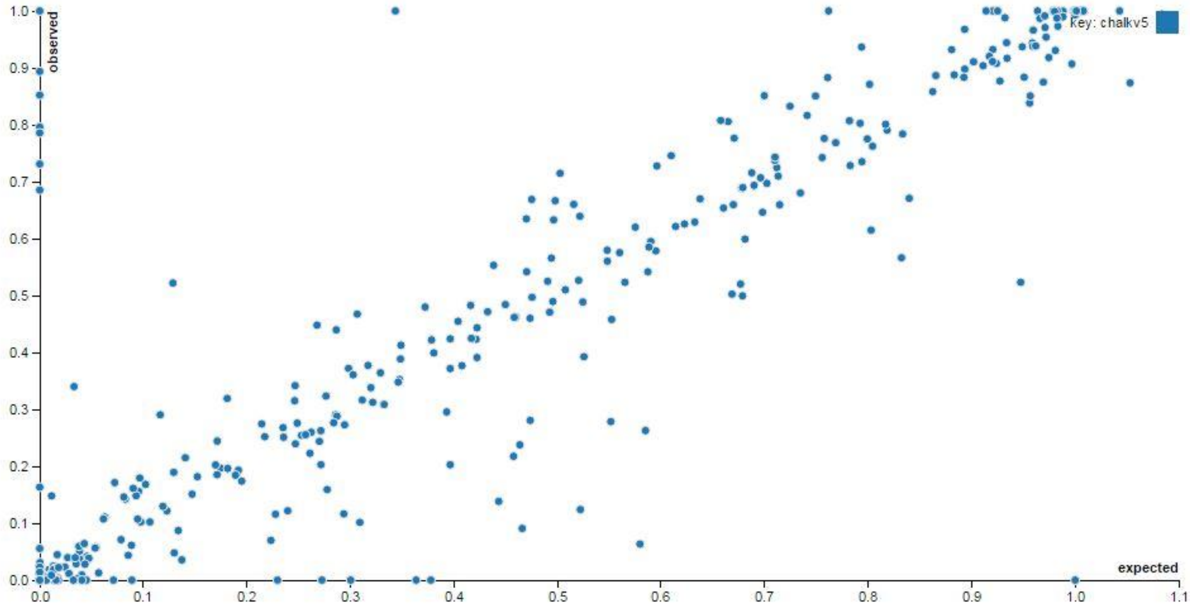


Figure 40 Proportion of chalk RICT category in a catchment at calibration sites (horizontal axis) and as calculated based on accumulation of the BGS Bedrock Geology Map v5 (vertical axis) in Great Britain. Values out of range not shown.

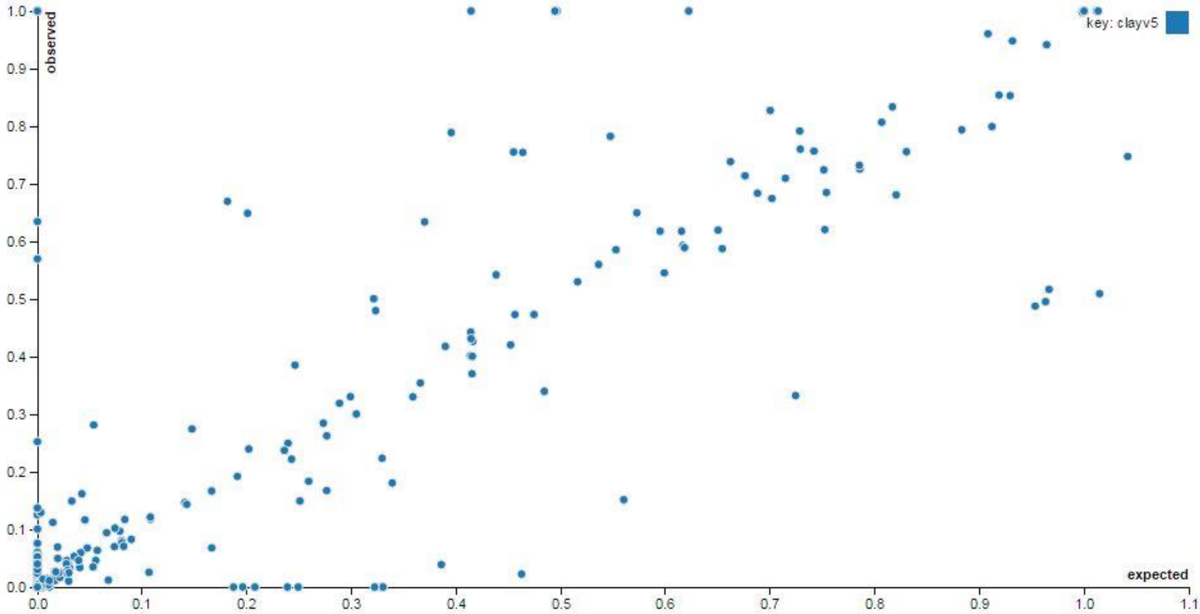


Figure 41 Proportion of clay RICT category in a catchment at calibration sites (horizontal axis) and as calculated based on accumulation of the BGS Bedrock Geology Map v5 (vertical axis) in Great Britain. Values out of range not shown.

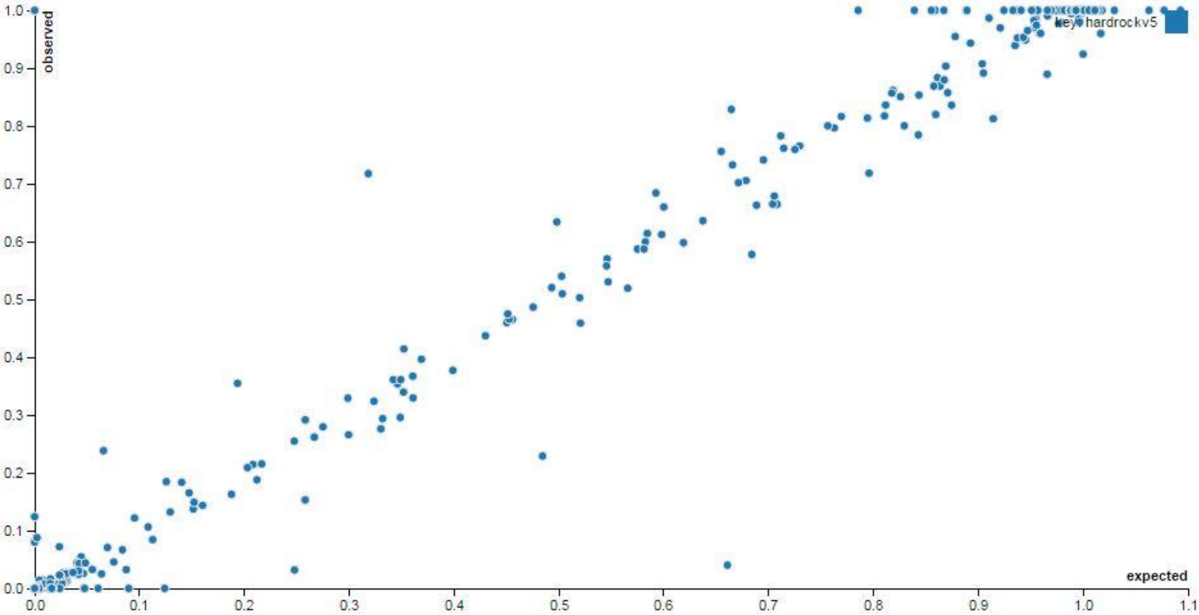


Figure 42 Proportion of hard rock RICT category in a catchment at calibration sites (horizontal axis) and as calculated based on accumulation of the BGS Bedrock Geology Map v5 (vertical axis) in Great Britain. Values out of range not shown.

