Exploratory classification and cluster analyses of New Zealand coastal hydrosystems

Timothy Jones

July 2014

A report prepared for the Department of Conservation (DOC) and the National Institute of Water and Atmospheric Research (NIWA)

Executive Summary

The aim of this report is to examine potential statistical grouping methods for the classification of New Zealand coastal hydrosystems. Classification schemes are a key tool for environmental management and to inform conservation decisions, helping to ensure that the most appropriate measures are taken to manage and preserve each system type. This report examines three alternative statistical clustering methods and their application to the grouping of NZ coastal hydrosystems based on a dataset of physical and environmental variables. The three methods examined are (1) Hierarchical clustering based on Euclidean distance measures, (2) Hierarchical clustering based on Random Forests distance measures and (3) Model-based clustering. The first method is well accepted and perhaps the most frequently used, and determines group structure based on absolute distances between objects in multivariate space. The second method is broadly similar to the first, but the distances between objects used to identify the group structure is determined via a machine learning algorithm. The third method differs from the first two in that a statistical mixture model is developed and then optimised to match the distribution of the data points. The results of the cluster analyses are compared to the classification system identified in Hume et al. (2007) to identify similarities and dissimilarities between the results of the cluster analyses and a classification system identified by expert knowledge. In addition, the strengths and weaknesses of each method are discussed.

The statistical clustering methods were initially applied to a reduced dataset (R12/TEV – river input relative to volume, STP/TEV – tidal input relative to volume, SCI – structural complexity and MCI – mouth closure) of physical variables to provide a direct comparison with the rule-based classification developed in Hume et al. (2007). For the Euclidean hierarchical method, analyses were performed on scaled and unscaled datasets and a range of data transformations were investigated, which were required to standardise the datasets prior to analysis. The resulting group structures distinguished primarily closed systems, such as lakes and lagoons (Hume class A), from completely open systems such as bays (Hume class D), which were also differentiated from river mouths (Hume class B). However, a large proportion of partially closed systems (classes E-H) were placed into a single group or multiple groups of mixed class. The random forests hierarchical method successfully differentiated between bays (Hume class D), which were split into high and low STP/TEV groups, river mouths (Hume class B), which were split into high and low R12/TEV groups, and lakes and lagoons (Hume class A). However, the remaining systems (classes E-H, including harbour systems, sounds, fiords and estuaries) were placed into groups containing multiple class types. The model-based cluster

analysis successfully differentiated bays (Hume class D), lakes and lagoons (Hume class A), and combined large complex hydrosystems, such as harbour systems, sounds and fiords into a single group, accounting for classes G and H. However, a large proportion of classes B, E and F were split among a series of different groups with different characteristics. In general, each analysis method bore some similarities to the Hume et al. (2007) classification, but in most cases the similarities were for systems with physical variables at the extremes of their ranges (e.g. river mouths, bays and closed systems exist at the extremes of R12/TEV, SCI and MCI, respectively, and were most commonly distinguished by the statistical methods). Systems with intermediate values for these physical variables, which make up those in Hume classes E-H, were often grouped together or split into different groupings than those identified by Hume et al. (2007).

The analyses were subsequently expanded to include additional physical variables (addition of % IA – intertidal area, mean depth and CLA/EWA – catchment land area relative to system water area) to further distinguish among systems. Between seven and eleven groups were identified from this dataset, but the optimal number of groups for the two hierarchical methods was difficult to ascertain. Each analysis method performed similarly to each other in that groups were created for bays, river mouths, closed systems and partially closed systems, but the exact breakdown of groups differed among methods, particularly for partially closed systems, as detailed in the table below.

Table summarising classification schemes identified by the three clustering methods based on SCI, MCI, STP/TEV, R12/TEV, CLA/EWA, % intertidal area and mean depth

Method	Closed	Open		Partially Closed	
		River Mouths	Bays	Harbours, Inlets, Estuaries	Fiords and Sounds
Euclidean (9 groups)	1 Group	1 – Low % IA 2 - High % IA	1 - Open and simple 2 - Recessed and complex	1 – Low structural complexity 2 – High structural complexity, low R12/TEV 3 – High structural complexity, high R12/TEV	1 Group
Random Forests (9 groups)	1 Group	1 Group	1 - Deep 2 - Shallow	1 – High R12/TEV 2 – Low structural complexity, high % IA 3 – High structural complexity, moderate % IA 4 – High structural complexity, high % IA 5 – High structural complexity, predominantly subtidal	NA
Model Based (10 groups)	1 Group	1 - Low R12/TEV 2 - High R12/TEV	1 - Open and simple 2 - Recessed and complex	1 – High structural complexity, low R12/TEV 2 – High structural complexity, moderate R12/TEV 3 – High structural complexity, high R12/TEV	1 Group

With the addition of intertidal area and mean depth all three methods seem to primarily distinguish among prevailing inflow/outflow (R12 and STP) regimes, and then further split systems based on morphology in terms of depth, intertidal area and structural complexity. In all cases the resulting groups were relatively easy to identify in terms of specific variables, and the addition of mean depth and intertidal area enabled several valid distinctions to be made among systems. These group schemes also compared favourably to the Hume classification with many of the Hume classes uniquely identified as a single group, or multiple groups split along depth or intertidal area axes. The exception was for Hume classes E and F, which were placed into multiple groups containing mixtures of E and F systems, with each group characterised by a different set of physical variables. This suggests that statistically there are alternate, more statistically robust, ways to group these systems than that specified in Hume et al. (2007).

Comparing methods, there is no single "Gold-standard" analysis method for clustering data, and all three methods used in this report provide a valid assessment of the grouping structure of the data. However, these analyses highlighted three major analysis decisions across methods that can affect the resultant classification scheme. These consist of (1) data pre-treatment, (2) hierarchical linkage type and (3) the number of groups, summarised in the table below. The Euclidean hierarchical analysis is demonstrated to be sensitive to all three analysis decisions, with data transformation and scaling having a large influence on group structure, whilst there is no clearly defined method for identifying the best number of groups. The random-forests method is not sensitive to data scale or differences in the magnitude of physical variables and so is not affected by data pre-treatment but is sensitive to linkage type and there is no clearly defined method for identifying the best number of groups. Model-based clustering is sensitive only to data pre-treatment as no linkage type is utilised, and the number of groups is determined using an information criterion which identifies the best number of groups.

Major analysis decisions that can affect group structure summarised by analysis type.

Cluster analysis type	Source of uncertainty/Analysis decision		
	Data pre-treatment	Linkage type	Number of Groups
Euclidean hierarchical	Y	Y	Y
Random forests	N	Y	Y
Model-based	Y	N	N

Model-based and random forests therefore require the least number of analysis decisions, and resultantly there are fewer potential group schemes to evaluate. However, the number of possible permutations of physical variable transformations can be high, which can lead to many potential "valid" datasets to evaluate when performing model-based clustering. By comparison the random forests method is not governed by the scale of the variables (i.e. splits are chosen based on the relative differences among points, rather than the absolute differences) and pre-treatment or variable type has no influence on the results. In addition, the remaining analysis decisions for the random forests method (linkage type and number of groups) are both fewer and easier to evaluate based on the resulting classification schemes. Therefore, this report recommends the use of the random forests methodology given the large disparity in the scales and variances of the physical variables. This method combined with a suitable choice of linkage criterion and number of groups, as judged by expert opinion, is best suited to identify the classification scheme that best reflects the variability of the data, without adding unnecessary groupings.

Table of Contents

1.0 – Introduction and Rationale

The aim of this report is to examine potential grouping and classification schemes of New Zealand coastal hydrosystems. Classifications are necessary to aid in management and conservation decisions, such that the most appropriate measures are taken to manage and preserve each system type. They are also valuable in identifying systems that play an important role in the life cycle of particular species, which may show a preference to certain system attributes.

There are a wide variety of statistical methods for the clustering or unsupervised classification (the act of grouping individuals or systems into approximately similar groups) of multivariate data. The following is a short, non-exhaustive, list of the methods that are currently available:

- Hierarchical clustering
- K-means clustering
- Partitioning around medoids (PAM)
- Model-based clustering
- Fuzzy clustering

Within each of these methods there are also a range of options that can affect the grouping structure of the data, such as distance measure (the measure chosen to represent the similarity/dissimilarity among observations), clustering type (the specific algorithm used to group observations) and most importantly the number of groups. In particular hierarchical clustering is a broad reaching term with a multitude of options for the grouping of data.

This report examines three alternative clustering methods and their application to the grouping of NZ coastal hydrosystems based on a dataset of physical and environmental variables. The three methods examined are (1) Hierarchical clustering based on Euclidean distance measures, (2) Hierarchical clustering based on random forests distance measures and (3) Model-based clustering. The first method is perhaps the most widely used clustering method (although distance measures may vary depending on the type of data being examined) and has been utilised in a broad range of studies. The second and third methods are relatively recent additions to the catalogue of clustering methods, and each addresses a particular shortcoming of the first method, namely scale and transformation invariance (i.e. group

structures are not affected by changing the spread or skew of the data) and a data-driven selection of an optimal number of groups, respectively.

The remainder of this report is split into four sections. Section 2 examines the application of the three clustering methods to a reduced dataset (four physical variables) of physical variables to provide a direct comparison with the rule-based classification developed in Hume et al. (2007). This section also provides an exploration of the alternative approaches to clustering the data regarding data transformation, scaling and alternative hierarchical clustering algorithms. Section 3 expands on Section 2, incorporating additional physical variables and examining how grouping structures are affected by the incorporation of these additional factors (a third dataset was also examined, and brief results from applying the three methods to this dataset are provided as an appendix). Sections 4 and 5 summarise the results in Sections 2 and 3 and provide recommendations based on the advantages and disadvantages of each method and analysis type.

2.0 – Examination of statistical methods and comparison to Hume et al. (2007)

2.1 – Introduction

The aim of this section is to highlight the three statistical methods and to provide a direct comparison of the resultant classification schemes with those identified by Hume et al. (2007) (henceforth this will be referred to as "The Hume classification", for brevity). The Hume classification is based on a decision tree rule set regarding the primary discriminatory variables of 443 coastal hydrosystems. These variables are:

- R12/TEV – total mean volume of water flowing into the system over a tidal cycle (R12) divided by the total estuary volume at high water (TEV)
- STP/TEV – the volume of water entering the system on the flood or incoming tide (STP) divided by the total estuary volume at high water (TEV)
- SCI – shoreline complexity index calculated as the reciprocal of the length of the system perimeter divided by the circumference of a circle that has the same area as the system. SCI varies from 1.0 (a simple circular basin) to <0.1 (very complex shoreline with multiple arms).
- MCI – mouth closure index calculated as the width of the system mouth, divided by the length of the perimeter of the estuary shoreline. This ratio is always <0.4. MCI is a measure of the openness of the estuary mouth and varies from ~0.3 (wide mouth) to <0.01 (very narrow and constricted entrance).

Based on these variables (and an additional variable that wasn't available for this analysis) a classification system with eight distinct classes was identified ranging from A-H, with the following characteristics:

- A – defined by zero STP/TEV, typically coastal lakes or lagoons
- B, C – defined by high STP/TEV and high R12/TEV, typically tidal river mouths
- D – defined by high STP/TEV and low R12/TEV, with high SCI and MCI, typically coastal embayments
- E – defined by high STP/TEV and low R12/TEV, with high SCI, but low MCI, typically tidal or barrier enclosed lagoons
- F – defined by high STP/TEV and low R12/TEV, with low SCI, typically barrier enclosed lagoons or drowned valleys

- G – defined by lower STP/TEV and higher R12/TEV, typically fiords or sounds
- H – defined by lower STP/TEV and low R12/TEV, typically drowned valleys, sounds or fiords

For this analysis groups B and C are pooled into a single group "B", as the variable that distinguishes among them (elongation index – EE) was not available for all systems, and therefore could not be included because of the potential for bias toward selecting B and C systems as a separate artificial group based on EE.

In this section the dataset of the four variables for each system are subjected to the three clustering methodologies and the resulting classifications are compared to the Hume classification. For each method the appropriate number of groups is explored using either cross-validation or model selection criteria, and ~ 2 grouping schemes (i.e. number of groups) are examined for each method. The characteristics of each of the resultant clustering schemes are examined via plots of the physical variables, calculation of variable ranges (maximum, minimum, mean and standard deviation), examination of systems within each group (i.e. are systems named bays, sounds, lakes or lagoons restricted to particular classes) and examination of potential rule-based schemes that replicate the observed grouping structure.

2.2 – Method 1: Hierarchical clustering based on Euclidean distance measures

The hierarchical clustering method is based on grouping objects into a tree of clusters. At any "height" of the tree there are a number of groups, with objects within each group displaying a greater similarity to each other than to objects in other groups. There are two main methods of hierarchical clustering. Agglomerative clustering begins with each observation in its own group, and observations are subsequently joined based on the distance between observations/clusters and some linkage criteria. Divisive clustering is the opposite and begins with each observation in a single group that is subsequently split based on similarity measures and criterion (Hastie et al. 2009). In this report only agglomerative clustering methods are used.

In order to decide on how to split or join groups, a measure of similarity/dissimilarity is required. The choice of similarity/dissimilarity metric is an important one and will partly be driven by the type of data being examined. Some similarity metrics include

- Euclidean distance – straight line distance between objects in multivariate space
- Manhattan distance – shortest path length between objects where paths must be parallel or perpendicular to multivariate axes

- Bray-Curtis dissimilarity – metric of compositional similarity commonly used in ecological studies

For physical, or environmental variables, the Euclidean distance measure is commonly used and is given by

$$d_{A,B} = \sqrt{\sum_{i}^{N} (A_i - B_i)^2}$$

(eqn. 1)

Where $d_{A,B}$ is the Euclidean distance between objects *A* and *B*, A_i and B_i are the physical co-ordinates of *A* and *B* on the i^{th} multivariate axis, and *N* is the number of dimensions. It has the quality of representing the shortest possible distance between two objects in multivariate space.

Clusters are identified by combining these distance measures with the calculation of some linkage criterion, which usually corresponds to finding the maximum or minimum possible value across all potential joins (Hastie et al. 2009). Some linkage criteria include:

- Complete linkage – at each stage groups are combined by finding the two groups that have the minimum of the maximum distance between any two members of the constituent groups (i.e. minimises the furthest "neighbour" within each combined group)
- Single linkage – at each stage groups are combined by finding the two groups that have the minimum distance between any two observations that are in different groups, (i.e. minimises the nearest "neighbour" within each combined group)
- Average linkage – aims for a compromise between single and complete linkage methods
- Linkage based on Ward's criterion – at each stage groups are combined that minimise the sum of the within-cluster variances, identifying compact, approximately spherical clusters

Agglomerative clustering proceeds by meeting these criterion at each join to construct a hierarchy of clusters. This hierarchy can then be "cut" at any height (where height is defined via the distance metric and linkage criterion) to identify a potential grouping scheme. Deciding on the number of groups is difficult, and is often left up to the investigator's aims and knowledge of the data being examined. For some metrics (i.e. Bray-Curtis) a meaningful height

can be identified which relates to the compositional similarities of the clusters (i.e. identify clusters that have communities whose compositional similarity is greater than 50%), and a cut can be made at this height. However, in other cases the height doesn't relate to a meaningful measure that is easily interpreted.

In this section hierarchical clustering, based on Euclidean distance measures and Ward's linkage criterion, is applied to the dataset of R12/TEV, STP/TEV, SCI and MCI for the 443 NZ coastal hydrosystems and the resulting grouping schemes are discussed and analysed.

2.2.1 – Initial treatment of data

The variables to be examined have markedly different ranges, spread and skew (Table 1 and Figure 1). In particular STP/TEV and R12/TEV have a much larger range of values than SCI and MCI and display a great deal of skew with the majority of values at the lower end of the range, and a few values that are significantly higher (Figure 1 and Table 1). Calculation of Euclidean distances among observations are therefore likely to be heavily biased towards the distances among STP/TEV and R12/TEV, with very little contribution from SCI and MCI. Furthermore the resultant clustering schemes are likely to isolate each extreme value in a group of its own, with the bulk of the data in a single group.

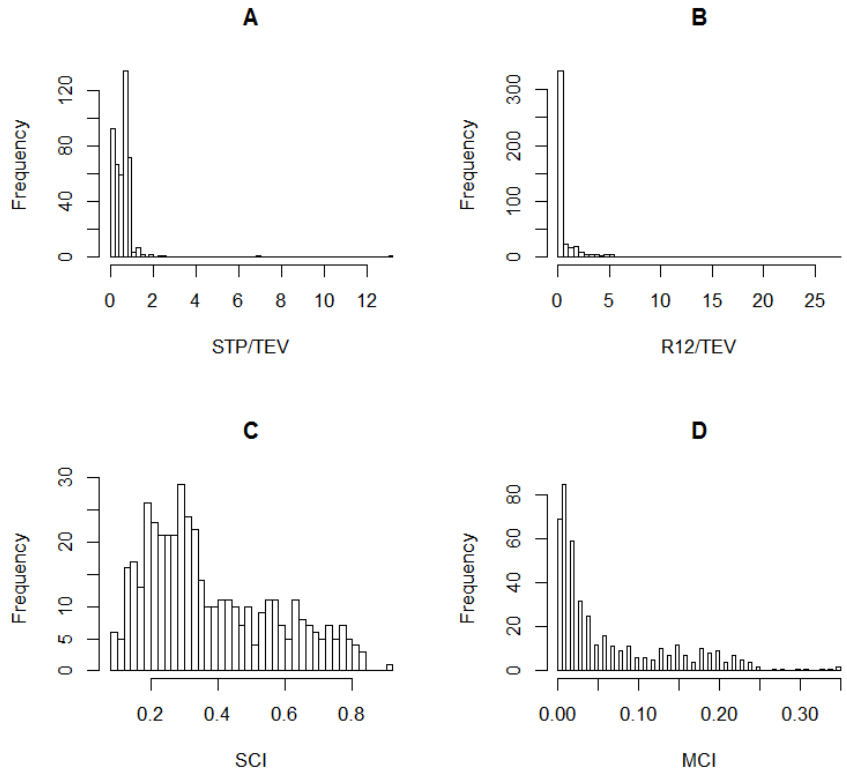

Figure 1. Histograms of (A) STP/TEV, (B) R12/TEV, (C) SCI and (D) MCI for the coastal hydrosystems dataset.

To address this the data can be transformed to attempt to reduce the influence of extreme values and reduce the influence of STP/TEV and R12/TEV on the calculation of distances among observations. For each variable a total of three transformations are trialled, namely square-root, fourth-root and log, or log (X+1) when data contains zero's, and the resultant data distributions are examined. A suitable transformation should ideally reduce the influence of extreme values, and approximately centre the distribution of values (i.e. reduce skew). Distribution plots of these transformations and the original data are contained within Appendix A. Final datasets were based on data transformed according to the fourth-root transformation for STP/TEV and R12/TEV, whilst SCI and MCI were untransformed. As a caveat, alternative transformations are also appropriate for these data, and are likely to influence group structure (see Appendix B).

These transformations have approximately centred the data, but the range of values for STP/TEV and R12/TEV still exceeds that of MCI and SCI, and will therefore carry more weight in any cluster analysis based on Euclidean distance. To address this each of the variables are scaled such that the standard deviations are all equal (SD=1) and with mean values of zero (Table 1).

Table 1. Summary statistics of untransformed, transformed and scaled physical variables. N/A indicates these variables weren't transformed.

Variable	Untransformed		Transformed		Transformed and Scaled	
	Mean (SD)	Range	Mean (SD)	Range	Mean (SD)	Range
STP/TEV	0.59 (0.76)	0 – 13.14	0.78 (0.28)	0 – 1.90	0 (1)	-2.74 – 3.96
R12/TEV	0.83 (2.57)	0 – 27.13	0.57 (0.44)	0 – 2.28	0 (1)	-1.31 – 3.92
SCI	0.38 (0.19)	0.08 – 0.92	N/A	N/A	0 (1)	-1.57 – 2.81
MCI	0.06 (0.08)	0 – 0.35	N/A	N/A	0 (1)	-0.84 – 3.76

Cluster analyses based on transformed and transformed and scaled datasets were carried out to examine the influence of scaling on the resultant group structures. Each of these datasets was then used to calculate a distance matrix comprised of the Euclidean distances between each system and every other system. These distance matrices were subsequently used to construct a hierarchical cluster scheme (using the **hclust** function in R), based on Ward's linkage criterion, with an example cluster dendrogram shown in Figure 2. All analyses were carried out in R version 3.0.0 (R Core Team 2013).

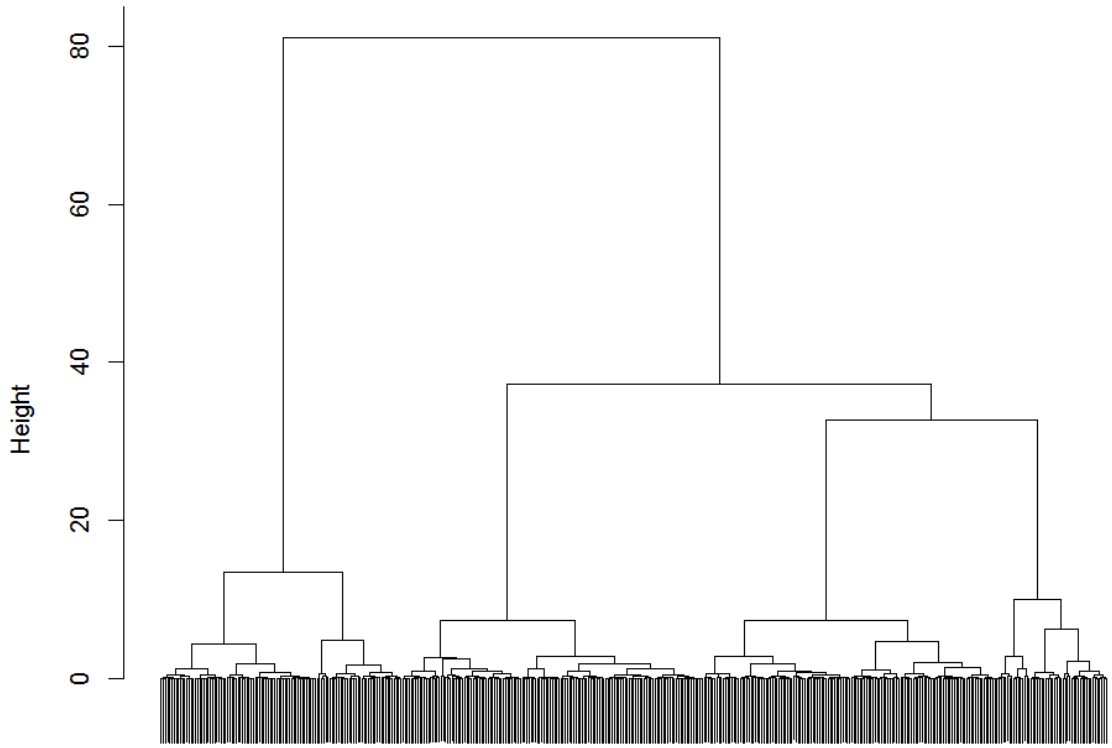

Figure 2. Dendrogram illustrating the resulting structure of a hierarchical cluster analysis based on Euclidean distances (transformed, but not scaled data) and Ward's linkage criterion.

2.2.2 – Deciding on the "best" number of groups

From the resulting cluster analysis it is not immediately obvious what an appropriate number of clusters would be. There is a large difference in height between four groups (height of ~ 32, Figure 2) and five groups (height ~ 16, Figure 2), which may suggest something about the grouping structure, but this is not immediately obvious. An alternative way to determine the number of groups would be to examine the predictive accuracy of the grouping structure through cross-validation. A routine for cross-validation was constructed in R that performs the following steps:

1) Using the full dataset group labels are identified through hierarchical cluster analysis, identifying *k* separate groups, whose group labels are stored
2) The dataset is split into a training and test set at random, with test sets containing approximately a tenth of the entire dataset (40 systems were chosen at random for the test set)
3) A hierarchical cluster analysis is applied to the training dataset and is similarly cut to form *k* groups

4) The group labels for the training dataset are matched to those labels identified from the full dataset analysis. The percent difference in tree structure (caused by leaving out certain datapoints that may be important in creating certain groupings) is calculated as the proportion of non-matching entities between the training and full group labels
5) For each of the training groups, the group centroid is calculated as the multivariate mean of the system variables assigned to each group
6) Each system in the test dataset is then assigned a group label for the group to which the Euclidean distance between the test datapoint and the training group centroid is minimised
7) The labels for the test dataset are then compared to the labels given to those datapoints in the full analysis, and the percent classification accuracy is the proportion of systems assigned to the correct (i.e. full classification) group
8) Steps 2-7 are repeated 500 times to build up a distribution of classification accuracy and the proportional difference in tree structure
9) The mean classification accuracy and proportional difference in tree structure is then plotted against a range of values for *k* (number of groups)
10) To account for correct classifications that could have occurred by chance, Cohens-κ (Cohen 1960) is also calculated;

$$\kappa = \frac{\Pr(Obs) - \Pr(Chance)}{1 - \Pr(Chance)}$$

(eqn. 2)

Where *Pr(Obs)* is the observed proportion of correct classifications and *Pr(Chance)* is the proportion that would be expected by chance, which is taken to be the reciprocal of the number of groups. General rules for Cohens-κ are that 0-0.2 indicates slight agreement, 0.2-0.4 indicates fair agreement, 0.4-0.6 indicates moderate agreement, 0.6-0.8 indicates substantial agreement and > 0.8 is very strong agreement (Landis & Koch 1977). Values for Cohens-κ are also plotted against the number of groups.

A good group structure should have high classification accuracy, indicating that group membership is strong, and low sensitivity to data being left out (i.e. the tree structure doesn't

vary much when a small proportion of the data is omitted). This routine was carried out on transformed (Figure 3) and transformed and scaled datasets (Figure 4).

