# RICT (River Invertebrate Classification Tool) Functional Spec v1.1

**November 2019**

Contents

[RICT (River Invertebrate Classification Tool) Functional Spec v0.8 1](#_Toc536799258)

[1. Introduction 5](#_Toc536799259)

[1.1 Background 5](#_Toc536799260)

[1.2 Document Version History 6](#_Toc536799261)

[1.3 Purpose of the document 6](#_Toc536799262)

[2. Functional Overview 7](#_Toc536799263)

[3. Supporting Data Tables and R functions 11](#_Toc536799264)

[4. Starting the Project Experiment 13](#_Toc536799265)

[4.1 Introduction 13](#_Toc536799266)

[4.2 Default values 13](#_Toc536799267)

[4.3 Future Experiments 13](#_Toc536799268)

[5. Acceptance and validation of input data 15](#_Toc536799269)

[5.1 Environmental Variable Values and Validation 15](#_Toc536799270)

[5.2 Observed Biological Data Values and Validation 21](#_Toc536799271)

[6. Conversion of Input Environmental Variables 22](#_Toc536799272)

[6.1 Deriving Latitude and Longitude from National Grid Reference for Great Britain Model sites 22](#_Toc536799273)

[6.2 Deriving Latitude and Longitude from National Grid Reference for Northern Ireland Model Sites 22](#_Toc536799274)

[6.3 Calculating Long Term Average Mean and Range of Air Temperature from Latitude and Longitude 23](#_Toc536799275)

[6.4 Calculating Mean Substratum Composition - MSUBST (in phi units) 23](#_Toc536799276)

[6.5 Estimating Discharge Category (if unavailable) from Stream Velocity 24](#_Toc536799277)

[6.6 Estimating Alkalinity (if unavailable) from Hardness, Calcium or Conductivity 24](#_Toc536799278)

[6.7 Transform Environmental Variables to Logarithm Base 10 Form 25](#_Toc536799279)

[6.8 Validation of Derived Environmental Variables 26](#_Toc536799280)

[7. Predict 27](#_Toc536799281)

[7.1 Predict Process 27](#_Toc536799282)

[7.2 Prediction Overview 28](#_Toc536799283)

[7.3 Environmental Variables 28](#_Toc536799284)

[7.4 Discriminant Functions (DFs) and the Co-efficients 28](#_Toc536799285)

[7.5 Calculate the Discriminant Function Scores (Step 4 in diagram above) 28](#_Toc536799286)

[7.6 Calculate Mahanalobis Distance (Step 5 in diagram above) 29](#_Toc536799287)

[7.7 Use Mahanalobis Distance to Determine Probability of End Group Membership (Step 6 in diagram above) 30](#_Toc536799288)

[7.8 Determine Environmental Suitability of the Test Site for Prediction (Step 7 in diagram above) 30](#_Toc536799289)

[7.9 Determine the Predicted Values of Indices (Step 8 in diagram above) 31](#_Toc536799290)

[7.10 Determine the Taxa Predicted to be Present at the Site 33](#_Toc536799291)

[7.11 Completion and Outputs of Prediction Processing 35](#_Toc536799292)

[8. Classify 36](#_Toc536799293)

[8.1 Classify Process Summary 36](#_Toc536799294)

[8.2 Classification overview 36](#_Toc536799295)

[8.1 Expected Values have been calculated by the Prediction Process during the Current Run (Step 1 in process diagram above) 37](#_Toc536799296)

[37](#_Toc536799297)

[8.2 Adjust Expected Values to Standardise against the High/Good Boundary (Step 2 in process diagram above) 37](#_Toc536799298)

[8.3 EQR Conversion Factor (Step 3 in process diagram above) 38](#_Toc536799299)

[8.4 Convert adjusted expected to reference value (Step 4 in process diagram above) 39](#_Toc536799300)

[8.5 Simulate uncertainty in expected (Step 5 in process diagram above) 39](#_Toc536799301)

[8.6 User input of observed index value and bias; calculation of bias adjustment (step 6 in process diagram above) 41](#_Toc536799302)

[8.7 Simulate uncertainty in observed values (step 7 in process diagram above) 41](#_Toc536799303)

[8.8 Simulate variation in Bias (step 8 in process diagram above) 42](#_Toc536799304)

[8.9 Correct observed values for bias (Step 9 in process diagram above) 43](#_Toc536799305)

[8.10 Calculate Environmental Quality Ratio (EQR) (Step 10 in process diagram above) 44](#_Toc536799306)

[8.11 Combine spring and autumn EQRs (Step 11 in process diagram above) 45](#_Toc536799307)

[8.12 Determine the probability of the class being High, Good, Moderate, Poor and Bad status (Step 12 in process diagram above) 45](#_Toc536799308)

[8.13 Determine the worst class indicated by WHPT ASPT and WHPT Ntaxa using MINTA (step 13 in process diagram above) 46](#_Toc536799309)

[8.14 Determine overall status class for a water body based on samples collected from more than one site 47](#_Toc536799310)

[9. Multi Year Classification 48](#_Toc536799311)

[10. Production of Output files 49](#_Toc536799312)

[11. Compare Project Experiment 50](#_Toc536799313)

[12. Model 44 Prediction and Classification 53](#_Toc536799314)

[13. Non-Functional Requirements 56](#_Toc536799315)

[13.1 Operational Support 56](#_Toc536799316)

[13.2 Look and Feel of User Interface 56](#_Toc536799317)

[13.3 Input and Output Files 56](#_Toc536799318)

[13.4 Hosting 56](#_Toc536799319)

[13.5 Accessibility 56](#_Toc536799320)

[13.6 Performance 57](#_Toc536799321)

[13.7 System Administration 57](#_Toc536799322)

[13.8 Data Protection 57](#_Toc536799323)

[Annex A: Input File 58](#_Toc536799324)

[Annex B: Output File 60](#_Toc536799325)

[Annex C: Critical Success Factors 65](#_Toc536799326)

## Introduction

### Background

The UK Regulatory Agencies have a requirement to classify ecological quality in rivers using data from invertebrate samples. The RIVPACS (River Invertebrate Prediction and Classification Tool) model and software tool, which was developed on behalf of the agencies by the Freshwater Biological Association (FBA) and subsequently by the Centre for Ecology and Hydrology (CEH), was originally used for this purpose. It incorporates a set of summarised reference data from sites assumed to be the best available of their type.

The EU Water Framework Directive (WFD) introduced in 2000, included modifications to the regulatory requirements in relation to classification and therefore the RIVPACS software processing tool had to be amended.

The original RIVPACS software tool was written in FORTRAN and had been developed in stages over a number of years. As a result, it was not written in a modern programming language and did not have flexibility to allow changes without significant cost. Therefore, a new tool was developed that allowed future changes to be made with minimal cost in many cases (e.g. the addition of a new index). This was the River Invertebrate Classification Tool (RICT) which went live in 2007. This version of RICT was a web-served application which was accessed via a website hosted by SEPA. The SEPA website also hosted copies of the manual, guidance and background documents.

RICT has had some updates since 2007, including the addition of ‘split season’ classification to enable WHPT (Walley, Hawkes, Paisley and Trigg) calculations using seasonal average EQRs (Ecological Quality Ratio). RICT was managed and hosted by Scottish Environment Protection Agency (SEPA) but is also used by Environment Agency (EA), Natural Resources Wales (NRW) and Northern Ireland Environment Agency (NIEA) as part of their regulatory and statutory requirements as well as public users such as academic institutions.

However, a number of issues were identified with this version of RICT which can be found in the outline business case produced by the scoping project. These include improving the performance, support arrangements and resilience as well as adding the new RIVPACS model (RIVPACS version IV) to the RICT tool, resolving critical bugs and other additional functionality. The full list of Critical success factors can be found in Annex C.

A shortlist of options to resolve the issues was discussed which resulted in the recommendation that RICT should adopt a Platform as a Service arrangement and run a modified application.

A new cloud-based open source version of RICT has been developed which will incorporate the current RIVPACS IV model and WFD status classification and uses the same basic algorithms to provide the same final results (other than correction of errors). Code and data tables for the new version will be more accessible to developers who wish to further develop RICT and their own project experiments. This new version of RICT is called RICT2 (version 1).

The platform chosen for RICT2 is Microsoft Azure Machine Learning (ML) Studio with open source R programming language as the development choice. Users will need to create a Microsoft Azure ML Studio account to allow them to access the RICT2 software tool. The majority of users will be able to use the free account subscription provided which allows adequate processing time for the number of runs required for a typical user. For more intensive RICT2 users, a monthly subscription account can be used on an ad-hoc basis, i.e. a user could subscribe for a month during the month when intensive processing is required, then unsubscribe the following month. This will keep costs minimal. Further, advanced users can create their own experiments and edit the R code or the pipeline process to suit their own computing needs.

The RICT2 service is accessible via ML with the latest documentation including the technical specification and user guide being hosted on a SharePoint site containing the link to the RICT2 tool along with information and guidance.

### Document Version History

|  |  |  |
| --- | --- | --- |
| **Version No.** | **Purpose** | **Date** |
| (previous functional spec 1.0) | Initial version of the 2007 detailed functional specification issued to Steering Group and CEH | 06/08/2007 |
| RICT\_New functional spec v0.1 | Updated and amended version of the previous functional specification | 02/07/2018 |
| RICT\_New functional spec v0.2 | Updated and amended version of the new functional specification following review by John Murray Bligh. | 18/07/2018 |
| RICT\_New functional spec v0.3 | Updated and amended version of the new functional specification following review by Technical Group members, decisions made during Technical group discussion and clarification via emails. | 16/08/2018 |
| RICT\_New functional spec v0.4 | Updated and amended version of the new functional specification following discussions held on 30th August to confirm the classification process in RICT2. | 07/09/2018 |
| RICT\_New functional spec v0.5 | Further amended version with questions to highlight areas that need input for meeting on 12/12/18 | 11/12/2018 |
| RICT\_New functional spec v0.6 | Further amended version following some clarifications at meeting on 12/12/18 | 24/12/18 |
| RICT\_New functional spec v0.7 | Updated and amended version following clarification meetings held over w/c 14th Jan. Sent out for review on 23/01/19. | 23/01/19 |
| RICT\_New functional spec v0.8 | Updated and amended version following review. | 01/02/19 |
| RICT\_New functional spec v0.9 | Update and amended following developer review | 12/02/19 |
| RICT\_New functional spec v0.10 | Update following comments from SEPA Lead | 23/04/2019 |
| RICT\_New functional spec v1.1 | Updated to reflect the addition of Summer calculations | 28/11/2019 |

**Note**: Yellow highlight within the document indicates functionality that is not yet developed but has been identified after the initial Project experiment for WFD classification has been released.

### Purpose of the document

The purpose of this document is to define the business, functional and non-functional capabilities of RICT2 to support the design and implementation of RICT2 software tool.

The details will be sufficient to enable the software tool to be designed and built without having to continually refer queries to business representatives.

Further, it will be of sufficient detail to provide supporting documented information for any future RICT development.

## Functional Overview

RICT2 is an application that implements the RIVPACS IV predictive model. The primary requirement of the tool is to be able to allocate WFD (Water Framework Directive) compliant classifications to rivers with regard to the quality of the river invertebrate community.

The classifications are calculated and described as H = high, G = good, M = moderate, P = poor or B = bad. This is to be achieved by comparing sample data from a river site against sample data previously taken from river sites considered to be the best available of their type.

The primary functions of the RICT2 tool are:

1. **Predict** - Predict the value of various biotic indices and the presence and abundance of invertebrate species, genera or families that you would expect at any place on any stream or river in the UK if it was not environmentally disturbed.
2. **Classify** - Determine the Water Framework Directive (WFD) quality status class and statistical information about its confidence.
3. **Compare** - Compare the statistical significance of differences between classifications.

It is to be capable of doing this for one or more sites at a time, with the results provided in an output file for subsequent processing by the relevant user.

Each operation of the tool is carried out within a project experiment which may cover anything from a single site up to all sites within the UK that require classification.

Standard experiments with default settings will be provided and made available on Microsoft Azure ML studio. RICT2 experiments will be collated as a “Collection” in Microsoft Azure Learning Studio to help users find all the experiments relevant to RICT2 as per the link below:

<https://gallery.azure.ai/Collection/River-Invertebrate-Classification-Tool-RICT-Current-Versions>

Users can also download the R software to their own work area and run the experiment locally. The user can amend the R software code to create a new experiment to suit their own requirements.

However, it must be noted that the version of R software studio on Microsoft Azure ML studio, at the time of writing this document, are currently older (latest being 3.4.4) than the latest version of 3.6.1 available. Therefore, if the experiment is downloaded and the user does some amendments to the R code using their latest version of the R studio software, it may have compatibility issues if this experiment is subsequently uploaded back into Microsoft Azure ML studio.

It should be noted that this document is written from the expectation that the user will be running the experiment in Microsoft Azure ML studio.

For initial release, the GB Single year, GB multi-year and NI single year for WFD classification have been produced. These three experiments are available in the Azure ML Studio gallery.

The diagram below shows a diagrammatical representation of the main functionality of the RICT2 tool with numbered links to more detailed information.



**LOGIN TO RICT:** A new user will need to create a Microsoft account and log in to be able to access to the RICT tool as the tool uses the Microsoft Azure Machine Learning Studio as its platform.

How to access RICT is outlined in the RICT user guide.

**COMPLETE DATA INPUT:** User will need to complete a CSV (comma delimited) input file containing the data for the run required. A standard template excel spreadsheet will be provided to enter this data which can be downloaded via a SharePoint link from the website.

Validation will be done on the spreadsheet as the user enters the data via conditional formatting to warn the user if the data input will not meet certain validation limits.

Annex A of this document has a copy of the data input table in more detail.

**PROVIDE DATA:** User will need to upload data into the RICT2 Tool and link the data to the project experiment in Microsoft Azure ML Studio.

Support file data tables and R code support functions will be contained within Microsoft Azure ML Studio and will be linked to the project experiment.

Section 3 of this document describes the support file data tables and R code support functions in more detail.

**TRIGGER RUN:** User can choose to run a project experiment for Predict and Classify experiment using the GB (Great Britain) model or a Predict and Classify experiment using the NI (Northern Ireland) model using default values.

Users can also choose to download and amend the standard default experiments, including amending the support file data table values and create their own experiments in their own work area to run.

Section 4 of this document describes the default values used in more detail.

**VALIDATE DATA:** A series of validation checks to ensure data provided meets specified requirements will be carried out using R code in RICT2.

Section 5 of this document describes the validation steps in more detail.

**PREDICTION:** Tool will perform a series of steps in the project experiment to calculate Prediction:

* Convert Environmental Data
* Calculate Probability of End Group Membership
* Calculate Environmental Suitability
* Calculate Predicted Index Values
* Calculate Predicted Taxa

Section 6 and 7 of this document describe the conversion and prediction steps in more detail.

**CLASSIFICATION:** Tool will perform a series of steps in the project experiment to calculate Classification:

* Calculate reference values from expected values
* Apply bias to Observed Values
* Run simulations to take account of errors in observed and expected
* Calculate EQR for all simulations
* Average spring and autumn EQR for all simulations
* Calculate probability of class from the simulations

Section 8 of this document describe the classification steps in more detail.

**COMPARE RUN:** Tool runs processing steps to compare data outputs from two previous predict and classify runs to produce a compare output file report.

This functionality has been considered but is not yet implemented in RICT.