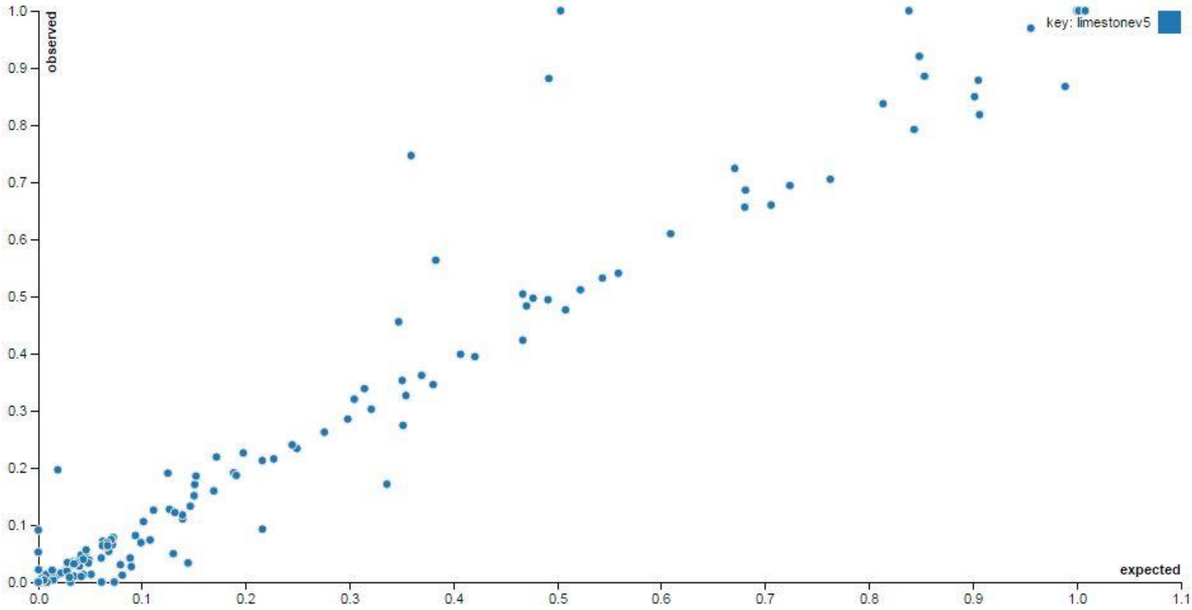


Figure 43 Proportion of limestone RICT category in a catchment at calibration sites (horizontal axis) and as calculated based on accumulation of the BGS Bedrock Geology Map v5 (vertical axis) in Great Britain. Values out of range not shown.

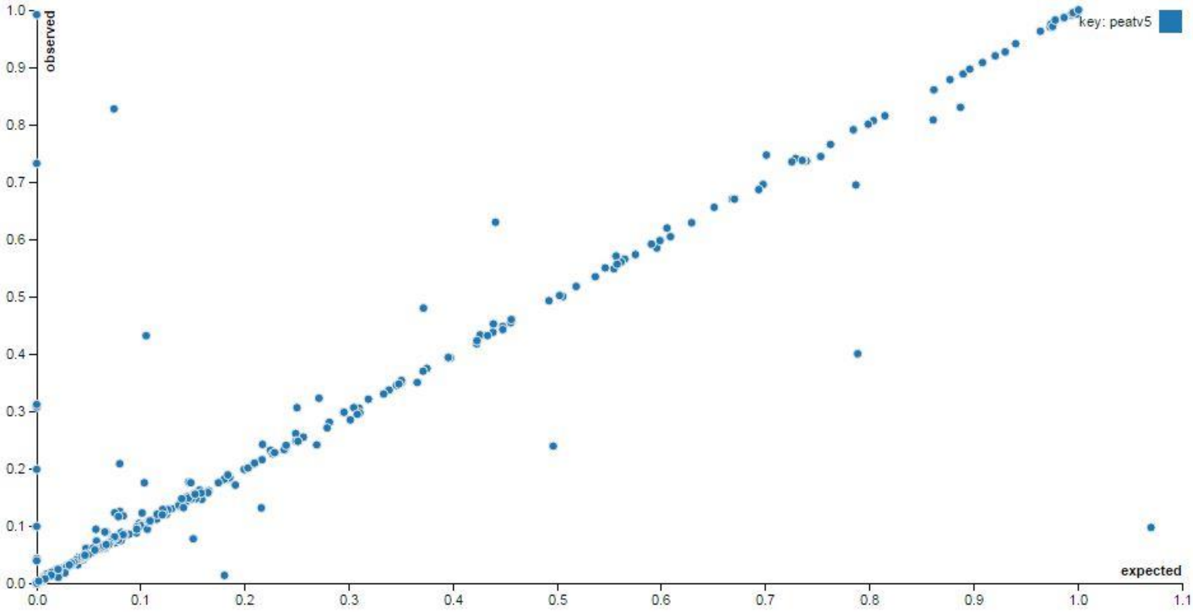


Figure 44 Proportion of peat RICT category in a catchment at calibration sites (horizontal axis) and as calculated based on accumulation of the BGS Superficial Geology Map v5 (vertical axis) in Great Britain. Values out of range not shown.

Appendix 3 Slope calculated using different elevation data

Scatterplots of slope calculated along flow segments using different elevation datasets against slope from WFD119 calibration dataset. There were two calibration datasets available for slope. One was the Slope field in the Sites table included in the RICT Access database (Figure 45 to Figure 49), which was considered the reference source of calibration data for slope. Second was the Slope field in IRN_SiteInfo sheet of RICT_Sites_IRN_fitted_XY_and_IRN_information_CLaize_October_2010.xls (Figure 50 to Figure 53). Notice that the overall match between results based on OS Landform-PANORAMA and the calibration data from the spreadsheet is remarkably good.