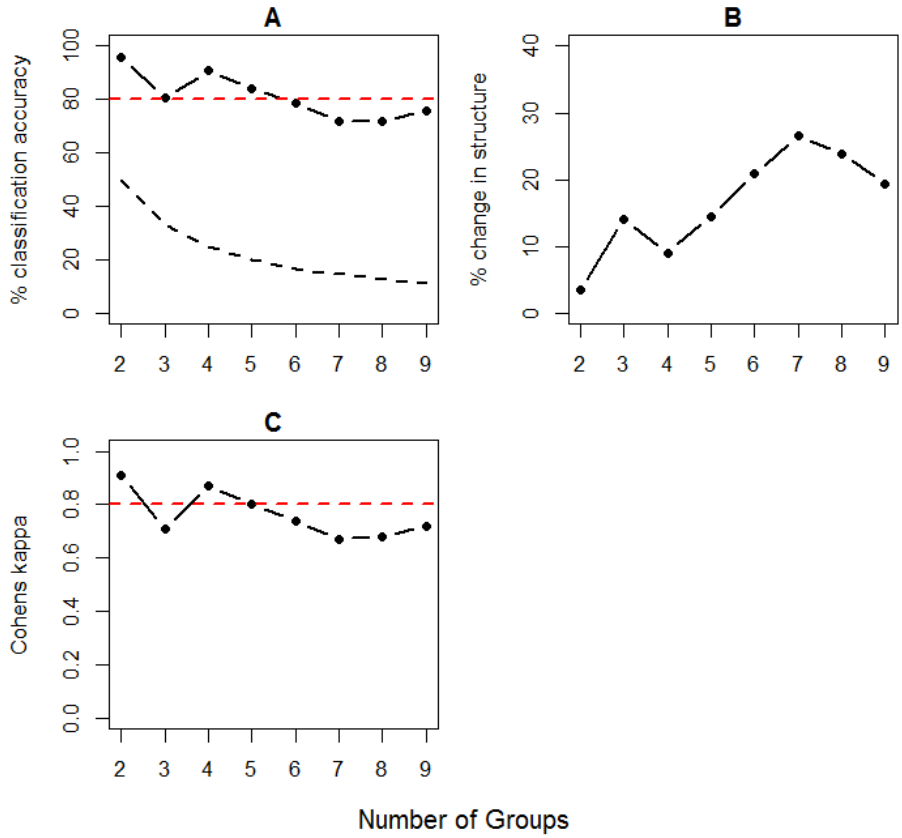

Figure 3. Cross validation metrics of (A) % classification accuracy, (B) % change in group structure and (C) Cohens-κ plotted against number of groups. Based on Euclidean hierarchical clustering of transformed, but not scaled data.

Examining the cross-validation metrics for the unscaled data, the classification accuracy is higher than 80% for five or fewer groups, and is marginally below 80% for six groups. In addition Cohens-κ is above 0.8 for two and four groups, and is equal to 0.8 for five groups (Figure 3). Based on these metrics four to six groups are reasonably well supported and group structures for five and six groups will be investigated. For the scaled data the classification accuracy is higher than 80% for four or fewer groups, with five groups having a classification accuracy marginally lower than 80% (79.7%) (Figure 4). Cohens-κ is also above 0.8 for four or fewer groups, but measured 0.74 for five groups (Figure 4). Therefore only four groups are supported by cross validation, but five groups will also be investigated to provide a direct comparison between scaled and unscaled classifications.

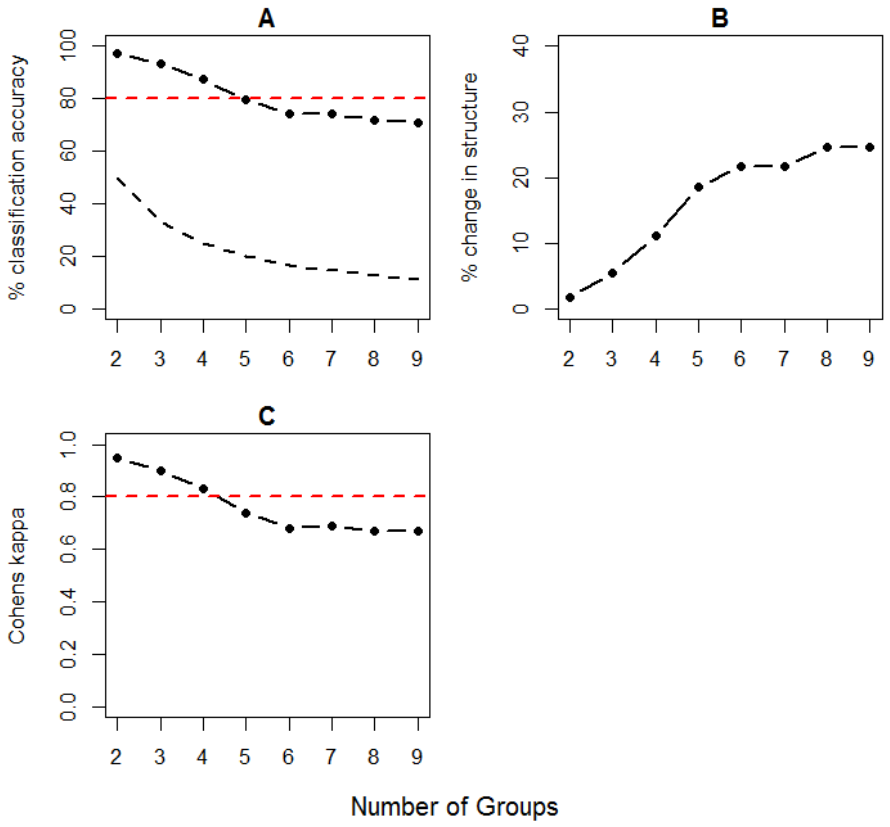

Figure 4. Cross validation metrics of (A) % classification accuracy, (B) % change in group structure and (C) Cohens-κ plotted against number of groups. Based on hierarchical clustering of transformed and scaled data.

2.2.3 – Grouping scheme based on unscaled data for five groups

The five group analysis of the unscaled data is illustrated in Figures 5 and 6. The groupings are best illustrated along the STP/TEV and R12/TEV axes, as the majority of the splits were along these axes. Groups 1, 4 and 5 are separated from the other groups primarily along the R12/TEV axis, whilst group 2 is primarily separated along the STP/TEV axis. The exception is group 3, which is primarily split along the SCI and MCI axes (Figure 5). The greater prominence of splits along the R12/TEV axis is likely associated with the greater variability (the standard deviation of the transformed R12/TEV is 1.5 to 5.5 times higher than the other variables) along this axis (Table 1).

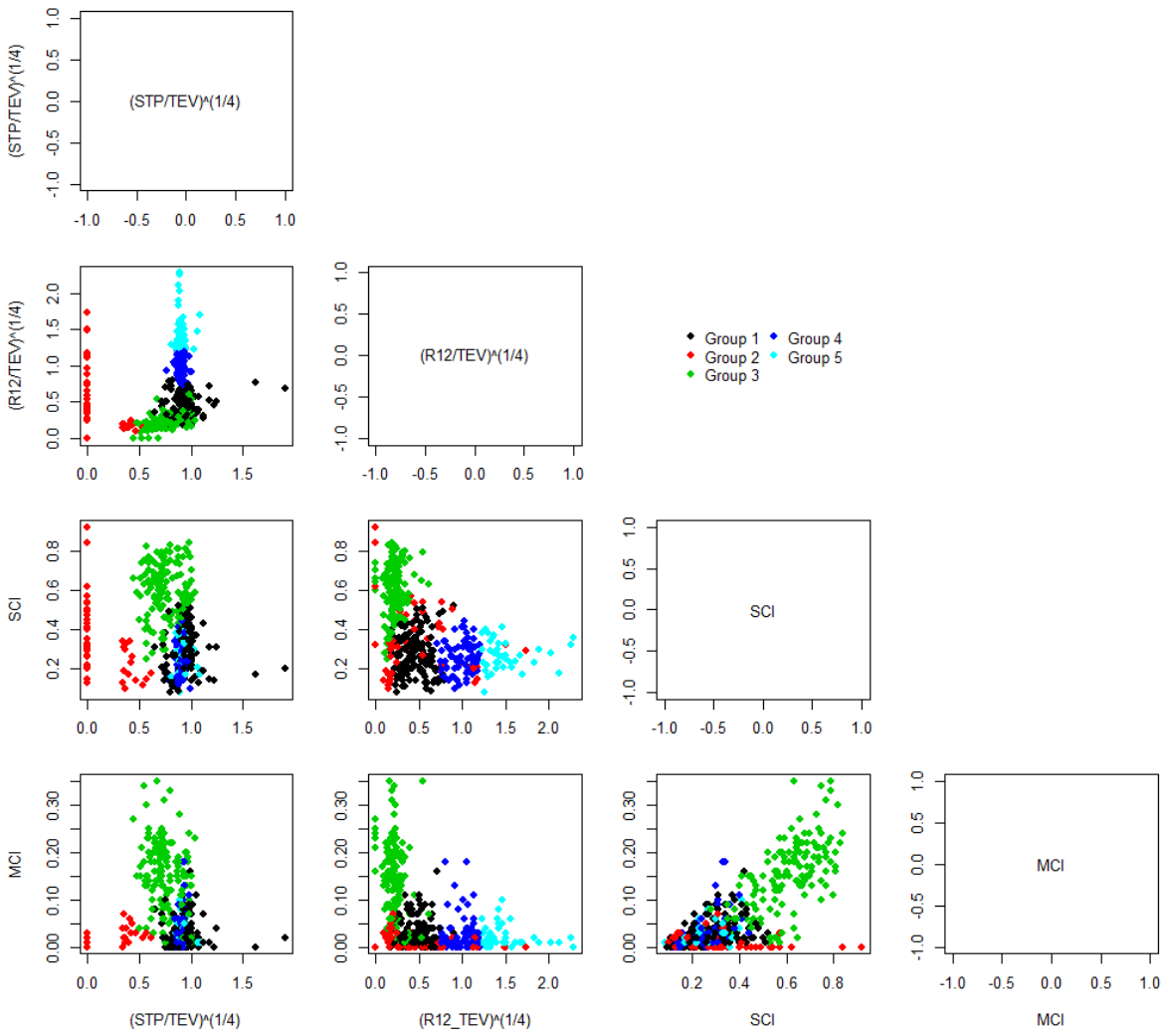

Figure 5. Biplots of the four physical variables colour-coded by group label as identified by the hierarchical clustering of the unscaled data, with $k = 5$. STP/TEV and R12/TEV are fourth-root transformed, whilst SCI and MCI are untransformed.

Examining the groups according to their physical characteristics (Figure 6) and inspecting the names of systems within each group (see separate Classifications Appendix for full group labelling's), the groups can be characterised as:

- 1 – physical factors are not particularly separate from any other group, mix of harbours/harbour systems, inlets and river mouths
- 2 – low STP/TEV, mix of lakes, lagoons, fiords and sounds
- 3 – high SCI and MCI, primarily bays and harbours
- 4 – moderate R12/TEV, primarily river mouths
- 5 – high R12/TEV, primarily river mouths

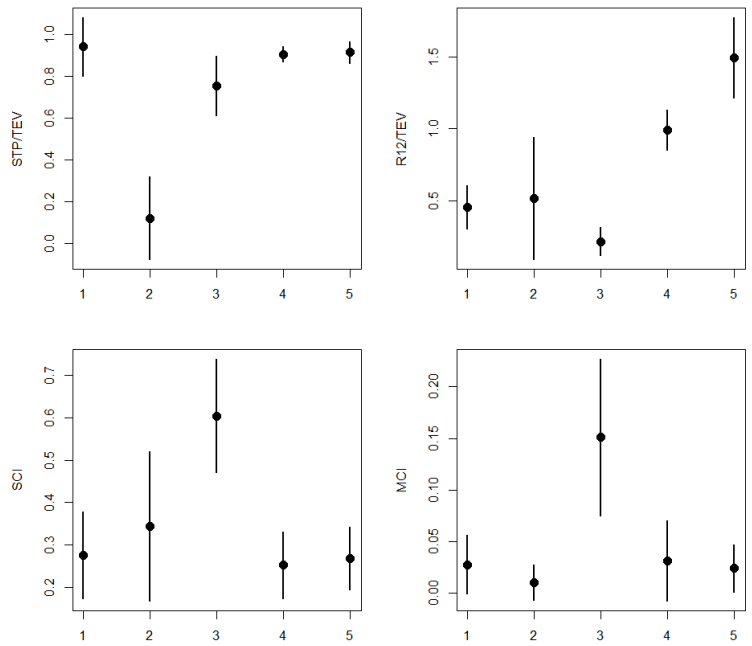

Figure 6. Means (±1 SD) of the four physical variables by group based on group labels identified by Euclidean hierarchical clustering on unscaled variables with $k = 5$. STP/TEV and R12/TEV are fourth-root transformed, whilst SCI and MCI are untransformed.

Comparing these labels to the Hume classification (Table 2):

- 1 – corresponds to a mix of the majority of F classed systems and two thirds of class E systems
- 2 – corresponds to all of class A and the majority of class G systems
- 3 – corresponds closely to class D, with some E class systems
- 4 – corresponds to class B systems with low-moderate river input
- 5 – corresponds to class B systems with highest river input

There is fair agreement between the Hume classification and this grouping scheme as the majority of hydro classes belong to one or two separate groups. The major differences are that classes D, E and F are confined to two groups (groups 1 and 3), rather than three, with class D and F belonging almost entirely to groups 1 and 3, respectively, and class E overlapping with both. Class A has been classified in its own group, along with some of class G and H, with which it shares some characteristics (low STP/TEV). Elsewhere class B is split into low (group 4) and high (group 5) R12/TEV representations, most likely due to the over importance of R12/TEV in the unscaled dataset (Table 2).

Table 2. Table illustrating the number of systems in each group that were of the different hydro classes outlined in the Hume et al. (2007) classification. This is based on Euclidean hierarchical clustering of unscaled data, with $k = 5$.

Hume Hydro Class	Hierarchical Group				
	1	2	3	4	5
A	0	37	0	0	0
B	8	0	0	73	40
D	9	0	104	0	0
E	44	0	24	0	0
F	80	2	2	0	0
G	0	9	2	0	0
H	0	4	5	0	0

2.2.4 – Grouping scheme based on unscaled data for six groups

The six group analysis of the same data is similar to the five group analysis, but with an additional split along the R12/TEV axis at zero STP/TEV (Figure 7). The groups can be characterised as (Figure 8):

- 1 – physical factors are not particularly separate from any other group, mix of harbours/harbour systems, inlets and river mouths
- 2 – low STP/TEV, mix of lakes, lagoons, fiords and sounds
- 3 – high SCI and MCI, primarily bays and harbours
- 4 – moderate R12/TEV, primarily river mouths
- 5 – high R12/TEV, primarily river mouths
- 6 – zero STP/TEV, primarily lakes and lagoons

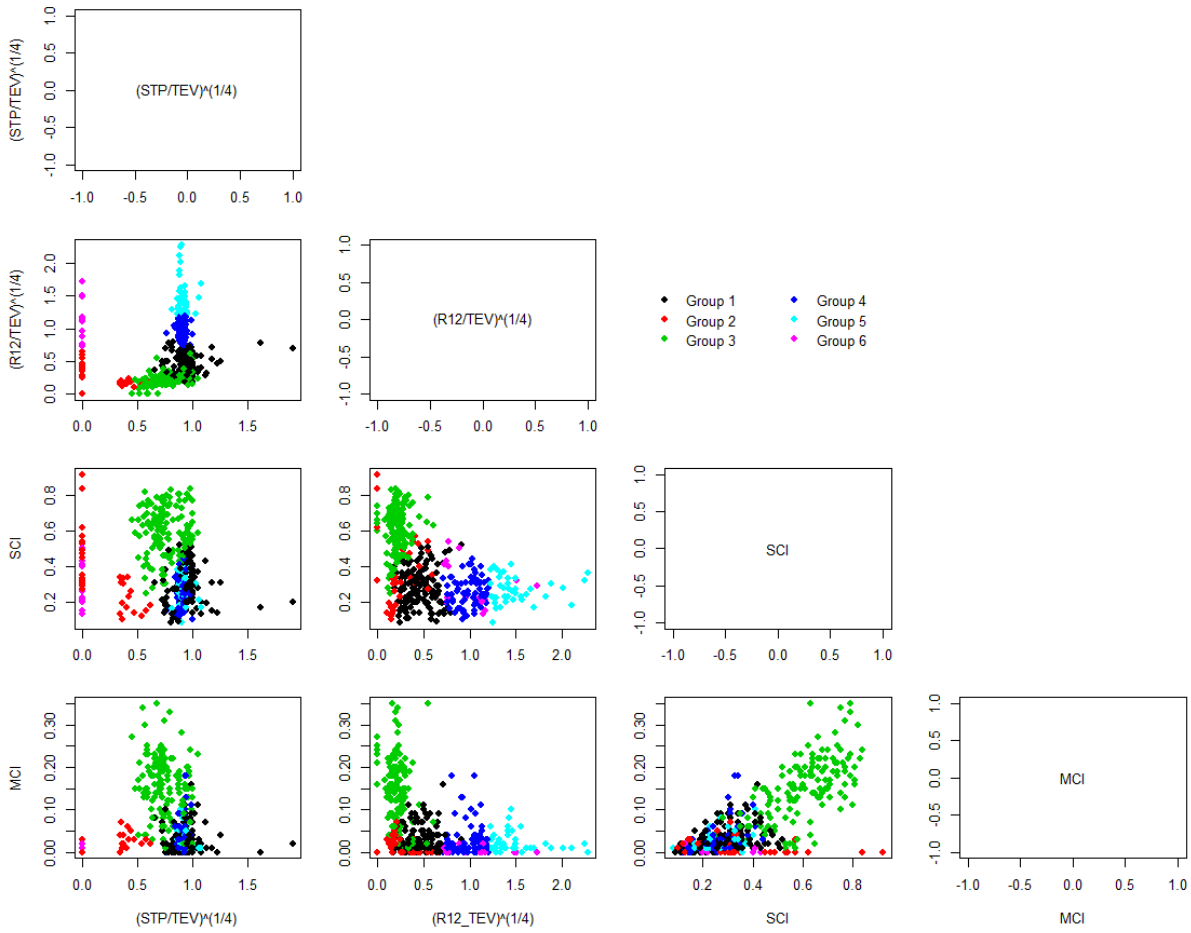

Figure 7. Biplots of the four physical variables colour-coded by group label as identified by the Euclidean hierarchical clustering of the unscaled data, with $k = 6$. STP/TEV and R12/TEV are fourth-root transformed, whilst SCI and MCI are untransformed.

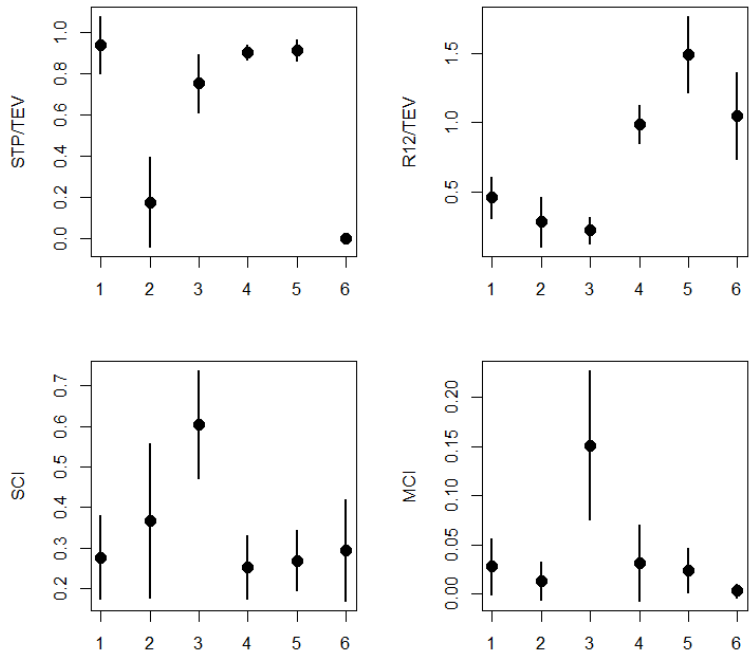

Figure 8. Means (±1 SD) of the four physical variables by group as identified by Euclidean hierarchical clustering of unscaled data with k = 6. STP/TEV and R12/TEV are fourth-root transformed, whilst SCI and MCI are untransformed.

Comparing these labels to the Hume classification (Table 3):

- 1 – corresponds to a mix of the majority of F and two thirds of the class E systems
- 2 – corresponds to class A systems with low river input and the majority of class G systems
- 3 – corresponds closely to class D, with some E class systems
- 4 – corresponds to class B systems with low-moderate river input
- 5 – corresponds to class B systems with the highest river input
- 6 – corresponds to class A systems with high river input

This is similar to the five group scheme, only class A has been split into high (group 6) and low river input (group 2) variants, with the majority of class G remaining in group 2 mixed with the low river input systems in class A (Table 3).

Table 3. Table illustrating the number of systems in each group that were of the different hydro classes outlined in the Hume et al. (2007) classification. This is based on hierarchical clustering of the unscaled data, with k=6.

Hume Hydro Class	Hierarchical Group					
	1	2	3	4	5	6
A	0	21	0	0	0	16
B	8	0	0	73	40	0
D	9	0	104	0	0	0
E	44	0	24	0	0	0
F	80	2	2	0	0	0
G	0	9	2	0	0	0
H	0	4	5	0	0	0

2.2.5 – Grouping scheme based on scaled data for four groups

Scaling the data and running the hierarchical analysis resulted in very different classifications (Figures 9 and 10). Group 2, 3 and 4 are prominently defined along STP/TEV, MCI/SCI and R12/TEV axes, respectively, whilst group 1 overlaps with the other groups along multiple axes, but primarily seems to occupy the intermediate space between group 3 and group 4 (Figure 9). Groups can be characterised as:

- 1 – high STP/TEV but prominent overlap with other groups, mix of harbours/harbour systems, inlets and river mouths
- 2 – zero STP/TEV and MCI, primarily lakes and lagoons
- 3 – high SCI and MCI and low R12/TEV, primarily bays
- 4 – high R12, primarily river mouths

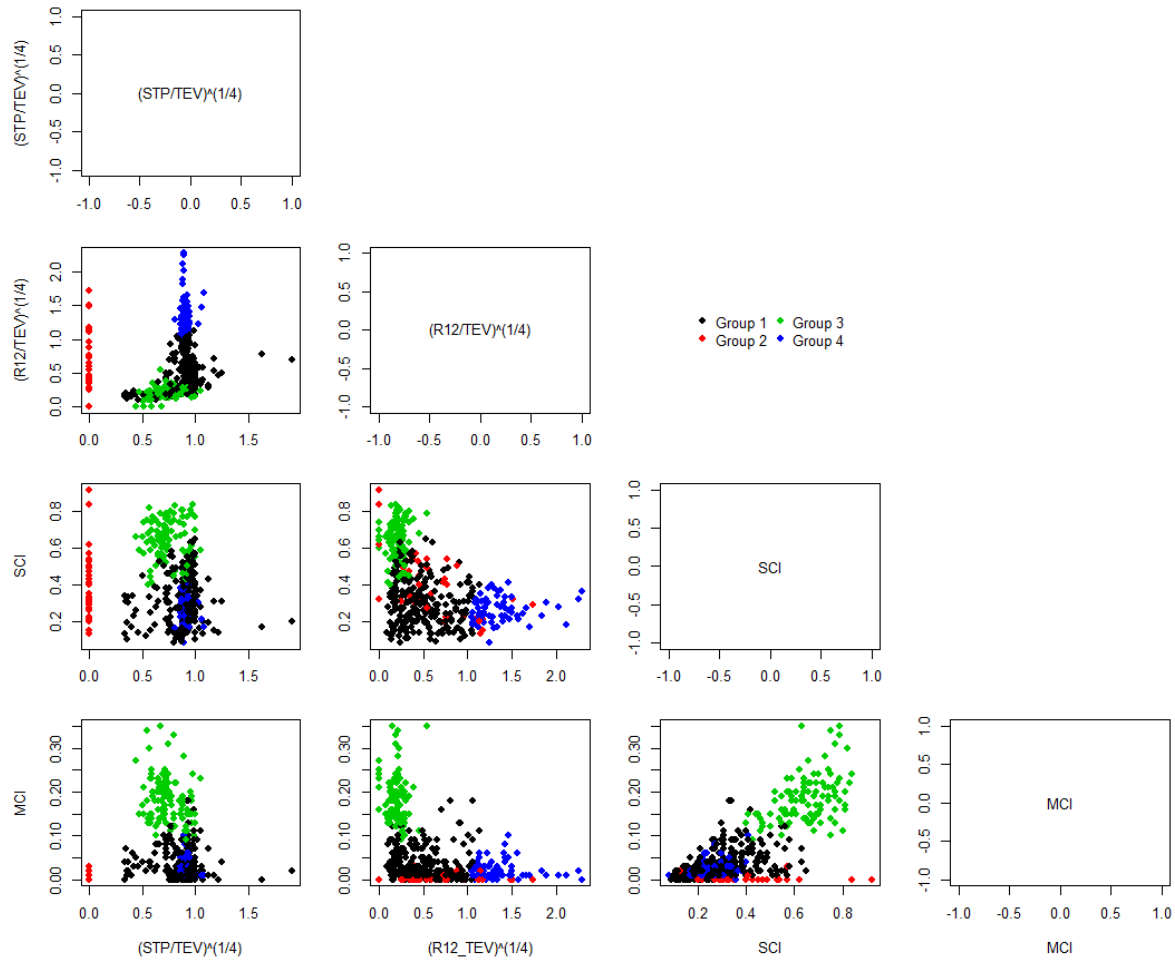

Figure 9. Biplots of the four physical variables colour-coded by group label as identified by the Euclidean hierarchical clustering of the scaled dataset with $k = 4$. STP/TEV and R12/TEV are fourth-root transformed, whilst SCI and MCI are untransformed.

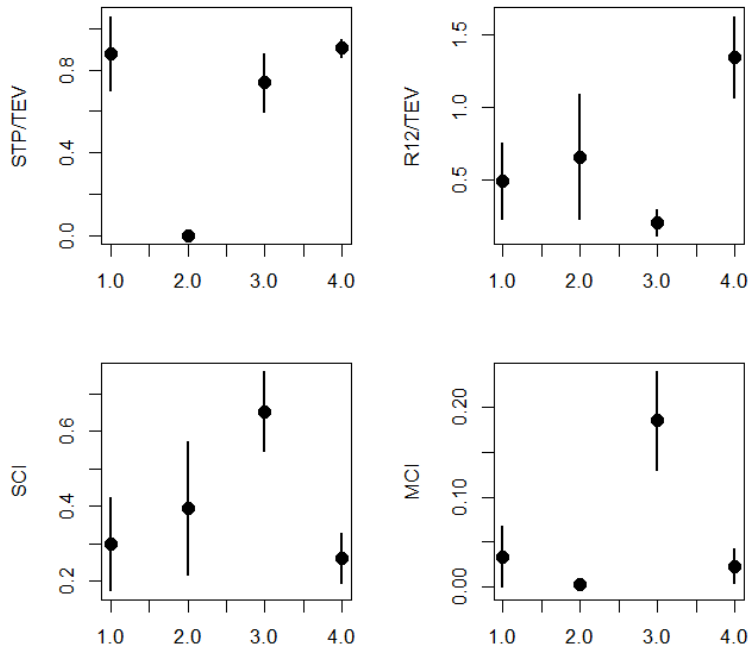

Figure 10. Means (±1 SD) of the four physical variables by group with group labels identified by Euclidean hierarchical clustering of the scaled dataset, with *k* = 4. STP/TEV and R12/TEV are fourth-root transformed, whilst SCI and MCI are untransformed.