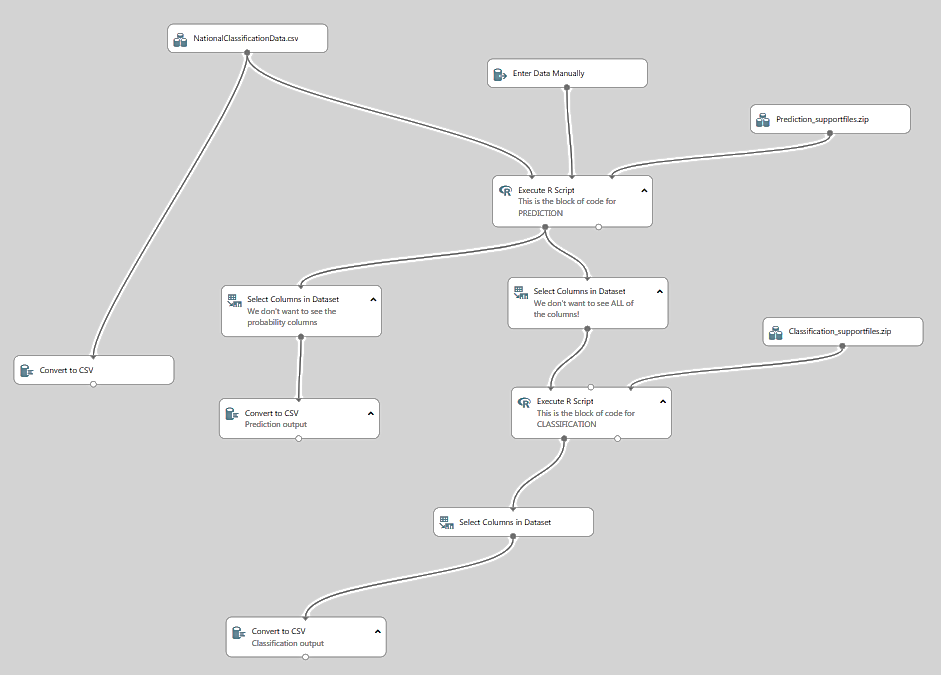
**OUTPUT FILE:** RICT2 will generate an output file with processed data for Predict and Classify in csv format for the user to download. This data extract can then be used for analysis and reporting.

Annex B of this document has a copy of the data output table in more detail.

Users can also choose the R Device dataset option in Machine studio to Visualise the data and if required could create code to create the output file in the format they choose.

**RICT Screenshot**

A screenshot of the RICT2 Project Experiment in Microsoft Azure ML Studio being used with input file from the National Classification is shown below:



## Supporting Data Tables and R functions

The tool uses RIVPACS reference data which contain sites that are considered to be the best available of their type. The current RIVPACS IV data model has been used for this version of RICT2 and will be used for WFD river status classification.

Here is a link to the complete RIVPACS reference database:

<https://www.ceh.ac.uk/services/rivpacs-reference-database>

Within RICT2, the RIVPACS reference data is not used at site level but is used at the End Group level, where an End Group contains combined data for sites considered to be of a similar biological composition. End Groups are grouped up to End Group Sets (e.g. New GB – 43 Group level).

The End Group table used in RICT2 is called *x103 End Group Means (FORMATTED)* and can be found in the Prediction\_supportfiles folder in the RICT2 experiment.

However, for the initial release of RICT2, only relevant data from this table is being used for the standard WFD classification. This being the data for Spring, Summer and Autumn (Season ID 1, 2 and 3), RIVPACS model ID, End Group ID and the data in the columns for TL2 WHPT NTAXA (AbW.DistFam) and TL2 WHPT ASPT (AbW,Dist

Fam).

As well as this data table, RICT2 needs other data which will be held and used as data input tables and R software code functions in the project experiment. These can be found in the folder identified as Prediction support files and Classification support files in the project experiment.

A list of the tables and functions required for the **GB model experiment** is as below.

**Prediction support files:**

* Rnrfa\_1.4.0 folder – zip folder which contains a set of utility functions to retrieve data from the UK National River Flow Archive (NRFA). The package contains R wrappers to the UK NRFA temporary API. There are functions to retrieve stations falling in a bounding box, to generate map and extracting time series and general information.
* AirTempGrid – Excel csv which contains previously derived Environmental data values for the Mean Air Temperature (TEMPM) and Temperature range (TEMPR) for points in the centres of 5km interval grid squares. Values are used after converting the supplied NGR to latitude and longitude to identify the relevant values required.
* DFCOEFF\_GB685 – DAT file used for Discriminant score as described in section 7.5 of this document.
* DFMEAN\_GB685 – DAT file used for calculating Mahanalobis distance as described in section 7.6 of this document.
* EndGrp\_AssessScores – Excel csv used for calculating proportion during prediction.
* Helperfunctionsv1 – R code functions to carry out the Environmental Variables validation described in Section 5.
* MeanAirTempAirTempRangeASFunction - R code function used when calculating Mean Air Temperature and Air Temperature Range as described in Section 6.3.
* PredictionfunctionsV1 – A large number of R code functions used during the prediction process as described in Section 7.
* TAXAAB – Excel csv containing abundance data for each taxon used in calculating Prediction of taxa. This functionality has not yet been developed but is kept in the support files in readiness for development.
* TAXAPRAB - Excel csv containing probability category data for each taxon used in calculating Prediction of taxa. This functionality has not yet been developed but is kept in the support files in readiness for development.
* x103EndGroupMeans(FORMATTED) – Excel csv containing End group data as described above

**Classification support files:**

* adjustParams\_ntaxa\_aspt - Excel csv containing reference adjustment values which are used as described in Section 8.2.
* ClassificationfunctionsV2 - A large number of R code functions used during the prediction process as described in Section 8.
* EndGrp\_AssessScores - Excel csv used for adjusting the expected values during classification as described in Section 8.2 and for calculating proportion.

A list of the tables and functions required for the **NI model experiment** is as below:

**Prediction support files:**

* classInt\_0.3-1 – zip folder containing geo-spatial information for calculating latitude and longitude
* DFCOEFF\_NI – DAT file used for Discriminant score as described in section 7.5 of this document.
* DFMEAN\_NI\_RALPH – DAT file used for calculating Mahanalobis distance as described in section 7.6 of this document.
* EndGrp\_AssessScores – Excel csv used for calculating proportion during prediction.
* HelperfunctionsNIv1 – R code functions to carry out the Environmental Variables validation described in Section 5.
* PredictionfunctionsNIV1 – A large number of R code functions used during the prediction process as described in Section 7.
* Rnrfa\_1.4.0 folder – zip folder which contains a set of utility functions to retrieve data from the UK National River Flow Archive (NRFA). The package contains R wrappers to the UK NRFA temporary API. There are functions to retrieve stations falling in a bounding box, to generate map and extracting time series and general information.
* Sf\_0.7-2 - zip folder containing the SF package in R. The SF package contains support for simple features which is a standardized way to encode spatial vector data. It is used in RICT2 for calculating latitude and longitude for the Northern Ireland experiment.
* Units\_0.6-2 – zip folder containing geo-spatial information for calculating latitude and longitude
* x103EndGroupMeans(FORMATTED) – Excel csv containing End group data as described above

**Classification support files:**

* adjustParams\_ntaxa\_aspt - Excel csv containing reference adjustment values which are used as described in Section 8.2.
* ClassificationfunctionsV2 - A large number of R code functions used during the prediction process as described in Section 8.
* EndGrp\_AssessScoresNI - Excel csv used for adjusting the expected values during classification as described in Section 8.2.

The support files folders will be accessible by users of RICT2 who wish to amend and create their own experiments and may also wish to alter the data tables. The tables for standard experiments will be locked down but can be downloaded and amended by the user in their own work area experiments.

## Starting the Project Experiment

### Introduction

The “settings” will be determined by the project experiment which will use default values relevant for that project experiment. For initial release of RICT2, the default settings values provided will be those relevant for the WFD Classification and are as described below.

However, the tool will allow a user to download this experiment and change the settings and values by amending the standard experiment to create their own experiments.

Standard published experiments can only be amended by an administrator user.

A run of the tool can only be carried out against one Reference End Group model. Users can choose to run an experiment for either:

* RIVPACS IV GB (M1) (Set ID = 1)
* RIVPACS IV NI (M1) (Set ID = 2)

This standard WFD experiment for both of these will be a “locked down” version displayed in the Azure Studio gallery.

### Default values

The list below describes the default values used:

* Number of Monte-Carlo simulations – 10,000
* Index sets - WHPT ASPT and WHPT NTAXA indices (from end group means support file (i.e. the x103EndGroupMeans(FORMATTED) file)
* Season type – Spring, Summer and Autumn or combined
* Classification boundary limits – Coded as per WFD compliant values found in the classification support files “ClassificationfunctionsV2” (user would need to amend the function in this file to alter values for their own experiments)
* Bias - Bias values can be provided in the input file by the user and will be used in the experiment if they have been provided. However, if no data values for Bias has been entered in the input file, then the default value of 1.62 (GB) or 0 (NI) will be used and has been embedded in the code for the experiment.

Note that 1.62 is the default value currently being used for Bias as this was input as the default value in the previous version of RICT2 as used in SEPA for BMWP NTAXA. However, the EA uses a value of 1.68 as the Bias value for WHPT which was based on an audit of the analytical quality of Environment Agency laboratories. The Bias value for WHPT should be greater than the value for BMWP for the same level of analytical quality, because WHPT uses more taxa than BMWP. Plus WHPT taxa also includes additional *Diptera* families which are less easy to identify.

### Future Experiments

Future development will allow a user to choose to carry out a standard experiment for **“All Indices”** which will calculate output using all indices for prediction only. If this **“All indices”** option is chosen the output file will contain all calculations which could be amended to only show the specific index set required by the user. For example, if the user only wanted the HEV index set, the user could delete all the other indices output except for ASPT; NTAXA; TL2 LIFE(Fam) (DistFam) and TL3 PSI(Fam).

Additional standard default experiments to only calculate the output for more specific Index sets (e.g. HEV) may be added at a future date by a system administrator if the need arises.

Future RICT2 release may allow a user to run predictions and classifications against previous RICT2 models primarily to allow BMWP (Biological Monitoring Working Party) classification. These previous models are:

* RIVPACS III GB (M1)
* RIVPACS III NI (M1)

## Acceptance and validation of input data

In order to be able to predict and classify, the tool requires input data in one csv file which includes:

* Site Identifier
* Values for a number of Environmental Variables (e.g. NGR, width, depth, etc)
* Values for a number of Observed Tool Indices (e.g. ASPT, NTAXA, etc) (only required for classification)

An excel spreadsheet template will be provided for the user to enter input data into the relevant column headings. Validation rules will be applied to the data that has been input for each column to help the user meet the validation requirements detailed below.

The input data spreadsheet template can be found in Annex A of this document.

Once completed, this spreadsheet will need to be saved as a csv format and uploaded by the user to link to the required project experiment.

### Environmental Variable Values and Validation

Environmental Variables (EV) are needed to enable the prediction process to calculate the probability of end group membership. The values of some Environmental Variables are provided and input by the user. The values of other Environmental Variables have to be derived using the values of one or more provided Environmental Variables and then converted. The complete set of input and derived Environmental Variables used by RICT2 are called Predictive Environmental Variables (PEVs).

The table below refers to the Environmental Variable data required and the data field validation for the RIVPACS IV Great Britain Model M1 which requires input of 13 Environmental Variables and for the RIVPACS IV Northern Ireland Model M1 which requires input of 11 Environmental Variables.

Input file screening will be carried out using the rules described in the table below when a user enters data into the input file spreadsheet. This will be done by the conditional formatting tool within the MS Excel ribbon to check and highlight to the user where data does not meet validation requirements. It will highlight failures in red and warnings in orange. However, it will not stop the user from supplying that data to RICT2 for processing. When RICT2 processes the data, the appropriate Fail and Warn errors will be generated as described below.

Validation has also been coded into RICT2 for validation of the value ranges, i.e. “Overall valid range”, “GB model range” and “NI model range”.

Note that in the table below, where the value of an environmental input variable is input as a “Warn” limit, RICT2 carries-on with prediction but flags the transgression in a warning report for that particular site.

Note that in the table below, where the value of an environmental input variable is input as a “Fail” limit, RICT2 does not proceed with prediction for that particular site and flags the transgression in a warning report.

Note that at least one set of the following data will be required for each site but there may be up to three sets of data where the ‘raw’ data is being provided for up to three years for multi-year classification.