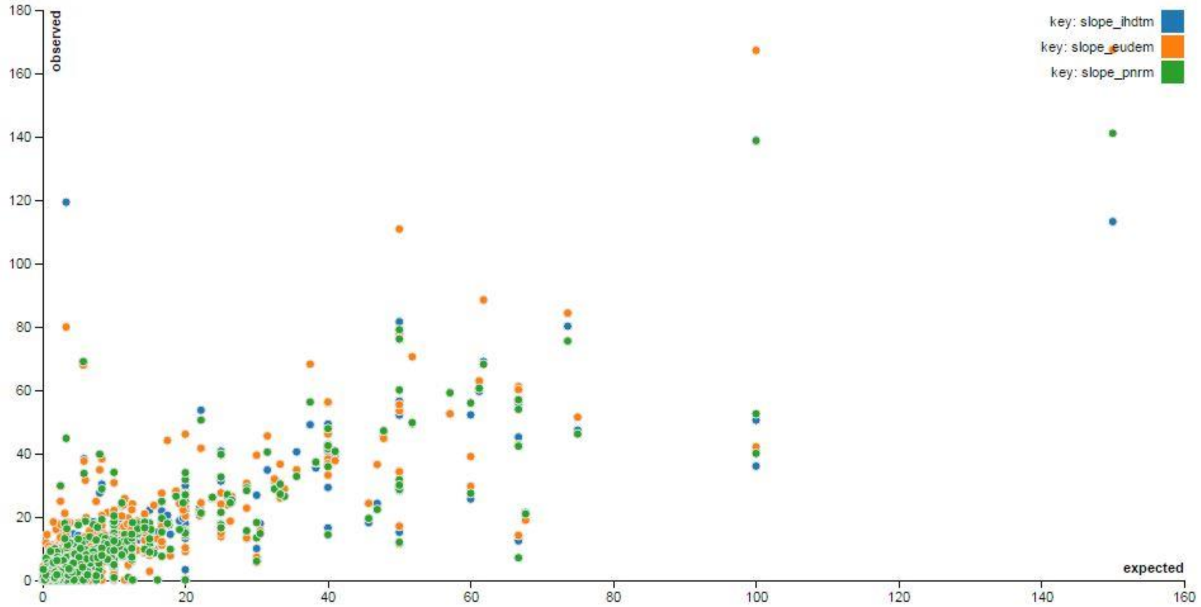


Figure 45 Slope at calibration sites (horizontal axis) and slope calculated along drainage direction grid segments (vertical axis) based on IHDTM HGHT grid (blue), EU-DEM (orange), and OS Landform-PANORAMA (green) in Great Britain.

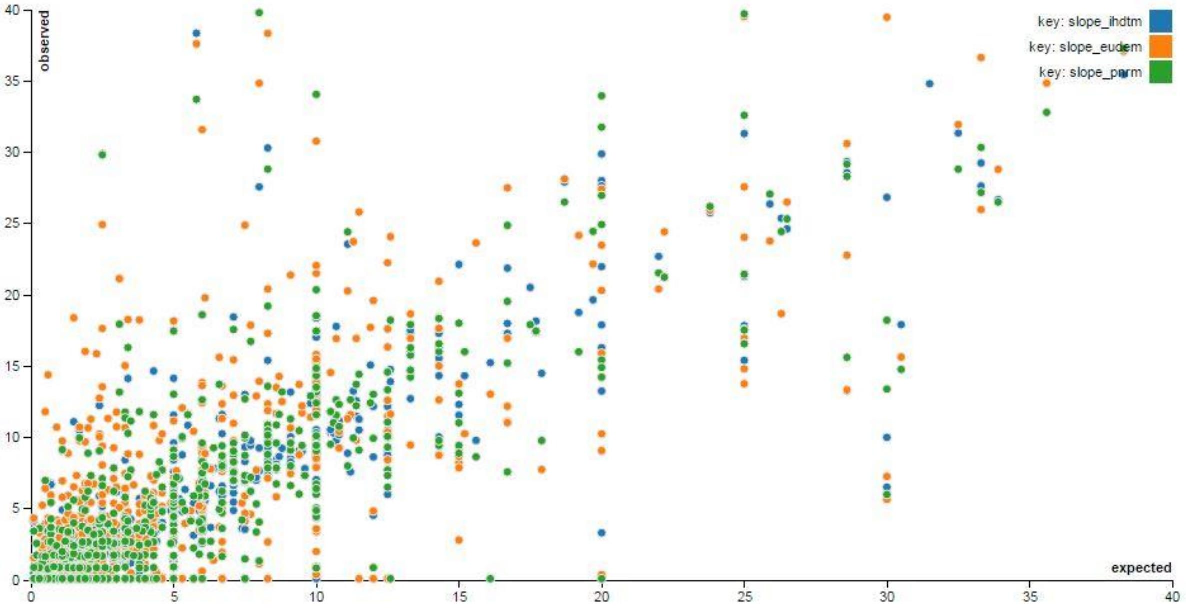


Figure 46 Slope at calibration sites (horizontal axis) and slope calculated along drainage direction grid segments (vertical axis) based on IHDTM HGHT grid (blue), EU-DEM (orange), and OS Landform-

PANORAMA (green) in Great Britain. This plot shows only sites where either slope was less than 40 m per km.

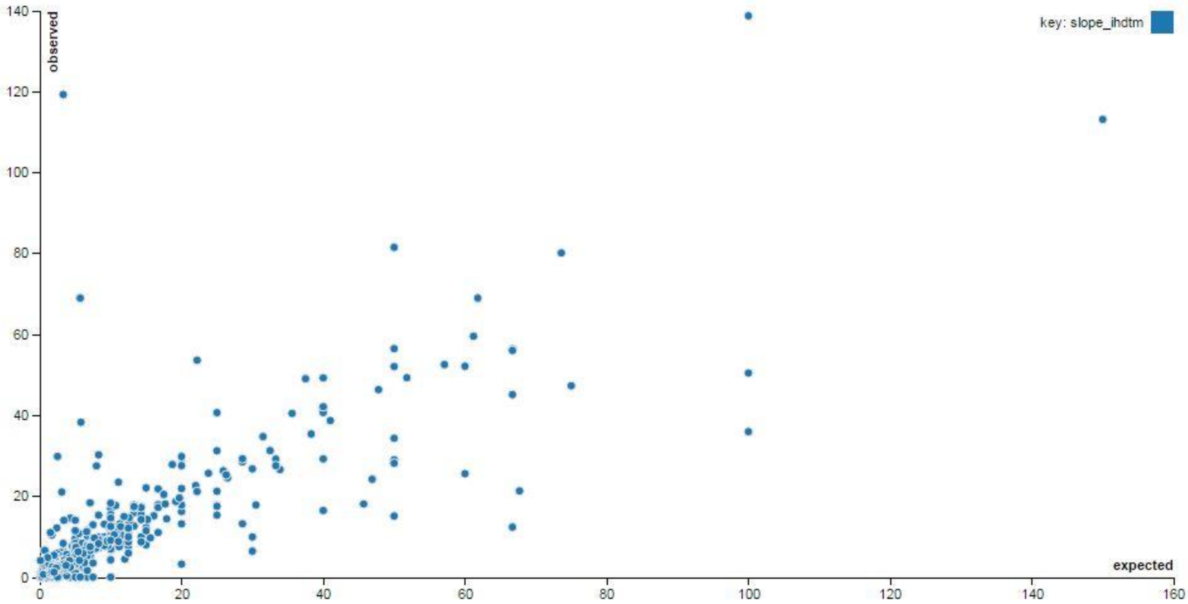


Figure 47 Slope at calibration sites (horizontal axis) and slope calculated along drainage direction grid segments based on IHDTM HGHT grid (vertical axis) in Great Britain.