Comparing these labels to the Hume classification (Table 4):

- 1 – corresponds to a mix of classes, containing almost half of B and nearly all of E-G systems
- 2 – corresponds exactly to class A
- 3 – corresponds closely to a selection of class D systems
- 4 – corresponds to a selection of class B systems

Table 4. Table illustrating the number of systems in each group that were of the different hydro classes outlined in the Hume et al. (2007) classification. This is based on Euclidean hierarchical clustering of the scaled dataset, with *k* = 4.

Hume Hydro Class	Hierarchical Group			
	1	2	3	4
A	0	37	0	0
B	54	0	0	67
D	18	0	95	0
E	68	0	0	0
F	84	0	0	0
G	10	0	1	0
H	4	0	5	0

There is some agreement between the Hume classification and this grouping scheme for some groups, but the vast majority are classified into a single group. Class A is likely separated out as group 2 due to those data-points having STP/TEV equal to zero. Class D is split between group 3, which is characterised by high SCI and MCI, and group 1, which has lower values for these parameters. Therefore group 3 is likely coastal embayments that are structurally simple (high SCI), and are almost entirely open (high MCI), whilst the remaining bays are partially closed (i.e. narrower opening, or partially recessed) and are grouped with harbours and drowned valleys in group 1. Class B is split among groups 1 and 4, with group 4 representing those systems with the highest river input relative to their volume, and the remaining class B systems with lower R12/TEV combined with those other systems in group 1. Group 1 therefore most likely represents systems that are intermediate along all axes between open river mouths, bays and completely closed lakes and lagoons. It therefore represents systems that are partially closed.

2.2.6 – Grouping scheme based on scaled data for five groups

The five group analysis is similar to the four group analysis, with an additional split primarily along the STP/TEV axis, splitting the previous group 1 into two groups (Figure 11).

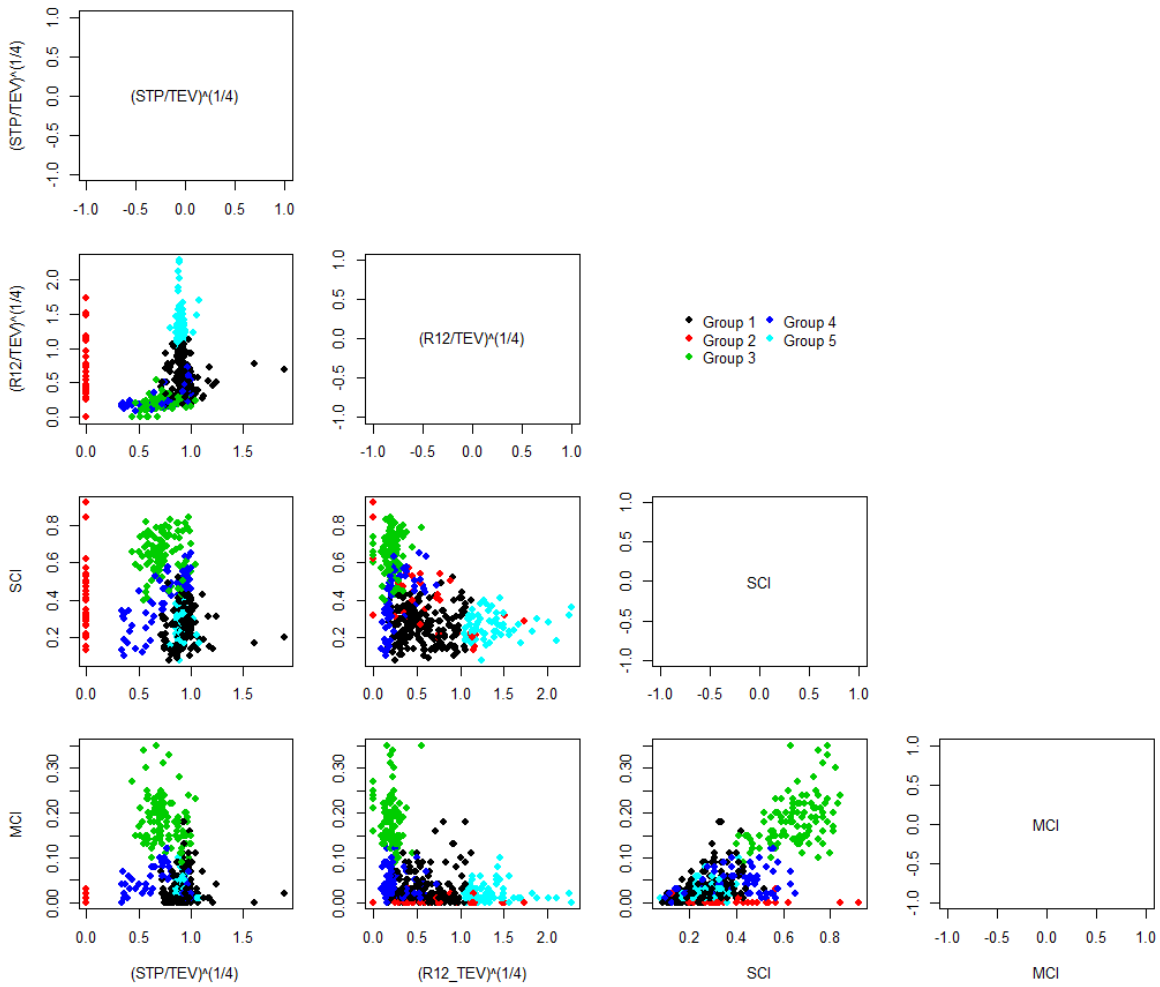

Figure 11. Biplots of the four physical variables colour-coded by group label as identified by the Euclidean hierarchical clustering of the scaled dataset, with k = 5. STP/TEV and R12/TEV are fourth-root transformed, whilst SCI and MCI are untransformed.

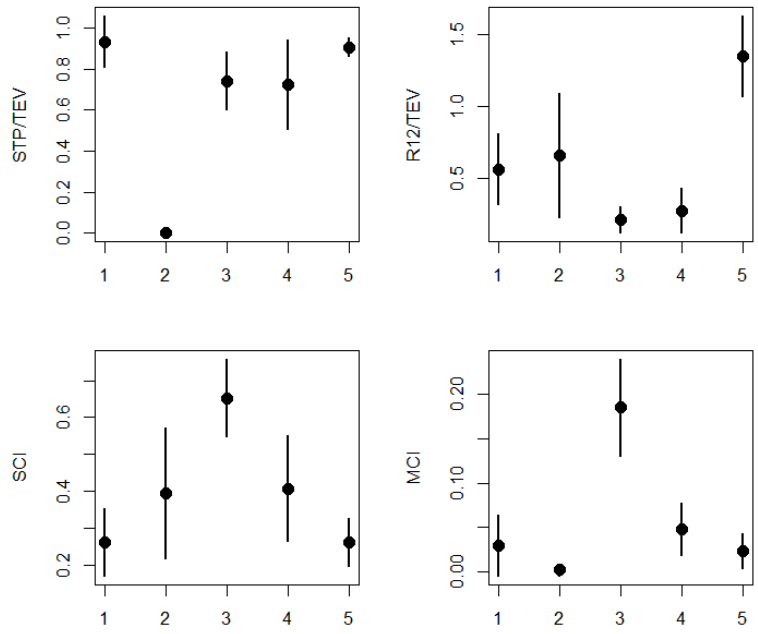

Figure 12. Means (±1 SD) of the four physical variables by group with group labels identified by Euclidean hierarchical clustering of the scaled dataset, with *k* = 5. STP/TEV and R12/TEV are fourth-root transformed, whilst SCI and MCI are untransformed.

Groups can be characterised as:

- 1 – high STP/TEV but overlapping with other groups, mix of estuaries, inlets/harbours and river mouths
- 2 – zero STP/TEV and MCI, primarily lakes and lagoons
- 3 – high SCI and MCI and low R12/TEV, primarily bays
- 4 – is not distinguished from the other groups along any particular single axis, consists of harbours, inlets, fiords and sounds
- 5 – high R12/TEV, primarily river mouths

Comparing these labels to the Hume classification (Table 5):

- 1 – mix of B, E and the majority of class F systems
- 2 – corresponds exactly to class A
- 3 – corresponds strongly to class D
- 4 – mix of D-H class systems, but primarily D, E and G
- 5 – corresponds exactly to a selection of class B systems

Table 5. Table illustrating the number of systems in each group that were of the different hydro classes outlined in the Hume et al. (2007) classification. This is based on Euclidean hierarchical clustering of the scaled dataset, with $k = 5$.

Hume Hydro Class	Hierarchical Group				
	1	2	3	4	5
A	0	37	0	0	0
B	53	0	0	1	67
D	7	0	95	11	0
E	39	0	0	29	0
F	80	0	0	4	0
G	0	0	1	10	0
H	0	0	5	4	0

There is some agreement between this grouping scheme and the Hume classification and is particularly strong for classes A and D. There is considerable overlap in classes between groups 1 and 4, although 1 contains mostly B, E and F systems, whilst 4 contains mostly D, E and G systems. The primary differences between groups 1 and 4 are R12/TEV which is higher for group 1 than 4, and SCI, which is higher for group 4 than 1. Therefore group 1 is more closely aligned with structurally complex harbours, perhaps containing multiple arms, and river inputs, whilst group 4 contains less structurally complex systems with lower riverine inputs.

Comparing the scaled and unscaled five group schemes, the unscaled analysis had greater success in defining class B from the remaining data, which it placed into two separate groups (high and low R12/TEV), but also grouped lakes and lagoons, with fiords and sounds, as both have low R12/TEV. The scaled analysis had a greater ability to define class A from the remaining data, but split class B into one mixed group (group 1) and one pure group (group 5). Both performed similarly regarding classes E-F, with both identifying two groups for these classes separated into low and high structural complexity cases.

2.2.7 – Rule-based summary of grouping schemes

At present the information presented provides some idea of how groups are separated along each of the physical variable axes, but this information is particularly vague for some of the groups (i.e. group 1 in the four group scaled analysis overlaps with at least on other group along each axis). To provide greater interpretability to these groupings, an approximate rule-based classification, similar to that used in Hume et al. (2007), is developed for each grouping scheme.

For each group the range of each of the four variables (not transformed or scaled) is firstly examined to identify any groups that are distinct from the other groups (i.e. with no overlap along a particular axis). For groups which are completely distinct (based on one or more physical variable) a binary rule (i.e. Value > 5: Y – group 1, N – next rule) is created, which separates this groups from the remaining groups. Where no distinct groups are present, or all distinct groups are accounted for, the remaining data is analysed using a classification tree using the **rpart** function in R (Therneau et al. 2014). Classification, or decision trees are a method in the machine learning literature for supervised classification of data (i.e. where the group labels are available), and work by partitioning data at certain points in multivariate space, such that the data either side of the partition belong to fewer classes than prior to the partition. Each partition is then subsequently partitioned until each partition contains only data of a single class (De'ath & Fabricius 2000). Classification trees are therefore a direct equivalent of a rule based scheme, but based on statistical analysis to identify the appropriate locations to partition the data. However, because Classification trees can only partition data along, or perpendicular to multivariate axes, absolute classification accuracy is not assured, particularly when classes are partitioned along a line which measures 45° to any two axes. In addition, classification trees can also contain more partitions and terminal nodes (i.e. end points) than there are groups in order to maximise classification accuracy. In this case this is not desirable, as multiple rule sets for each group would make interpretation more difficult. Therefore each classification tree is restricted such that the simplest tree containing a terminal node for each group is obtained. This ensures that in most cases there is a single rule set for each group. Along with the rules identified by the classification tree, the overall classification accuracy of the tree (relative to the group scheme being examined) is recorded as a measure of how well the tree reproduces the group scheme.

Five groups – unscaled data

None of the groups are fully distinct from the remaining groups, with overlap between groups along one or more axes (Table 6). Group 2 is perhaps the most distinct along the STP/TEV axis as its upper value of 0.144 overlaps only with group 3, which has a minimum value of 0.039 (Table 6).

Table 6. Physical variable ranges for each group based on the five group scheme developed from the Euclidean hierarchical clustering of the unscaled dataset.

Group	R12/TEV	STP/TEV	SCI	MCI
1	0.001, 0.675	0.180, 13.135	0.08, 0.52	0, 0.16
2	0, 8.968	0, 0.144	0.10, 0.92	0, 0.07
3	0, 0.136	0.039, 1.182	0.25, 0.84	0.01, 0.35
4	0.256, 2.116	0.341, 1.000	0.10, 0.44	0, 0.18
5	2.258, 27.130	0.422, 1.357	0.08, 0.41	0, 0.10

Applying a classification tree analysis to the data resulted in a rule set with an overall classification accuracy of 86%, but was considerably lower for group 2 (Table 7). The rule set contains a single split along the SCI axis (defining group 3), two splits along the R12/TEV axis, and a single split along the STP/TEV axis (Figure 13). The rule set indicates that groups are characterised by:

(1) SCI < 0.445, R12/TEV < 0.2745, STP/TEV ≥ 0.3012

(2) SCI < 0.445, R12/TEV < 0.2745, STP/TEV < 0.3012

(3) SCI ≥ 0.445

(4) SCI < 0.445, R12/TEV ≥ 0.2745, R12/TEV < 2.187

(5) SCI < 0.445, R12/TEV ≥ 2.187

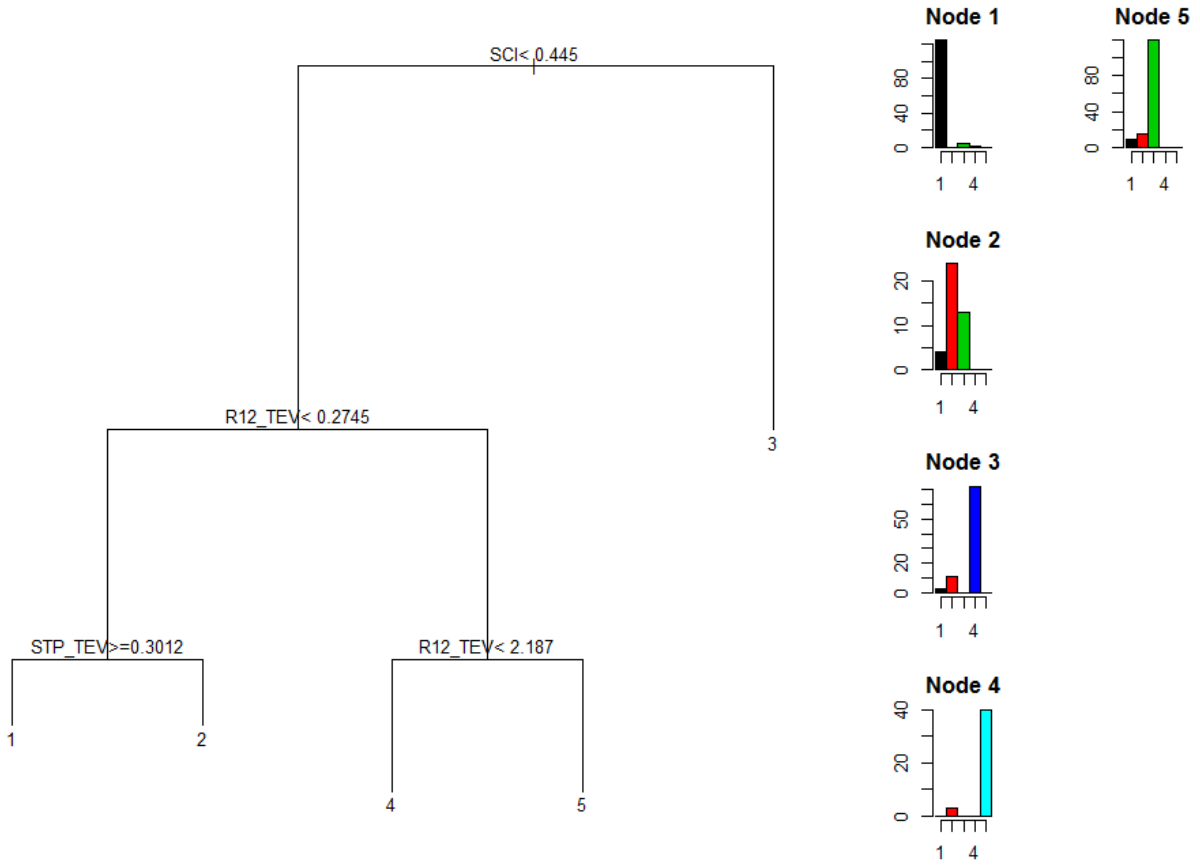

Figure 13. Classification tree for the five group scheme identified using Euclidean hierarchical clustering on the unscaled dataset. Histograms to the right of the classification tree indicate the class make-up of each terminal node, where node numbering proceeds from left to right along the base of the classification tree.

Table 7. Classification accuracy of the classification tree developed for the five group scheme identified using Euclidean hierarchical clustering of the unscaled dataset.

Group	Classification accuracy (%)	No. in Group
1	88.6	141
2	46.1	52
3	87.5	137
4	98.6	73
5	100	40
Total	86.0	443

Six groups – unscaled data

None of the groups are fully distinct from the remaining groups, with overlap between groups along one or more axes (Table 8). Group 6 is perhaps the most distinct along the STP/TEV axis as all members of group 6 have STP/TEV = 0, but so do some systems in group 2 (Table 8).

Table 8. Physical variable ranges for each group based on the six group scheme identified by Euclidean hierarchical clustering of the unscaled dataset.

Group	R12/TEV	STP/TEV	SCI	MCI
1	0.001, 0.675	0.180, 13.135	0.08, 0.52	0, 0.16
2	0.000, 0.179	0.000, 0.143	0.10, 0.92	0, 0.07
3	0.000, 0.136	0.0386, 1.182	0.25, 0.84	0.01, 0.35
4	0.256, 2.116	0.341, 1.000	0.10, 0.44	0, 0.18
5	2.258, 27.130	0.422, 1.357	0.08, 0.41	0, 0.10
6	0.281, 8.968	0.000, 0.000	0.13, 0.54	0, 0.02

The subsequent classification tree had an overall classification accuracy of 88%, but was considerably lower for group 2 and group 6 (Table 9). The rule set contains a single split along the SCI axis (defining group 3), two splits along the R12/TEV axis, and two splits along the STP/TEV axis (Figure 14). The rule set indicates that groups are characterised by

(1) SCI < 0.445, R12/TEV < 0.2745, STP/TEV ≥ 0.1722

(2) SCI < 0.445, R12/TEV < 0.2745, STP/TEV < 0.1722

(3) SCI ≥ 0.445

(4) SCI < 0.445, R12/TEV ≥ 0.2745, R12/TEV < 2.187, STP/TEV ≥ 0.5609

(5) SCI < 0.445, R12/TEV ≥ 2.187

(6) SCI < 0.445, R12/TEV ≥ 0.2745, R12/TEV < 2.187, STP/TEV < 0.5609

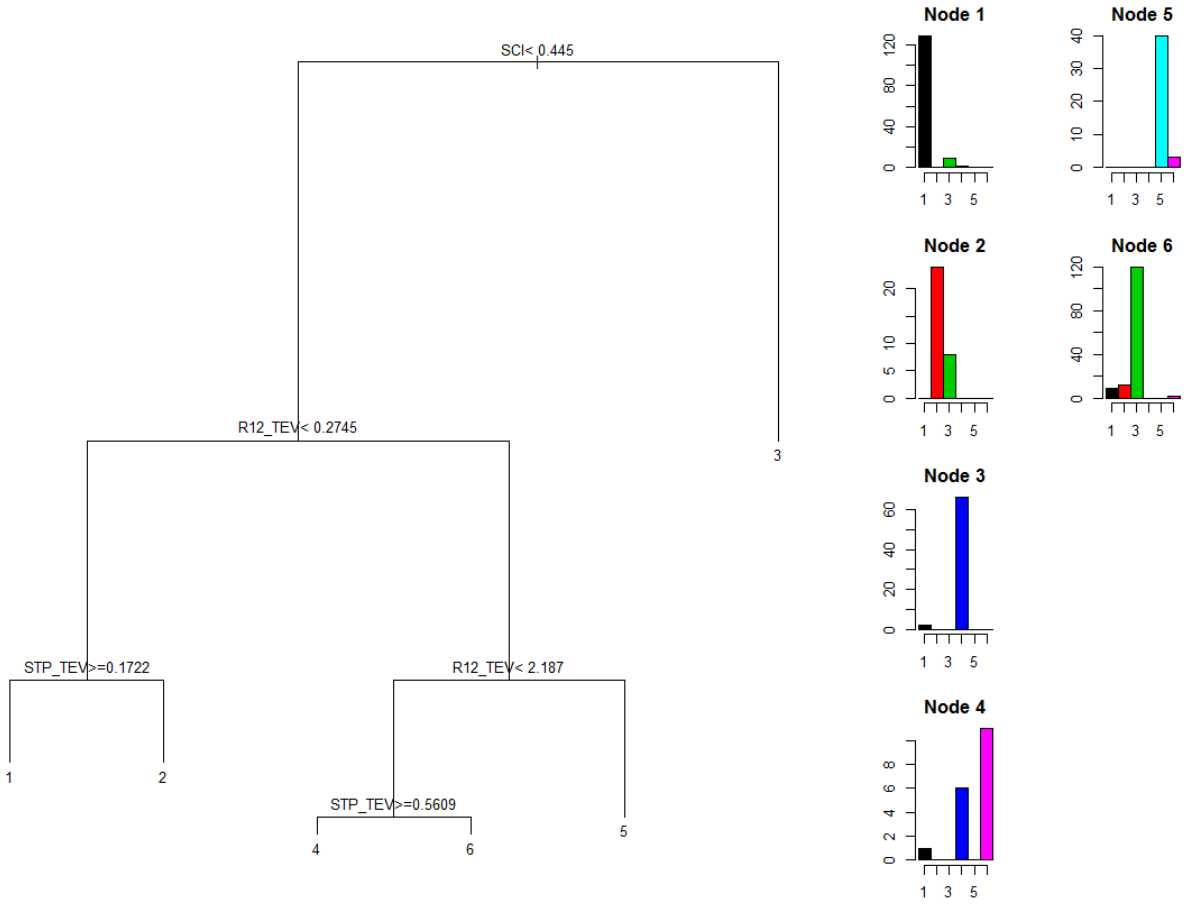

Figure 14. Classification tree for the six group scheme identified by Euclidean hierarchical clustering of the unscaled dataset. Histograms to the right of the classification tree indicate the class make-up of each terminal node, where node numbering proceeds from left to right along the base of the classification tree.

Table 9. Classification accuracy of the classification tree developed for the six group scheme identified by Euclidean hierarchical clustering of the unscaled dataset.

Group	Classification accuracy (%)	No. in Group
1	91.5	141
2	66.7	36
3	87.6	137
4	90.4	73
5	100	40
6	68.8	16
Total	88.0	443

Four groups – scaled data

Group 2 is completely distinct from the remaining groups along the STP/TEV axis as its range (0 – 0) does not overlap with any other group (Table 10). The remaining groups overlap along one or more axes. Therefore a suitable rule would be:

- STP/TEV < 0.0067
    - Y – Group 2
    - N – see decision tree (Figure 15)

Table 10. Physical variable ranges for each group based on the four group scheme identified by Euclidean hierarchical clustering of the scaled dataset.

Group	R12/TEV	STP/TEV	SCI	MCI
1	0.000, 1.613	0.013, 13.135	0.08, 0.65	0, 0.18
2	0.000, 8.968	0.000, 0.000	0.13, 0.92	0, 0.03
3	0.000, 0.090	0.0386, 1.182	0.4, 0.84	0.09, 0.35
4	1.184, 27.130	0.422, 1.357	0.08, 0.41	0, 0.10

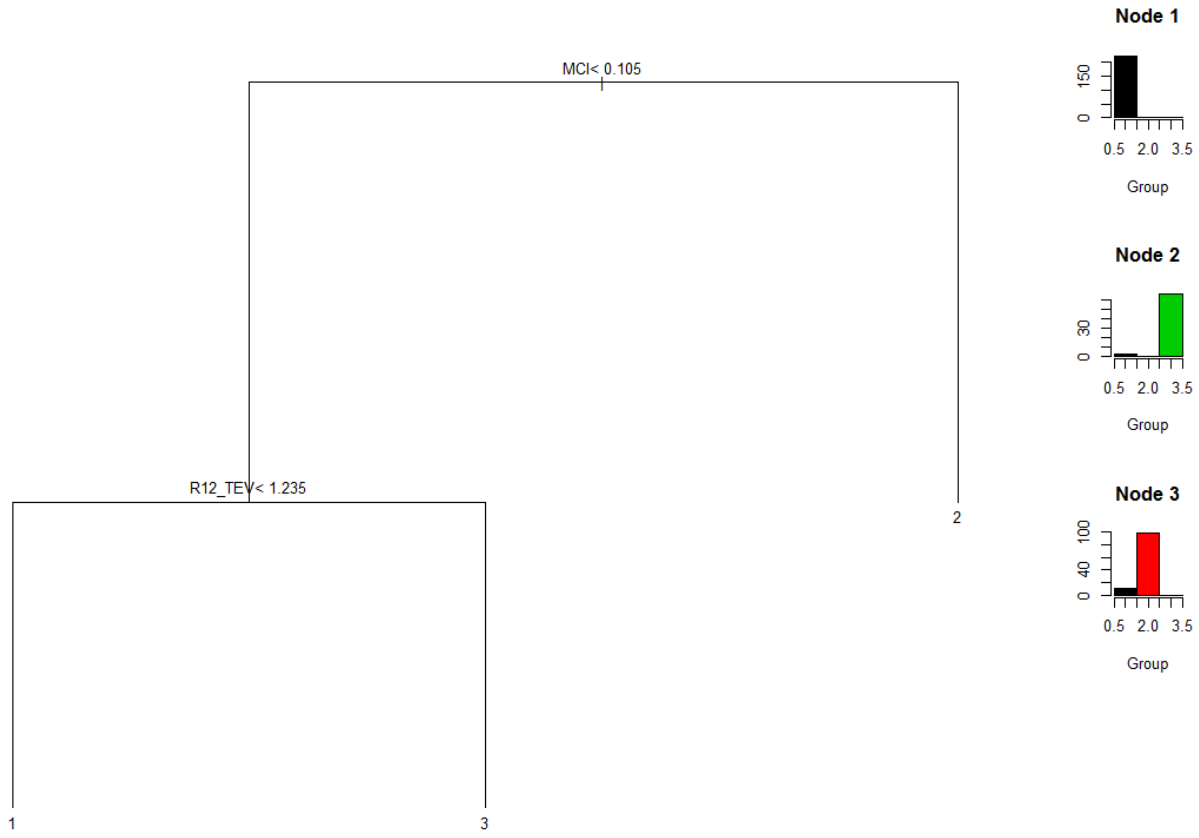

Figure 15. Classification tree for the four group scheme identified by Euclidean hierarchical clustering of the scaled dataset. Group 2 has been removed from the analysis and groups 3 and 4 are labelled as 2 and 3 in the classification tree panel, respectively. Histograms to the right of the classification tree indicate the class make-up of each terminal node, where node numbering proceeds from left to right along the base of the classification tree.