|  |  |
| --- | --- |
| **Data Item** | **Validation Comment** |
| Site | Identifies the site that the sample data represents. Note this ID is only meaningful to the organisation providing the data.  Must be present (input file screening) |
| Waterbody | Only required for multi-site classification.  No validation required. |
| Year | Must be present (input file screening)  Must not be more than 7 years in future or 7 years in past (input file screening)  Must be format “NNNN” where N is a number (input file screening) |
| NGR | Must be present (input file screening)  For GB: This must be in the format “AA” where “AA” is the alpha reference value for that NGR square, e.g. “SD” (input file screening)  For NI: This much be in the format “A” where “A” is the alpha reference value for that NGR square, e.g. “H” (input file screening) |
| Easting | Must be present (input file screening)  Must be 5 digits; text not number, to retain leading zeros (input file screening)  In R Code: If less than 5 digits leading zeros have been added |
| Northing | Must be present (input file screening)  Must be 5 digits; text not number, to retain leading zeros (input file screening)  In R Code: If less than 5 digits leading zeros have been added |
| Latitude\* (LAT)  (\*internally derived from NGR by algorithm)  Units = o North | Calculated within RICT2 code from the grid reference data entered above (not a column in the user input spreadsheet)  Overall Valid Range: If the value is less than 49 or greater than 71 = “Fail”  GB model range: If the value is less than 50.8 or greater than 62 = “Warn”  NI model range: If the value is less than 54 or greater than 55.2 = “Warn” |
| Longitude\* (LONG)  (\*internally derived from NGR by algorithm)  Units = o South | Calculated within RICT2 code from the grid reference data entered above (not a column in the user input spreadsheet)  Overall Valid Range: If the value is less than -11 or greater than 2 = “Fail”  GB model range: If the value is less than -8 or greater than 1.4 = “Warn”  NI model range: If the value is less than -8.1 or greater than -5.7 = “Warn” |
| Altitude (ALTITUDE)  Units = m | Must be present (input file screening)  Must be numeric (input file screening)  Overall Valid Range : If the value is less than 0 or greater than 1345 = “Fail”  GB model range: If the value is less than 1 or greater than 590 = “Warn”  NI model range: If the value is less than 3 or greater than -180 = “Warn”  Validation: if user enters 0, then replace with 1. |
| Slope at site (SLOPE)  Units = m km-1 | Must be present (input file screening)  Must be numeric (input file screening)  Overall Valid Range : If the value is less than 0 = “Fail”  GB model range: If the value is less than 0.1 or greater than 150 = “Warn”  NI model range: If the value is less than 0.1 or greater than 50 = “Warn”  Validation: if value is 0 then replace with 0.1 |
| Discharge category (DISCHARGE) | Must be provided if Velocity category value is not provided (input file screening)  Must be numeric (input file screening)  Overall Valid Range : If the value is less than 0 or greater than 10 = “Fail”  GB model range: If the value is less than 1 or greater than 9 = “Warn”  NI model range: If the value is less than 1 or greater than 8 = “Warn”  Validation: if value is 0 then replace with 1. |
| Velocity category (VELOCITY) | Must be provided if Discharge category value is not provided (input file screening)  Must be numeric (input file screening)  Overall Valid Range : If the value is less than 0 or greater than 10 = “Fail”  GB model range: If the value is less than 1 or greater than 5 = “Warn”  NI model range: If the value is less than 1 or greater than 5 = “Warn” |
| Distance from Source (DIST\_FROM\_SOURCE)  Units = km | Must be present (input file screening)  Must be numeric (input file screening)  Overall Valid Range : If the value is less than 0 = “Fail”  GB model range: If the value is less than 0.1 or greater than 202.8 = “Warn”  NI model range: If the value is less than 2.2 or greater than 75 = “Warn”  Validation: if less than 0.1 then value replaced with 0.1 |
| Stream Width (MEAN\_WIDTH)  Units = m | Must be present (input file screening)  Must be numeric (input file screening)  Overall Valid Range : If the value is less than 0 = “Fail”  GB model range: If the value is less than 0.4 or greater than 117 = “Warn”  NI model range: If the value is less than 2 or greater than 37 = “Warn”  Validation: if less than 0.1 then replace with 0.1 |
| Stream Depth (MEAN\_DEPTH)  Units = cm | Must be present (input file screening)  Must be numeric (input file screening)  Overall Valid Range : If the value is less than 0 = “Fail”  GB model range: If the value is less than 1.7 or greater than 300 = “Warn”  NI model range: If the value is less than 15 or greater than 183 = “Warn”  Validation: if less than 1 then replace with 1. |
| Boulder Cobble Percentage (BOULDED\_COBBLES) | Combined substrate percentage must be between 97% and 103%  Must be numeric (input file screening)  Overall Valid Range : If the value is less than 0 = “Fail” over 100 – “Fail” |
| Pebbles Gravel Percentage (PEBBLES\_GRAVEL) | Combined substrate percentage must be between 97% and 103%  Must be numeric (input file screening)  Overall Valid Range : If the value is less than 0 = “Fail” over 100 – “Fail” |
| Sand Percentage (SAND) | Combined substrate percentage must be between 97% and 103%  Must be numeric (input file screening)  Overall Valid Range : If the value is less than 0 = “Fail” over 100 – “Fail” |
| Silt Clay Percentage (SILT\_CLAY) | Combined substrate percentage must be between 97% and 103%  Must be numeric (input file screening)  Overall Valid Range : If the value is less than 0 = “Fail” over 100 – “Fail” |
| Mean substratum (MSUBST)  Units = phi | Calculated within RICT2 code from the percentage data entered above (not a column in the user input spreadsheet)  GB model range: If the value is less than -7.71 or greater than 8 = “Warn”  NI model range: If the value is less than -7.75 or greater than 6.61 = “Warn” |
| Alkalinity (ALKALINITY)  Units = mg l-1 CaCo3 | Hardness, Calcium concentration or Electrical conductivity can be provided if unable to provide Alkalinity – One of the values must be present (input file screening)  Must be numeric (input file screening)  Overall Valid Range : If the value is less than 0 = “Fail”  GB model range: If the value is less than 1.2 or greater than 366 = “Warn”  NI model range: If the value is less than 2.5 or greater than 194 = “Warn”  Validation: if less than 0.1 then replace with 0.1 |
| Hardness (HARDNESS) | Can be provided instead of Alkalinity and follows the same validation rules (input file screening)  Must be numeric (input file screening) |
| Calcium concentration (CALCIUM) | Can be provided instead of Alkalinity and follows the same validation rules (input file screening)  Must be numeric (input file screening) |
| Electrical Conductivity (CONDUCTIVITY) | Can be provided instead of Alkalinity and follows the same validation rules (input file screening)  Must be numeric (input file screening) |
| Mean air temperature (TEMPM)  Units = oC | Calculated within RICT2 code from the grid reference data entered above (not a column in the user input spreadsheet)  GB model range: If the value is less than 7.5 or greater than 11.5 = “Warn”  Not required for NI experiment so no validation checks required. |
| Air temperature range (TEMPR)  Units = oC | Calculated within RICT2 code from the grid reference data entered above (not a column in the user input spreadsheet)  GB model range: If the value is less than 8.3 or greater than 13.9 = “Warn”  Not required for NI experiment so no validation checks required. |

#### Environmental Variables Validation Process

The validation process is as described below. Some validation is done as the user enters data into the input spreadsheet for the Input Environment Variables. Validation checks on minimum and maximum values is embedded in the Helperfunctionsv1 support file within RICT2

* Spreadsheet validation is carried out to check that the Environmental Values are present and they are the correct format as user enters data. If any values are missing when data is linked to RICT2 the program fails.
* Check against ‘Overall’ Valid Max/Min Values
  + Check that each provided EV value is greater than or equal to the ‘overall’ minimum value and less than or equal to the ‘overall’ maximum value defined in the Environmental Variable parameter table above.
  + Within RICT2 code, if the values do not meet the validation criteria the site is flagged as a failure and that particular site is removed from the experiment.
* Check against ‘Warning’ Max/Min Values
  + Check that each provided EV value is greater than or equal to the ‘warning’ minimum value and less than or equal to the ‘warning’ maximum value defined in the Environmental Variable parameter table above.
  + If not, then this should be recorded so that it can be flagged as a ‘warning’ in the output file but the experiment will proceed as normal.
* Replace with values if applicable –
  + If EV value is less than the ‘overall’ minimum value for the EV values below, replace with the EV value given:
    - If Altitude = 0 then replace value with 1
    - If Distance from source is <0.1 then replace value with 0.1
    - If Width is <0.1 then replace value with 0.1
    - If Depth is <1 then replace value with 1
    - If Discharge category is 0 then replace value with 1
    - If Alkalinity is <0.1 then replace value with 0.1
    - If Slope is =0 then replace value with 0.1
  + If replacement values have been used, then this should be recorded so that it can be flagged in the output file.

### Observed Biological Data Values and Validation

Biological data is needed to enable the classification process. The values of the Biological data is provided by the user.

The table below refers to the data items required and the data field validation for the RIVPACS IV Great Britain Model 1 and for the RIVPACS IV Northern Ireland Model 1.

Conditional formatting validation will be carried out for all of the validation rules described in the table below when a user enters data into the input file spreadsheet. There is no validation for the Observed Biological data within RICT2.

However, RICT2 will validate that relevant values have been provided from the prediction process of the experiment prior to the classification process. This is to check that there are no null values if these values are required for the classification to be successful.

Note, this data will need to be entered for each season relevant for the project experiment. For WFD, this will be data for Spring, Summer and Autumn.

|  |  |
| --- | --- |
| **Data Item** | **Validation** |
| Season ID  (Season\_ID) | Must be numeric single digit code, i.e. from list:   * 1 = spring * 2 = summer * 3 = autumn * 4 = spring + summer * 5 = spring + autumn * 6 = summer + autumn * 7 = spring + summer + autumn |
| WHPT ASPT Abundance  (TL2\_WHPT\_ASPT (AbW,DistFam)) | Must be numeric |
| WHPT NTAXA Abundance  (TL2\_WHPT\_NTaxa (AbW,DistFam)) | Must be numeric integer (i.e. whole number) |
| Bias  (Ntaxa\_Bias) | Must be numeric |

## Conversion of Input Environmental Variables

The values for some input environmental variable data needs to be converted to particular values that are required for the prediction process. These are as follows:

* NGR converted to latitude and longitude
* Calculate Mean Substratum Composition
* Estimating Discharge Category (if unavailable) from Stream Velocity
* Estimating Alkalinity (if unavailable) from Hardness, Calcium or Conductivity

In addition, some ‘converted’ EVs then need further conversion to obtain values required for the prediction process. However, this is only required for the GB model experiment:

* Latitude and Longitude converted to Long Term Average Mean and Long Term Range of Air Temperature

The required conversion rules and some validation rules that need to be applied are described below.

### Deriving Latitude and Longitude from National Grid Reference for Great Britain Model sites

National Grid reference (BNG) is eight characters long consisting of:

NGRLET = 2 letters indicating the National Grid 100km x 100km grid square

NGREAST = 3 digit Easting within the 100km grid square (to nearest 100m)

NGRNORTH = 3 digit Northing within the 100km grid square (to nearest 100m)

To derive latitude and longitude, the NGR input by the users is run through Helperfunctionsv1.r within the prediction support files folder.

This contains code that calls the WGS84 system (which is contained in the rNRFA folder\_1.4.0) to derive Latitude and Longitude that is required to calculate Long Term Average Mean Temperature and Range of Air Temperature environmental variables.

The specific function code used within Helperfunctionsv1.r is as below:

getLatLong <- function (nat\_grid\_ref, easting, northing, coordsys\_latlon)

where:

nat\_grid\_ref, is the NGR, easting and northing from the user input file

coordsys\_latlon uses the WGS84 system as described above

### Deriving Latitude and Longitude from National Grid Reference for Northern Ireland Model Sites

Grid references in Norther Ireland are different and consist of:

NGR = 1 letter indicating the National Grid 100km x 100km grid square

EASTING = 3 digit Easting within the 100km grid square (to nearest 100m)

NORTHING = 3 digit Northing within the 100km grid square (to nearest 100m)

To derive latitude and longitude, the grid reference input by the users is run through HelperfunctionsNI\_v1.r within the Northern Ireland prediction support files folder.

This contains code that calls the relevant function within the SF package (which is also within the Northern Ireland prediction support files folder) to derive Latitude and Longitude. The specific function called is:

getLatLong\_NI <- function (EASTING, NORTHING)

This algorithm has been used in RICT2 as found in the HelperfunctionsNI\_v1.

### Calculating Long Term Average Mean and Range of Air Temperature from Latitude and Longitude

This conversion process is only applicable to GB model sites.

When converting Latitude and Longitude to Long Term Average Mean Temperature and Range of Air Temperature, RICT2 uses a data file containing previously derived values for Mean air temperature and Air temperature range for points in the centres of 5km interval grid squares. This data file can be found in the Prediction support files and is called AirTempGrid.

To calculate Long Term Average Mean Temperature and Range of Air Temperature, the Latitude and Longitude previously derived is run through MeanAirTempAirTempRangeASFunction.r within the prediction support files folder.

The specific function code used within MeanAirTempAirTempRangeASFunction.r.is as below:

calc.temps <- function(coordinates)

where:

coordinates are the eastings and northings (lat, long) calculated values as described in section 6.1

### Calculating Mean Substratum Composition - MSUBST (in phi units)

MSUBST is derived from the following user-supplied environmental data (i.e. from the input file) for the river bed substratum composition at the test site:

BOLDCOBB = percentage cover of Boulders/Cobbles

PEBBGRAV = percentage cover of Pebbles/Gravel

SAND = percentage cover of sand

SILTCLAY = percentage cover of silt/clay

The Algorithm used is:

TOTSUB = BOLDCOBB + PEBBGRAV + SAND + SILTCLAY

MSUBST = ( -7.75\*BOLDCOBB - 3.25\*PEBBGRAV + 2\*SAND + 8\*SILTCLAY ) / TOTSUB

During this calculation RICT2 will check that TOTSUB is between 97 and 103. If not, an error will be recorded against the site and prediction and classification will not then take place for that site.

The specific function code used can be found within Helperfunctionsv1.r (within the prediction support files) and is as below:

getSubstrate <- function(bould\_cob, pebbles\_gr, snd, silt\_cl, lower\_b, upper\_b)

### Estimating Discharge Category (if unavailable) from Stream Velocity

If a value for discharge category (DCH) is not supplied by the user for the test site, then it is to be assumed that discharge category is to be estimated from a measurement of water velocity obtained for the site and supplied by the user.

If neither value is provided then an error will be recorded against the site and prediction and classification should not then take place for that site.

The algorithm for estimating DCH from Stream Velocity is:

RDCH = DEPTH/100. \* WIDTH \* VELC(VEL)/100

K=10

REPEAT

DCH=K

K=K-1

UNTIL RDCH>CDCH(K)

The specific function code used can be found within Helperfunctionsv1.r (within the prediction support files) and is as below:

getLogDischarge <- function (depth, width, discharge, velocity, lower\_b, upper\_b)

### Estimating Alkalinity (if unavailable) from Hardness, Calcium or Conductivity

If a value for alkalinity (ALK) is not supplied by the user for the test site, then it is to be assumed that alkalinity is to be estimated from, in order of preference, the user-supplied values for either water hardness (HARD), calcium concentration (CALCIUM) or conductivity (CONDUCT).

If none of these values are provided then an error should be recorded against the site and prediction and classification should not then take place for that site.

The algorithm for estimating Alkalinity from the other values is:

IF HARD is provided THEN

ALK = 4.677 + 0.6393\*HARD

ELSE

IF CALCIUM is provided THEN

ALK = 14.552 + 1.7606\*CALCIUM

ELSE

IF CONDUCT is provided THEN

ALK = 0.3201\*CONDUCT -8.0593

The specific function code used can be found within Helperfunctionsv1.r (within the prediction support files) and is as below:

getLogAlkalinity <- function (hardness, calcium, conduct, alkal, lower\_b, upper\_b)

### Transform Environmental Variables to Logarithm Base 10 Form

Some environmental predictor variables are used in discriminant functions in a logarithm to base10 transformed form (denoted Log10). They, therefore, need to be converted prior to prediction.

The discriminant functions for the Great Britain (GB) model (Reference end group set 1) use 13 environmental predictor variables (i.e. vN = 13), as detailed below.

The Northern Ireland (NI) model (Reference end group set 2) uses the same set of variables apart from air temperature mean (TEMPM) and air temperature range (TEMPR), which were not available for Northern Ireland and probably would not add discriminatory power within this relatively small geographic region.

For the purposes of specifying the discriminant functions and inputting their coefficients (DFCoeffv,d), RICT2 will have a separate Helperfunctionv1.r folder (within the prediction support files) for the Northern Ireland model experiment.

Some of the environmental predictor variables are used in the discriminant functions in a logarithm to base 10 transformed form (denoted Log10 ), as detailed below.

Specific functions within the Helperfunctionv1.r will do the conversions as described in Sections 6.1 – 6.6 above and will also calculate to log base 10 for any Environmental variables that require Log Base 10 form.

The precise order and form of the environmental variables to be used as Env 1– EnvvN in the discrimination functions equations used to calculate the discriminant function axis scores (DFScore) is as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ENV v** | **Environmental predictor variable** | **User input** | **Internally derived** | **Form used in DFs** |
| ENV 1 | Latitude | LAT |  | LAT |
| ENV 2 | Longitude | LONG |  | LONG |
| ENV 3 | Altitude | ALT |  | LGALT = Log10(ALT) |
| ENV 4 | Distance from source | DIST |  | LGDIST = Log10(DIST) |
| ENV 5 | Stream width | WIDTH |  | LGWIDTH = Log10(WIDTH) |
| ENV 6 | Stream depth | DEPTH |  | LGDEPTH = Log10(DEPTH) |
| ENV 7 | Mean substratum |  | MSUBST | MSUBST |
| ENV 8 | Discharge category | DCH |  | DCH |
| ENV 9 | Alkalinity\* | ALK |  | ALK |
| ENV 10 | Alkalinity\* | ALK |  | LGALK = Log10(ALL) |
| ENV 11 | Slope at site | SLOPE |  | LGSLOPE = Log10(SLOPE) |
| ENV 12 | Mean air temperature |  | TEMPM | TEMPM |
| ENV 13 | Air temperature range |  | TMPR | TMPR |

\* Alkalinity is involved in the discriminant functions in both untransformed form (as Env9) and in log transformed form (as Env) to represent its non-linear impact on the biota.