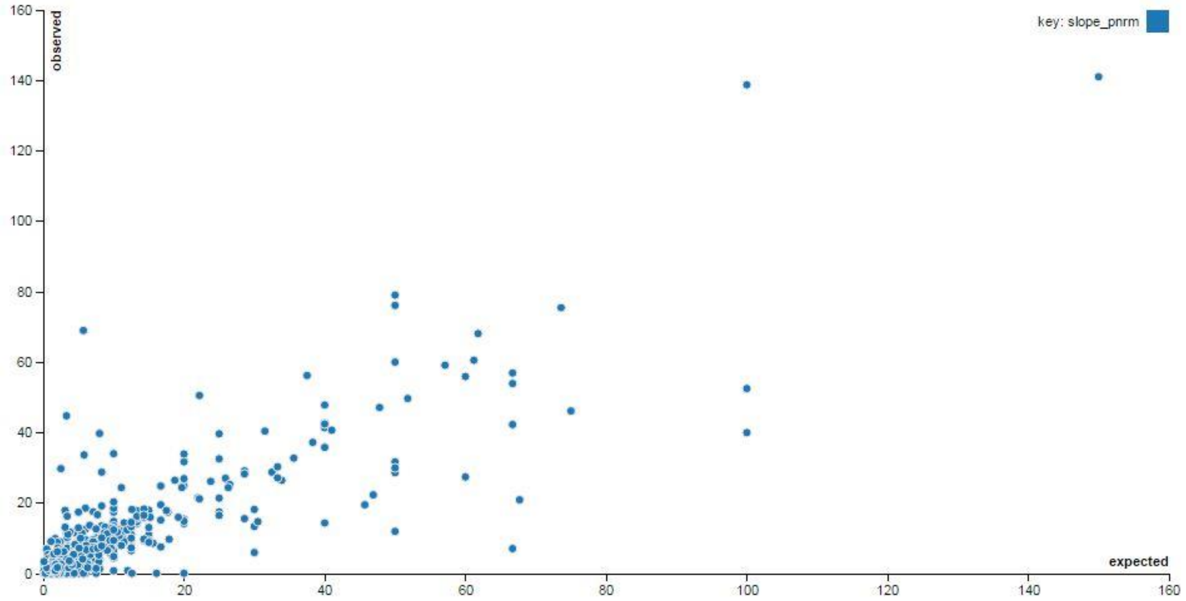


Figure 48 Slope at calibration sites (horizontal axis) and slope calculated along drainage direction grid segments based on OS Landform-PANORAMA grid (vertical axis) in Great Britain.

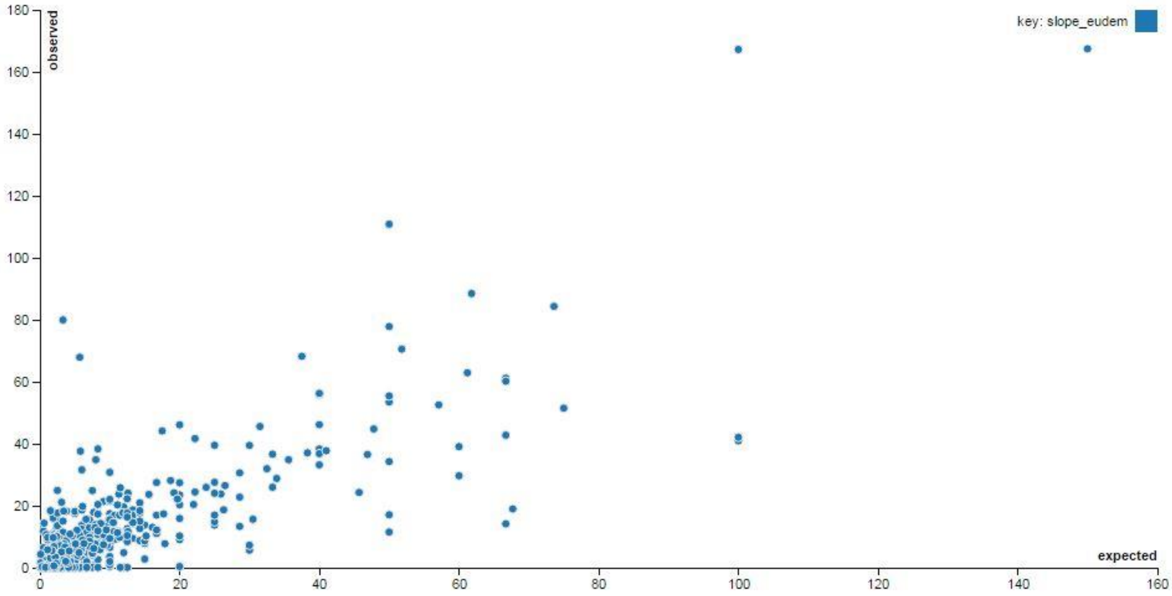


Figure 49 Slope at calibration sites (horizontal axis) and slope calculated along drainage direction grid segments based on EU-DEM grid (vertical axis) in Great Britain.

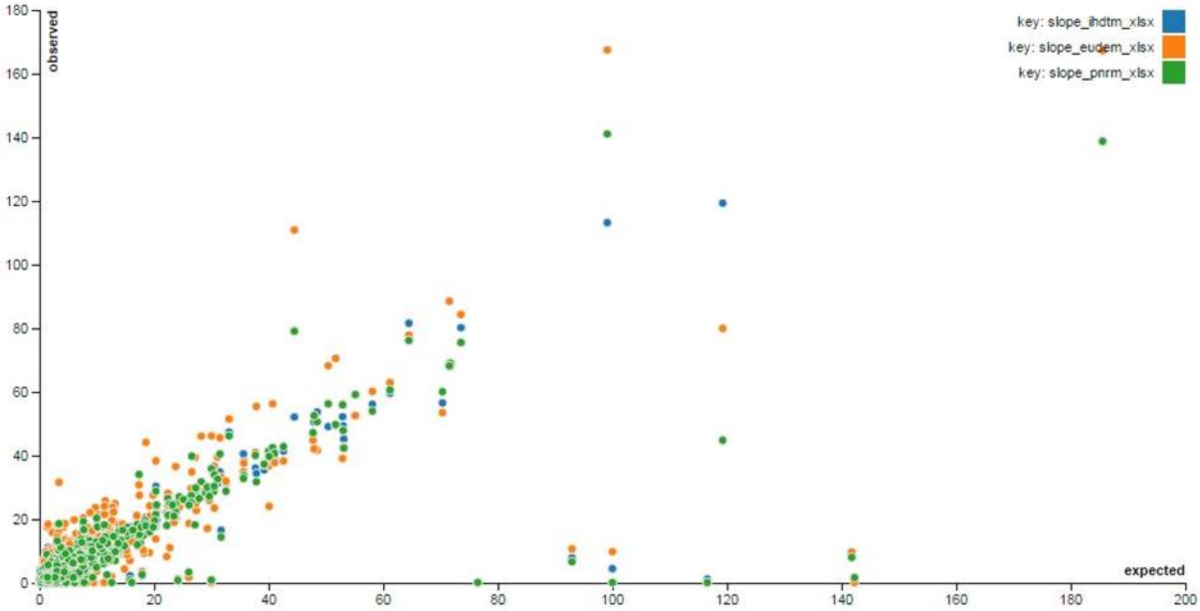


Figure 50 Slope at calibration sites in an Excel spreadsheet (horizontal axis) and slope calculated along drainage direction grid segments (vertical axis) based on IHDTM HGHT grid (blue), EU-DEM (orange), and OS Landform-PANORAMA (green) in Great Britain.