Table 11. Classification accuracy of the classification tree developed for the four group scheme identified by Euclidean hierarchical clustering of the scaled dataset.

Group	Classification accuracy (%)	No. in Group
1	94.5	238
2	100	37
3	97.0	101
4	98.5	67
Total	96.2	443

The resulting classification tree has a very high classification accuracy of 96.2%, and is above 90% for all four classes (Table 11). The entire tree (including the split for group 2) has splits along STP/TEV, MCI and R12/TEV axes (Figure 15). Groups are characterised by

(1) STP/TEV $\geq$ 0.0067, MCI < 0.105, R12/TEV < 1.235

(2) STP/TEV < 0.0067

(3) STP/TEV $\geq$ 0.0067, MCI $\geq$ 0.105

(4) STP/TEV $\geq$ 0.0067, MCI < 0.105, R12/TEV $\geq$ 1.235

Five groups – scaled data

Group 2 is completely distinct from the remaining groups along the STP/TEV axis as its range (0 – 0) does not overlap with any other group (Table 11). The remaining groups overlap along one or more axes. Therefore a suitable rule would be:

- STP/TEV < 0.0067
    - Y – Group 2
    - N – see decision tree (Figure 16)

Table 12. Physical variable ranges for each group based on the five group scheme identified by Euclidean hierarchical clustering of the scaled dataset.

Group	R12/TEV	STP/TEV	SCI	MCI
1	0.001, 1.613	0.270, 13.135	0.08, 0.52	0, 0.18
2	0.000, 8.968	0.000, 0.000	0.13, 0.92	0, 0.03
3	0.000, 0.090	0.0386, 1.182	0.40, 0.84	0.09, 0.35
4	0.000, 0.288	0.0134, 1.000	0.10, 0.65	0, 0.12
5	1.184, 27.130	0.422, 1.357	0.08, 0.41	0, 0.10

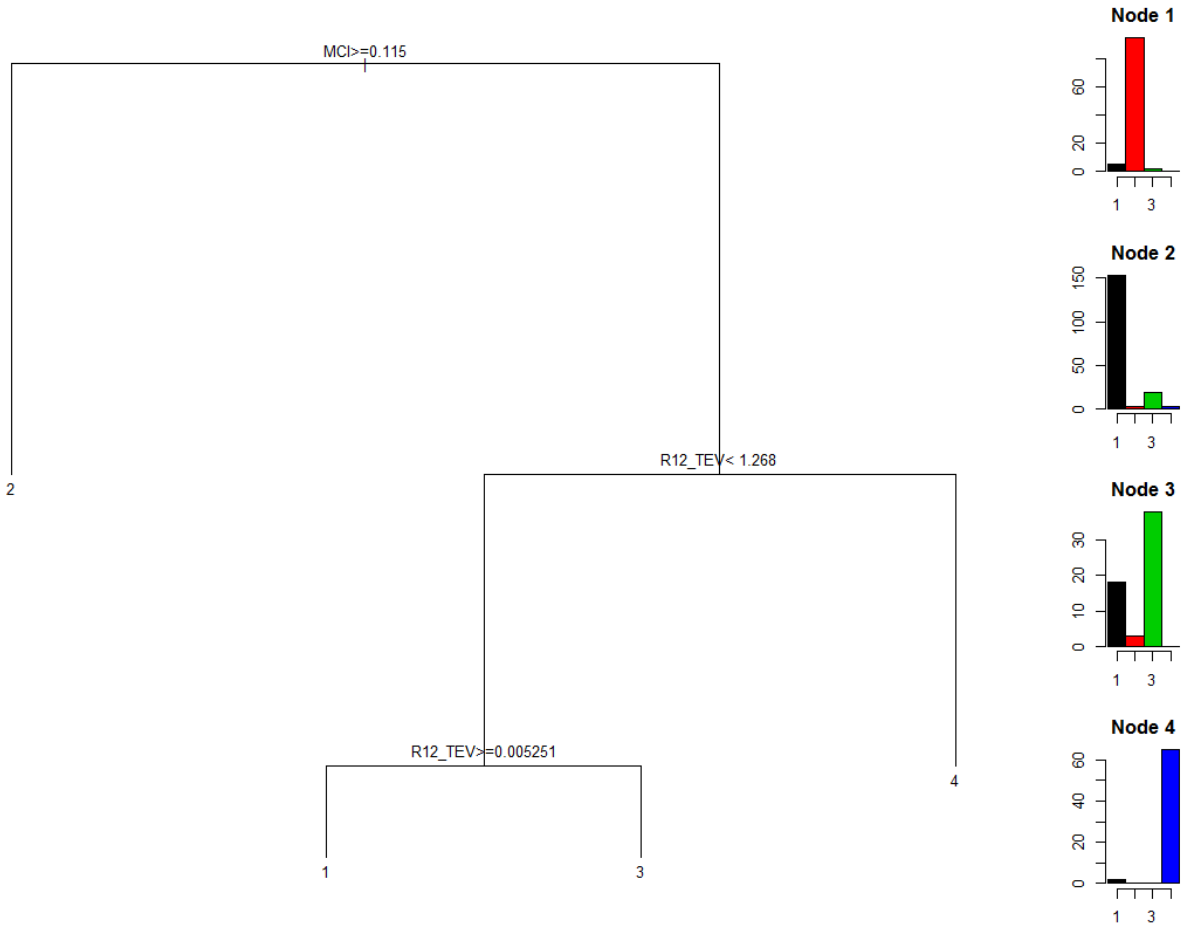

Figure 16. Classification tree for the five group scheme identified by Euclidean hierarchical clustering of the scaled dataset. Group 2 has been removed from the analysis and groups 3-5 are labelled as 2- 4 in the classification tree panel, respectively. Histograms to the right of the classification tree indicate the class make-up of each terminal node, where node numbering proceeds from left to right along the base of the classification tree.

Table 13. Classification accuracy of the classification tree developed for the five group scheme identified by Euclidean hierarchical clustering of the scaled dataset.

Group	Classification accuracy (%)	No. in Group
1	86.0	179
2	100.0	37
3	94.1	101
4	64.4	59
5	97.0	67
Total	87.8	406

The resulting classification tree had good classification accuracy of 87.8%, but misclassified a large proportion of group 4 (Table 13). Group 4 seems to overlap with many of the other groups, and its boundary seems to run at an angle to both STP/TEV and SCI axes

(Figure 11), which would lead to misclassifications due to the constraints of the classification tree method. The entire tree (including the split for group 2) has splits along STP/TEV, MCI and R12/TEV axes, although the R12/TEV axis is split twice, compared to only once for the other factors (Figure 16). Groups are characterised by

(1) STP/TEV $\geq$ 0.0067, MCI < 0.115, R12/TEV < 1.268, R12/TEV $\geq$ 0.00525
(2) STP/TEV < 0.0067
(3) STP/TEV $\geq$ 0.0067, MCI $\geq$ 0.115
(4) STP/TEV $\geq$ 0.0067, MCI < 0.115, R12/TEV < 0.00525
(5) STP/TEV $\geq$ 0.0067, MCI < 0.115, R12/TEV $\geq$ 1.268

2.2.8 – Summary, advantages and disadvantages

In summary the hierarchical clustering method using Euclidean distances produces groupings which reflect some of the physical differences between hydrosystems. The analysis of unscaled data successfully differentiated bays and river mouths from the remaining systems, but grouped a large proportion of partially closed systems into a single group and grouped lakes and lagoons with fiords and sounds. In comparison the analysis of scaled data successfully differentiated lakes and lagoons, bays and some river mouths from the other systems, but other river mouths were grouped with harbours, which were split into two groups, those with higher structural complexity and higher riverine inputs, and those with lower structural complexity and lower riverine inputs.

The advantages of this method are:

- Euclidean distance is a sensible concept for distinguishing among systems
- Individual parameters (e.g. SCI, MCI, R12/TEV or STP/TEV) can be weighted in the calculation of Euclidean distance (weighted Euclidean distances) to increase the likelihood of splits along those axes that are deemed to be more important in distinguishing among systems
- With the hierarchical structure, observations can be pooled up the tree to create super-groups, and equally groups can be split into finer scale groupings
- Tried and trusted method

The disadvantages are

- For data on very different scales, the data needs to be transformed, and potentially scaled

- Group structure is sensitive to changing transformations (see Appendix B), and scaling of data, leading to uncertainty over what's the most appropriate way to treat the data prior to clustering
- Different linkage methods can produce alternative clustering, adding another level of uncertainty (see Appendix B)
- Deciding on the number of groups can be difficult, and is often arbitrarily chosen, although this report presents a method that can guide these choices

2.3 – Method 2: Hierarchical clustering based on Random Forests distance measures

The second method applied to the coastal hydrosystem dataset is similar to the first, in that it is based on hierarchical clustering, but uses a different distance metric to discern between observations. The Random Forests method is a machine learning method that builds upon the classification tree method described earlier (Breiman 2001, Cutler et al. 2007). It was primarily developed to improve predictive accuracy in supervised classification analyses, but a by-product of this in the form of the random forests proximity measure can be altered to form a distance metric that can be used for unsupervised classification (i.e. where the labels are unknown).

The random forests method works in the following way. Datasets containing predictor variables (in this case the physical variables of each of the coastal hydrosystems) and classification labels are split randomly into training and test datasets, usually with two thirds of the data reserved in the training dataset and one third in the test dataset. The training dataset is then used to construct an overfitted classification tree, which partitions multivariate space until every partition, and hence terminal node of the tree, contains a single observation. The test dataset objects are then passed down the tree (i.e. subjected to all of the rules in the tree in a sequential order) until it reaches a terminal node. The group label of the object forming the terminal node is then used as a vote for the class of the test dataset object. This process is repeated creating a number of trees (usually thousands, but depends on dataset size) which are built using a random subset of the data (Breiman 2001, Cutler et al. 2007). Every time an observation is in the test dataset a vote for a class type is assigned to it based on the terminal node it ended up in. This can also be used to determine the proximity between any two points, as the frequency at which any two objects are assigned to the same terminal node (Shi & Horvath 2006). To increase predictive accuracy and to reduce correlation among trees (i.e. trees voting similar ways each time), each partition in any particular tree is formed based on a randomly chosen subset of the predictor variables. For example at a particular node a number

of predictors approximately equal to the square root of the total number of predictors are chosen, which in this case would involve selecting two of the four physical variables. The partition that best classifies data at that point in the tree along those two axes only, is chosen. Thus the tree structure changes with each random tree constructed based partly on the random selection of the training dataset and the random selection of axes to partition at each stage (Breiman 2001).

For unsupervised classification this method can be manipulated to obtain proximities between points when no group labels exist (Shi & Horvath 2006). A dataset consisting of the real dataset, and a simulated dataset with similar properties and size, are labelled according to whether the data is real or simulated. This combined dataset is then subjected to the Random Forests classification analysis to distinguish between the real and simulated data. The proximity measures for the real data, obtained through the Random Forests voting process, are extracted after all trees have been constructed and then scaled to remove all votes where real data was assigned to a simulated data terminal node. This proximity matrix between all possible real observations, has the quality of representing the differences between points that is invariant to monotonic transformations of the data, and isn't adversely affected when variables have ranges and/or variances of different scales (Shi & Horvath 2006). It also represents the similarity between points as determined through a statistical process of classifying the data, and therefore may be better at distinguishing meaningful differences and clusters than the Euclidean distance measure. The proximity measure between any two points, $\rho_{A,B,}$ can be converted to a distance metric, $\delta_{A,B}$, via:

$$\delta_{A,B} = \sqrt{1 - \rho_{A,B}}$$

(eqn. 3)

The resulting random forests distance measure can then be used as a dissimilarity measure in a hierarchical cluster analysis. However, the construction of the proximity measure is stochastic in that it is determined by random selections of the real dataset and the construction of the trees. Therefore proximity measures can vary between alternate runs, which may result in different grouping structures. However, by performing multiple runs of the same analysis, the change in group structure between runs can be used to identify how strongly observations belong to specific groups, and which observations are intermediate between groups.

In this section Hierarchical clustering, based on random forests distance measures and Ward's linkage criterion, is applied to the dataset of R12/TEV, STP/TEV, SCI and MCI for the 443 NZ coastal hydrosystems and the resulting grouping schemes are discussed and analysed further.

2.3.1 – Initial treatment of data

As stated in the previous section the random forests method and the resulting distance measure are scale and transformation invariant with regard to the predictor or physical variables (Shi & Horvath 2006). To illustrate this four parallel analyses are performed. The untransformed dataset and the scaled and transformed dataset (detailed in Section 2.2.1) are both used to calculate Euclidean distances and random forest distances. The resulting distance matrices are then visualised using non-Metric Dimensional Scaling (nMDS), which aims to visualise differences between multivariate objects in a lower dimensional space.

The resulting plots of data based on Euclidean distance measures vary considerably between untransformed and transformed datasets, whereas for the random forests distance measure the resulting distributions are almost identical between untransformed and transformed datasets (Figure 17). Consequently all analyses were based on untransformed data. The random forests routine was performed with 20,000 trees to calculate proximity measures using the **RandomForest** function in R (Liaw & Wiener 2002), which were subsequently converted to distance measures. The distances were then used to perform a hierarchical cluster analysis based on Ward's linkage criterion using the **hclust** function in R.

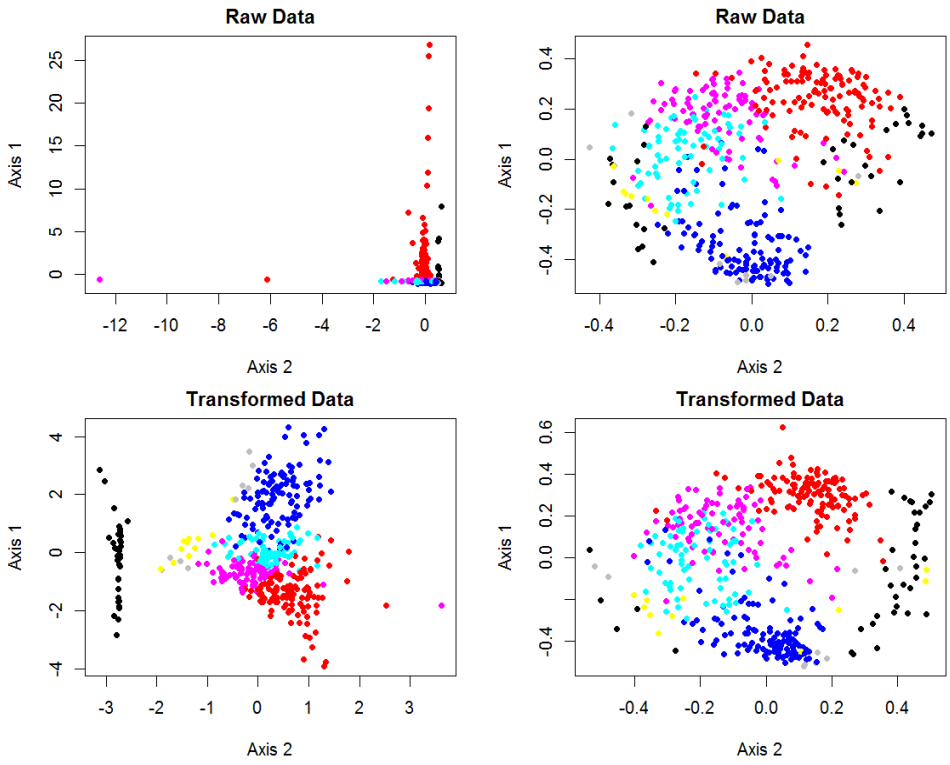

Figure 17. nMDS plots illustrating the multivariate dispersion of the coastal hydrosystem dataset according to Euclidean (left panels) and random forests distance measures (right panels) calculated from untransformed (top panels) and transformed and scaled datasets (bottom panels). Points are colour coded according to the Hume et al. (2007) classification.

2.3.2 – Deciding on the "best" number of groups

As for the previous hierarchical cluster analysis, it is difficult to identify the correct number of groups. A similar cross-validation routine can be used to identify how classification accuracy changes with increasing numbers of groups, but this can be done taking advantage of the random forest analysis method. Based on the hierarchical cluster analysis of the random forests distance measure, a classification scheme with *k* groups can be obtained. This classification scheme can then be analysed using the random forests method for **supervised** classification to discern between classes, with the physical variables as predictors, and the out-of-bag (OOB) error rate (i.e. the rate at which test dataset observations are misclassified) can be obtained. This can be performed multiple times across a range of values for *k* to examine how OOB error varies with *k*. This routine was implemented using a function written in R.

The resulting plot of OOB error against *k* indicates that OOB error rates of 5% or less are satisfied by four or fewer groups, error rates of 7.5% or less are satisfied by seven or fewer groups, and error rates of 10% are satisfied by 21 or fewer groups (Figure 18). The OOB error

for six groups is lower than for five groups, which may suggest that this number of groups is more strongly supported, however it is difficult to identify which number of groups is suitable. Despite this six and seven groups will be investigated, corresponding to the maximum number of groups with error rates lower than 7.5%.

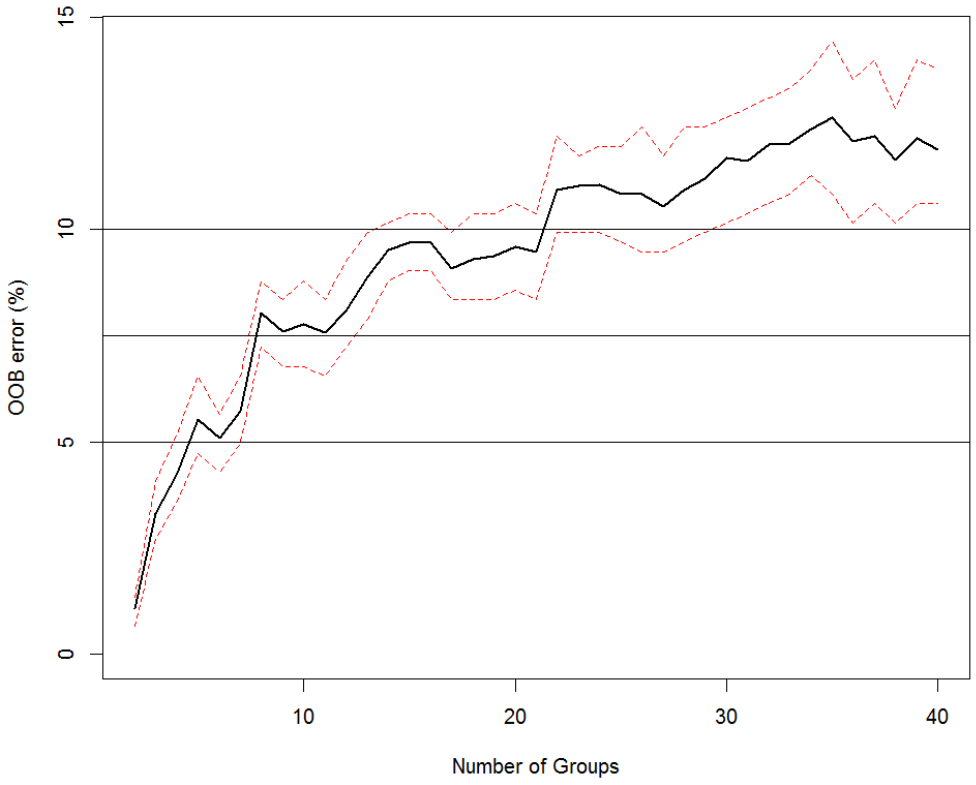

Figure 18. Out of bag error rates for random forests hierarchical clustering as a function of the number of groups. The thick black line is the mean OOB error, whilst red dotted lines are the 95% confidence interval. Black horizontal lines are placed at OOB error rates of 5, 7.5 and 10%.

2.3.3 – Grouping scheme for six groups

The six group analysis is illustrated in Figure 19. Splits are apparent along multiple axes with group 2 separated along STP/TEV, group 3 and 5 along SCI and MCI axes and groups 4 and 6 along the R12/TEV axis (Figure 19).

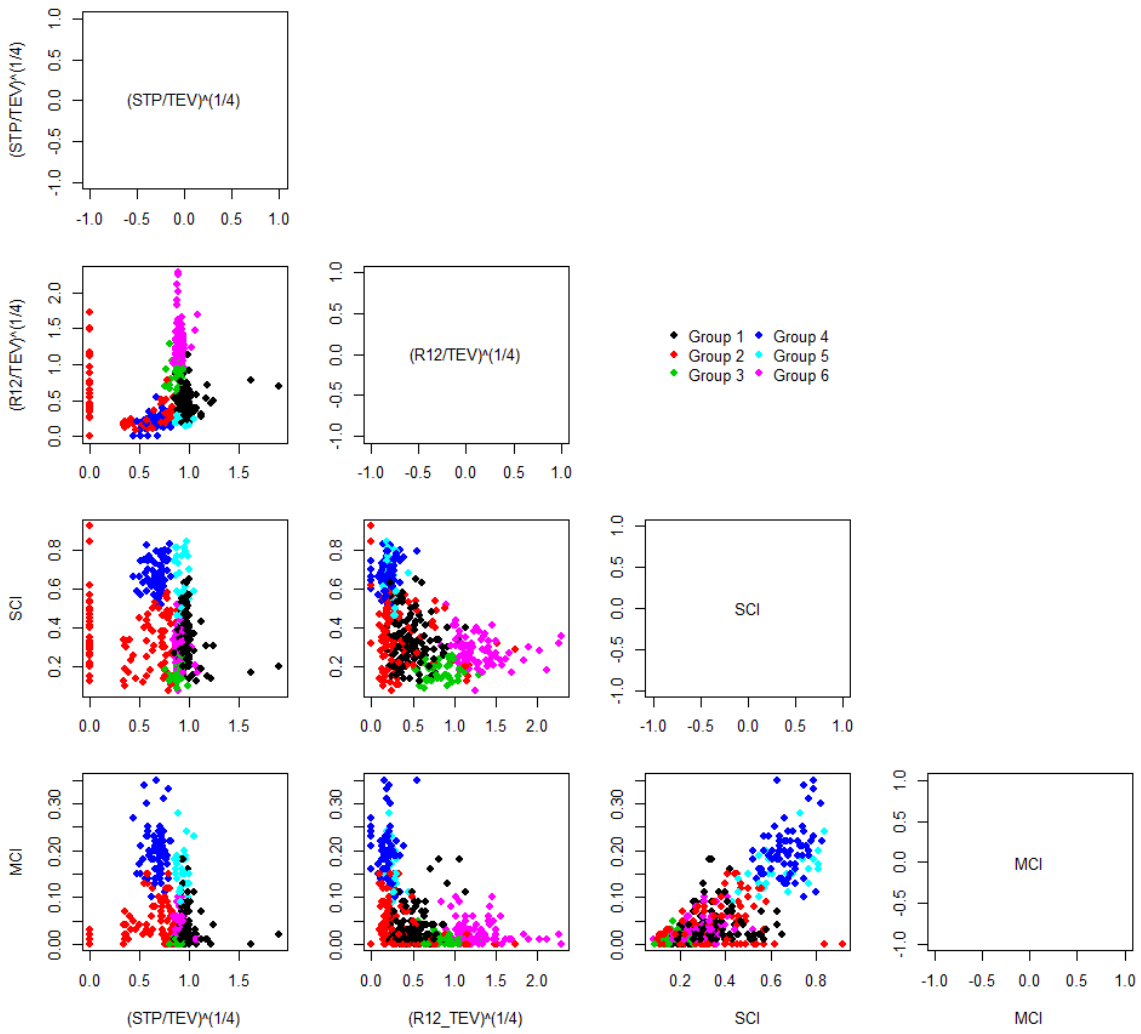

Figure 19. Biplots of the four physical variables colour-coded by group label as identified by the random forests hierarchical clustering method, with k = 6. STP/TEV and R12/TEV are fourth-root transformed, whilst SCI and MCI are untransformed.

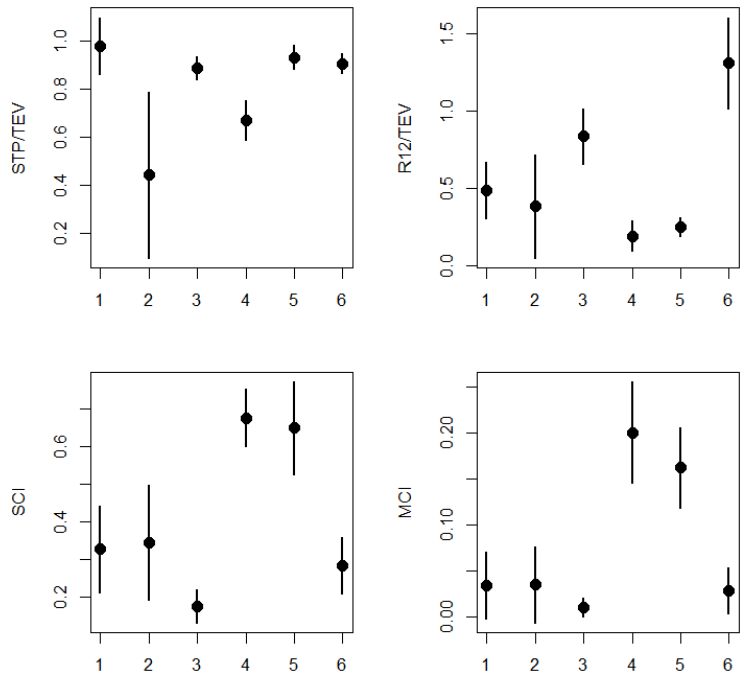

Figure 20. Means (±1 SD) of the four physical variables by group with group labels identified by the random forests hierarchical clustering method, with k = 6. STP/TEV and R12/TEV are fourth-root transformed, whilst SCI and MCI are untransformed.