This table also shows which EVs are supplied by the user in the input file and which EVs are internally derived by RICT2 conversions.

### Validation of Derived Environmental Variables

Whenever the value of an Environmental variable is derived from any conversion applied, it needs to be validated against the various max/min values. There are two distinct max/min values; Fail and Warn and the required validation is as detailed previously in section 5.

Once all the conversion processing has been carried out, then this should result in a value being present for every required Environmental variable. This complete set of Environmental variables is called the Predictive Environmental Variables (PEVS).

## Predict

### Predict Process

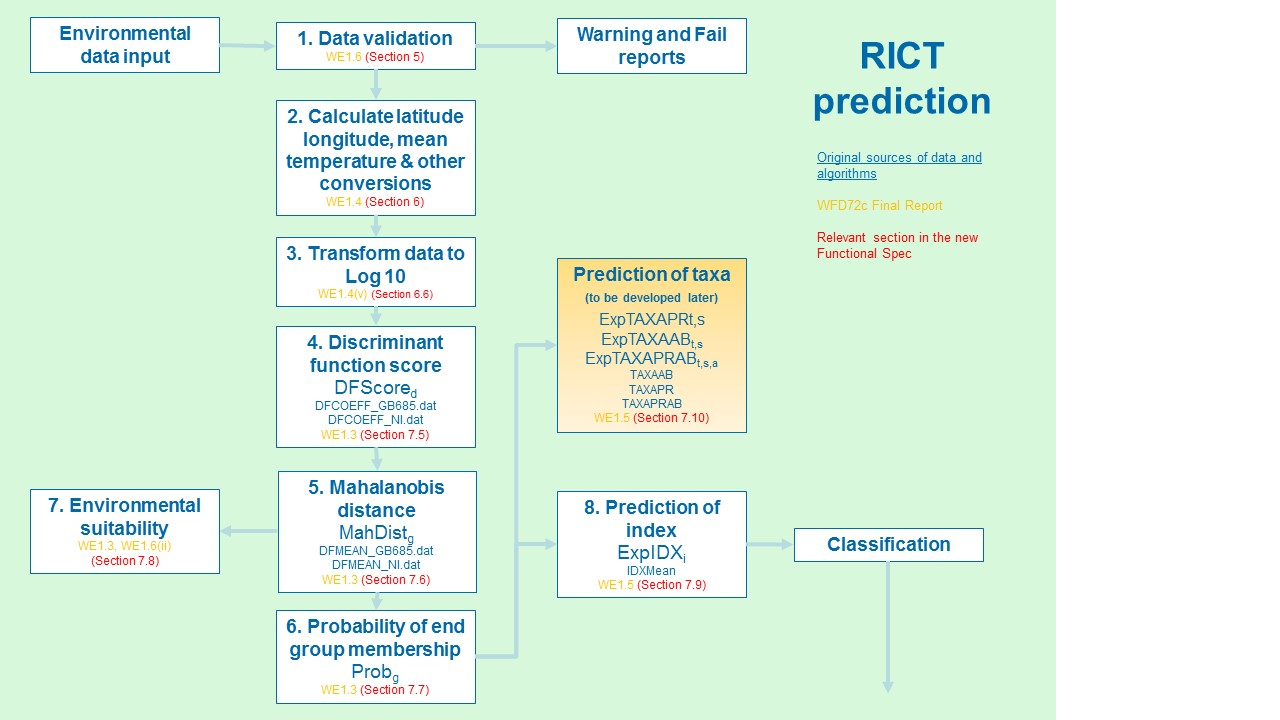
The prediction functionality involves four distinct steps:

1. determine the probability of end group membership
2. determine Environmental Suitability
3. based on a), determine the predicted index values for the indices relevant to the run
4. based on a), determine the predicted taxa to be present at the site)

The most complex part of this is a), which involves the use of multiple discriminant analysis and specialised statistical techniques (e.g. Mahanalobis Distance).

It needs to be carried out for every set of site data applicable to the run (a site may be present more than once where more than one years’ worth of environmental data has been provided).

The flow diagram below summarises the prediction process and calculations (including the data input, validation, conversion and transform steps described above) as well as identifying the relevant section for further information from the WFD72C final report:



\*\* Are there any changes to this slide that needs to be copied back into the tech spec? \*\*

### Prediction Overview

The probability of a site belonging to an End Group is determined by the similarity of its Environmental Variables (EVs) to those of the relevant End Group.

Due to the multivariate nature of the environmental variable data, Multiple Discriminant Analysis (MDA) is used to turn the environmental variable data into a format that can be used to discriminate between the end groups.

MDA depends on multiple Discriminant Functions (DF). A Discriminant Function has the general format:-

Z=constant+c1EV1+c2EV2+c3EV3+…….+cnEVn

where Z is the Discriminant Score (DS) for that particular discriminant function, c1…cn are the coefficients of that discriminant function and EV1…EVn are the values of the environmental variables.

A DF is a linear combination of the EVs and the coefficients, which maximises the separation in data space of the end groups. In other words, if the discriminant function was solved for the mean values of the environmental variables of each end group and the results plotted on a line, the value from the solved formula for each end group would be as far apart from its nearest neighbour as could be achieved by using the available environmental variables to separate the end groups

### Environmental Variables

The values of the EVs relevant to the prediction run will have been determined during the acceptance, validation and subsequent conversion processing of the input data as described in section 5 and 6 of this document. These are referred to in the algorithm below as EnvvN.

### Discriminant Functions (DFs) and the Co-efficients

The RIVPACS model used will determine the DFs that have to be run. For the GB model with 13 EVs the DFs values can be found in the prediction support files folder for GB in a file called DFCOEFF\_GB685.

For the NI model with 11 EVs, the DFs values can be found in the prediction support files folder for Northern Ireland in a file called DFCOEFF\_NI

These are separate occurrences of DF. Each of these has a single ‘constant’ and, for each component EV, a co-efficient that is to be applied to the EV value. The co-efficients are referred to in the algorithm below as DFCoef.

### Calculate the Discriminant Function Scores (Step 4 in diagram above)

**4. Discriminant function score**

DFScored

DFCOEFF\_GB685.dat

DFCOEFF\_NI.dat

WE1.3 (Section 7.5)

The functions that RICT2 uses for the calculations in the steps below (steps 4 - 8) can be found in the PredictionfunctionsV1 in the prediction support file folder. The algorithms and definitions provided below are copied from WFD72C and have been used to create the functions.

For each DF, the Discriminant Score needs to be calculated as follows:

DFScored = DFCoef1,d \* Env1 + … + DFCoefvN,d \* EnvvN; for d = 1. … dN

**Definitions** – these have been copied from WFD72C:

v = id of current environmental predictor variable

vN = number of environmental predictor variables

d = id of current discriminant function axis

dN = number of discriminant function axes in current Reference end group set

DFCoefv,d = discriminant function coefficient for predictor variable v on discriminant function d

(Obtained from the values in the DFCOEFF\_GB685 file for Great Britain or DFCOEFF\_NI file for Northern Ireland which can be found within the prediction support files folder for either experiment)

Envv = value of environmental predictor variable v for the current test site

DFScored = discriminant function score on axis d for the current test site

### Calculate Mahanalobis Distance (Step 5 in diagram above)

**5. Mahalanobis distance**

MahDistg

DFMEAN\_GB685.dat

DFMEAN\_NI.dat

WE1.3 (Section 7.6)

For a particular End Group g, the probability of End Group Membership is calculated as follows:

* Calculate Mahanalobis distance (MahDistg) of test site from each End Group

MahDistg = (DFScore1 - DFMeang,1)2 + ... + (DFScoredN - DFMeang,dN )2 ; for g = 1,...,gN

**Definitions** – these have been copied from WFD72C:

g = id of current end group (set ID: 1 = GB, 2 =NI)

gN = number of end groups in current Reference end group set

NRefg = number of reference sites in end group g

(Obtained from the table x103EndGroupMeans(FORMATTED) which is within the prediction support files folder)

DFMeang,d = mean discriminant function score of end group g on axis d

(Obtained from the ‘target’ values in the DFMEAN\_GB685 file for Great Britain or DFMEAN\_NI file for Northern Ireland which can be found within the prediction support files folder for either experiment)

MahDistg = Mahalanobis distance of test site from end group g

### Use Mahanalobis Distance to Determine Probability of End Group Membership (Step 6 in diagram above)

**6. Probability of end group membership**

Probg

WE1.3 (Section 7.7)

* Calculate the PDistg for each End Group

PDistg = NRef g \* EXP(-MahDist g/2)

where NRef g = the number of reference sites in the End Group (from the End Group Reference Data table) and EXP is the natural exponential function

* Calculate the PDistTot

PDistTot = PDist 1 to PDist gN

where gN is the number of end groups in the relevant end group set

* Calculate the Probability of End Group Membership (Probg)

Probg = PDistg /PDistTot

where Probg is the Probability test site belongs to end group g

At the end of this process we end up with a set of probabilities of End Group Membership, one for each End Group. These are referred to below as Prob1 to Probn.

### Determine Environmental Suitability of the Test Site for Prediction (Step 7 in diagram above)

**7. Environmental suitability**

WE1.3, WE1.6(ii)

(Section 7.8)

In certain cases it may be calculated that there is a very low probability of the site being in any of the end groups based on the data provided.

In order to highlight this to the user, a calculation is required to allocate an Environmental Suitability code to the site data indicating the probability the site belongs to any end group. The potential codes are:

* 1 OK
* 2 < 5%
* 3 < 2%
* 4 < 1%
* 5 < 0.1%

The required calculation uses the Chi-square values (CQ1 to CQ4) which vary per End Group Set as below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **End Group set (ID)** | **CQ1** | **CQ2** | **CQ3** | **CQ4** |
| RIVPACS IV GB (set 1) | 21.02606 | 24.05393 | 26.21696 | 32.90923 |
| RIVPACS IV NI (set 2) | 18.307 | 21.1608 | 23.2093 | 29.5883 |

PredictionfunctionsV1 (within the prediction support files folder) has a function called getSuitabilityCode() which will do this calculation

The detailed calculation is as follows:

* Calculate the minimum Mahanalobis distance (MahDistmin)

MahDistmin = minimum of (MahDist1,….,MahDistn)

* Calculate the Environmental Suitability Code and Suitability text

If the MahDistmin < value for CQ1 then Suitability Code = 1 and suitability text is >5%

If the MahDistmin >= value for CQ1 and < value for CQ2 then Suitability Code = 2 and suitability text is <5%

If the MahDistmin >= value for CQ2 and < value for CQ3 then Suitability Code = 3 and suitability text is <2%

If the MahDistmin >= value for CQ3 and < value for CQ4 then Suitability Code = 4 and suitability text is <1%

If the MahDistmin >= value for CQ4 then Suitability Code = 5 and suitability text is >0.1%

The output file will provide the user with the Suitability Code and suitability text. The Suitability text gives the user information and warnings about the suitability code calculated.

### Determine the Predicted Values of Indices (Step 8 in diagram above)

**8. Prediction of index**

ExpIDXi

IDXMean

WE1.5 (Section 7.9)

The indices that require to be processed will be identified from the project experiment settings. For the initial release of RICT2 this will be for WHPT ASPT and WHPT NTAXA indices which are the indices required for the WFD classification.

PredictionfunctionsV1 (within the prediction support files folder) contains the following functions which will do this calculation:

getSeasonIndexScores(). - getSeasonIndexScores <- function (data\_to\_bindTo, season\_to\_run, index\_id, end\_group\_IndexDFrame)

and the: end\_group\_IndexDFrame uses the x103EndGroupMeans(FORMATTED) within the prediction support files folder to get values

**Algorithm**: (copied from WFD72C)

ExpIDXi = Prob1 \* IDXmeani,s,1, + ... + ProbgN \* IDXmeani,s,gN

**Definitions: (**these have been copied from WFD72C)

g = id of current end group

gN = number of end groups in current Reference end group set (set: 1 = GB, 2 =NI)

Probg = Probability test site belongs to end group g

i = id of current biological index

iN = total number of biological indices

s = id of selected season(s) combination (referred to as ‘season *s*’); (1 = spring, 2 = summer, 3 = autumn, 4 = spring+summer, 5 = spring+autumn, 6 = summer+autumn, 7 = all three seasons)

IDXMeani,s,g = Mean value of index *i* for season *s* for reference sites in end group *g*

ExpIDXi = Expected value of index *i* for selected season *s* for current test site

**Internally-supplied Data files:**

A separate EXCEL file (called x103EndGroupMeans(FORMATTED) which can be found within the prediction support files) will be provided with the values for:

IDXMeani,s,g = Mean value of index *i* for season *s* for reference sites in end group *g*

### Determine the Taxa Predicted to be Present at the Site

**Prediction of taxa**

**(to be developed later)**

ExpTAXAPRt,s

ExpTAXAABt,s

ExpTAXAPRABt,s,a

TAXAAB

TAXAPR

TAXAPRAB

WE1.5 (Section 7.10)

This functionality has not yet been developed. The information given below is from previous RICT2 technical spec and extracts from WFD72C. When this functionality is fully developed, this information will need amending to reflect what was developed.

Although not required for classification, an important requirement of the tool is to be able to predict the taxa expected to be present at the site. Two values require to be output for each appropriate taxon: ‘expected probability of occurrence’ and ‘expected log abundance category’. The steps involved are described below.

#### Identify the Relevant Taxa for the run

Due to a potentially overwhelming amount of data in the output files, the user will specify within the project experiment at what taxonomic level the taxon prediction is to be carried out. Therefore, the taxa relevant to the run are those identified where the taxonomy level = Y for the taxonomy for the run. This identifies a list of NBN Codes to be subject to prediction.

The process to predict the occurrence of taxa at a particular sample can be carried out at a variety of taxonomic levels. The five levels (TL1 – TL5) that a user can select from are:

* TL1 = BMWP taxa
* TL2 = WHPT taxa
* TL4 = all RIVPACS families
* TL4 = all RIVPACS species
* TL5 = mixed taxon

#### For each Applicable Taxon Calculate Expected Probability of Occurrence

For each of the NBN Codes identified, calculate the expected probability of occurrence as follows:

* Prob1\*EG1TPR + Prob2\*EG2TPR +……..+ Probn\*EG1TPR

where Prob1 to Probn are the probabilities of End Group membership and EG1TPR to EGnTPR are the Probability of Capture values of the relevant taxon for End Groups 1 to n. These are obtained from the End Group reference table but note that the value for an End Group/Taxon is also dependent on Season, which has been provided in the project experiment settings. Note that the taxon may not be present against each End Group/Season in which case Probability of Capture = 0.

Once the expected probability has been calculated for a taxon then the value has to be compared against the Minimum Taxon Reporting Value to see whether it is to be output in the prediction or not.

#### For each Applicable Taxon Calculate Expected Log Abundance Category

For each of the NBN Codes identified as having probability of occurrence > Minimum Taxon Reporting Value, calculate the expected log abundance category as follows:

* Prob1\*EG1TLA + Prob2\*EG2TLA +……..+ Probn\*EG1TLA

where Prob1 to Probn are the probabilities of End Group membership and EG1TLA to EGnTLA are the Expected Log Abundance values of the relevant taxon for End Groups 1 to n. These are obtained from the End Group reference table, but note that the value for an End Group/Taxon is also dependent on Season, which has been provided in the project experiment settings.