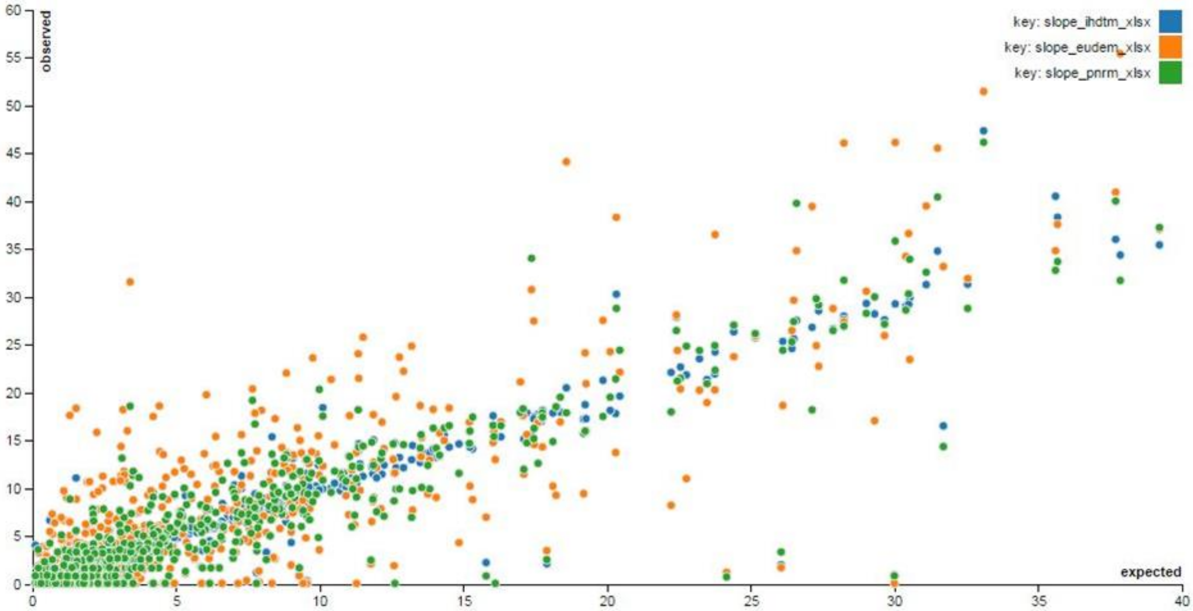


Figure 51 Slope at calibration sites in the Excel spreadsheet (horizontal axis) and slope calculated along drainage direction grid segments (vertical axis) based on IHDTM HGHT grid (blue), EU-DEM (orange), and OS Landform-PANORAMA (green) in Great Britain. This plot shows only sites where either slope was less than 40 m per km.

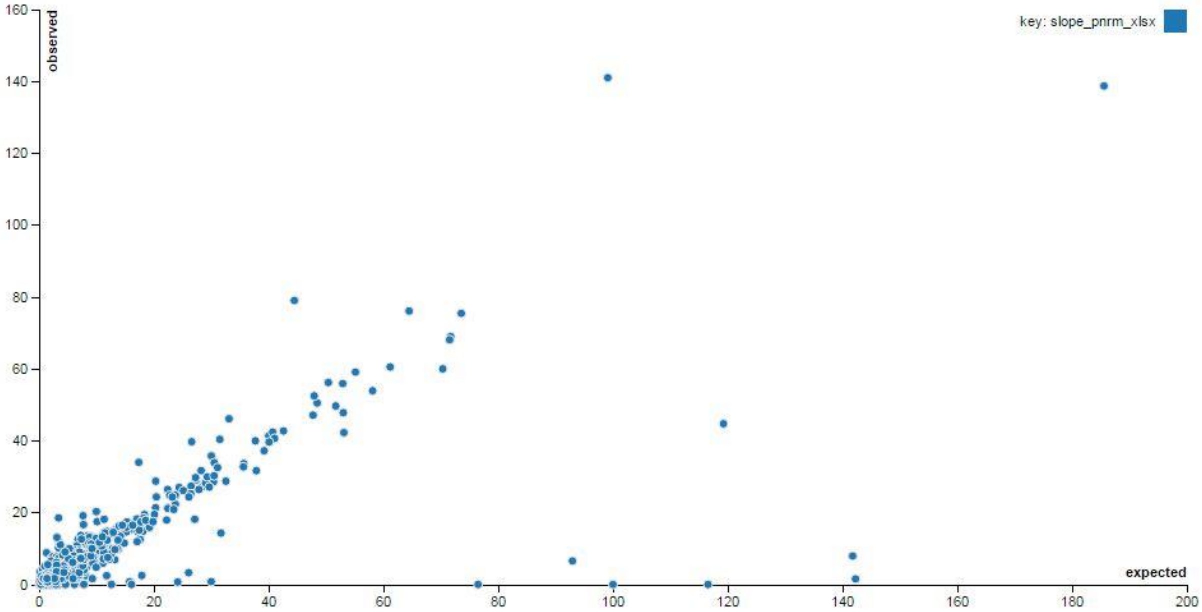


Figure 52 Slope at calibration sites (horizontal axis) and slope calculated along drainage direction grid segments based on IHDTM HGHT grid (vertical axis) in Great Britain.

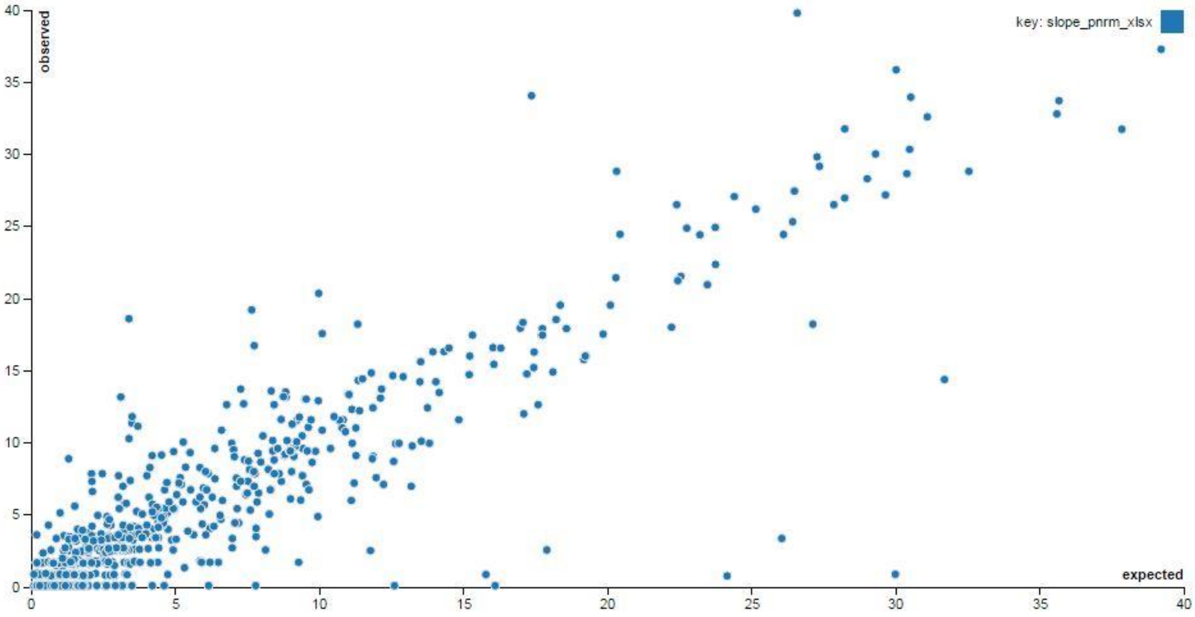


Figure 53 Slope at calibration sites (horizontal axis) and slope calculated along drainage direction grid segments based on IHDTM HGHT grid (vertical axis) in Great Britain. This plot shows only sites where either slope was less than 40 m per km.