Groups can be characterised as (Figure 20):

- 1 – high STP/TEV but overlapping with other groups, mix of estuaries, inlets/harbours and river mouths
- 2 – low-moderate STP/TEV, but largely overlapping with other groups, primarily harbour systems, lakes, lagoons and sounds
- 3 – moderate R12/TEV, low SCI and MCI, primarily river mouths
- 4 – high SCI/MCI and moderate STP, primarily bays
- 5 – high SCI/MCI, high STP/TEV, primarily bays and harbours
- 6 – high R12/TEV, primarily river mouths

Comparing these labels to the Hume classification (Table 14):

- 1 – mix of B, E and F class systems
- 2 – mix of all classes, but predominantly A, F and D systems
- 3 – corresponds strongly to a selection of B, with some F class systems
- 4 – corresponds to the majority of D class systems
- 5 – corresponds to a selection of class D systems
- 6 – corresponds to the majority of class B systems

Table 14. Table illustrating the number of systems in each group that were of the different hydro classes outlined in the Hume et al. (2007) classification. This is based on random forests hierarchical clustering, with $k = 6$.

Hume Hydro Class	Hierarchical Group					
	1	2	3	4	5	6
A	0	37	0	0	0	0
B	17	1	30	0	0	73
D	7	18	0	61	27	0
E	52	16	0	0	0	0
F	50	22	12	0	0	0
G	0	10	0	1	0	0
H	0	4	0	5	0	0

There is some agreement between this classification and the Hume classification, particularly regarding B and D class systems, but many classes are grouped together in groups 1 and 2. Group 1 is primarily harbours, with some river mouths and is characterised by high STP and moderate complexity. This could indicate that group 1 is typical of moderately complex partially open bodies with high tidal influence, including harbours and tidal river mouths. Group 2 consists of lakes, lagoons, harbour systems, sounds and fiords. This is the most mixed group and is perhaps characterised as those hydrosystems with the lowest tidal influence relative to their volume. Group 3 is primarily river mouths, but a selection of river mouths with moderate R12/TEV and high structural complexity. Group 3 also contains some drowned valleys or harbour systems, and so this group could be characterised is structurally complex harbour systems and partially closed river mouths. Group 4 and 5 both correspond to bays, with 4 corresponding to bays with lower tidal influence and also includes several sounds and fiords, and 5 corresponding to bays with large tidal influence relative to their volume. Group 6 are those river mouths with large river input relative to their volume, which separates them from group 3.

2.3.4 – Grouping scheme for seven groups

The seven group structure is the same as the six group structure, but with the previous group 2 split into a very low STP/TEV group (Group 2) and a group with moderate STP/TEV (Group 3) (Figure 21).

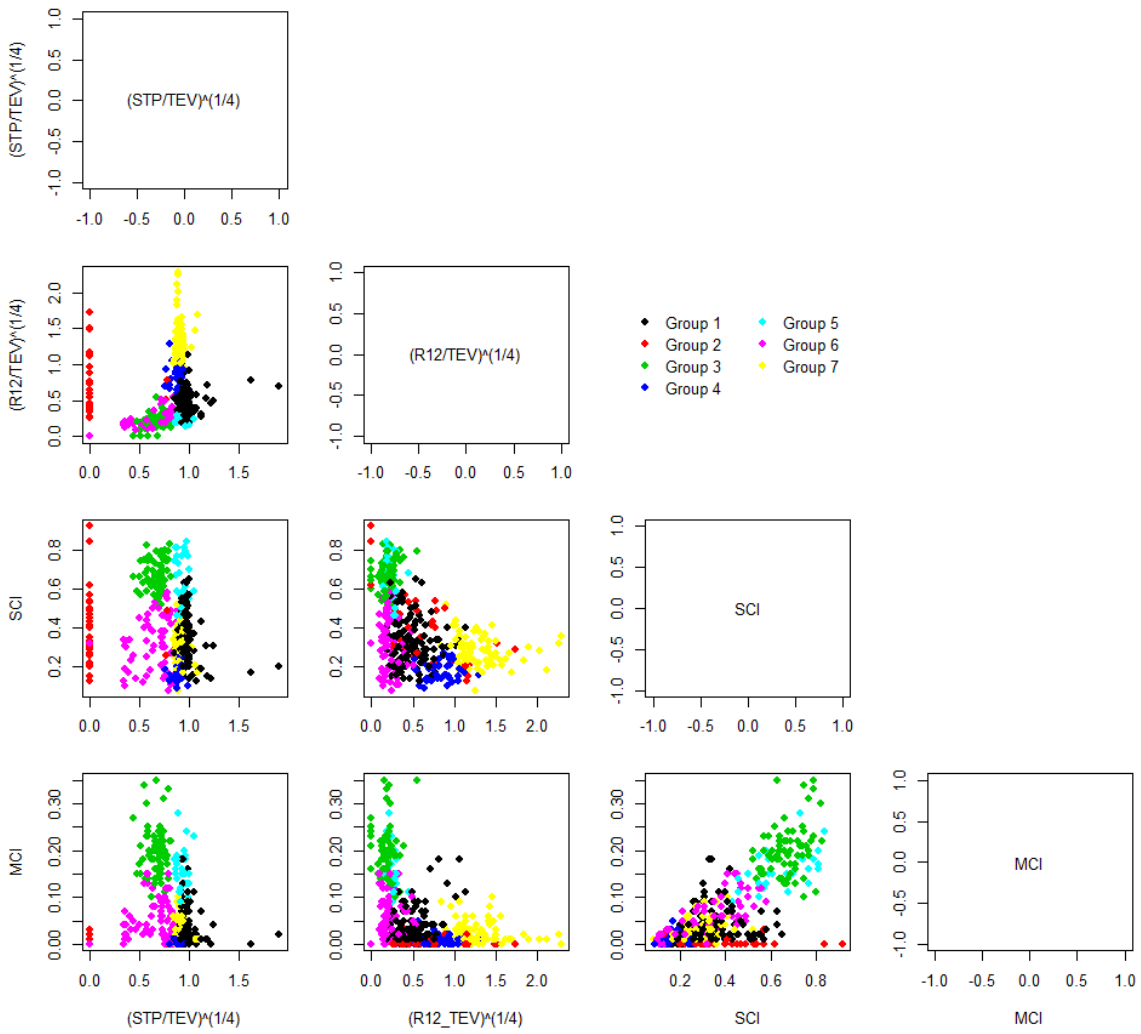

Figure 21. Biplots of the four physical variables colour-coded by group label identified by the random forests hierarchical clustering method, with k = 7. STP/TEV and R12/TEV are fourth-root transformed, whilst SCI and MCI are untransformed.

Groups can be characterised as (Figure 22):

- 1 – high STP/TEV but overlapping with other groups, mix of estuaries, inlets/harbours and river mouths
- 2 – very low STP/TEV, primarily lakes and lagoons
- 3 – high SCI/MCI and moderate STP/TEV, primarily bays
- 4 – low SCI, primarily river mouths
- 5 – high SCI/MCI, high STP/TEV, primarily bays and harbours
- 6 – moderate STP/TEV, low SCI/MCI, primarily harbours, harbour systems and sounds
- 7 – high R12/TEV, primarily river mouths

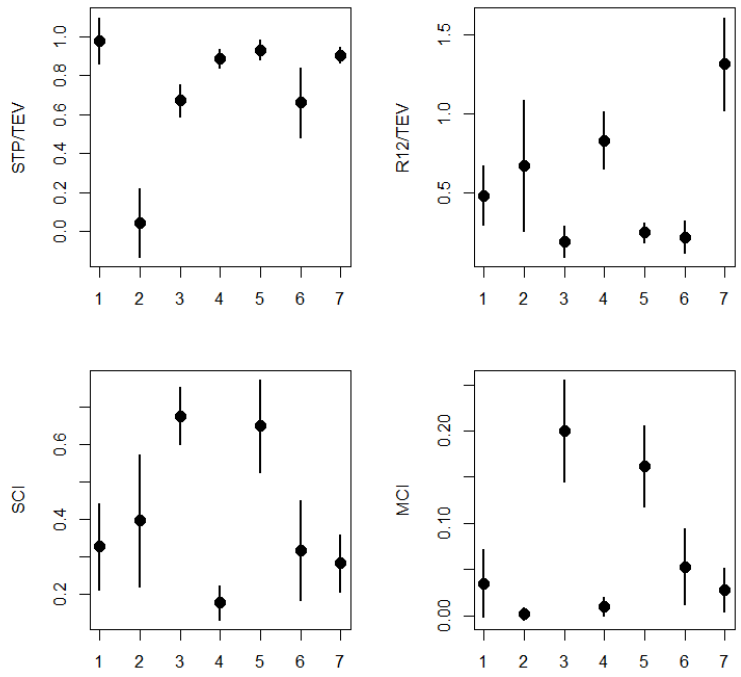

Figure 22. Means (±1 SD) of the four physical variables by group with group labels identified by random forests hierarchical clustering, with k = 7. STP/TEV and R12/TEV are fourth-root transformed, whilst SCI and MCI are untransformed.

Table 15. Table illustrating the number of systems in each group that were of the different hydro classes outlined in the Hume et al. (2007) classification. This is based on random forests hierarchical clustering, with k = 7.

Hume Hydro Class	Hierarchical Group						
	1	2	3	4	5	6	7
A	0	36	0	0	0	1	0
B	17	1	0	31	0	0	72
D	7	0	61	0	27	18	0
E	52	0	0	0	0	16	0
F	50	1	0	12	0	21	0
G	0	0	1	0	0	10	0
H	0	0	5	0	0	4	0

Comparing these labels to the Hume classification (Table 15):

- 1 – mix of B, E and F class systems
- 2 – corresponds strongly to class A
- 3 – corresponds strongly to class D with some H class systems
- 4 – corresponds to a mix of classes B and F
- 5 – corresponds to a selection of class D systems

- 6 – corresponds to a mix of D-H systems
- 7 – corresponds to the majority of class B systems

This grouping scheme builds on the six group scheme by separating lakes and lagoons from the other hydrosystems with low STP/TEV. The newly formed group 2 corresponds almost exactly to all of class A and corresponds to isolated lakes and lagoons. The remaining group, group 6, is not very distinct from the other groups in that it occupies the middle ground between many groups in terms of physical characteristics, and contains primarily harbours, harbour systems and sounds, which have a lower STP/TEV due to their significantly larger volumes.

2.3.5 – Rule-based summary of grouping schemes

Six groups

All groups displayed significant overlap along one or more axis (Table 16).

Table 16. Physical variable ranges for each group based on the six group scheme identified using random forests hierarchical clustering.

Group	R12/TEV	STP/TEV	SCI	MCI
1	0.001, 1.613	0.543, 13.135	0.13, 0.65	0, 0.18
2	0.000, 8.968	0.000, 0.592	0.08, 0.92	0, 0.15
3	0.074, 2.789	0.327, 0.971	0.09, 0.26	0, 0.05
4	0.000, 0.090	0.039, 0.459	0.52, 0.83	0.10, 0.35
5	0.000, 0.041	0.527, 1.182	0.43, 0.84	0.09, 0.28
6	0.492, 27.130	0.498, 1.357	0.08, 0.52	0, 0.10

The resulting rule set determined via classification tree analysis achieves good classification accuracy overall (87.8%), but is less accurate at classifying group 3 (50%) (Table 17). Splits are located along STP/TEV (1), R12/TEV (2) and MCI (2) axes (Figure 23). The rule set (Figure 23) indicates that groups are characterised by:

(1) STP/TEV $\geq$ 0.5224, R12/TEV < 0.9548, MCI < 0.095, R12/TEV < 0.2976

(2) STP/TEV < 0.5224, MCI < 0.125

(3) STP/TEV $\geq$ 0.5224, R12/TEV < 0.9548, MCI < 0.095, R12/TEV $\geq$ 0.2976

(4) STP/TEV < 0.5224, MCI $\geq$ 0.125

(5) STP/TEV $\geq$ 0.5224, R12/TEV < 0.9548, MCI $\geq$ 0.095

(6) STP/TEV $\geq$ 0.5224, R12/TEV $\geq$ 0.9548

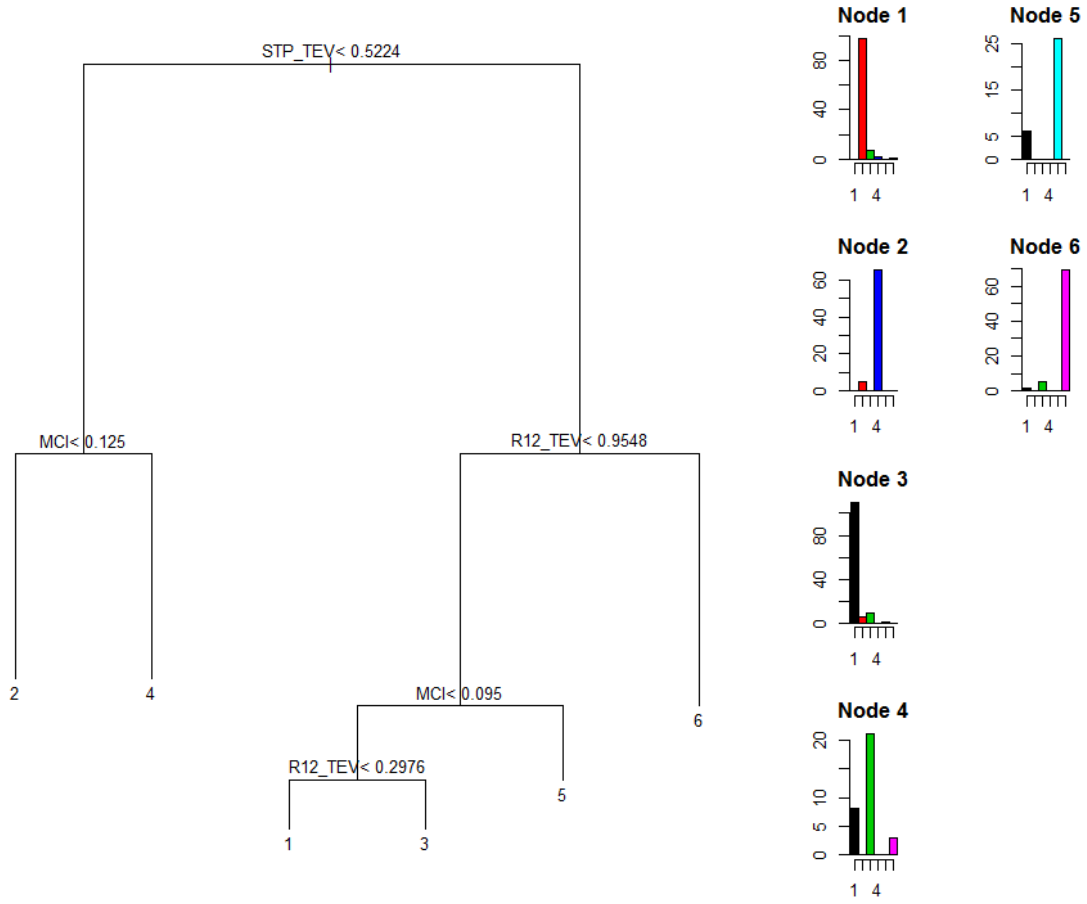

Figure 23. Classification tree for the six group scheme identified by random forests hierarchical clustering. Histograms to the right of the classification tree indicate the class make-up of each terminal node, where node numbering proceeds from left to right along the base of the classification tree.

Table 17. Classification accuracy of the classification tree developed for the six group scheme identified by random forests hierarchical clustering.

Group	Classification accuracy (%)	No. in Group
1	87.3	126
2	90.7	108
3	50.0	42
4	97.0	67
5	96.3	27
6	94.5	73
Total	87.8	443

Seven groups

All groups displayed significant overlap along one or more axis (Table 18).

Table 18. Physical variable ranges for each group based on the seven group scheme identified by random forests hierarchical clustering.

Group	R12/TEV	STP/TEV	SCI	MCI
1	0.001, 1.613	0.543, 13.135	0.13, 0.65	0, 0.18
2	0.000, 8.968	0.000, 0.373	0.13, 0.92	0, 0.03
3	0.000, 0.090	0.0386, 0.459	0.52, 0.83	0.1, 0.35
4	0.073, 2.789	0.327, 0.971	0.09, 0.27	0, 0.05
5	0.000, 0.041	0.527, 1.182	0.43, 0.84	0.09, 0.28
6	0.000, 0.087	0.000, 0.592	0.08, 0.58	0, 0.15
7	0.492, 27.130	0.498, 1.357	0.08, 0.52	0, 0.1

The resulting rule set determined via classification tree analysis achieves good classification accuracy overall (86.9%), but is less accurate at classifying group 4 (51.2%) (Table 19). The rule set (Figure 24) indicates that groups are characterised by:

(1) STP/TEV $\geq$ 0.5412, R12/TEV < 1.061, MCI < 0.095, R12/TEV < 0.2976

(2) STP/TEV < 0.5412, MCI < 0.125, STP/TEV < 0.0067

(3) STP/TEV $\geq$ 0.5412, MCI $\geq$ 0.125

(4) STP/TEV $\geq$ 0.5412, R12/TEV < 1.061, MCI < 0.095, R12/TEV $\geq$ 0.2976

(5) STP/TEV $\geq$ 0.5412, R12/TEV < 1.061, MCI $\geq$ 0.095

(6) STP/TEV < 0.5412, MCI < 0.125, STP/TEV $\geq$ 0.0067

(7) STP/TEV $\geq$ 0.5412, R12/TEV $\geq$ 1.061

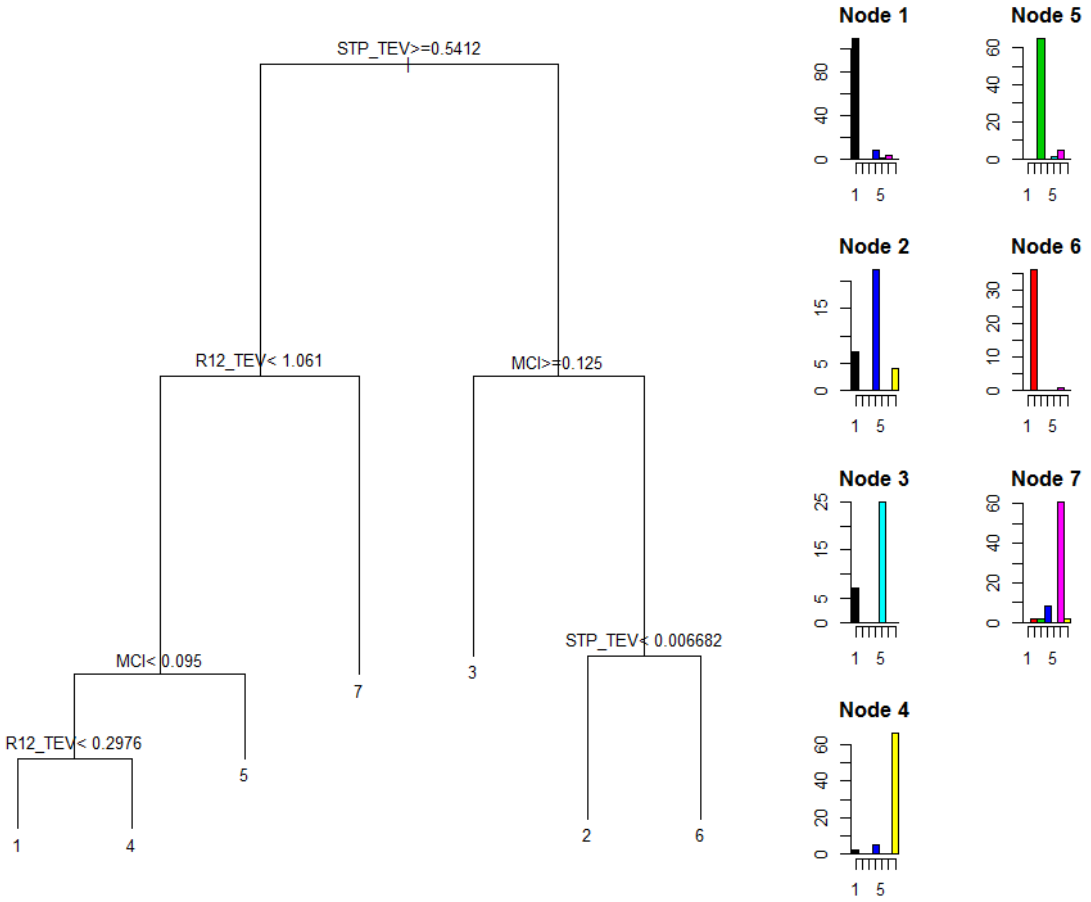

Figure 24. Classification tree for the seven group scheme identified using random forests hierarchical clustering. Histograms to the right of the classification tree indicate the class make-up of each terminal node, where node numbering proceeds from left to right along the base of the classification tree.

Table 19. Classification accuracy of the classification tree developed for the seven group scheme identified using random forests hierarchical clustering.

Group	Classification accuracy (%)	No. in Group
1	87.3	126
2	94.7	38
3	97.0	67
4	51.2	43
5	92.6	27
6	87.1	70
7	91.7	72
Total	86.9	443

2.3.6 – Summary, advantages and disadvantages

In summary the hierarchical clustering method using random forests distances produces groupings that highlight some of the physical differences between hydrosystems. The analysis

successfully differentiated between bays (Hume class D), river mouths (Hume class B) and lakes and lagoons (Hume class A) but a large proportion of the remaining classes were undifferentiated. In a similar way to the Euclidean distance analysis the remaining groups were those that were partially closed such as harbour systems, sounds, fiords, estuaries and partially closed river mouths. In addition class B (river mouths) systems were split into high and moderate R12/TEV groups that also had differences in structural complexity, suggesting more than one class might be required to classify river mouths. Similarly class D (bays) were split into moderate and high STP/TEV classes. In contrast classes E and F were predominantly grouped together, indicating that the physical differences (at least as determined by this method) between high and low variants of bays and river mouths, are much more prominent than the differences displayed between classes E and F.

The advantages of this method are:

- Distance measure does not depend on the spread or scale of the data and is invariant under transformations of the data
- The distance measure is based upon a machine learning algorithm that tries to find points that are similar and is less influenced by extreme values
- The importance of physical variables in identifying group structure is determined by the data alone
- With the hierarchical structure, observations can be pooled up the tree to create super-groups, and equally groups can be split into finer scale groupings, which can be useful from a management perspective
- The variability among repeat trials can be used to identify how strongly an observation belongs to a group, as well as group membership across all groups for each observation, including the identification of systems that are intermediate among classes (see Appendix C)

The disadvantages are

- Group structure is stochastic and can vary from analysis to analysis
- Less control over which physical axis is the most important for determining group structure
- Different linkage methods can produce alternative clustering (see Appendix B)
- Deciding on the number of groups can be difficult, and is often arbitrarily chosen, although this report presents a method that can guide these choices

- It is a more complicated concept than Euclidean distance, and has been less widely adopted.

2.4 – Method 3: Model-based clustering

Model-based clustering differs from the previous two methods in that it is not based on a hierarchical construction but is based on the identification of clusters in the data using mixture modelling (Fraley & Rafferty 2002). Mixture modelling consists of identifying separate components within a mixed dataset based on the creation of probabilistic models for each component, concurrent with the labelling of individual systems that relate to the model components.

The entire dataset can be considered as a mixture of *k* groups, where the labelling and the number of groups is unknown. Each group can be considered to occupy a portion of multivariate space, in that it has a mean location $\underline{\mathbf{X}} = (\overline{x_1}, \overline{x_2}, \overline{x_3} \ldots \overline{x_N})$, where $\overline{x_i}$ is the mean location of that group along the i^{th} axis, and a variance-covariance matrix $\underline{\underline{\mathbf{V}}}$, that determines the extent along each axis, in addition to its orientation relative to the physical axes. Therefore each group can be considered as a probabilistic component and likelihood methods (i.e. maximising the global probability of that group number and group type across all observations) can be used to identify the best fitting parameters for $\underline{\mathbf{X}}$ and $\underline{\underline{\mathbf{V}}}$. However, in order to do this a pre-existing group labelling structure is required to assign observations to a specific group so that the correct probability of that observation coming from that group is used in the likelihood maximisation process. Therefore a process known as the expectation maximisation (EM) algorithm is used. The EM algorithm alternates between an expectation (E) step and a maximisation (M) step. In simple terms, for a specific number of groups some initial label structure for the data is created (usually by running a hierarchical or k-means cluster analysis on the same data), such that each observation has an initial label. In the maximisation step the means ($\underline{\mathbf{X}}$) and variance-covariance matrix ($\underline{\underline{\mathbf{V}}}$) of each group is estimated so that the overall likelihood of that label structure is maximised. These components are then used in the expectation step to re-estimate and re-define the group labelling structure. This new labelling structure is passed to the maximisation step, where $\underline{\mathbf{X}}$ and $\underline{\underline{\mathbf{V}}}$ are re-estimated. This process repeats until there is little to no change in the overall likelihood between stages, indicating that the most likely group structure and its associated model components, in terms of mean and variance, have been estimated (Fraley & Rafferty 2002). This method has the advantage that an overall likelihood (i.e. how well this model fits the data) is assigned to this model structure, which can be used to select the best number of groups. This is usually performed using the

Bayesian Information Criterion (BIC), which combines likelihood (a measure of goodness of fit) penalised by the number of model components (i.e. greater number of groups are penalised more than fewer groups due to added complexity) to identify the most parsimonious model, which adequately describes the variability of the data without overfitting.

$$BIC = 2\ln(L) - k.\ln(n)$$

(eqn. 4)

Where L is the (maximised) likelihood of the model, k is the number of model components and n is the size of the dataset. The model with the highest BIC is usually selected as the model which provides the best explanation of the grouping structure present in the data (Fraley et al. 2012).

There are several different types of model of varying complexity that can be fitted to the data, corresponding to different cluster shapes, sizes and orientations (as determined via restrictions on the variance covariance matrix) (Table 20).

Table 20. Types of model-based clusters used for multivariate unsupervised cluster analyses. Model complexity increases from top to bottom. Adapted from Fraley et al. (2012).