Although Log Abundance codes are held in Reference Data against taxa, the table below shows what ranges each of these codes refers to:

|  |  |
| --- | --- |
| Log10 category | Numerical abundance |
| 1 | 1 – 9 |
| 2 | 19 - 99 |
| 3 | 100 - 999 |
| 4 | 1000 – 9999 |
| 5 | 10000 + |

**Algorithm**: (copied from WFD72C)

ExpTAXAPRt,s = Prob1 \* TAXAPRt,s,1, + ... + ProbgN \* TAXAPRt,s,gN

ExpTAXAABt,s = Prob1 \* TAXAABt,s,1 + ... + ProbgN \* TAXAABt,s,gN

ExpTAXAPRABt,s,a = Prob1 \* TAXAPRABt,s,1,a + ... + ProbgN \* TAXAPRt,s,gN,a (optional)

**Definitions**

TAXAPRt,s,g = Proportion of reference sites in end group *g* with taxon *t* for season *s*

TAXAABt,s,g = Average of log abundance categories for taxon *t* for season *s* amongst reference sites in end group *g*

TAXAPRABt,s,g,a = Proportion of reference sites in end group *g* with taxon *t* at abundance category *a* for season *s* (optional)

ExpTAXAPRt,s = Expected probability of occurrence of taxon *t* for season *s* for the current test site

ExpTAXAABt,s = Expected log abundance category values of taxon *t* for season *s* for the current test site

ExpTAXAPRABt,s,a = Expected probability of occurrence of taxon *t* at abundance category *a* for season *s* for the current test site (optional)

**Internally-supplied Data files: (contained with prediction support files folder)**

A separate EXCEL file will be provided with the values of each of:

TAXAPRt,s,g = Proportion of reference sites in end group *g* with taxon *t* for season *s*

TAXAABt,s,g = Average of log abundance categories for taxon *t* for season *s* amongst reference sites in end group *g*

TAXAPRABt,s,g,a = Proportion of reference sites in end group *g* with taxon *t* at abundance category *a* for season *s* (optional)

### Completion and Outputs of Prediction Processing

Once the prediction processing has been completed, the following data has been calculated for each occurrence of site data:

* Probabilities of End Group Membership
* Expected Index Values
* Predicted Taxa occurrence details

The expected index data is then passed to the classification process.

If the user wishes to only view the prediction results, a csv output will be generated in the project experiment which can be downloaded and used for reports.

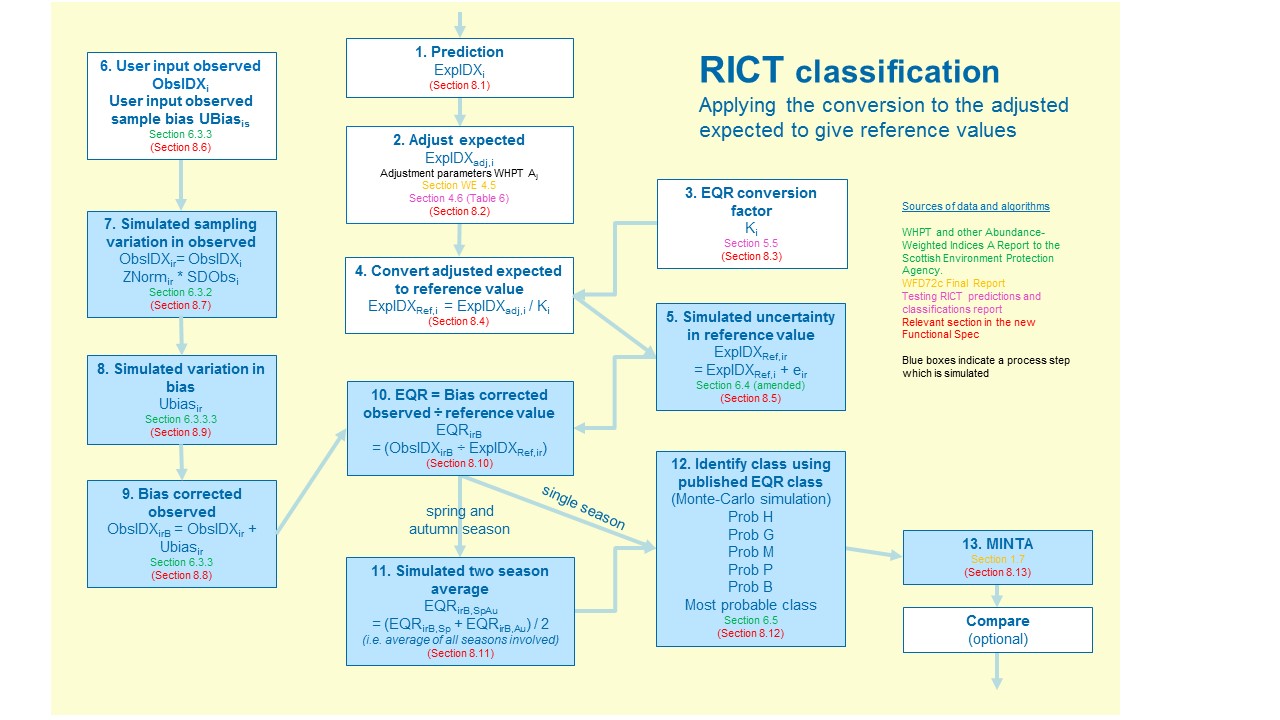
An example of the prediction output file and the columns that are included in the output file can be found in Annex B.

### 

## Classify

### Classify Process Summary

The flow diagram below summarises the classification process and calculations as well as identifying the relevant section for further information from the WFD72C final report and other sources of information:



**WWho**

Note that the blue boxes indicate processes that are simulated.

\*\* my assumption here is that this slide will change to reflect the inclusion of summer, if that’s the case, I will need to get the updated slide to insert here \*\*

### Classification overview

For every site in the experiment, the input to the classification process consists of Observed and Expected Index Values for the ‘current year’ for each applicable index.

Multiyear classification will include Observed and Expected Index Values for up to two previous years for each applicable index.

The indices that need to be classified will relevant to the project experiment. For the WFD classification default experiment, RICT2 will use TL2 WHPT Abundance weighted with distinct taxa for ASPT and NTAXA indices.

Any functions that RICT2 uses for the calculations described in the steps below (steps 1 - 13), can be found in the ClassificationfunctionsV2 within the classification support files folder. The algorithms and definitions provided below are copied from WFD72C plus other sources and have been used to create the functions.

### Expected Values have been calculated by the Prediction Process during the Current Run (Step 1 in process diagram above)

### 

**1. Prediction**

ExpIDXi

(Section 8.1)

This is the predicted value calculated by previous RIVPACS prediction, based on environmental variable data input by the user. It is the prediction of the value of the index expected at the site if it was in the best quality available for that type of site, based on samples collected from best available reference sites.

### Adjust Expected Values to Standardise against the High/Good Boundary (Step 2 in process diagram above)

**2. Adjust expected**

ExpIDXadj,i

Adjustment parameters WHPT Aj

Section WE 4.5

Section 4.6 (Table 6)

(Section 8.2)

The aim of this step is to standardise the raw predictions so that they relate to the same environmental quality (the high/good boundary quality) using algorithms first developed in WFD72b. This step is necessary to take account of variation in the environmental quality of RIVPACS reference sites.

There are a number of processes involved in this step as described below:

* The computeScoreProportions () function within the prediction support files folder is used to calculate proportions (Qij) for each end group using a CSV file (also within the prediction support files folder) called EndGrp\_AssessScores.
* The getWeighted\_proportion\_Rj () function within the prediction support files folder is used to calculate weighted proportions (Rj) using the probabilities from prediction process by multiplying with the Qij proportions.
* The compute\_RjAj () function within the prediction support files folder is used to calculate RjAj by multiplying the weighted proportions by the adjustment factors. See below for details regarding the adjustment factors.
* Finally, the RjAj value is used to divide by for all the predictions (NTAXA and ASPT) to get the adjusted expected value (ExpIDXadj,i) for each site and each index.

The reference adjustment values referred to above are default values for the WFD classification experiment and can be found in adjustParams\_ntaxa\_aspt csv file within the classification support files folder.

This supporting data table csv file contains the number of reference sites by Biologist assessment score (1-5) for each End Group together with the five adjustment factors (Q1, Q2, Q3, Q4, Q5) for each index.

The source for this csv file is based on the Table 6 from the Testing RICT2 predictions and classifications draft report 2018 which is as below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Index | Q1 | Q2 | Q3 | Q4 | Q5 |
| WHPT NTAXA | 1 | 1 | 1 | 0.967 | 0.926 |
| WHPT ASPT | 1 | 1 | 1 | 0.977 | 0.945 |

Table 6. Adjustment factors for reference site quality (Q1, Q2, Q3, Q4, Q5) for abundance-weighted WHPT NTAXA and WHPT ASPT

However, further work has been done which has resulted in slightly different values for these adjustment factors which are reflected in the adjustParams\_ntaxa\_aspt csv file used by RICT2, It is expected that the Testing RICT2 predictions and classifications draft report 2018 will be shortly updated to include a Table 7 to show these changes.

There is one set of adjustment factors for each index. Within current RICT2, adjustment factors for WHPT NTAXA and WHPT ASPT have been calculated for the default WFD experiment. If a user wants to try a different value then they will need to amend and create their own experiment.

### EQR Conversion Factor (Step 3 in process diagram above)

**3. EQR conversion factor**

Ki

Section 5.5

(Section 8.3)

This is the conversion factor used to convert the adjusted predictions that relate to high/good boundary quality to reference values that relate to reference condition, which is somewhere in high status (reference values are the average value of the metric at reference sites in reference condition = high status)

Each index has a different conversion factor, based on the average value of the index at all RIVPACS Reference Sites divided by the value at RIVPACS reference sites that are in WFD high status (reflected in Biologists’ Assessment values for each reference site).

For the default WFD classification in RICT2 the conversion factor value will be (as copied from Section 5.5 of the Testing RICT2 predictions and classifications draft report 2018):

* 1.0049 for WHPT NTAXA
* 0.9921 for WHPT ASPT

### Convert adjusted expected to reference value (Step 4 in process diagram above)

**4. Convert adjusted expected to reference value**

ExpIDXRef,i  = ExpIDXadj,i / Ki

(Section 8.4)

The aim of this step is to convert the predictions from high/good boundary to reference values, which is somewhere in high status (actually the average value of the index at reference sites in high status), by applying the conversion factor.

The conversion factor is applied to the adjusted expected value to give an expected reference value which will be used to calculate the EQRs expected by the Water Framework Directive.

Specifically, ExpIDXRef,i is calculated as:

ExpIDXRef,i = ExpIDXAdj,i / K

Where:

ExpIDXAdj,i = the adjusted expected value which is calculated as described in section 8.2 above.

K = WFD conversion factor (i.e. WHPT NTAXA: K = 1.0049, and WHPT ASPT K = 0.9921 as described in section 8.3 above.

IDX = do this for each Index, i.e. for WHPT NTAXA and WHPT ASPT

i = do this for each specific site

This avoids the need for the intermediary step of calculating and referring to EQI (Environmental Quality Indices) values (which are calculated as EQI = Obs / ExpAdj ) and means that the site classifications can use the published class limits for index EQR values directly.

This calculation is done directly within RICT2 code and not via any function.

### Simulate uncertainty in expected (Step 5 in process diagram above)

**5. Simulated uncertainty in reference value** ExpIDXRef,ir

= ExpIDXRef,i + eir

Section 6.4 (amended)

(Section 8.5)

The aim of this step is to take account of error in measurement of environmental variables used to predict the reference values.

It should be noted that the “ +eir “ shown in the algorithm in the process step above is a summary of the algorithm values of ZExpir \* SDExpi / √(NExpyear) as described in further detail below. It is the error term used to adjust the expected value.

The simulation of uncertainty in ExpRef is obtained using the algorithms in Section 6.4 of the Clarke & Davy-Bowker WHPT indices RICT2 Report of March 2014 report, but now where uncertainty in the Expected values is measured about the WFD Reference-adjusted value ExpRef., this has been amended as described below.

Main algorithm used:

ExpIDXir = ExpIDXi + ZExpir \* SDExpi / √(NExpyear)

Where (using the same notation as section 6.4 of the 2014 WHPT report):

ExpIDXi = WFD Reference-adjusted Expected value (ExpRef) of index *i*

ExpIDXir = WFD Reference-adjusted Expected value of index *i* in simulation *r*

SDExpi = Error SD for expected value of index *i*

NExpyear = 1 for single-year run

= number of years (1, 2 or 3) for which a separate estimate of the E value was involved in the estimate of average E value (for multi-year run)

ZExpir = Random number deviate from a standard Normal distribution with a mean of 0.0 and SD of 1.0, for use in simulation *r* for index *i*

ExpIDXir = ExpIDXi + ZExpir \* SDExpi / √(NExpyear)

Where:

SDExp7 = 4.3 = Measurement error SD of Expected values of weighted WHPT Score (index 7)

SDExp8 = 0.53 = Measurement error SD of Expected values of weighted WHPT NTAXA (index 8)

SDExp9 = 0.081 = Measurement error SD of Expected values of weighted WHPT ASPT (index 9)

Within RICT, these calculations are done using functions which can be found in the ClassificationfunctionsV2 (within the classification support files folder).

It should be noted that it is important to do this for each index and also be aware of the numbering for each index. As indicated above, each index has a number, i.e.:

* WHPT Score (index 7)
* WHPT NTAXA (index 8)
* WHPT ASPT (index 9)

### User input of observed index value and bias; calculation of bias adjustment (step 6 in process diagram above)

**6. User input observed**

**ObsIDXi**

**User input observed sample bias UBiasis**

Section 6.3.3

(Section 8.6)

RICT2 classification requires users to input an observed value for an index (ObsIDXi), based on samples collected in the field. This value is provided by the user completing the relevant data field in the input file as described in section 5 and Annex B of this document. Values are required for each index and for all three seasons, spring, summer and autumn.

Users may also input a bias value (UBiasis) for NTAXA, which reflects the impact of analytical error (measured by auditing) on the observed index values. This value is provided by the user completing the relevant data field in the input file as described in section 5 and Annex B of this document. RICT2 calculates bias for ASPT from bias for NTAXA and the observed NTAXA. If no input value for Bias is provided, RICT2 will use a default value of 1.62.

If bias correction is not wanted, users should enter a bias value of zero into the relevant data field in the input file. The zero value must be entered and not left blank or RICT2 will use a default value.

### Simulate uncertainty in observed values (step 7 in process diagram above)

**7. Simulated sampling variation in observed**

ObsIDXir= ObsIDXi ZNormir \* SDObsi

Section 6.3.2

(Section 8.7)

These simulations take account of sampling error, estimated in replicated sampling studies, including the Biological Assessment Methods Study (BAMS) which is needed so that statistical confidence of the status classification can be estimated.