Appendix 4 Adjustments to the drainage grid

SEPA identified 21 locations in Scotland where the IHDTM should be improved to better reflect drainage direction as observed by SEPA practitioners (Figure 56). We reviewed these locations and modified the flow direction grid to better match SEPA suggestions, except for four locations where more extensive editing would be required, and four others for which the existing IHDTM flow direction seemed already in line with SEPA suggestions (no change was made).

Once the drainage direction grid had been corrected, cumulative catchment area grid and several ancillary datasets were recalculated, such as a raster of cells downstream from river sources and a feature class of flow segments representing rivers. RICT variables were then calculated using the updated datasets.

Differences in CCAR before and after the edits were checked visually. The biggest change considering upstream catchment area was at Inverie Burn (348666.0, 703788.0, catchment area around 12 km^2, Figure 55). Area upstream from the changes was generally less than 12 km^2, but often a significant length of flow paths downstream from the changes was affected. The change near Raecleugh (360536.0, 651368.0) was the largest considering the length of downstream flow path (Figure 54). It included calibration sites 4913, 4915, 4917, 4979, 4983, 4987, 4991, and 4995. However, the original catchment area was 0.75 km^2 and the update created two even smaller catchments so the impact on the results was relatively small. Other affected calibration sites were FO01, 4017, 4009. It seemed that the issues we were not able to fix would not have major impact because the change in upstream catchment area would be relatively small.

The impact of any change in flow direction can be different in each case. Altitude should not be affected at all. Changes in drainage direction may affect slope in the immediate vicinity of the change, but no further than 500 m away from it. Distance from source is affected, although the length of the new flow route was not always significantly different from the original route. Discharge category is unlikely to be severely affected since the area upstream from the changes was generally less than 12 km^2. The same logic applies to mean catchment altitude (although value of LOGALTBAR for site FO01 66.2 before corrections, 131.3 after corrections and value used for RICT calibration was 66.0; several other variables at FO01 changed significantly too), catchment area, and to some extent to proportion of key geological types.

It was possible to recalculate all RICT variables using the updated flow direction grid, except for PROPWET which was taken directly from the original FEH data.

In summary, the updates in drainage direction grid introduced larger differences from the calibration data in some variables at a few sites and PROPWET could not be updated to reflect the new drainage grid. Despite that, the updated flow direction grid better represents real flow direction and the delivered results are based on the adjusted flow direction grid.

The scatter plots included in this report currently show differences between calibration data and results based on the original IHDTM flow direction grid, not the one with SEPA adjustments.

In addition to the adjustments suggested by SEPA, we attempted to infill estuaries of rivers where the IHDTM was missing far inland (e.g. River Ouse, River Trent, etc.). We removed, added, or changed specific flow direction cells near the edges of the IHDTM to allow semi-automatic infilling of estuaries. The method was semi-automatic in that it required significant manual input to create and validate flow segments for missing cells and we concluded that filling in estuaries across the UK was not achievable in this project.

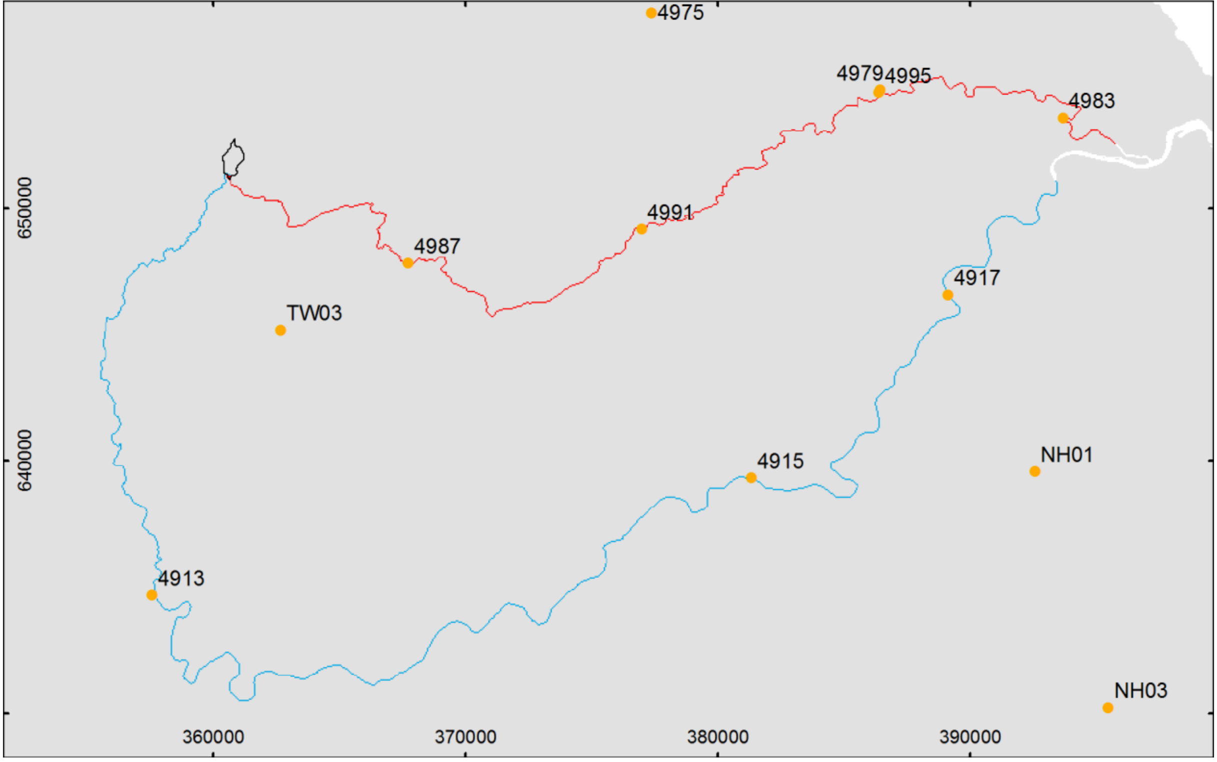


Figure 54 Illustration of changes to introduce corrections suggested by SEPA near Raecleugh. The original catchment area (black outline) was less than 1 sq km. Red cells are cells where cumulative catchment area of the new raster was higher than what the original IHDTM suggested, blue cells are cells where cumulative catchment area of the new raster was lower than the original IHDTM suggested. Orange points are RICT calibration sites labelled with RICT_ID.