Type	Distribution	Volume	Shape	Orientation
EII	Spherical	Equal	Equal	N/A
VII	Spherical	Variable	Equal	N/A
EEI	Diagonal	Equal	Equal	Coordinate Axes
VEI	Diagonal	Variable	Equal	Coordinate Axes
EVI	Diagonal	Equal	Variable	Coordinate Axes
VVI	Diagonal	Variable	Variable	Coordinate Axes
EEE	Ellipsoidal	Equal	Equal	Equal
EEV	Ellipsoidal	Equal	Equal	Variable
VEV	Ellipsoidal	Variable	Equal	Variable
VVV	Ellipsoidal	Variable	Variable	Variable

The distribution (spherical, diagonal or ellipsoidal) refers to the shape of the multivariate normal distribution, with spherical clusters having the same extent along each multivariate axis (i.e. a spherical shape), diagonal clusters are those with a distribution that is ovoid and the major axes are parallel and perpendicular to the multivariate axes and ellipsoidal clusters can have any shape and orientation (Table 20). Volume refers to the physical size of the cluster in multivariate space (determined via the variance-covariance matrix) and those models with equal volume have clusters that are of equal volume, whereas variable volume models can have clusters of different volumes. Shape refers to the ratio of the cluster axes relative to each other

and whether they are equal or variable among clusters. Orientation refers to the orientation of the clusters relative to the cluster axes, with types EEI, VEI, EVI and VVI having clusters whose major axes are parallel and perpendicular to coordinate axes, whereas the remaining models have orientations that can be at an angle to coordinate axes that is either equal (EEE) or variable (EEV, VEV, VVV) among clusters (Table 20) (Fraley et al. 2012). Models of each type can be fitted to the data for a range of group numbers and the "best" number of groups, and group type can be selected by BIC.

In this section a model-based clustering analysis is applied to the dataset of R12/TEV, STP/TEV, SCI and MCI for the 443 NZ coastal hydrosystems and the resulting grouping schemes are discussed and analysed further.

2.4.1 – Initial treatment of data

In contrast to the random forests analysis, model-based clustering is sensitive to data transformations and scaling of variables. Both unscaled and scaled (data transformations and scaling were performed as detailed in Section 2.2.1) analyses were performed, with the scaled results presented here and the unscaled results presented in Appendix D.

2.4.2 – Deciding on the "best" number of groups

Deciding on the best number of groups using this method is more intuitive and less open to interpretation than for the hierarchical cluster analysis methods. Models with group numbers ranging from 2-20 of all model types detailed in Table 20, were fitted to the data using the **mclustBIC** function in R (Fraley et al. 2012) and the BIC statistics obtained (Table 21). The VEV model with 7 groups had the maximum BIC, but a 6 group model (also VEV) differed by less than 0.1 indicating almost identical goodness of fit (Table 21). No other models were within $\Delta BIC < 10$, indicating that these models had far greater support than any other model (Table 21).

Table 21. BIC statistics of model-based clustering types for group numbers ranging from 2-20. NA indicates models failed to fit the data due to singularities in variance-covariance matrices. BIC statistics in **bold** indicate the two best fitting models based on BIC.

No. Groups	Model Type									
	EII	VII	EEI	VEI	EVI	VVI	EEE	EEV	VEV	VVV
2	-4639.6	-4538.4	-4416.4	-4412.4	-4101.2	-4077.1	-4348.7	-3957.2	-3904.5	-3882.4
3	-4241.8	-4218.5	-4017.1	-4017.0	-3578.1	-3525.5	-3961.2	-3366.0	-3285.4	NA
4	-3970.0	-3926.3	-3946.0	-3888.1	NA	NA	-3859.4	-3179.7	-3171.4	NA
5	-3939.8	-3830.8	-3879.5	-3780.6	NA	NA	-3780.9	-3124.3	-3114.4	NA
6	-3917.2	-3751.0	-3873.7	-3726.9	NA	NA	-3750.9	-3128.9	**-3068.7**	NA
7	-3832.4	-3701.6	-3774.3	-3659.7	NA	NA	-3761.1	-3101.5	**-3068.6**	NA
8	-3831.6	-3686.4	-3707.0	-3615.8	NA	NA	-3616.3	-3116.9	-3103.1	NA
9	-3855.9	-3630.6	-3732.6	-3578.0	NA	NA	-3639.7	-3147.1	-3145.1	NA
10	-3803.8	-3624.3	-3668.4	-3595.5	NA	NA	-3710.0	-3134.5	-3161.4	NA
11	-3795.9	-3590.6	-3653.9	-3549.9	NA	NA	-3673.8	-3120.1	-3114.1	NA
12	-3820.1	-3573.8	-3570.0	-3525.8	NA	NA	-3699.8	-3137.0	-3186.9	NA
13	-3660.1	-3537.1	-3457.5	-3484.7	NA	NA	-3610.2	-3177.1	-3175.9	NA
14	-3661.7	-3547.4	-3454.0	-3455.9	NA	NA	-3434.2	-3225.4	-3232.2	NA
15	-3571.5	-3532.8	-3470.2	-3432.0	NA	NA	-3461.6	-3276.5	-3218.8	NA
16	-3585.6	-3547.2	-3474.0	-3449.0	NA	NA	-3469.1	-3301.4	-3244.4	NA
17	-3567.0	-3556.0	-3414.8	-3422.5	NA	NA	-3413.8	-3386.0	-3321.7	NA
18	-3582.1	-3557.4	-3429.0	-3426.8	NA	NA	-3423.7	-3392.2	-3435.6	NA
19	-3587.1	-3595.8	-3453.6	-3432.9	NA	NA	-3451.3	-3435.3	-3464.9	NA
20	-3493.9	-3576.2	-3404.5	-3417.2	NA	NA	-3400.5	-3454.6	-3468.7	NA

2.4.3 – Grouping scheme for six groups (VEV)

The six group structure displays splits along STP/TEV, R12/TEV and SCI axes (Figures 25 and 26). Group 2 is the most distinct, and is separated from the remaining groups along the STP/TEV axis (Figure 25).

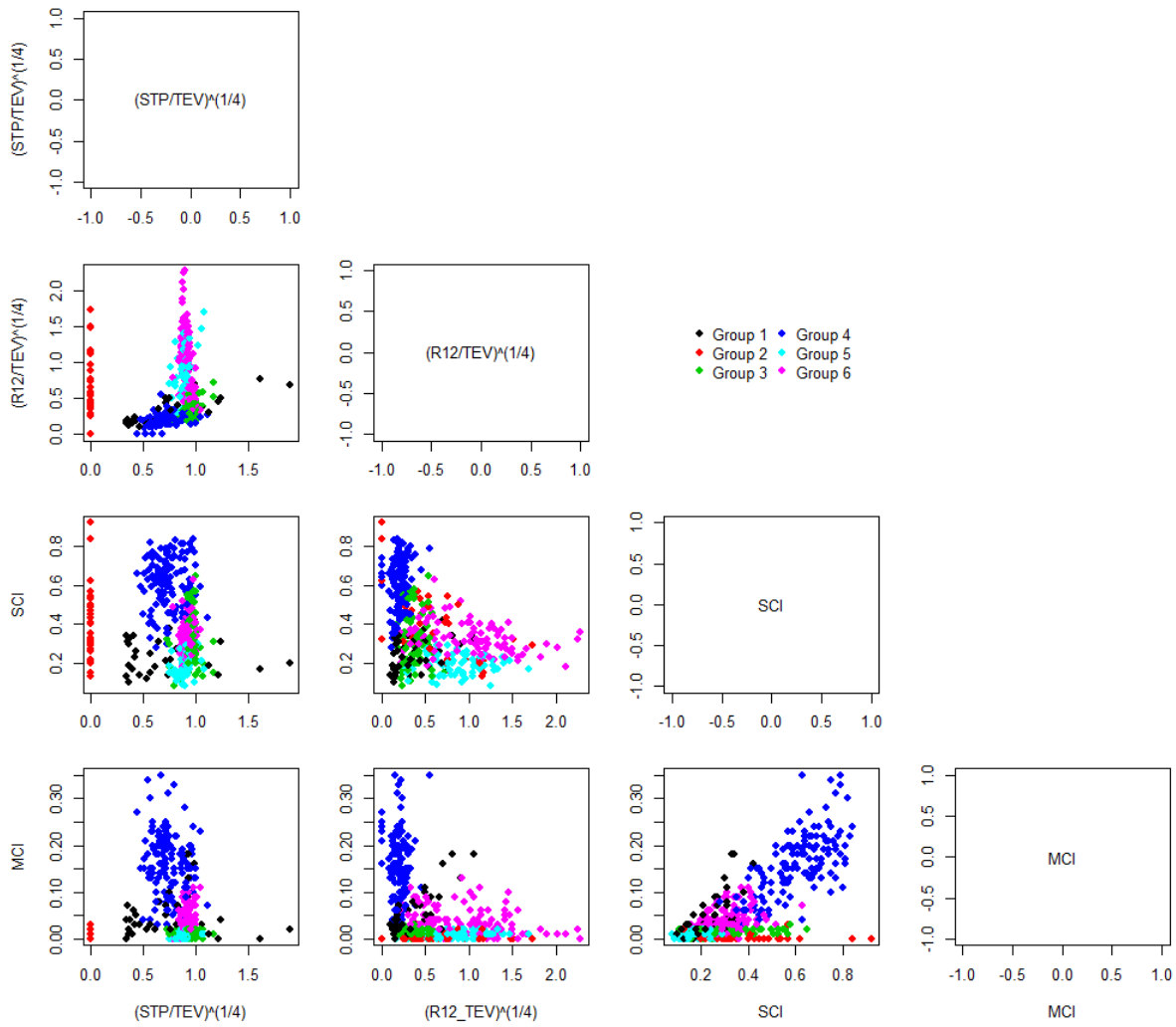

Figure 25. Biplots of the four physical variables colour-coded by group label identified by model-based clustering (type VEV) with k = 6. STP/TEV and R12/TEV are fourth-root transformed, whilst SCI and MCI are untransformed.

Groups can be characterised as (Figure 26):

- 1 –prominent overlap with other groups, mix of harbours/harbour systems, sounds, inlets and river mouths
- 2 – zero STP/TEV and low MCI, primarily lakes and lagoons
- 3 – high STP/TEV but overlapping with many other groups, primarily harbours, estuaries and inlets
- 4 – high SCI and MCI, primarily bays and harbours
- 5 – high R12/TEV, low SCI and MCI, but considerable overlap with other groups, primarily river mouths and harbour systems
- 6 – high R12/TEV, moderate SCI, but overlaps with other groups, primarily river mouths and harbours

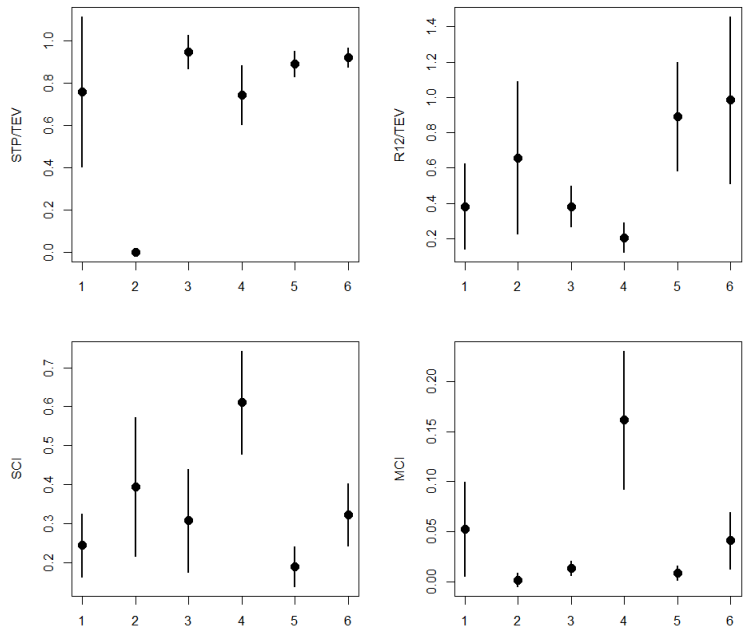

Figure 26. Means (±1 SD) of the four physical variables by group with group labels identified by model-based clustering (type VEV), with k = 6. STP/TEV and R12/TEV are fourth-root transformed, whilst SCI and MCI are untransformed.

Comparing these labels to the Hume classification (Table 22):

- 1 – mix of all classes, except A, but predominantly class F systems
- 2 – corresponds exactly to class A
- 3 – corresponds predominantly to a mix of E and F class systems
- 4 – corresponds to the majority of D, with some E and H systems
- 5 – corresponds to a mix of some of class B and F systems
- 6 – corresponds to a mix of some of class B and some E and F systems

Table 22. Table illustrating the number of systems in each group that were of the different hydro classes outlined in the Hume et al. (2007) classification. This is based on model-based clustering (type VEV), with k = 6.

Hume Hydro Class	Group					
	1	2	3	4	5	6
A	0	37	0	0	0	0
B	5	0	1	0	49	66
D	3	0	0	104	0	6
E	2	0	32	15	0	19
F	17	0	37	1	20	9
G	9	0	0	2	0	0
H	4	0	0	5	0	0

There is some agreement between this classification and the Hume classification, particularly regarding A and D classes. The majority of classes E and F were pooled into a single group characterised by high STP/TEV. Elsewhere, B class systems were split into two groups. The first of which, group 5, is characterised by high R12/TEV and low SCI and MCI, and groups a selection of group B with some of class F, and so these systems can be considered as structurally complex river mouths and harbours, whereas group 6 also has high R12/TEV but higher SCI, and thus these would be river mouths and partially open harbour systems with lower structural complexity. Group 4 corresponds almost entirely to class D, and therefore incorporates the majority of coastal embayments as well as some structurally simple, predominantly open harbours. Group 1 is difficult to interpret as it covers a range of physical variables, and has no clear defining feature.

2.4.4 – Grouping scheme for seven groups (VEV)

The seven group structure displays splits along STP/TEV, R12/TEV and SCI axes (Figures 25 and 26). Group 2, 4 and 7 are the most distinct, and are separated from the remaining groups along the STP/TEV and SCI/MCI axes (Figure 27). Groups can be characterised as (Figure 28):

- 1 – high STP/TEV, moderate MCI, mix of river mouths, estuaries and streams
- 2 – zero STP/TEV and low MCI, primarily lakes and lagoons
- 3 – high STP/TEV, low MCI, primarily harbours, harbour systems and inlets
- 4 – high SCI and MCI, low R12/TEV, primarily bays and harbours
- 5 – high R12/TEV, low SCI and MCI, primarily river mouths and harbours
- 6 – high R12/TEV, moderate SCI, but overlaps with other groups, primarily river mouths
- 7 – very low R12/TEV, moderate STP/TEV, primarily sounds and harbour systems

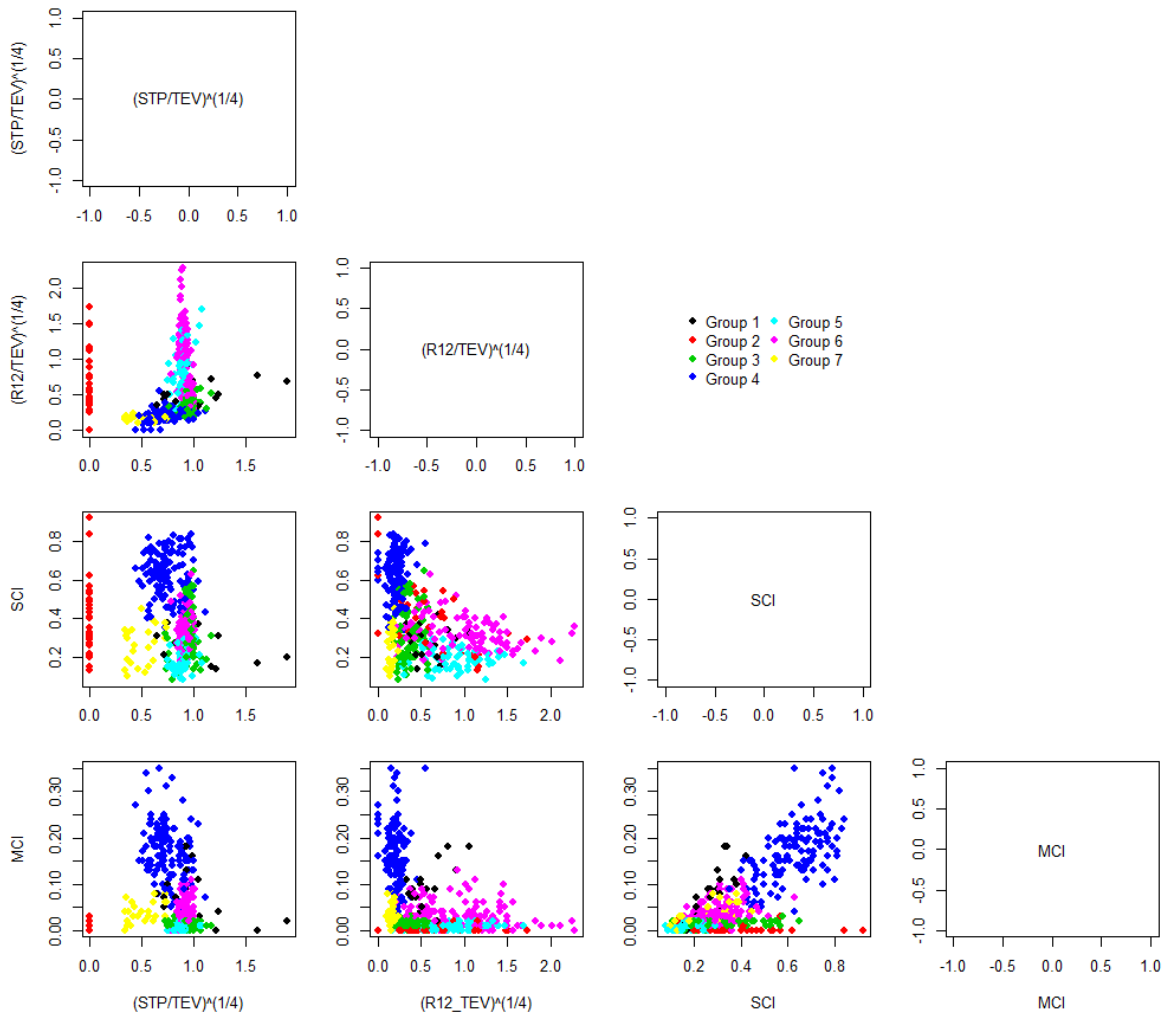

Figure 27. Biplots of the four physical variables colour-coded by group label as identified by model-based clustering (type VEV) with k = 7. STP/TEV and R12/TEV are fourth-root transformed, whilst SCI and MCI are untransformed.

Comparing these labels to the Hume classification (Table 23):

- 1 – mix of B-F class systems
- 2 – corresponds exactly to class A
- 3 – corresponds to a mix of E and F class systems
- 4 – corresponds to the majority of D, with some E and H systems
- 5 – corresponds to a mix of some of class B and F systems
- 6 – corresponds to a mix of some of class B and some E and F
- 7 – corresponds to a mix of classes E-H, but primarily G class systems

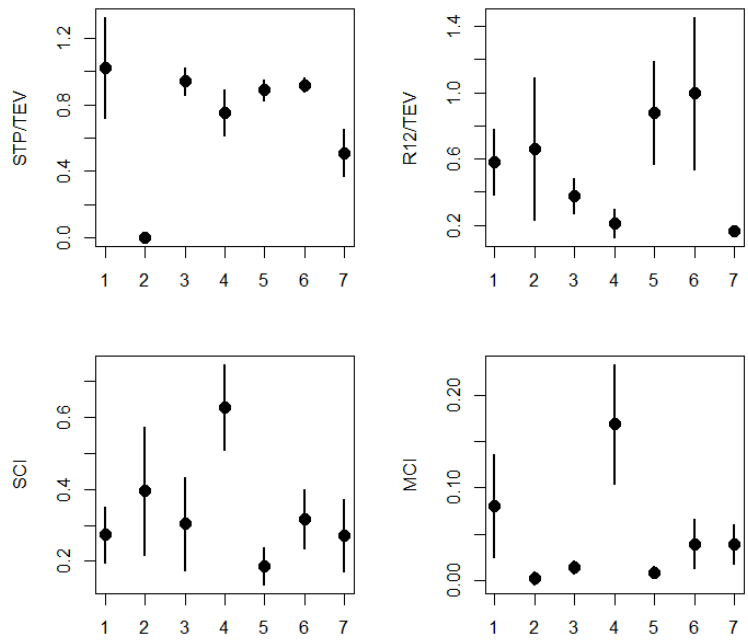

Figure 28. Means (±1 SD) of the four physical variables by group with group labels identified by model-based clustering (type VEV), with k = 7. STP/TEV and R12/TEV are fourth-root transformed, whilst SCI and MCI are untransformed.

Table 23. Table illustrating the number of systems in each group that were of the different hydro classes outlined in the Hume et al. (2007) classification. This is based on model-based clustering (type VEV), with k = 7.

Hume Hydro class	Group						
	1	2	3	4	5	6	7
A	0	37	0	0	0	0	0
B	6	0	0	0	44	71	0
D	5	0	0	103	0	4	1
E	1	0	32	10	0	19	6
F	8	0	42	0	20	10	4
G	0	0	0	1	0	0	10
H	0	0	0	5	0	0	4

This classification performs similarly to the six group labelling scheme, with groups 2-6 in the six group analysis closely matching groups 2-6 in the seven group analysis. The main distinction offered by the seven group analysis is between dynamic systems contained in group 1 that are characterised by higher STP/TEV and R12/TEV and were typically estuaries and river mouths, and static or less dynamic systems contained in group 7, characterised by lower STP/TEV, R12/TEV and MCI, and were typically larger bodies of water including sounds and harbour systems.

2.4.5 – Rule-based summary of grouping schemes

Six groups (VEV)

Evaluation of the ranges of variables for each group reveals that there is an obvious split of group 2 from the remaining data based on STP/TEV (Table 24). A reasonable rule to model this split would be

- STP/TEV < 0.0067
    - Y – Group 2
    - N – decision tree (Figure 29)

Table 24. Physical variable ranges for each group based on the six group scheme identified by model-based clustering.

Group	R12/TEV	STP/TEV	SCI	MCI
1	0.000, 1.238	0.013, 13.135	0.1, 0.42	0, 0.18
2	0.000, 8.968	0.000, 0.000	0.13, 0.92	0, 0.03
3	0.001, 0.267	0.281, 1.921	0.08, 0.65	0, 0.03
4	0.000, 0.090	0.039, 1.558	0.28, 0.84	0.03, 0.35
5	0.001, 8.253	0.320, 1.357	0.08, 0.30	0, 0.02
6	0.002, 27.130	0.373, 1.212	0.16, 0.63	0, 0.13

The resulting rule set determined via classification tree analysis performs less well than previous rule sets, achieving only 77% classification accuracy and performs particularly poorly for group 1 and 6 (Table 25). Splits are arranged along all four axes, with two along STP/TEV, and one each for R12/TEV, SCI and MCI (Figure 29). The rule set indicates that groups are characterised by the following rules (Figure 29):

(1) STP/TEV ≥ 0.0067, SCI < 0.445, MCI ≥ 0.025, STP/TEV < 0.5432

(2) STP/TEV < 0.0067

(3) STP/TEV ≥ 0.0067, SCI < 0.445, MCI < 0.025, R12/TEV < 0.1233

(4) STP/TEV ≥ 0.0067, SCI ≥ 0.445

(5) STP/TEV ≥ 0.0067, SCI < 0.445, MCI < 0.025, R12/TEV ≥ 0.1233

(6) STP/TEV ≥ 0.0067, SCI < 0.445, MCI ≥ 0.025, STP/TEV ≥ 0.5432

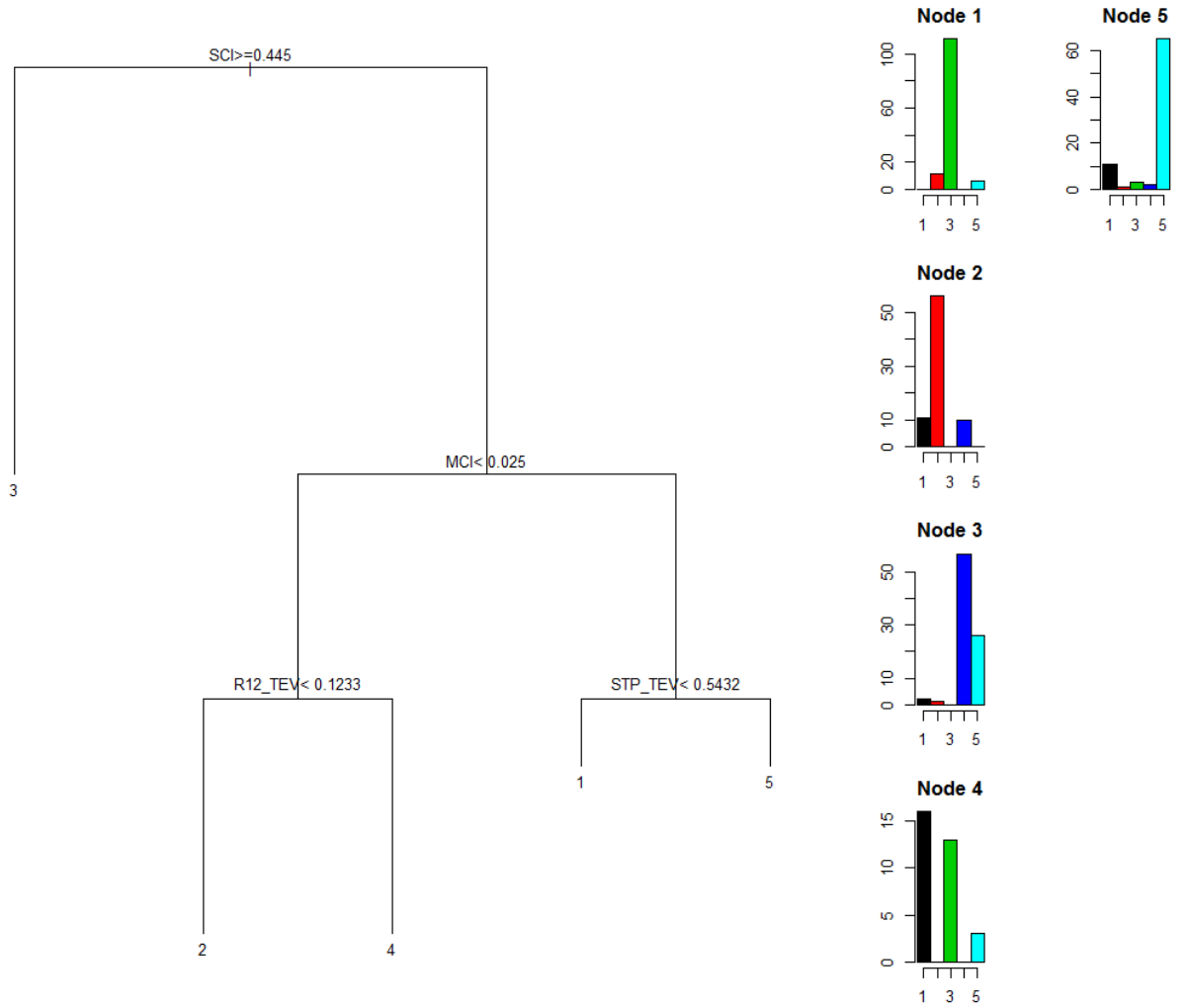

Figure 29. Classification tree for the six group scheme identified using model-based clustering. Group 2 has been removed from the analysis and groups 3-6 are labelled as 2-5 in the classification tree, respectively. Histograms to the right of the classification tree indicate the class make-up of each terminal node, where node numbering proceeds from left to right along the base of the classification tree.