The main algorithm used:

ObsIDXir = ObsIDXi ZNormir \* SDObsi

Where:

ObsIDXi  = input value provided by the user for observed value for an index as described in section 8.6 above

ZNormir = Random number deviate from a standard Normal distribution with a mean of 0.0 and SD of 1.0 for index *i* in simulation *r*

SDObsi *For single-year runs:*  SDObsi = √( (SDRepi)2 + (SDTSeasi)2)

Where values for SDRep = 0.269 and SDTSeas = 0.279

Definitions for each index used within RICT2 (which are only the abundance-weighted WHPT indices are as below copied from Section 6.3.3.2 (b) of the WHPT and other Abundance-Weighted Indices SEPA Report:

ObsIDX8r = (√(ObsIDX8) + ZObs8r)2 = rth simulated value for observed weighted WHPT NTAXA

ObsIDX9r = ObsIDX9 + ZObs9r = rth simulated value for observed weighted WHPT ASPT

ObsIDX7r = ObsIDX8r \* ObsIDX9r = rth simulated value for observed weighted WHPT Score

Within RICT2, these calculations are done using functions which can be found in the ClassificationfunctionsV2 (within the classification support files folder).

### Simulate variation in Bias (step 8 in process diagram above)

**8. Simulated variation in bias**

Ubiasir

Section 6.3.3.3

(Section 8.9)

This is simulation of user input bias which is calculated in RICT2 using the getUbias8r () function found in the ClassificationfunctionsV2.r (within the classification support files folder).

The algorithms used within this function are as below which is copied from Section 6.3.3.3 of the WHPT and other Abundance-Weighted Indices SEPA Report.

Definitions:

ObsIDXir = Simulation *r* Observed sample value of index *i* for current test site (uncorrected for bias)

Ubias8 = estimate of average net under-estimation of WHPT NTAXA for the observed sample

Ubias8 is either:

1. input by the user of the RICT2 software
2. estimated as 36% higher than the user-input bias (Ubias2) for number of BMWP taxa

i.e. Ubias8 = 1.36 Ubias2

Ubiasir = Estimate of bias (net under-estimation) of index *i* for simulation *r*

ObsIDXirB  = Bias-corrected observed value of index *i* for simulation r

Special case : when no WHPT taxa were recorded in the sample (i.e. ObsIDX8 = 0), assume none were missed (i.e, set Ubias8 = 0)

Ubias8r = bias (net under-estimate of number of WHPT taxa) for simulated sample r, estimated as a random deviate from a Poisson distribution with a mean of Ubias8

Zbias9r = Random number deviate from a standard Normal distribution with a mean of 0.0 and SD of 1.0

Ubias9r = abundance-weighted WHPT ASPT of the Ubias8r missed WHPT taxa for simulated sample r

= u9a + u9b \* ObsIDX9 + Zbias9r \* (u9c / √Ubias8r)

where u9a = 4.35 , u3b = 0.271 , u9c = 2.5

Then:

Ubias7r = Ubias8r \* Ubias9r = bias of abundance-weighted WHPT score for simulated sample r

### Correct observed values for bias (Step 9 in process diagram above)

**9. Bias corrected observed**

ObsIDXirB = ObsIDXir + Ubiasir

Section 6.3.3

(Section 8.8)

The aim of this step is to use the values provided in the previous steps, add them together to give an overall calculated observed value corrected for Bias.

Bias is a measure of the impact of laboratory error on the value of observed indices. It is calculated from audits of laboratory performance and is input by users as described in Section 8.6 above.

The main algorithm used is:

ObsIDXirB = ObsIDXir + Ubiasir

Where:

ObsIDXirB = Bias corrected observed value

ObsIDXir = Simulated observed value from Section 8.7 above

Ubiasir = Simulated Bias value from Section 8.8 above

Within RICT2, these calculations are done using functions which can be found in the ClassificationfunctionsV2 (within the classification support files folder).

The algorithms used within this function are as below which is copied from Section 6.3.3.3 of the WHPT and other Abundance-Weighted Indices SEPA Report.

ObsIDX7rB  = ObsIDX7r + Ubias7r = bias-corrected observed abundance-weighted WHPT Score

for simulation r

ObsIDX8rB  = ObsIDX8r + Ubias8r = bias-corrected observed abundance-weighted WHPT NTAXA

for simulation r

ObsIDX9rB  = ObsIDX7rB / ObsIDX8rB = bias-corrected observed abundance-weighted WHPT ASPT

for simulation r

It should be noted as mentioned previously, that it is important to do this for each index and also be aware of the numbering for each index. As indicated above, each index has a number, i.e.:

* WHPT Score (index 7)
* WHPT NTAXA (index 8)
* WHPT ASPT (index 9)

### Calculate Environmental Quality Ratio (EQR) (Step 10 in process diagram above)

**10. EQR = Bias corrected observed ÷ reference value**

EQRirB

= (ObsIDXirB ÷ ExpIDXRef,ir)

(Section 8.10)

The aim of this step is to express the observed value of the index as an indication of human impact by removing the effect of natural environmental conditions.

Status class boundaries for WFD are expressed as EQRs.

Simulated EQRs, based on simulated observed and expected values, are needed to estimate probabilities of class.

The main algorithm used is:

(EQRirB = (ObsIDXirB ÷ ExpIDXRef,ir)

Where:

EQRirB = simulated bias-corrected EQR value for index *i*

ObsIDXirB = Bias-corrected observed value of index *i* for simulation *r* as calculated in section 8.9.

ExpIDXRef,ir = Reference-adjusted Expected value of index *i* in simulation *r* as calculated in section 8.5.

This calculation is done directly within RICT2 code and not via any function.

### Combine spring, summer and autumn EQRs (Step 11 in process diagram above)

Who updates the .ppt slides? I have managed to update this slide …

**11. Simulated two season average**

EQRirB,SpAu

= (EQRirB,Sp + EQRirB,Au) / 2

*(i.e. average of all seasons involved)*

(Section 8.11)

The aim of this step is to increase precision of classification and take account of seasonal variations in environmental quality.

The WFD status must be based on EQRs calculated from spring, summer and autumn samples. These values must be calculated and available within the experiment for this step to calculate the seasonal average.

The EQR values for spring, summer and autumn samples are divided by the number of season (i.e. 3) to calculate the seasonal average for each simulation.

The main algorithm is:

EQRirB,SpSuAu = (EQRirB,Sp + EQRirB,Su + EQRirB,Au) / 3

Where:

EQRirB,SpSuAu = Season average EQR

EQRirB,Sp = EQR value for spring season as calculated in section 8.10

EQRirB,Su = EQR value for summer season as calculated in section 8.10

EQRirB,Au = EQR value for autumn season as calculated in section 8.10

This is calculated in RICT2 using the getAvgEQR\_SprSumAut () function found in the ClassificationfunctionsV2.r (within the classification support files folder).

### Determine the probability of the class being High, Good, Moderate, Poor and Bad status (Step 12 in process diagram above)

**12. Identify class using published EQR class** (Monte-Carlo simulation)

Prob H

Prob G

Prob M

Prob P

Prob B

Most probable class

Section 6.5

(Section 8.12)

The aim of this step is to take account of imprecision in monitoring data in the classification.

The probability of each class is simply the sum of the simulations resulting in that class and is included in the output file.

As copied from Section 6.5 of the WHPT and other Abundance-Weighted Indices SEPA Report:

*“Confidence of class*

The likelihood of the true status class (i.e. averaged across all possible samples) of a test site being each of the five possible WFD classes is estimated simply by applying the ‘status classification method’ to each simulation sample *r* in turn. Thus the class for simulation *r* is based on the EQI/EQR values for simulation *r*, namely the set of EQIir.

For each index and overall, the proportion of simulations assigned to a status class estimates the probability that the true (average) quality of the test site for that time period was of that ecological status class (based on its macroinvertebrates).”

So to summarise, the probability of each class is simply the sum of the simulations resulting in that class and this data is calculated and then included in the output file.

### Determine the worst class indicated by WHPT ASPT and WHPT Ntaxa using MINTA (step 13 in process diagram above)

**13. MINTA**

Section 1.7

(Section 8.13)

The aim for this step is to ensure that the classification reflects ecological impacts caused by any type of environmental pressure.

The overall status class for a sample is defined to be the worst of classification based on number of taxa and ASPT. This is referred to as the MINTA class (ie MINimum of classes based on number of Taxa and ASPT).

The WFD status of a site or water body is based on the worst class indicated by either WHPT ASPT or WHPT Ntaxa, following the ‘one out all out’ rule.

The WFD requires that all water bodies, including rivers sites, are classified into one of five ecological status class. These classes in RICT2 are coded as classes 1-5 with 1= ‘high’, 2=‘good’, 3=‘moderate’, 4=‘poor’ and 5=‘bad’.

There are a number of processes involved in this step as described below:

* The EQR values calculated in section 10 above are used within the getClassFromEQR\_ntaxa () function found in the ClassificationfunctionsV2.r (within the classification support files folder) to find the class and relevant boundary limit values.
* These values are then used within the getClassFromEQR\_aspt () function found in the ClassificationfunctionsV2.r (within the classification support files folder).
* The getMostProbableClass () function found in the ClassificationfunctionsV2.r (within the classification support files folder) is used to find the most probable class.

* The getMINTA\_ntaxa\_aspt () function found in the ClassificationfunctionsV2.r (within the classification support files folder is used to find the maximum of the two classes and which is the worst class. The worst class in then recorded in the output file and displayed in the MINTA column

### Determine overall status class for a water body based on samples collected from more than one site

The aim of this step is to estimate the overall quality of a water body when more than one site is monitored and to improve the precision of the classification

The Environment Agency collects samples from some water bodies from more than one site.

The Environment Agency currently uses VISCOUS software to calculate WFD status for water bodies monitored from more than one site, as well as sites sampled in more than one year.

Adding new functionality to allow status to be based on data collected from more than one site would streamline the classification process for the Environment Agency.

## Multi Year Classification

The aim of this is to take account of temporary differences in quality and improve precision if a user has more than one year of sample data.

River Invertebrate status must be assessed every 3 years for WFD. Where sites are sampled in a 3 year monitoring cycle, WFD status can be based on samples collected in more than one year.

Multi-year status classification takes account of inter-year variation based on observations from specific rivers and streams.

When the multi-year classification experiment is developed it will need to follow the agreed principles and process as described below:

* This will be a separate experiment that can be downloaded for any user wanting to carry out a multi-year classification.
* Assessment is always on a 3 year period.
* Users may supply 1,2 or 3 years data.
* A complete years data must include Spring, Summer and Autumn – otherwise this year cannot be included in the assessment.
* Therefore minimum requirement is 1 year of data which includes Spring, Summer and Autumn.
* Each year will be added as a separate row on the input file and for each year of data will have the same site name
* RICT2 will check each row to see if it is same site name, if it is will store the simulations and include this in the assessment and will carry out this check for the next row to collect a maximum of 3 years of data.
* There will be no changes needed to the input file as by the user selecting the multi-year experiment, RICT2 will be expecting multi-year data.
* Ralph will be able to obtain some test data which can be used to clarify if the calculations are correct.
* Output file will need to be amended to display the multi-year data and will display the term “multi-year” instead of the actual year in the Year column.
* Apart from the differences highlighted above, the multi-year classification will follow the same process at the Prediction and Classification procedures specified in the previous sections of this document.

## Production of Output files

An output file will be created from running a project experiment to show the calculated data results of the experiment run for prediction and classification.

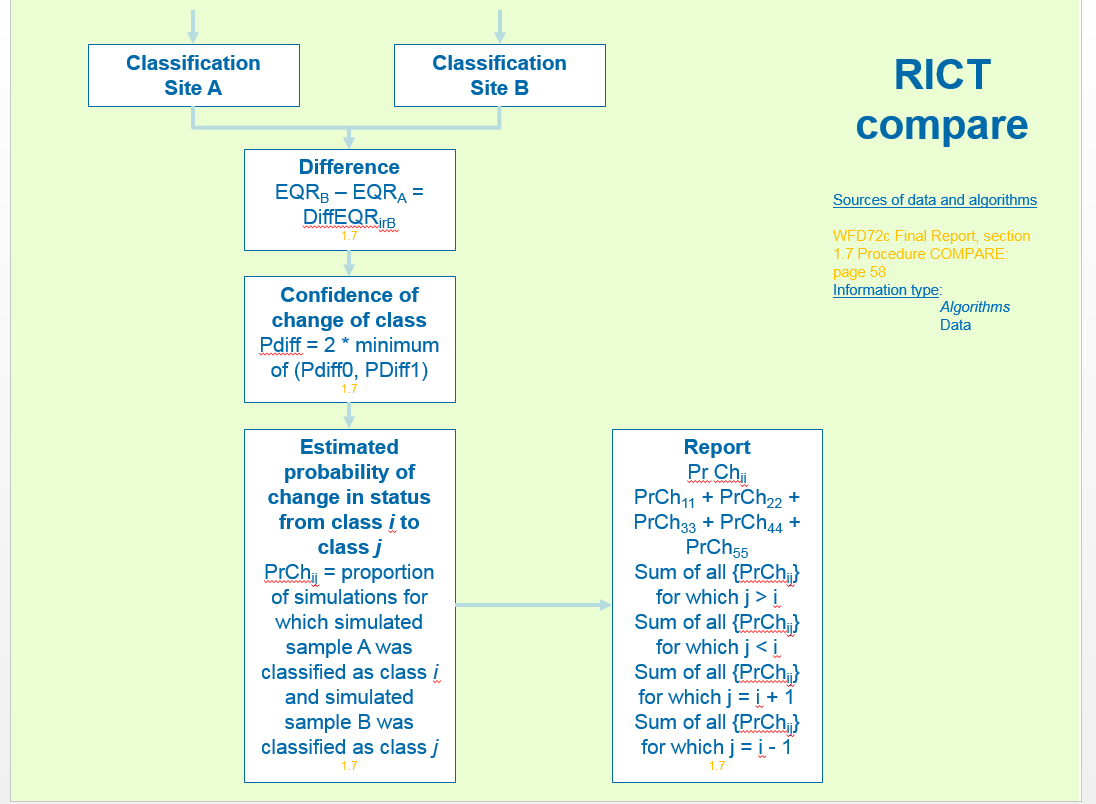
The output file will also indicate any validation errors found during the run of the project experiment and specify which values resulted in either a “fail” or “warn”.

The output file will be in csv format to be downloaded by the user and used for reporting. There will be an output file available for prediction and an output file available for classification.

Annex B contains an example table that lists all the data items that are included in the output file for each site.

## Compare Project Experiment

When the compare experiment is developed it will need to follow the agreed process as described in the flow diagram below:



Further information regarding the Compare function is given below and can be used for developing this function:

The Compare procedure allows the user to assess whether there is a real difference in EQR values and/or status class between a pair of samples and/or sites and/or time periods. This is done using an extension of the simulation techniques and algorithms described above that are used to assess the EQRs and ecological status class and the associated uncertainty for individual samples/sites and time periods.

The first sample in a pair is known as sample A and the second as sample B.

In each simulation r, independent simulated EQR values are calculated for each index i for each of the two samples using the methods described above for single samples. In this context, independent simulations means that the various random uncertainty terms are different and completely independent for the two simulated samples A and B for any simulation r.

For each simulation *r*, the difference in the two simulated EQR values (sample B value minus sample A value) is calculated for each index *i* to give DiffEQRirb.

**Confidence of change in EQR:**

Confidence limits for the difference in EQR values are obtained from the frequency distribution of the differences using the same approach as for single sample EQR values described above, namely, order the differences from smallest to largest and find the 2.5 and 97.5 percentile values of the order distribution.