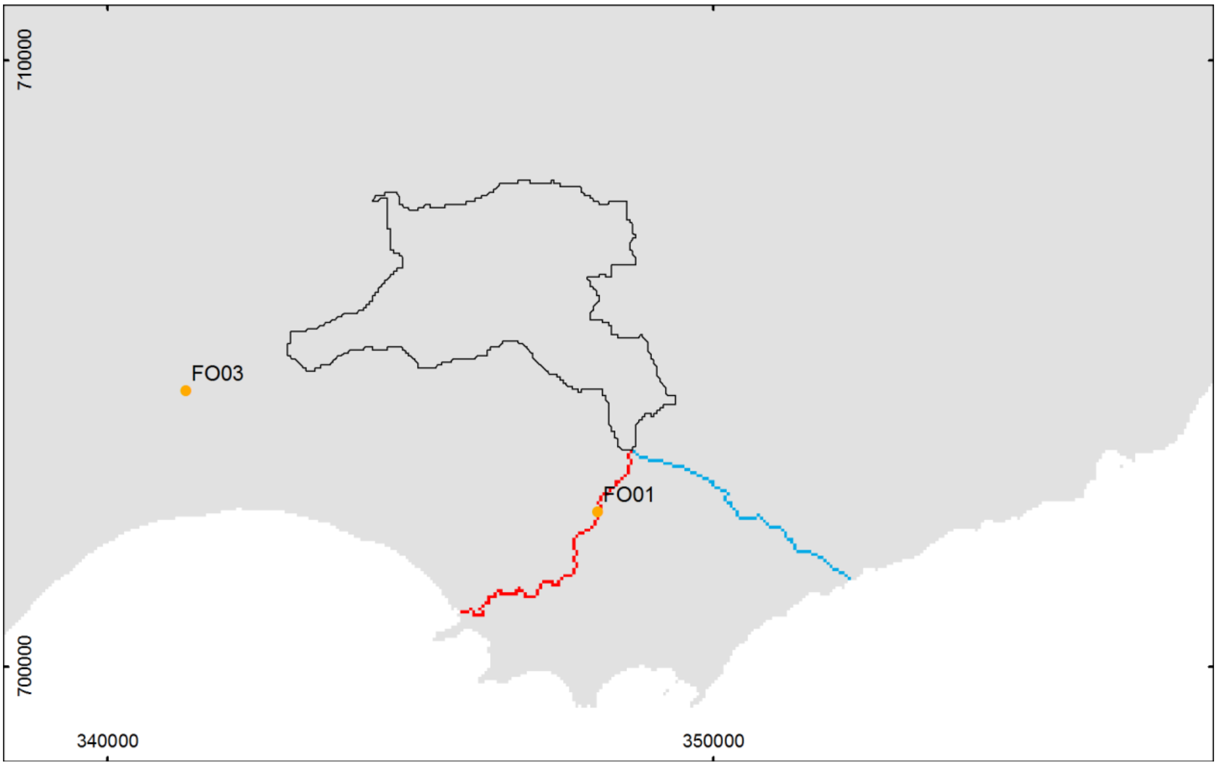


Figure 55 Illustration of changes to introduce corrections suggested by SEPA near Inverie Burn which seemed to have the highest impact on the calibration sites. The original catchment area (black outline) was 12.75 sq km. Red cells are cells where cumulative catchment area of the new raster was higher than what the original IHDTM suggested, blue cells are cells where cumulative catchment area of the new raster was lower than the original IHDTM suggested. Orange points are RICT calibration sites labelled with RICT_ID.

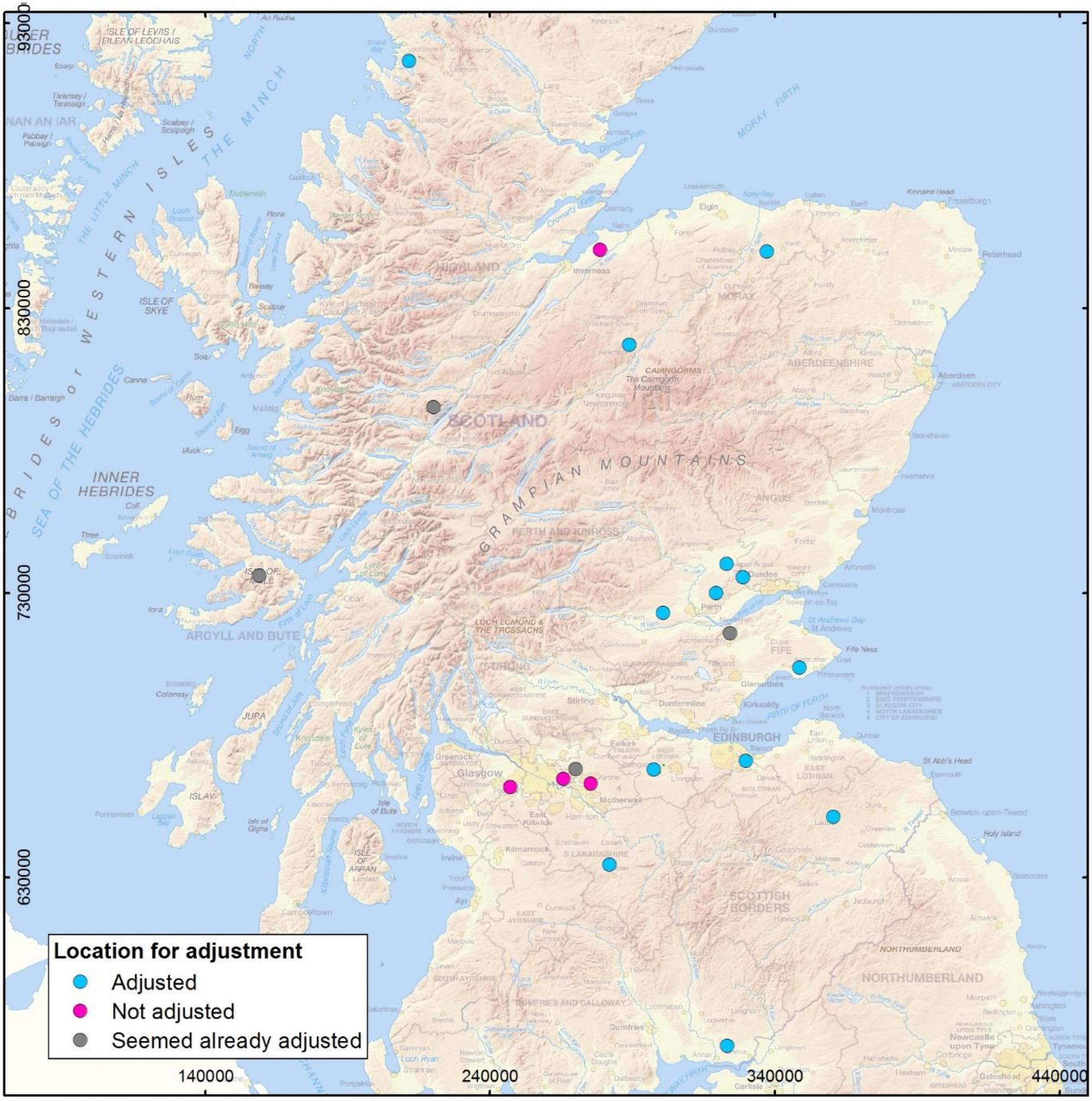


Figure 56 Locations for adjustment of the IHDTM drainage direction grid suggested by SEPA. Background map is Ordnance Survey Miniscale raster.

Appendix 5 Computational regions used for parallel processing

Results for the whole UK were obtained by processing multiple smaller regions at the same time. The regions were IHU Areas without Coastline (Kral et al., 2015), with the exception of area 104, which was split into smaller regions. Northern Ireland was treated as a single region.