Table 25. Classification accuracy of the classification tree developed for the six group scheme identified using model-based clustering.

Group	Classification accuracy (%)	No. in Group
1	40.0	40
2	100.0	37
3	80.0	70
4	87.4	127
5	82.6	69
6	65.0	100
Total	77.2	443

Seven groups (VEV)

Evaluation of the ranges of variables for each group reveals that there is an obvious split of group 2 from the remaining data based on STP/TEV (Table 26). A reasonable rule to model this split would be

- STP/TEV < 0.0067
    - Y – Group 2
    - N – decision tree (Figure 30)

Table 26. Physical variable ranges for each group based on the seven group scheme identified using model-based clustering.

Group	R12/TEV	STP/TEV	SCI	MCI
1	0.013, 1.238	0.18, 13.135	0.14, 0.42	0, 0.18
2	0, 8.968	0, 0	0.13, 0.92	0, 0.03
3	0.001, 0.123	0.281, 1.877	0.08, 0.65	0, 0.03
4	0, 0.09	0.039, 1.558	0.34, 0.84	0.04, 0.35
5	0.001, 8.253	0.32, 1.357	0.08, 0.3	0, 0.02
6	0.002, 27.13	0.373, 1.047	0.16, 0.63	0, 0.13
7	0, 0.003	0.013, 0.295	0.1, 0.45	0, 0.08

However, when performing the classification tree analysis on the remaining data, no rule-set containing a single terminal node for each group could be identified, likely due to the small group size of groups 1 and 7. Therefore, the entire dataset (including group 2) was evaluated using a classification tree analysis and a tree with nine terminal nodes was identified (Figure 30). The classification tree had reasonably high accuracy (85%) (Table 27), but contained more than a single rule set for group 4 and group 6, which each had two terminal nodes (Figure 30). In addition group 1 is not particularly well classified by this tree, with a classification accuracy of 35%. Nodes 5 and 6 shown in Figure 30 can be collapsed to form a single node corresponding to group 6, and wouldn't significantly reduce the accuracy of classifying group 4 (from 95 to 85%) and would create a simpler rule set. The resulting rules are (Figure 30):

(1) MCI ≥ 0.095, SCI < 0.43

(2) MCI < 0.095, STP/TEV < 0.0067

(3) MCI < 0.095, STP/TEV ≥ 0.0067, R12/TEV < 0.1233, MCI < 0.025

(4) MCI ≥ 0.095, SCI ≥ 0.43

(5) MCI < 0.095, STP/TEV ≥ 0.0067, R12/TEV ≥ 0.1233, SCI < 0.225

(6) 2 rules

    a. Primary: MCI < 0.095, STP/TEV ≥ 0.0067, R12/TEV ≥ 0.1233, SCI ≥ 0.225

b. Secondary: MCI < 0.095, STP/TEV ≥ 0.0067, R12/TEV < 0.1233, MCI ≥ 0.025, STP/TEV ≥ 0.3044

(7) MCI < 0.095, STP/TEV ≥ 0.0067, R12/TEV < 0.1233, MCI ≥ 0.025, STP/TEV < 0.3044

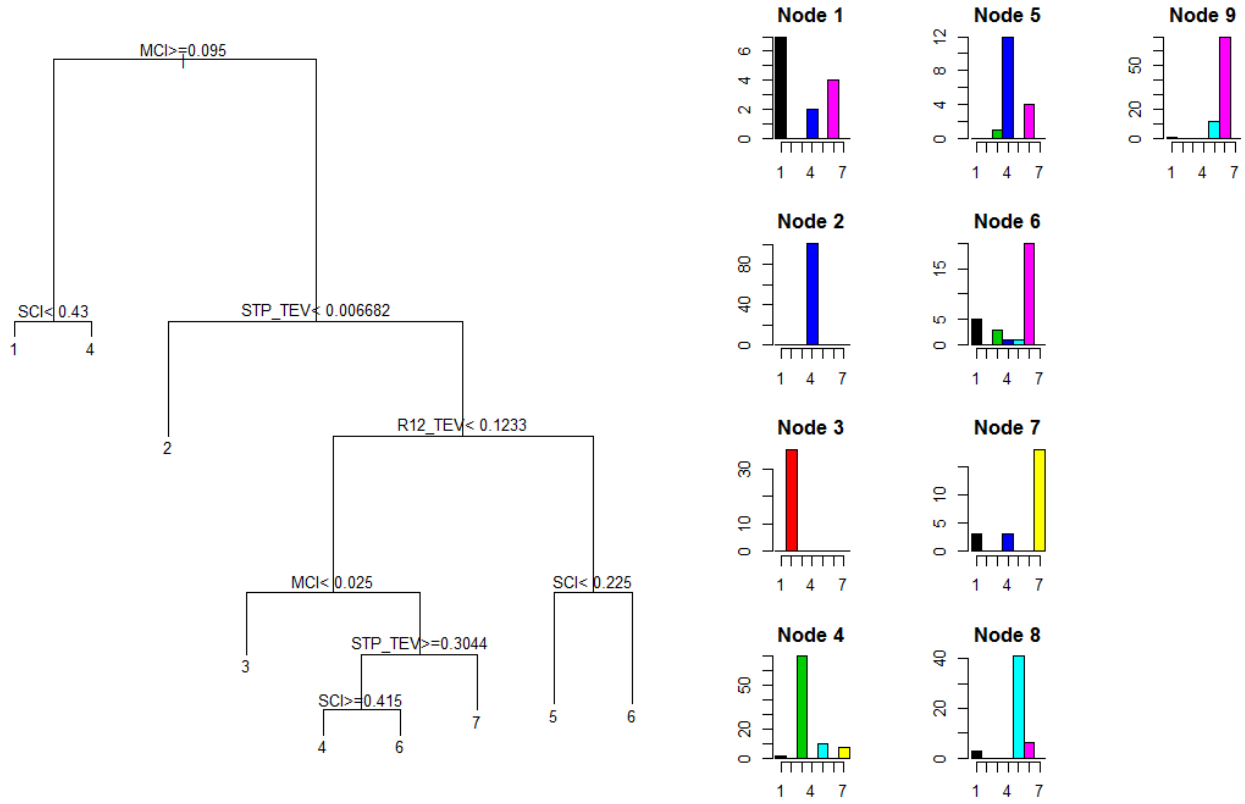

Figure 30. Classification tree for the seven group scheme identified using model-based clustering. Histograms to the right of the classification tree indicate the class make-up of each terminal node, where node numbering proceeds from left to right along the base of the classification tree.

Table 27. Classification accuracy of the classification tree developed for the seven group scheme identified using model-based clustering. The classification accuracy after merging nodes 5 and 6 is also given.

Group	Classification accuracy (%)	Classification accuracy after merge (%)	No. in Group
1	35.0	35.0	20
2	100.0	100.0	37
3	94.6	94.6	74
4	95.0	84.9	119
5	64.1	64.1	64
6	86.5	90.4	104
7	72.0	72.0	25
Total	84.9	83.1	443

2.4.6 – Summary, advantages and disadvantages

In summary model-based clustering produces groupings that illustrate meaningful differences among hydrosystems. The analysis successfully differentiated bays (Hume class D), lakes and lagoons (Hume class A), and combined large complex hydrosystems, such as harbour systems, sounds and fiords into a single group, accounting for classes G and H. However, a large proportion of classes B, E and F were split among a series of different groups with different characteristics. For example class B, which is predominantly river mouth's, was split into two groups, one which was aligned with a portion of class F and another group aligned predominantly with class E. Similarly to the previous analyses, classes E and F were not isolated from one another, but systems within these classes were assigned to multiple groups indicating that there are differences within E and F that lead to alternate groupings that do not match the split defined to separate class E systems from class F. In addition one group (group 1 in the seven group analysis) had no distinguishing characteristics and didn't match any particular hydro class, making it difficult to interpret.

The advantages of this method are:

- The "best" number of groups is selected as part of a robust data driven process
- The ratio of BIC statistics between any two models can be used as a measure of support for one grouping structure over another
- Probabilities of membership within particular groups can be identified from the model structure, giving strength of group membership, and identifying those systems that are of intermediate typology (see Appendix E)
- Model types can be used to control what type of clusters are developed. For example model types with clusters whose major axes are restricted to be parallel or perpendicular to component axes can be chosen, which would aid in identifying appropriate rule sets
- A wide variety of cluster types are available
- Can account for clusters of any size, including clusters with only a few observations within them if these are present within the data
- Methods exist to account for outliers and noise (i.e. those systems that exist far from any other observation in multivariate space)

The disadvantages are

- Clusters are restricted to be one of the supported types

- Sensitive to transformations and scaling of data (see Appendix D), and can be influenced by outliers
- Many of the more complicated cluster types failed to fit the data, and so group structures must be selected from those that did fit
- Makes more assumptions about the data, such as multivariate normality of observations about the cluster mean, which may not be appropriate to the data

2.5 – Conclusion

In summary each method has its merits, but random forests and model-based clustering address two of the major drawbacks of the Euclidean hierarchical clustering analysis, that of identifying the correct way to pre-treat the data, and the identification of the "best" number of groups. Fewer groups were chosen using the cross-validation method for the Euclidean cluster analyses than for the random forests or model-based clustering methods. However, the methods used to identify the best number of groups are not equivalent, and for the Euclidean and random forests analyses the number of groups chosen was an educated guess, despite efforts to identify an appropriate number of groups using cross-validation. This further highlights the advantages of the model-based system as it removes this uncertainty. The other major sources of uncertainty that affect the Euclidean hierarchical and the model-based approaches are the pre-treatment and scaling of the physical variables, whilst choice of linkage method affects Euclidean and random forests hierarchical analyses. As illustrated, both of these can greatly influence the grouping structure that results from these analyses. The susceptibility of each method to all these sources of uncertainty are summarised in Table 28. If a particular number of groups is desired then the random forests method has the advantage as only linkage type can alter group structure. If, however, the particular number of groups is unknown and the variables are similar in terms of variance and scale then model-based clustering holds the advantage.

Table 28. Major sources of uncertainty and analysis decisions that can affect group structure summarised by analysis type.

Cluster analysis type	Source of uncertainty/Analysis decision			
	Transformation of data	Scale data	Linkage type	Number of Groups
Euclidean hierarchical	Y	Y	Y	Y
Random forests hierarchical	N	N	Y	Y
Model-based	Y	Y	N	N

With regards to the Hume classification system all three methods performed similarly. The Euclidean analysis successfully separated bays and river mouths from the remaining systems, random forests separated bays, river mouths and lakes and lagoons from the other classes and the model-based analysis separated bays, lakes and lagoons, and large complex systems, such as fiords, sounds and harbour systems from the remaining classes. All of the systems that were distinguished in each analysis exist at the extremes of the physical variable space (e.g. lakes and lagoons have zero STP/TEV, and river mouths have high R12/TEV) and represent either completely open (bays, river mouths) or completely closed (lakes and lagoons) systems. However, systems that are intermediate in state were less successfully defined, perhaps due to there being a continuum of states among these classes, with no particular split or partition between them. In some analyses these intermediate systems were grouped with bays and/or river mouths based on similar R12/TEV and/or SCI among these systems. This suggests that perhaps an alternate means of classifying these systems might be more appropriate.

Examining the rule-based systems that were developed to approximate each grouping scheme, the splits occur at different locations in multivariate space for different analysis methods, but there are a few similarities (Table 29). For STP/TEV splits at 0.0067 were present in nearly all analyses, which closely matches the criterion used in the Hume classification (STP/TEV=0), whereas additional splits at ~ 0.3-0.55 were present in all random forests and model-based analyses (Table 29). For R12/TEV splits at ~ 1-1.3 were present in all hierarchical analyses (Table 29). A prominent split for MCI at 0.095-0.125 was present in all but one of the analyses, which is close to the value of 0.075 in the Hume classification which was used to distinguish between classes D and E, whilst no splits for SCI were replicated across methods.

The locations of these new split locations could be used to further develop the rules based system in line with the classifications determined via these statistical methods.

Table 29. Locations of classification tree partitions along the four physical variable axes for each of the analysis methods and group numbers investigated. Values in parentheses indicate the location in the classification tree where the rule was implemented. For Euclidean and model-based analyses only those splits identified from the transformed and scaled data are given.

Cluster analysis type	Split location			
	STP/TEV	R12/TEV	SCI	MCI
Euclidean hierarchical 4 groups	0.0067 (1)	1.235 (3)		0.105 (2)
Euclidean hierarchical 5 groups	0.0067 (1)	1.268 (3) 0.00525 (4)		0.115 (2)
Random Forests 6 groups	0.5224 (1)	0.955 (2) 0.2976 (4)		0.125 (2) 0.095 (3)
Random Forests 7 groups	0.541 (1) 0.0067 (3)	1.061 (2) 0.2976 (4)		0.125 (2) 0.095 (3)
Model-based 6 groups	0.0067 (1) 0.5432 (4)	0.1233 (4)	0.445 (2)	0.025 (3)
Model-based 7 groups	0.0067 (2) 0.3044 (5)	0.1233 (3)	0.43 (2) 0.225 (4)	0.095 (1) 0.025 (4)

3.0 – Incorporating additional physical variables to those used in Hume et al. (2007)

3.1 – Introduction

This section uses the methods described in section 2 and expands the analyses to include several additional physical attributes of the 443 NZ hydrosystems. In addition to the variables used in the previous analysis (R12/TEV, STP/TEV, SCI and MCI) the following variables are incorporated into the analyses performed in this section:

- Percent intertidal area (% IA) – the area of the system that is totally covered at high tide and uncovered at low tide expressed as a percentage of the entire system surface area
- Mean depth – the total estuary volume at high tide divided by the estuary area at high water spring tide
- CLA/EWA – the ratio of catchment land area (CLA) to the estuary water area at high water spring tide (EWA)

The aim of this section is to identify what additional groups, grouping structures and partitions in variable space are created when these additional variables are analysed using the three methods detailed in the previous section. Identified group schemes are also compared to the Hume classification to examine whether these additional variables increase the ability of the statistical methods to distinguish between the hydroclasses identified in Hume et al. (2007), and to examine whether these classes are split into two or more groups based on the new information provided by these variables.

3.2 – Method 1: Hierarchical clustering based on Euclidean distance measures

In this section Hierarchical clustering, based on Euclidean distance measures and Ward's linkage criterion, is applied to the dataset of R12/TEV, STP/TEV, SCI, MCI, % IA, mean depth and CLA/EWA for the 443 NZ coastal hydrosystems and the resulting grouping schemes are discussed and analysed further.

3.2.1 – Initial treatment of data

The distribution of % IA, mean depth and CLA/EWA were examined (Figure 31) and the means, standard deviations and ranges calculated (Table 30).

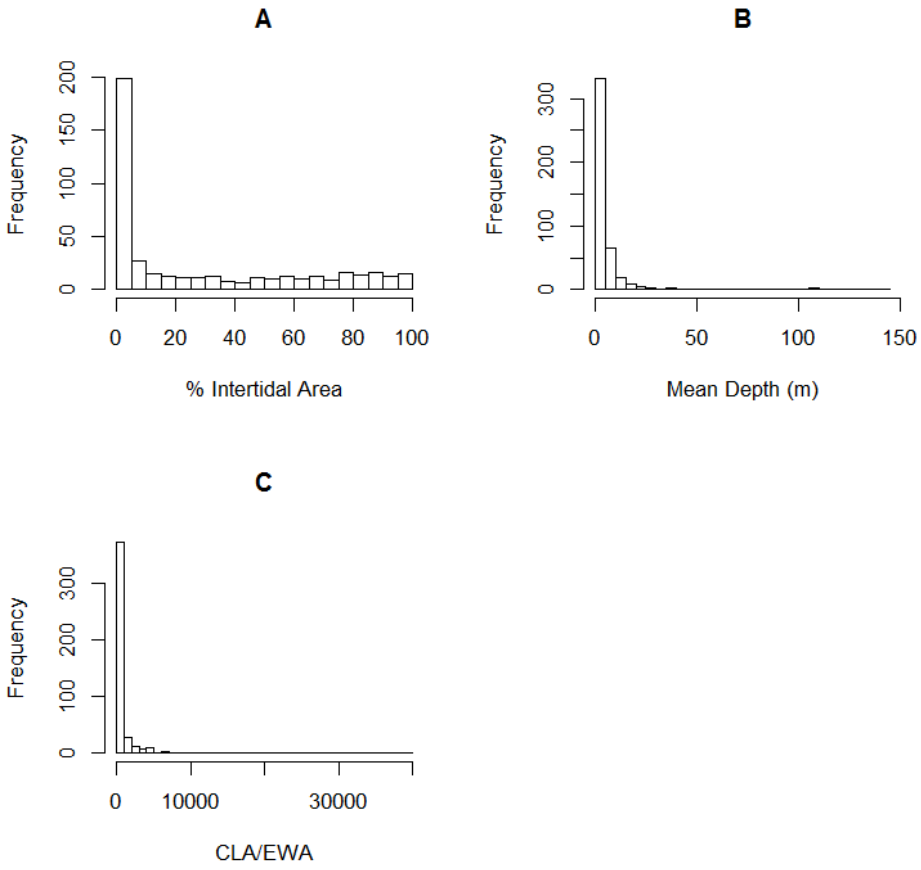

Figure 31. Histograms displaying the distributions of (A) percentage intertidal area, (B) mean depth and (C) CLA/EWA.

Table 30. Summary statistics of untransformed, transformed and scaled physical variables. N/A indicates these variables weren't transformed.

Variable	Untransformed		Transformed		Transformed and Scaled	
	Mean (SD)	Range	Mean (SD)	Range	Mean (SD)	Range
STP/TEV	0.59 (0.76)	0 – 13.14	0.78 (0.28)	0 – 1.90	0 (1)	-2.74 – 3.96
R12/TEV	0.83 (2.57)	0 – 27.13	0.57 (0.44)	0 – 2.28	0 (1)	-1.31 – 3.92
SCI	0.38 (0.19)	0.08 – 0.92	N/A	N/A	0 (1)	-1.57 – 2.81
MCI	0.06 (0.08)	0 – 0.35	N/A	N/A	0 (1)	-0.84 – 3.76
% IA	28.9 (33.7)	0 – 100	-2.31 (3.48)	-7.6 – 7.6	0 (1)	-1.52 – 2.85
Mean depth	6.7 (16.2)	0.1 – 141.1	1.39 (0.40)	0.56 – 3.45	0 (1)	-2.07 – 5.13
CLA/EWA	839 (2945)	0 – 39156.9	4.00 (2.40)	0 – 10.58	0 (1)	-1.67 – 2.74

All three variables display considerable left skew, and mean depth and CLA/EWA contain many extreme values that are likely to unduly influence the clustering of the data (Figure 31). In addition there is a large disparity between the variances of these variables (standard deviations vary by 4 orders of magnitude between variables, Table 30). Therefore each variable was transformed appropriately (% IA – logistic transform, mean depth – fourth root, CLA/EWA – log (X+1), see Appendix A) and scaled so that variances were equal across all physical variables (Table 30). Based on the previous analyses R12/TEV and STP/TEV were fourth-root transformed, whilst MCI and SCI were untransformed such that results would be comparable with those in Section 2. Biplots of the transformed and scaled variables revealed there is a close correlation between R12/TEV and CLA/EWA and between STP/TEV and mean depth (Figure 32).

The transformed and scaled dataset was then used to calculate a distance matrix comprised of the Euclidean distances between each system and every other system. These distance matrices were subsequently used to construct a hierarchical cluster scheme (using the **hclust** function in R), based on Ward's linkage criterion. As the difference between scaled and unscaled analyses was illustrated in the previous section, only analyses based on the scaled dataset were performed.

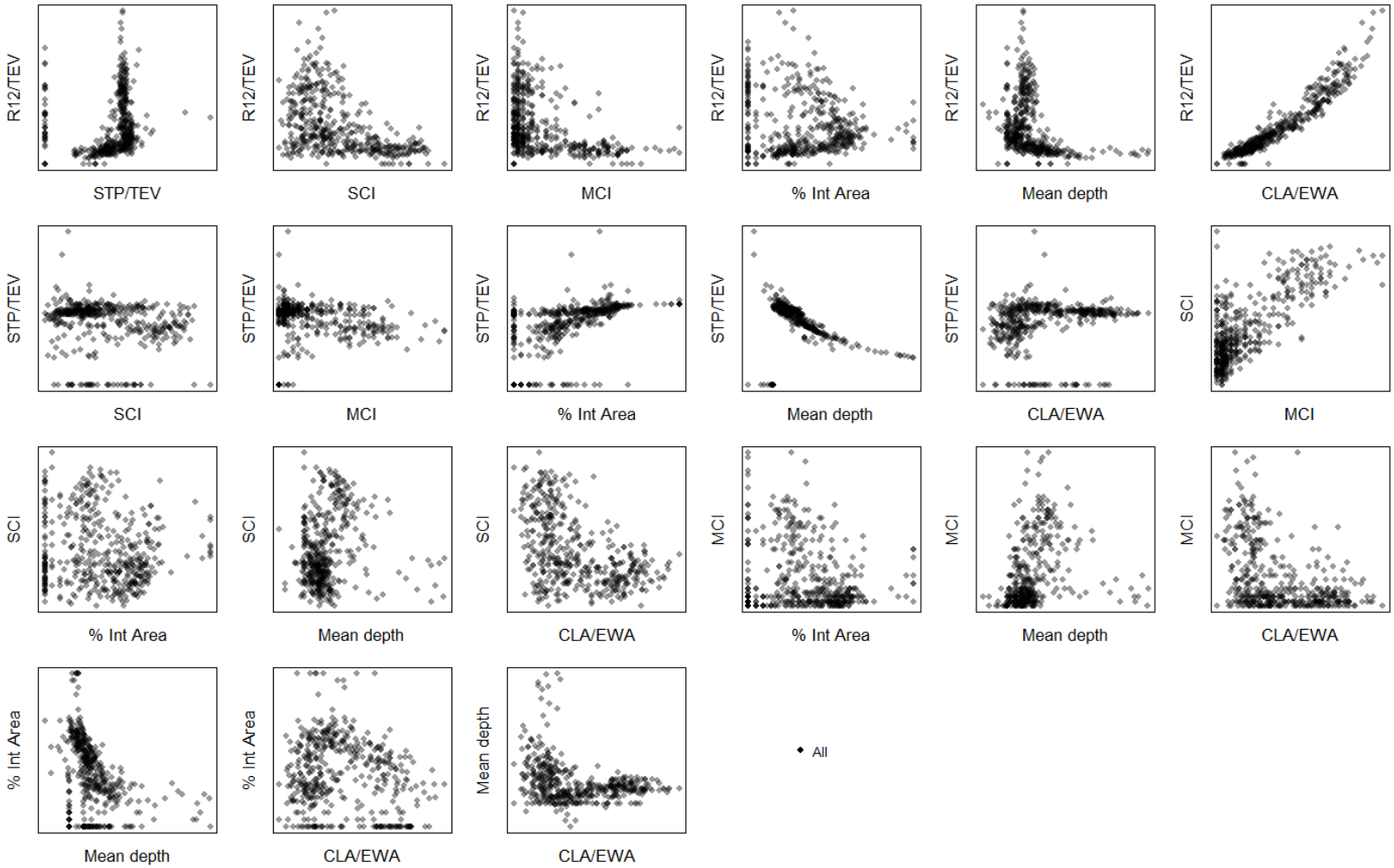

Figure 32. Biplots of all possible combinations of the seven physical variables, R12/TEV, STP/TEV, SCI, MCI, % IA, mean depth and CLA/EWA. All variables are transformed and scaled as detailed in section 3.2.1.

3.2.2 – Deciding on the "best" number of groups

The cross-validation routine described in section 2.2.2 was applied to this dataset to decide on the most appropriate number of groups. Cross validation was based on leaving a random subset of 40 observations (test dataset) out of the hierarchical cluster analysis and determining the classification accuracy of the resulting group structure for the left-out data. This was carried out for 2 – 10 groups (Figure 33).

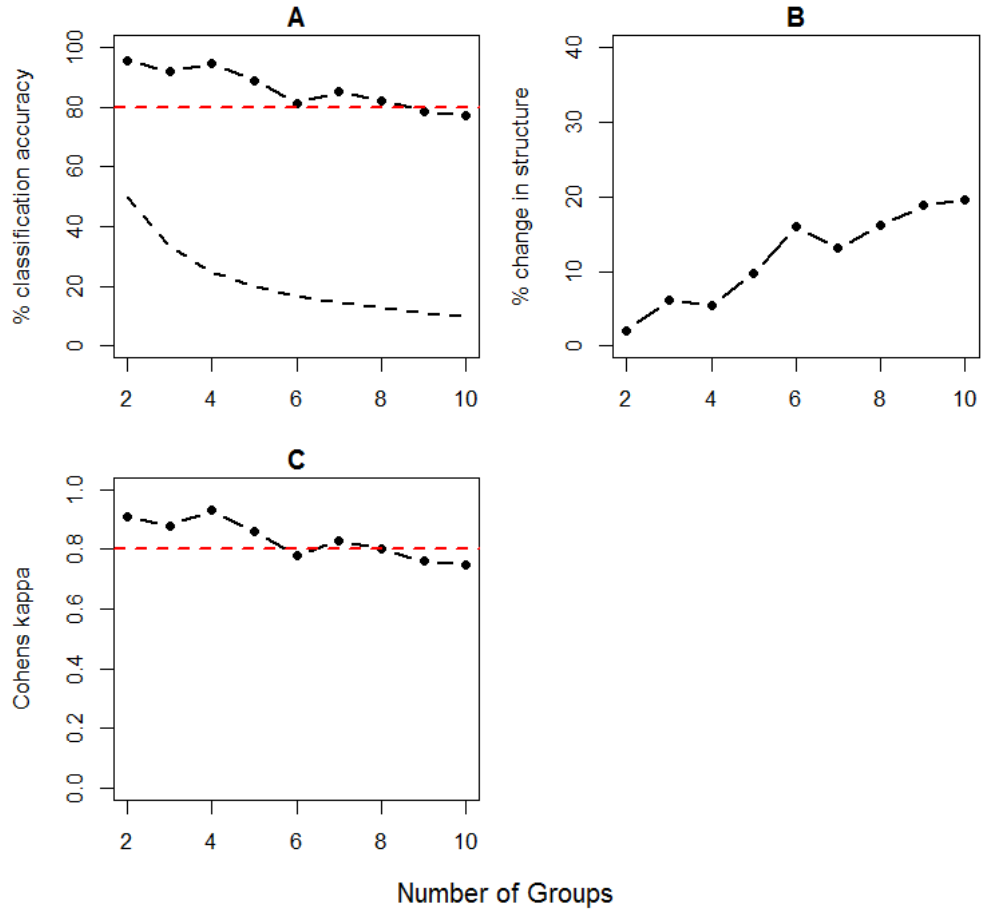

Figure 33. Cross validation metrics of (A) % classification accuracy, (B) % change in group structure and (C) Cohens-κ plotted against number of groups, with groups identified by Euclidean hierarchical clustering. Red dotted lines indicate (A) 80% and (C) 0.8, and the black dotted line in (A) corresponds to the classification accuracy that would be expected by chance.