**Confidence of change in status class:**

Each of the two simulated samples is assigned to a status class by applying the ‘status classification method’ independently to the simulated EQR values for each sample for that simulation. This is done repeatedly in *rN* simulations to build up a two-way frequency (and hence probability) distribution that the site (in its time period) from which sample A was taken was of status class 'x' (say) and the sites (in its time period) from which sample B was taken was of class 'y'.

From this two-way probability distribution, the probability that samples A and B are from the same quality band can be estimated, together with the probabilities that sample B is one band better than A, one band worse, two or more bands better, or two or more bands worse.

In terms of RICT2 software outputs, there is potentially one of these 5 x 5 probability tables of changes in class based on each index for which class limits are to be used and also for each multi-index classification and the overall status classification method. In each case, a wide range of outputs could be derived from this table as shown below.

If for coding purposes classes ‘high’ to ‘bad’ are referred to as classes 1-5 respectively, then:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| PrChij | | = | Proportion of simulations for which simulated sample A was classified as class *i* | | | | | | |
|  |  |  | and simulated sample B was classified as class *j*. | | | |  |  |  |
|  |  | =Estimated probability of change in status from class *i* to class *j* | | | | | |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  | Status class of sample B | | |  |  |
|  |  |  |  | high (1) | good (2) | mod (3) | poor (4) | bad (5) |  |
|  |  |  | high (1) | PrCh11 | PrCh12 | PrCh13 | PrCh14 | PrCh15 |  |
|  |  | Status | good (2) | PrCh21 | PrCh22 | PrCh23 | PrCh24 | PrCh25 |  |
|  | class of | | mod (3) | PrCh31 | PrCh32 | PrCh33 | PrCh34 | PrCh35 |  |
|  | sample A | | poor (4) | PrCh41 | PrCh42 | PrCh43 | PrCh44 | PrCh45 |  |
|  |  |  | bad (5) | PrCh51 | PrCh52 | PrCh53 | PrCh54 | PrCh55 |  |

Probability status class of site(s) from which samples A and B were taken is:

|  |  |
| --- | --- |
| Both status class *i* | = PrChii |
| Both the same class | = PrCh11 + PrCh22 + PrCh33 + PrCh44 + PrCh55 |
| Both B worse than A | = sum of all { PrChij } for which j > i |
| Both B better than A | = sum of all { PrChij } for which j < i |
| Both B one class worse than A | = sum of all { PrChij } for which j = i + 1 |
| Both B one class better than A | = sum of all { PrChij } for which j = i - 1 |

## Model 44 Prediction and Classification

The Centre for Ecology and Hydrology (CEH) have produced a database of GIS environmental variables data for every 50m of steam or river for use with the new RIVPACS GB Model M44.

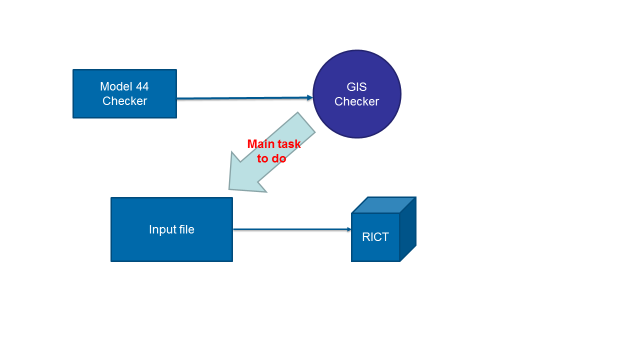
This model was produced by Ralph Clark and John Davey Bowker and is detailed in the report GIS Re-calibration of the hydromorphology-independent RIVPACS predictive model (M37): new model M44.

RICT2 will be used with the new RIVPACS Model 44 to give better predictions in streams impacted by flow alteration of siltation.

The development for this model has not been completed in phase 1 of this project but is planned for future development. The requirements for this development that have been captured to date are outlined below.

It is expected that the majority of the coding that has been done in RICT2 for prediction and classification will also be suitable when running the experiment for the Model 44 prediction and classification.

One of the main areas of development will be more up front work involving understanding how to extract the data from RIVPACS and to link to a location checker as outlined in the picture below:



One important change to understand is that the Environmental Variables for Model 44 experiment will be different as listed in the table below:

|  |  |
| --- | --- |
| SITE |  |
| Waterbody |  |
| Year |  |
| NGR |  |
| Easting |  |
| Northing |  |
| Alkalinity |  |
| Area\_CEH |  |
| AltBar\_CEH |  |
| Alt\_CEH |  |
| DFS\_CEH |  |
| Slope\_CEH |  |
| QCat\_CEH |  |
| Peat\_CEH | Value cannot exceed 100 or be less than 0 |
| Chalk\_O1\_CEH | Combined percentage figure for these 4 values cannot be more than 100 |
| Clay\_O1\_CEH | Combined percentage figure for these 4 values cannot be more than 100 |
| Hardrock\_O1\_CEH | Combined percentage figure for these 4 values cannot be more than 100 |
| Limestone\_O1\_CEH | Combined percentage figure for these 4 values cannot be more than 100 |

This will mean that a new version of the input spreadsheet may needed for these values. However, it should be noted that these values will be provided by the model and not entered by a user. Therefore validation requirements could be minimal as shown above. Validation may be required to check that values are numeric and not null.

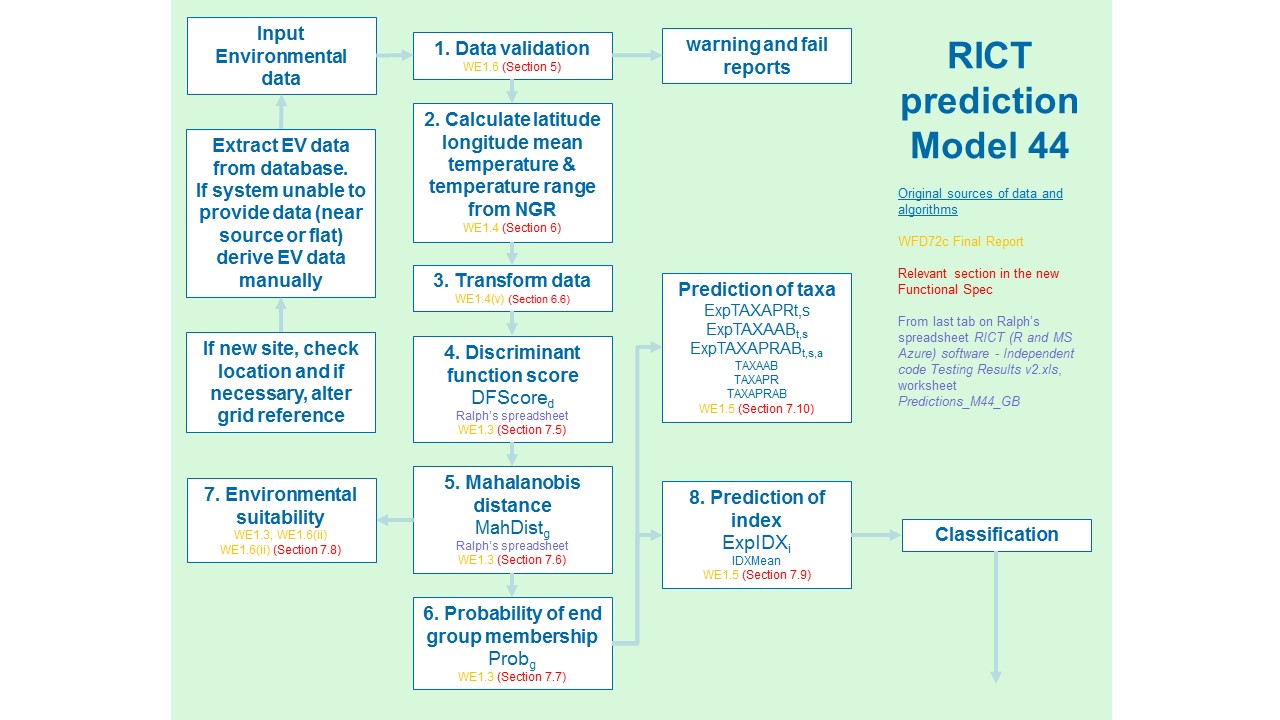
It is expected that the experiment will be built first and then checks will be made on results produced to determine if some validation may be required, i.e. warnings and failures based on a limits amount.

Replacement values will be required, especially for EVs that are not present in the model. A method for providing a replacement value will need to be determined. If a replacement value is available from the current RIVPACS IV GB or NI experiments, this can be used.

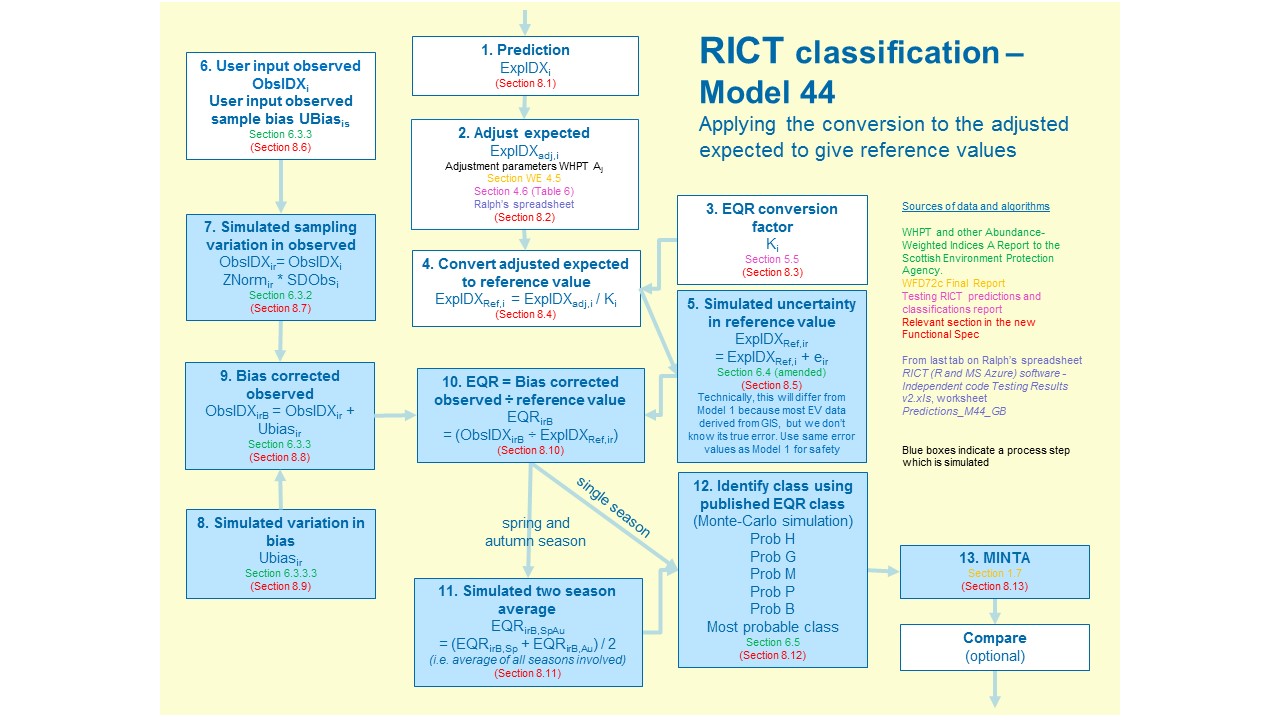
Another difference is that there are now 17 discriminant function scores and discriminant function end group means scores rather than 13. The values for these scores can be found in the Excel spreadsheet:

* RICT2 (R and MS Azure) software - Independent code Testing Results v2
  + Spreadsheet tab “Predictions\_M44\_GB” contains the data as produced by Ralph Clark. The Predictions\_M44\_GB contains the data and scrolling right will also show how the data is used and put together within prediction and classification by following the orange arrows.

The flow diagram below summarises the prediction process and calculations as well as identifying the relevant report for further information:



The flow diagram below summarises the classification process and calculations as well as identifying the relevant report for further information:



\*\* as per previous comment on this slide \*\*

## Non-Functional Requirements

### Operational Support

The tool needs to be freely available for anyone to access online.

User support for EA users will be via the Defra service desk (at Tier 4 service level) who will route the issue to the appropriate team to deal with. This routing could be to RICT2 support email or to the Environmental systems team depending on the nature of the issue. EA users will also be able to access an article in MyIT which will provide FAQs to help users with common issues.

User support for other UK regulatory agency users will be via the support email box called RICT2 support.

There will be no official contracted support for external users but they can send issues and comments to the RICT2 support inbox

### Look and Feel of User Interface

The user interface will be provided via Azure Studio which is intuitive and ‘user friendly’ and consistent with similar windows-based or web-based user interface applications.

Official RICT2 experiments will be collated in a designated “Collection” in Microsoft Azure Learning Studio to help users find all the experiments relevant to RICT2.

### Input and Output Files

Input files will be a standard template in Excel which can be downloaded, completed by the user and then uploaded as a CSV file to RICT2. A user guide will be available to help users understand and complete the input file.

The output file will be produced by RICT2 as a CSV file following the run. A CSV file can easily be opened in Excel or other applications to view the results.

### Hosting

RICT2 is cloud hosted within Microsoft Azure Learning Studio.

### Accessibility

Access to RICT2 functionality may be required by a number of different device types. RICT2 will be useable at a minimum by the following device types (model not specified):

* Desktop
* Laptop

Access to RICT2 functionality may be required by a number of different browser types. RICT2 will be useable in the following browser types (which must be updated to the latest version as required by Microsoft Azure):

* IE
* Google Chrome
* Firefox
* Edge

Users will only have access to their own data and cannot access data input or output from other users.

### Performance

RICT2 will be accessible by multiple concurrent users at any one time without degradation of performance or limitation of features. Users will be internal EA, other UK regulatory body users (SEPA, NRW and NIEA) and external users.

RICT2 will be capable of processing the prediction and classification of a very large number of sites (approx. 5,000) once per year.

RICT2 will meet the following rough order estimates of service usage and performance requirements:

* Service Availability: In excess of 97% at all times;
* Service Recovery: RTO (Recovery Time Objective) in 48 hours;
* Maximum Nos. of Users: <1000;
* Max number of Concurrent Users: 50.

RICT2 availability must be at an acceptable level. Planned downtime for maintenance must be advised in advance (minimum 48 hours’ notice). Unplanned downtime must be communicated regularly with up to date information on tools/services affected and projected downtime.

### System Administration

An “expert” user (system administrator) will be identified and given permission to publish new standard experiments using default values to the “official” Microsoft Azure ML Studio Studio gallery collection.

These experiments will be “locked down” for normal users and can only be amended by downloading to the users own work area to create their own copy of a project experiment.

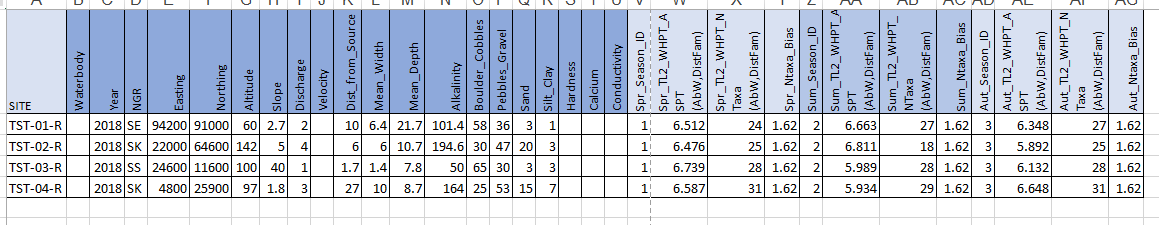
The System administrator will be permitted to amend the published standard experiments and the data tables used by the experiment (e.g. amend end group data) if required.