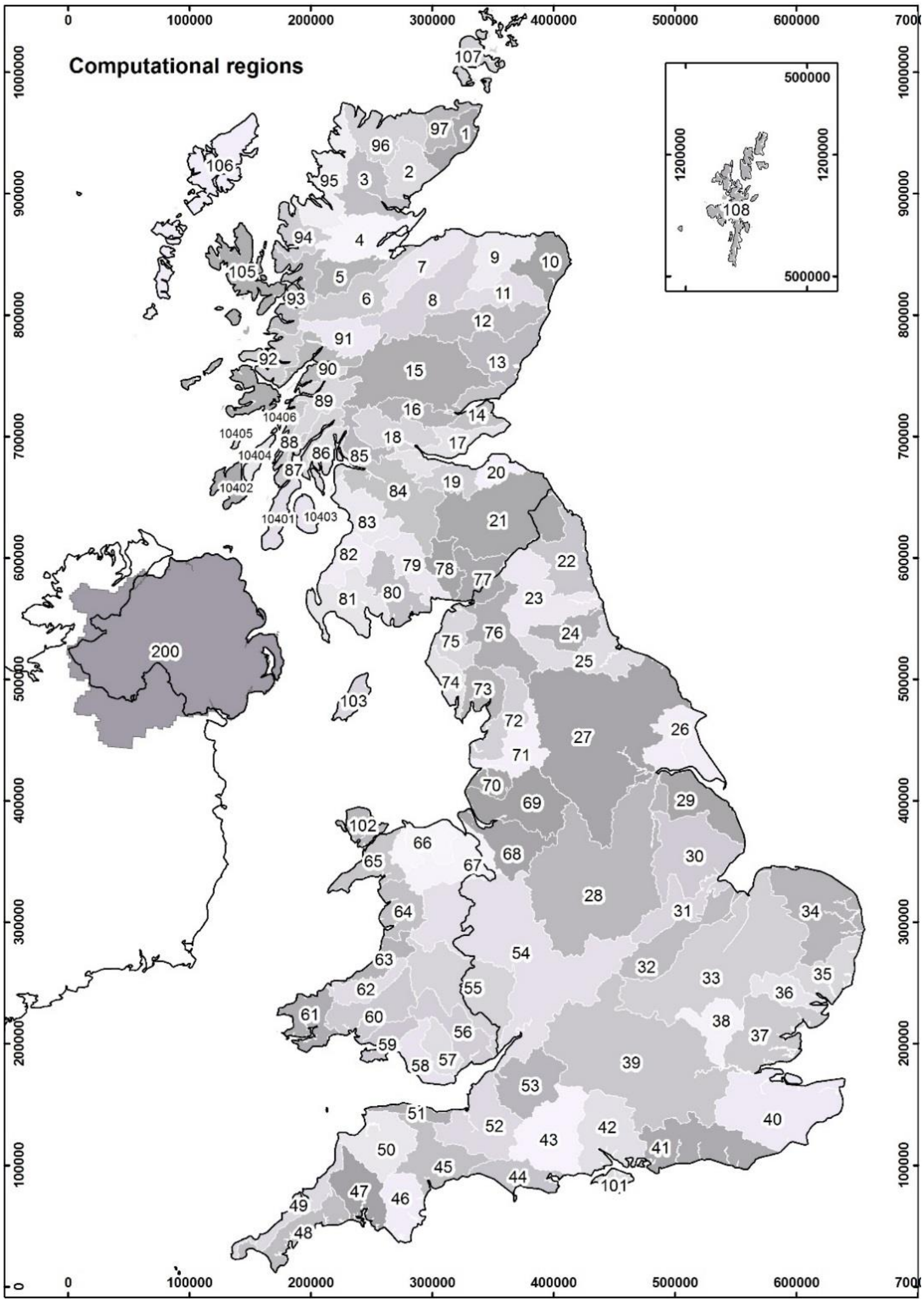


Figure 57 Computational regions used for parallel processing.

9. References

Bell VA, Kay AL, Jones RG, Moore RJ, Reynard NS. (2009) Use of soil data in a grid-based hydrological model to estimate spatial variation in changing flood risk across the UK. Journal of Hydrology, 377 (3-4), 335-350. doi: 10.1016/j.jhydrol.2009.08.031.

Bell VA, Kay AL, Davies HN, Jones RG. (2016) An assessment of the possible impacts of climate change on snow and peak river flows across Britain. Climatic Change, doi:10.1007/s10584-016-1637-x.

Burrough P, McDonnell R. (1998) Principles of Geographical Information Systems, Oxford University Press.

CEH (2016) RIVPACS reference database. Centre for Ecology and Hydrology. N.p., n.d. Web. 16 Aug. 2016. http://www.ceh.ac.uk/services/rivpacs-reference-database

Clarke R, Davy-Bowker J, Dunbar M, Laize C, Scarlett P, Murphy J. (2011) WFD119 Project Final Report: Enhancement of the River Invertebrate Classification Tool. Scotland & Northern Ireland Forum for Environmental Research.

Esri. (1998) Esri Shapefile Technical Description: An ESRI White Paper, Environmental Systems Research Institute, Inc.

Esri. (2016) Flow Length. Help | ArcGIS for Desktop. n.d. Web. 16 Aug. 2016. http://pro.arcgis.com/en/pro-app/tool-reference/spatial-analyst/flow-length.htm

Furse MT, Clarke RT, Winder JM, Symes KL, Blackburn JH, Grieve NJ, Gunn RJM. (1995) Biological assessment methods: Package 1 - The variability of data used for assessing the biological condition of rivers. NRA R&D Note 412. National Rivers Authority, Bristol.

Institute of Hydrology, 1999. Flood Estimation Handbook, 5 volumes and associated software. Institute of Hydrology.

Kral F, Fry M, Dixon H. (2015) Integrated Hydrological Units of the United Kingdom: Groups. NERC Environmental Information Data Centre. http://doi.org/10.5285/f1cd5e33-2633-4304-bbc2-b8d34711d902

Lewis (1994) USE OF MICROLOWFLOWS WITHIN QUASAR (http://nora.nerc.ac.uk/15385/1/N015385CR.pdf).

Morris DG, Flavin RW. (1990) A digital terrain model for hydrology. Proc 4th International Symposium on Spatial Data Handling. Vol 1 Jul 23-27, Zürich, pp 250-262.

OS OpenData. OS OpenData. N.p., n.d. Web. 16 Aug. 2016. (https://www.ordnancesurvey.co.uk/business-and-government/products/opendata-products.html).

SEPA (2016) River invertebrates classification tool. Scottish Environment Protection Agency. N.p., n.d. Web. 16 Aug. 2016. https://www.sepa.org.uk/environment/water/classification/river-invertebrates-classification-tool/

NERC SCIENCE OF THE ENVIRONMENT

PEER

INVESTORS IN PEOPLE

Athena SWAN Bronze Award

BANGOR
Centre for Ecology & Hydrology
Environment Centre Wales
Deiniol Road
Bangor
Gwynedd
LL57 2UW
United Kingdom
T: +44 (0)1248 374500
F: +44 (0)1248 362133

EDINBURGH
Centre for Ecology & Hydrology
Bush Estate
Penicuik
Midlothian
EH26 0QB
United Kingdom
T: +44 (0)131 4454343
F: +44 (0)131 4453943

LANCASTER
Centre for Ecology & Hydrology
Lancaster Environment Centre
Library Avenue
Bailrigg
Lancaster
LA1 4AP
United Kingdom
T: +44 (0)1524 595800
F: +44 (0)1524 61536

WALLINGFORD - Headquarters
Centre for Ecology & Hydrology
Maclean Building
Benson Lane
Crowmarsh Gifford
Wallingford
Oxfordshire
OX10 8BB
United Kingdom
T: +44 (0)1491 838800
F: +44 (0)1491 692424

enquiries@ceh.ac.uk