### Data Protection

Data protection requirements are handled by Microsoft as the tool is hosted on Microsoft ML Studio. It is freely available to access but requires the user to use an existing/or create a Microsoft account. This account requires an email address and password. These credentials are held solely by Microsoft and the usage is outlined in the Microsoft Privacy Statement (<https://privacy.microsoft.com/en-gb/privacystatement>).

## Annex A: Input File

A screenshot of the input file is as below:



This table shows test data which can be left in by the user to help understand format to use and run with the experiment.

The link below shows the complete input file excel spreadsheet and includes test data and validation rules:



A full description of each column for the input table and how it is completed can be found within the RICT2 user guide. For this functional specification, a summary list of the column headings is shown below:

|  |  |
| --- | --- |
| **Column Heading** | **Definition** |
| Site | Site reference number of the site the sample was taken |
| Waterbody | Name of the waterbody |
| Year | Year that the sample was taken |
| NGR | First two letters of the national grid reference for the site location |
| Easting | First five numbers of the national grid reference for the site location |
| Northing | Second five numbers of the national grid reference for the site location |
| Altitude | Measure of the Altitude at the sampling site |
| Slope | Measure of the Slope of the river or stream at the sampling site |
| Discharge | Measure of the Discharge Category at the sampling site |
| Velocity | Measure of the Velocity at the sampling site |
| Dist\_from\_Source | Measure of the Distance from the source of the river or stream at the sampling site |
| Mean\_Width | Measure of the Width of the river or stream at the sampling site |
| Mean\_Depth | Measure of the Depth of the river or stream at the sampling site |
| Alkalinity | Measure of the Alkalinity of the river or stream at the sampling site |
| Boulder\_Cobbles | Percentage of the stream substrate that would be classed as Boulders and Cobbles |
| Pebbles\_Gravel | Percentage of the stream substrate that would be classed as Pebbles and Gravel |
| Sand | Percentage of the stream substrate that would be classed as Sand |
| Silt\_Clay | Percentage of the stream substrate that would be classed as Silt and Clay |
| Hardness | Measure of the Hardness of the river or stream at the sampling site |
| Calcium | Measure of the Calcium of the river or stream at the sampling site |
| Conductivity | Measure of the Conductivity of the river or stream at the sampling site |
| Spr\_Season\_ID | ID number for Spring Season which is “1” |
| Spr\_TL2\_WHPT\_ASPT (AbW,DistFam) | Index value for WHPT ASPT for spring |
| Spr\_TL2\_WHPT\_NTaxa (AbW,DistFam) | Index value for WHPT NTaxa for spring |
| Spr\_Ntaxa\_Bias | Bias value for Spring |
| Sum\_Season\_ID | ID number for Summer Season which is “2” |
| Sum\_TL2\_WHPT\_ASPT (AbW,DistFam) | Index value for WHPT ASPT for summer |
| Sum\_TL2\_WHPT\_NTaxa (AbW,DistFam) | Index value for WHPT NTaxa for summer |
| Sum\_Ntaxa\_Bias | Bias value for Summer |
| Aut\_Season\_ID | ID number for Autumn Season which is “3” |
| Aut\_TL2\_WHPT\_ASPT (AbW,DistFam) | Index value for WHPT ASPT for autumn |
| Aut\_TL2\_WHPT\_NTaxa (AbW,DistFam) | Index value for WHPT NTaxa for autumn |
| Aut\_Ntaxa\_Bias | Bias value for Autumn |

Green rows above indicate input Environmental Values

Blue rows above indicate input Biological Values.

## Annex B: Output File

RICT2 will produce two output files; one for prediction, one for classification

**Prediction Output file**

An example of the output file for prediction is linked below:



A full description of each column for the prediction output table and what the values indicate can be found within the RICT2 user guide. For this functional specification, a summary list of the column headings is shown below:

|  |  |
| --- | --- |
| **Column Heading** | **Definition** |
| SITE | Site reference number |
| LATITUDE | Location point |
| LONGITUDE | Location point |
| LOG.ALTITUDE | Logarithmic value for Altitude |
| LOG.DISTANCE.FROM.SOURCE | Logarithmic value for Distance from source |
| LOG.WIDTH | Logarithmic value for Stream Width |
| LOG.DEPTH | Logarithmic value for Stream Depth |
| MEAN.SUBSTRATUM | Calculated mean substratum |
| DISCHARGE.CATEGORY | Discharge category |
| ALKALINITY | Alkalinity |
| LOG.ALKALINITY | Logarithmic value for Alkalinity |
| LOG.SLOPE | Logarithmic value for Slope |
| MEAN.AIR.TEMP | Calculated Mean Air Temperature |
| AIR.TEMP.RANGE | Calculated Air Temperature Range |
| p1 – p43 | Calculated probability of end group membership |
| SuitCode | Calculated Suitability code for test site |
| SuitText | Relevant text for the calculated suitability code |
| BelongsTo\_endGrp | Predicted probable end group the test site belongs to |
| TL2\_WHPT\_NTAXA\_AbW\_DistFam\_spr | Predicted index value for WHPT NTAXA for spring (used in the classification) |
| TL2\_WHPT\_ASPT\_AbW\_DistFam\_spr | Predicted index value for WHPT ASPT for spring |
| TL2\_WHPT\_NTAXA\_AbW\_DistFam\_sum | Predicted index value for WHPT NTAXA for summer (used in the classification) |
| TL2\_WHPT\_ASPT\_AbW\_DistFam\_sum | Predicted index value for WHPT ASPT for summer |
| TL2\_WHPT\_NTAXA\_AbW\_DistFam\_aut | Predicted index value for WHPT NTAXA for autumn |
| TL2\_WHPT\_ASPT\_AbW\_DistFam\_aut | Predicted index value for WHPT ASPT for autumn |
| TL2\_WHPT\_NTAXA\_AbW\_CompFam\_spr | Not used |
| TL2\_WHPT\_NTAXA\_ASPT\_CompFam\_spr | Not used |
| TL2\_WHPT\_NTAXA\_AbW\_CompFam\_sum | Not used |
| TL2\_WHPT\_NTAXA\_ASPT\_CompFam\_sum | Not used |
| TL2\_WHPT\_NTAXA\_AbW\_CompFam\_aut | Not used |
| TL2\_WHPT\_ASPT\_AbW\_CompFam\_aut | Not used |
| WATERBODY | Name of waterbody |
| YEAR | Year of sample |
| SPR\_SEASON\_ID | ID number for season |
| SPR\_TL2\_WHPT\_ASPT..ABW.DISTFAM. | Observed value for WHPT ASPT for spring |
| SPR\_TL2\_WHPT\_NTAXA..ABW.DISTFAM. | Observed value for WHPT NTAXA for spring |
| SPR\_NTAXA\_BIAS | Bias value used |
| SUM\_SEASON\_ID | ID number for season |
| SUM\_TL2\_WHPT\_ASPT..ABW.DISTFAM. | Observed value for WHPT ASPT for summer |
| SUM\_TL2\_WHPT\_NTAXA..ABW.DISTFAM. | Observed value for WHPT NTAXA for summer |
| SUM\_NTAXA\_BIAS | Bias value used |
| AUT\_SEASON\_ID | ID number for season |
| AUT\_TL2\_WHPT\_ASPT..ABW.DISTFAM. | Observed value for WHPT ASPT for autumn |
| AUT\_TL2\_WHPT\_NTAXA..ABW.DISTFAM. | Observed value for WHPT NTAXA for autumn |
| AUT\_NTAXA\_BIAS | Bias value used |

Green rows above indicate calculated Environmental Values.

Purple rows above indicate calculated Prediction output values.

Blue rows above indicate input Biological Values that go into classification process.

**Classification Output file**

An example of the output file for classification is linked below and includes data from the latest national classification run:



A full description of each column for the classification output table and what the values indicate can be found within the RICT2 user guide. For this functional specification, a summary list of the column headings is shown below:

|  |  |
| --- | --- |
| **Column Heading** | **Definition** |
| **SITE** | Site name |
| **YEAR** | Year that the sample was taken |
| **WATERBODY** | Name of the waterbody |
| **NTAXA\_eqr\_av\_spr** | Average EQR for Number of Taxa for spring from the monte-carlo simulations |
| **H\_NTAXA\_spr** | Probability of high status for spring |
| **G\_NTAXA\_spr** | Probability of good status for spring |
| **M\_NTAXA\_spr** | Probability of moderate status for spring |
| **P\_NTAXA\_spr** | Probability of poor status for spring |
| **B\_NTAXA\_spr** | Probability of bad status for spring |
| **mostProb\_NTAXA\_spr** | The status (H, G, M, P or B) for spring with the greatest probability |
| **NTAXA\_eqr\_av\_aut** | Average EQR for Number of Taxa for autumn from the monte-carlo simulations |
| **H\_NTAXA\_aut** | Probability of high status for autumn |
| **G\_NTAXA\_aut** | Probability of good status for autumn |
| **M\_NTAXA\_aut** | Probability of moderate status for autumn |
| **P\_NTAXA\_aut** | Probability of poor status for autumn |
| **B\_NTAXA\_aut** | Probability of bad status for autumn |
| **mostProb\_NTAXA\_aut** | The status (H, G, M, P or B) for autumn with the greatest probability |
| **H\_NTAXA\_spr\_aut** | Probability of high status for combined spring and autumn |
| **G\_NTAXA\_spr\_aut** | Probability of good status for combined spring and autumn |
| **M\_NTAXA\_spr\_aut** | Probability of moderate status for combined spring and autumn |
| **P\_NTAXA\_spr\_aut** | Probability of poor status for combined spring and autumn |
| **B\_NTAXA\_spr\_aut** | Probability of bad status for combined spring and autumn |
| **mostProb\_NTAXA\_spr\_aut** | The status (H, G, M, P or B) for combined spring and autumn with the greatest probability |
| **NTAXA\_aver\_spr\_aut** | Average EQR for Number of Taxa for spring and autumn from the monte-carlo simulations |
| **ASPT\_eqr\_av\_spr** | Average EQR for ASPT for spring from the monte-carlo simulations |
| **ASPT\_eqr\_av\_aut** |  |
| **H\_ASPT\_spr** | Probability of high status for spring |
| **G\_ASPT\_spr** | Probability of good status for spring |
| **M\_ASPT\_spr** | Probability of moderate status for spring |
| **P\_ASPT\_spr** | Probability of poor status for spring |
| **B\_ASPT\_spr** | Probability of bad status for spring |
| **mostProb\_ASPT\_spr** | The status (H, G, M, P or B) for spring with the greatest probability |
| **H\_ASPT\_aut** | Average EQR for Number of Taxa for autumn from the monte-carlo simulations |
| **G\_ASPT\_aut** | Probability of high status for autumn |
| **M\_ASPT\_aut** | Probability of good status for autumn |
| **P\_ASPT\_aut** | Probability of moderate status for autumn |
| **B\_ASPT\_aut** | Probability of poor status for autumn |
| **mostProb\_ASPT\_aut** | Probability of bad status for autumn |
| **H\_ASPT\_spr\_aut** | Probability of high status for combined spring and autumn |
| **G\_ASPT\_spr\_aut** | Probability of good status for combined spring and autumn |
| **M\_ASPT\_spr\_aut** | Probability of moderate status for combined spring and autumn |
| **P\_ASPT\_spr\_aut** | Probability of poor status for combined spring and autumn |
| **B\_ASPT\_spr\_aut** | Probability of bad status for combined spring and autumn |
| **mostProb\_ASPT\_spr\_aut** | Probability of high status for combined spring and autumn |
| **ASPT\_aver\_spr\_aut** | Average EQR for ASPT for combined spring and autumn from the monte-carlo simulations |
| **mintawhpt\_spr\_H** | Probability of high status for either ASPT or NTAXA depending on which is worse for spring |
| **mintawhpt\_spr\_G** | Probability of good status for either ASPT or NTAXA depending on which is worse for spring |
| **mintawhpt\_spr\_M** | Probability of moderate status for either ASPT or NTAXA depending on which is worse for spring |
| **mintawhpt\_spr\_P** | Probability of poor status for either ASPT or NTAXA depending on which is worse for spring |
| **mintawhpt\_spr\_B** | Probability of bad status for either ASPT or NTAXA depending on which is worse for spring |
| **mintawhpt\_spr\_mostProb** | The status (H, G, M, P or B) for spring for either ASPT or NTAXA depending on which is worse for spring |
| **mintawhpt\_aut\_H** | Probability of high status for either ASPT or NTAXA depending on which is worse for autumn |
| **mintawhpt\_aut\_G** | Probability of good status for either ASPT or NTAXA depending on which is worse for autumn |
| **mintawhpt\_aut\_M** | Probability of moderate status for either ASPT or NTAXA depending on which is worse for autumn |
| **mintawhpt\_aut\_P** | Probability of poor status for either ASPT or NTAXA depending on which is worse for autumn |
| **mintawhpt\_aut\_B** | Probability of bad status for either ASPT or NTAXA depending on which is worse for autumn |
| **mintawhpt\_aut\_mostProb** | The status (H, G, M, P or B) for spring for either ASPT or NTAXA depending on which is worse for autumn |
| **mintawhpt\_spr\_aut\_H** | Probability of high status for either ASPT or NTAXA depending on which is worse for combined spring and autumn |
| **mintawhpt\_spr\_aut\_G** | Probability of good status for either ASPT or NTAXA depending on which is worse for combined spring and autumn |
| **mintawhpt\_spr\_aut\_M** | Probability of moderate status for either ASPT or NTAXA depending on which is worse for combined spring and autumn |
| **mintawhpt\_spr\_aut\_P** | Probability of poor status for either ASPT or NTAXA depending on which is worse for combined spring and autumn |
| **mintawhpt\_spr\_aut\_B** | Probability of bad status for either ASPT or NTAXA depending on which is worse for combined spring and autumn |
| **mintawhpt\_spr\_aut\_mostProb** | The status (H, G, M, P or B) for spring for either ASPT or NTAXA depending on which is worse for combined spring and autumn |

## Annex C: Critical Success Factors

The CSF specified below denotes the minimum requirements as agreed with the project team and stakeholders as agreed in the outline business case:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | Score |
| 1 | Primary | RICT can complete the full national classification in June 2019 | 3 |
| 2 | Primary | Improved performance of RICT – end to end processing time | 3 |
| 3 | Primary | New RIVPACS model added to RICT | 3 |
| 4 | Primary | Platform and associated code accessible for future development | 3 |
|  |  |  |  |
| 5 | Secondary | Agree ongoing support and maintenance arrangement (Hosting) | 2 |
| 6 | Secondary | Agree ongoing support, maintenance and development arrangement (Application) | 2 |
| 7 | Secondary | Improved technical and scientific resilience | 2 |
|  |  |  |  |
| 8 | Tertiary | Critical bugs resolved | 1 |
| 9 | Tertiary | Critical functionality added | 1 |
| 10 | Tertiary | Non critical bugs resolved | 1 |
| 11 | Tertiary | Non critical functionality added | 1 |
| 12 | Tertiary | Produce an updated functional specification and architecture documentation | 1 |

Table 1: Critical Success Factors