The classification accuracy drops from 82.3 for eight groups to 78.6% for nine groups, whereas Cohen's kappa drops from 0.8 to 0.76 (Figure 33). Therefore eight groups seems to be the most appropriate, but nine groups will also be investigated.

3.2.3 – Grouping scheme for eight groups

Performing the hierarchical cluster analysis with eight groups resulted in the group structure illustrated in Figure 34. Several groups are distinct along particular variable axes. Group 2 is visually separate along the STP/TEV axis, group 4 along the SCI/MCI axes, group 5 along the % IA axis, groups 6 and 7 along the R12/TEV and CLA/EWA axes and group 8 along the mean depth axis (Figure 34). Examining group attributes (Figure 35) and systems contained in each group, groups can be characterised as:

- 1 – low R12/TEV and moderate MCI, but considerable overlap with other groups, primarily bays and harbours
- 2 – zero STP/TEV and MCI, primarily lakes and lagoons
- 3 – high STP/TEV and high % IA, primarily harbours and harbour systems
- 4 – highest SCI and MCI, primarily bays
- 5 – highest intertidal area, primarily inlets, creeks and streams
- 6 – high R12/TEV, moderate % IA, primarily river mouths
- 7 – high R12/TEV, low % IA, primarily river mouths
- 8 – greatest mean depth, moderate STP/TEV, predominantly fiords and sounds

Comparing these labels to the Hume classification (Table 31):

- 1 – mix of D, E and some of class F systems
- 2 – corresponds exactly to class A
- 3 – contains the vast majority of F and approximately half of the E class systems
- 4 – contains the majority of D, and half of H class systems
- 5 – corresponds strongly to a selection of class E systems
- 6 – corresponds to over half of the class B systems
- 7 – corresponds to just under half of the class B systems
- 8 – corresponds to the majority of G and some H class systems

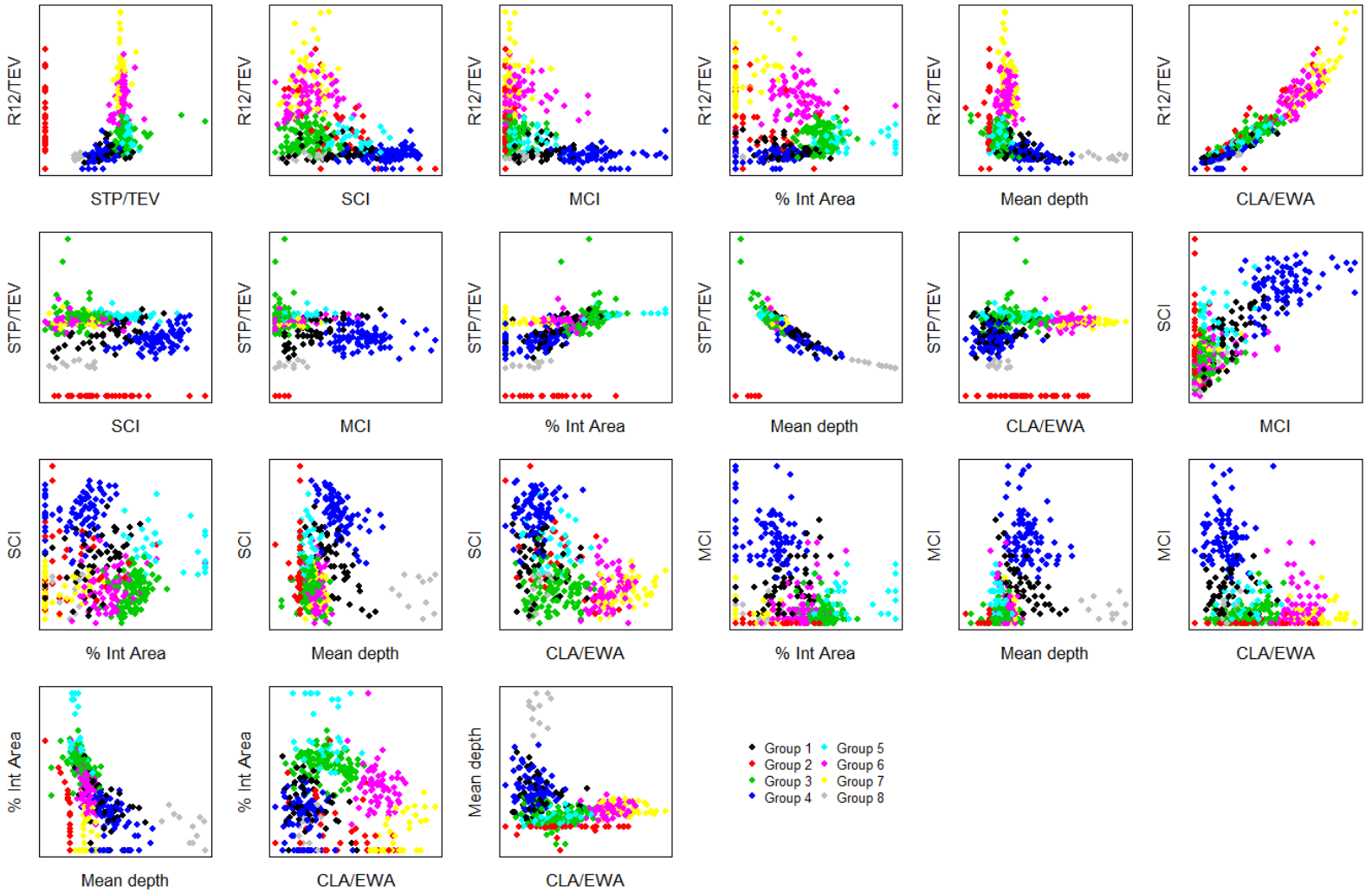

Figure 34. Biplots of all possible combinations of the seven physical variables, R12/TEV, STP/TEV, SCI, MCI, % IA, mean depth and CLA/EWA colour coded by group label identified by Euclidean hierarchical clustering with k = 8. All variables are transformed and scaled as detailed in section 3.2.1.

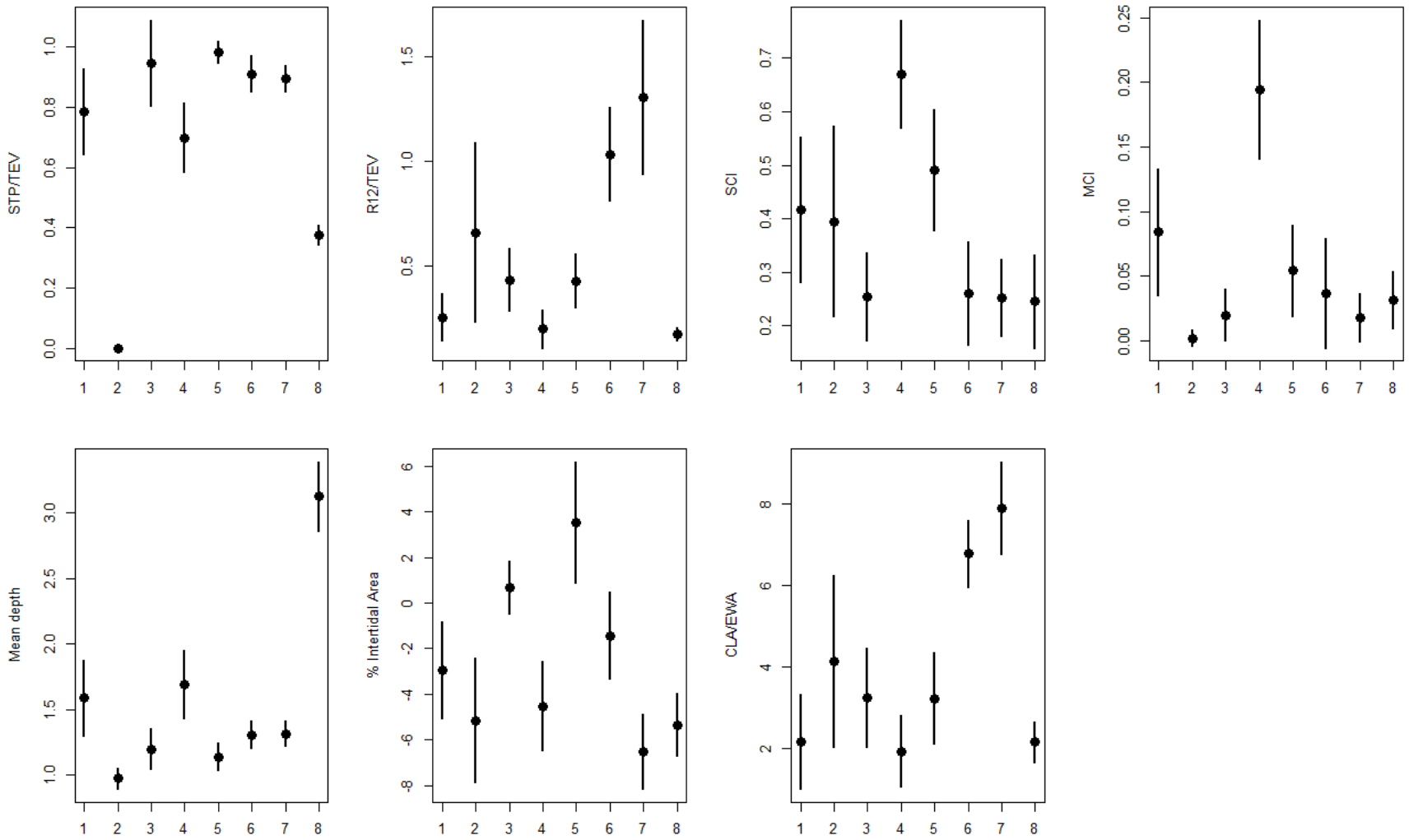

Figure 35. Means (±1 SD) of each of the seven physical variables for each group, based on the group structure identified by Euclidean hierarchical clustering, with k = 8. All representations are based on the transformed variables as detailed in section 3.2.1.

Table 31. Table illustrating the number of systems in each group that were of the different hydro classes outlined in the Hume et al. (2007) classification. This is based on Euclidean hierarchical clustering, with $k = 8$.

Hume Hydro Class	Group							
	1	2	3	4	5	6	7	8
A	0	37	0	0	0	0	0	0
B	0	0	4	0	1	66	50	0
D	27	0	1	78	7	0	0	0
E	15	0	30	0	23	0	0	0
F	7	0	74	0	0	3	0	0
G	1	0	0	1	0	0	0	9
H	2	0	0	5	0	0	0	2

There is considerable agreement between this grouping scheme and the Hume classification and is particularly strong for classes A, B (which is split into two groups), F and G. As identified from the previous analyses classes E and F, as well as some of D, are often assigned to the same group, suggesting that this method does not prioritise the difference between E and F and selects alternate groups. Examining the group attributes in greater detail there appears to be a gradient across groupings, moving from primarily D class systems (group 4), to a mix of D, E and F (group 1), followed by a group which is predominantly E, with some D (group 5) and finally a group which is primarily F with some E (group 3). Moving along this gradient (4 – 1 – 5 – 3) STP/TEV marginally increases, whilst MCI decreases, suggesting that these classes represent a gradient in harbours (or partially closed bodies) from fully open (bays), to almost closed with very narrow mouth relative to system size (e.g. typical of harbour systems). As with previous analyses class A is distinct due to having zero STP/TEV and river mouths (class B) are distinct from the other groups, but are split into two distinct groups. The main difference between groups 6 and 7 is % intertidal area, which is higher for systems in group 6 than for those in group 7 and R12/TEV which is higher for group 7 than group 6. Other classes were split into multiple groups. Class D was split into two groups, one with moderate SCI and MCI (group 1 – bays with moderately complex shorelines and narrower entrance, characteristic of recessed bays), whilst the other had high SCI and MCI (group 4 – bays with low structural complexity and are completely open, characteristic of an arcuate or circular shoreline). Class E was split among three groups, with the groups 1, 3, 5 displaying a gradient in intertidal area, STP/TEV and R12/TEV (low-high) and depth (high-low), representing a gradient from predominantly subtidal harbours, to harbours with large tidal flats. Finally, with the addition

of depth as a predictor variable, fiords and sounds (class G) were successfully differentiated from the remaining systems.

3.2.4 – Grouping scheme for nine groups

The group structure for nine groups is the same as eight group with the exception that the previous group 3 is split into two groups, the current group 3 and 5, which are primarily separated along the R12/TEV and CLA/EWA axes (Figure 36). Examining group attributes (Figure 37) and systems contained in each group, groups can be characterised as:

- 1 – low R12/TEV and moderate MCI, but considerable overlap with other groups, primarily bays and harbours
- 2 – zero STP/TEV and MCI, primarily lakes and lagoons
- 3 – High STP/TEV and high % IA, primarily harbour systems
- 4 – Highest SCI and MCI, primarily bays
- 5 – High STP/TEV and high % IA, primarily harbours
- 6 – Highest intertidal area, primarily inlets, creeks and streams
- 7 – High R12/TEV, moderate % IA, primarily river mouths
- 8 – High R12/TEV, low % IA, primarily river mouths
- 9 – Greatest mean depth, moderate STP/TEV, predominantly fiords and sounds

Comparing these labels to the Hume classification (Table 32):

- 1 – mix of D, E and some of class F systems
- 2 – corresponds exactly to class A
- 3 – consists of a mix of E and F class systems
- 4 – contains the majority of D, and half of H class systems
- 5 – consists predominantly of F class systems, with some E
- 6 – corresponds strongly to a selection of class E systems
- 7 – corresponds to over half of the class B systems
- 8 – corresponds to just under half of the class B systems
- 9 – corresponds to the majority of G and some of H

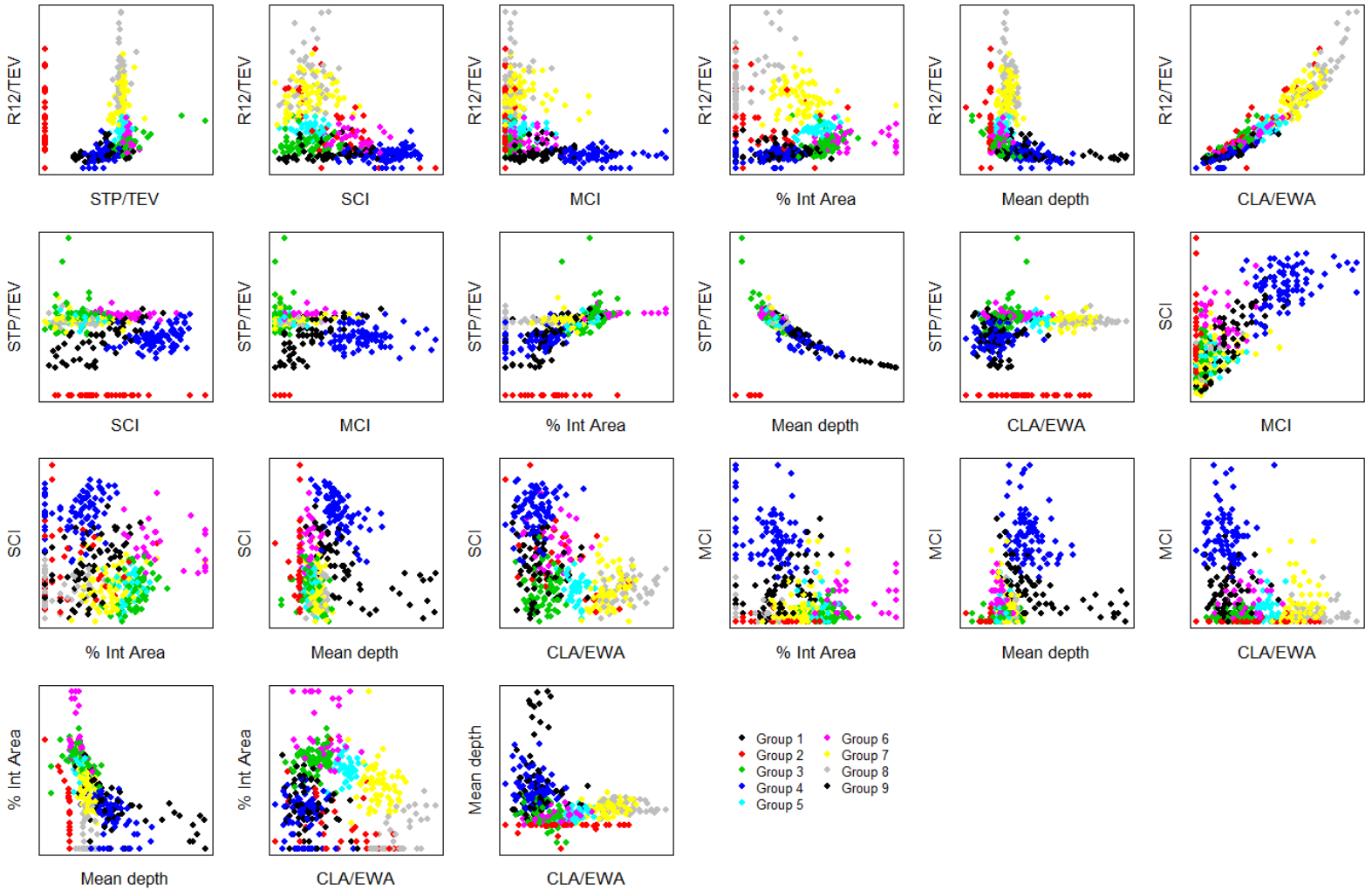

Figure 36. Biplots of all possible combinations of the seven physical variables, R12/TEV, STP/TEV, SCI, MCI, % IA, mean depth and CLA/EWA colour coded by group label identified by Euclidean hierarchical clustering with k = 9. All variables are transformed and scaled as detailed in section 3.2.1.

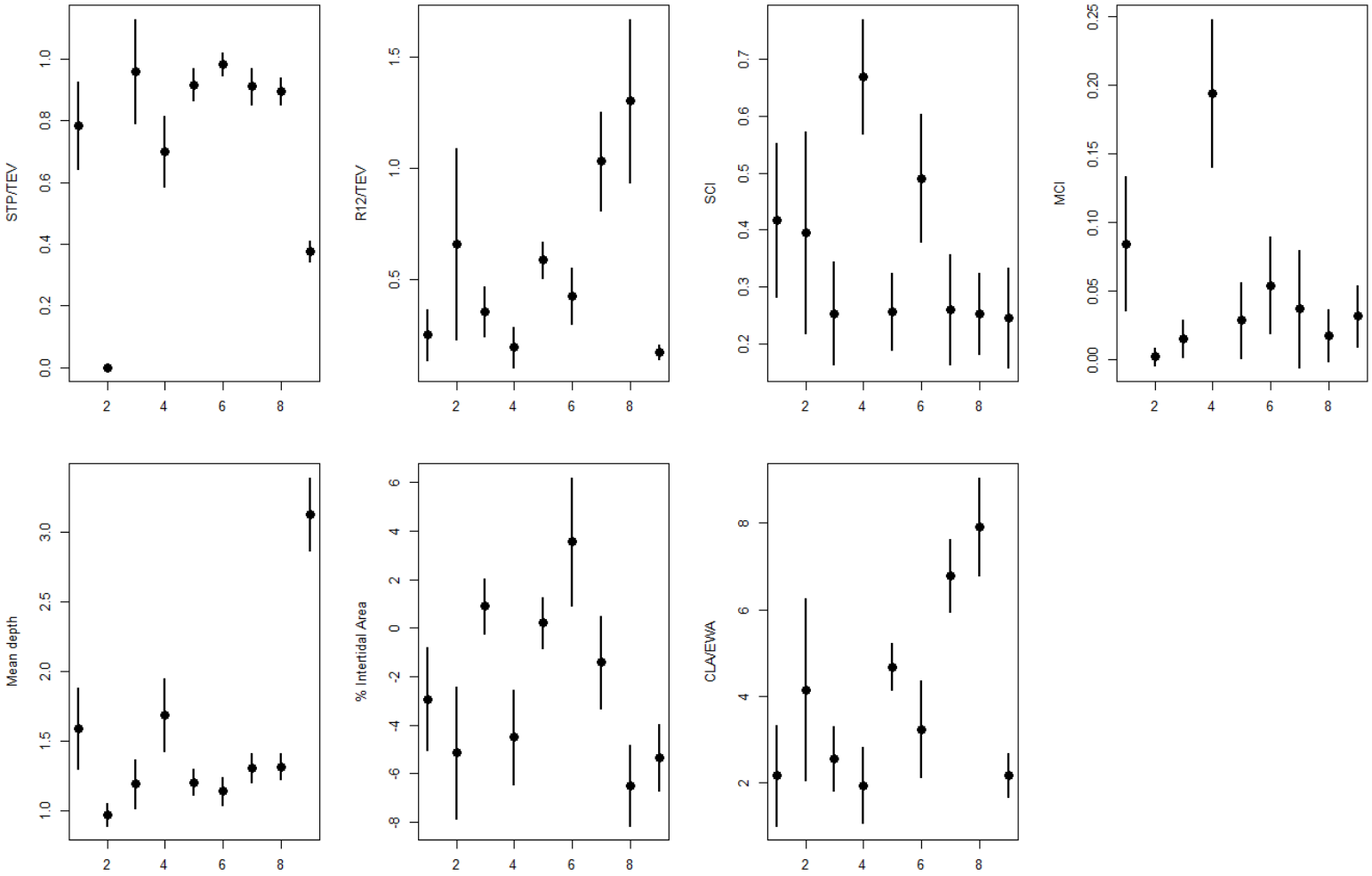

Figure 37. Means (±1 SD) of each of the seven physical variables for each group, based on the group structure identified by Euclidean hierarchical clustering with $k = 9$. All representations are based on the transformed variables as detailed in section 3.2.1.

Table 32. Table illustrating the number of systems in each group that were of the different hydro classes outlined in the Hume et al. (2007) classification. This is based on Euclidean hierarchical clustering, with $k = 9$.

Hume Hydro class	Group								
	1	2	3	4	5	6	7	8	9
A	0	37	0	0	0	0	0	0	0
B	0	0	1	0	3	1	66	50	0
D	27	0	0	78	1	7	0	0	0
E	15	0	24	0	6	23	0	0	0
F	7	0	48	0	26	0	3	0	0
G	1	0	0	1	0	0	0	0	9
H	2	0	0	5	0	0	0	0	2

This group structure builds upon the eight group structure adding an additional level of distinction for systems of classes E and F. In the eight group analysis a gradient moving from fully open (bays), to systems with narrow mouths relative to system size (e.g. typical of harbour systems) was identified. In the nine group analysis this is expanded upon with groups 4 – 1 – 6 – 3 – 5 exhibiting a continuous gradient from class D to F, and from high MCI to low MCI. Additional variables change over this gradient, including SCI (generally decreasing), mean depth (decreasing), R12/TEV (generally increasing), STP/TEV (generally increasing) and % IA (increases up to group 6 and decreases for the remaining groups). These relationships further strengthen the interpretation of these groups as a gradient in partially closed system types from largely open bays and harbours, to large, semi-closed harbour systems. This group scheme can essentially be simplified to a five class system, with sub-types in particular groups:

1. Fully closed, little to no tidal influence - Group 2 (or class A)
2. River mouths or river dominated systems
    a. with moderate-high intertidal area, lower river input – Group 7
    b. with little to no intertidal area, higher river input – Group 8
3. Fiords and sounds characterised by great depth - Group 9
4. Bays that are almost completely open, typified by arcuate or circular outline with low structural complexity – Group 4
5. Partially closed systems
    a. Bays that are typically recessed with narrow mouths – Group 1
    b. Partially closed harbours, with narrower mouths and expansive tidal flats but display low structural complexity, typical of a single enclosed basin – Group 6

c. Almost completely closed harbours/inlets with narrow mouths, high structural complexity but lower river inputs relative to their size and moderate coverage of intertidal flats – Group 3
d. Complex harbours and harbour systems with multiple arms, narrow mouths and higher river input relative to their size – Group 5

3.2.5 – Rule-based summary of grouping schemes

Eight groups

There are two groups that are distinct from the remaining groups based on individual variable ranges (Table 33). Group 2 has zero STP/TEV, whereas all other groups have a minimum STP of 0.01336, and group 8 ranges in depth from 50.2-141.1 m, whereas all other groups have a maximum depth of 38.3 m (Table 33). Therefore a suitable rule set would begin with:

- STP < 0.00667
    - Y – Group 2
    - N – Mean Depth > 44
        - Y – Group 8
        - N – Decision tree (Figure 38)

Table 33. Physical variable ranges for each group based on the eight group scheme identified by Euclidean hierarchical clustering.

	R12/TEV	STP/TEV	SCI	MCI	Mean Depth	% IA	CLA/EWA
1	0, 0.099	0.047, 1.182	0.12, 0.7	0.02, 0.23	1.5, 31.4	0, 80.5	0.6, 223.2
2	0, 8.968	0, 0	0.13, 0.92	0, 0.03	0.1, 1	0, 95	0, 3305.8
3	0.001, 0.357	0.29, 13.135	0.08, 0.53	0, 0.11	0.2, 6.1	9.2, 98.2	1, 345
4	0, 0.09	0.039, 0.917	0.4, 0.84	0.1, 0.35	2.1, 38.3	0, 30.4	0, 170.2
5	0.003, 0.288	0.662, 1.558	0.31, 0.77	0.01, 0.13	1, 3.9	46.6, 100	2.1, 296.6
6	0.167, 7.655	0.327, 1.921	0.08, 0.52	0, 0.18	1.1, 6.4	0.6, 100	116.1, 5309.9
7	0.307, 27.13	0.422, 1.357	0.1, 0.41	0, 0.1	1.5, 5.1	0, 11.7	492, 39156.9
8	0, 0.003	0.013, 0.036	0.1, 0.34	0, 0.07	50.2, 141.1	0, 3.7	3.7, 19

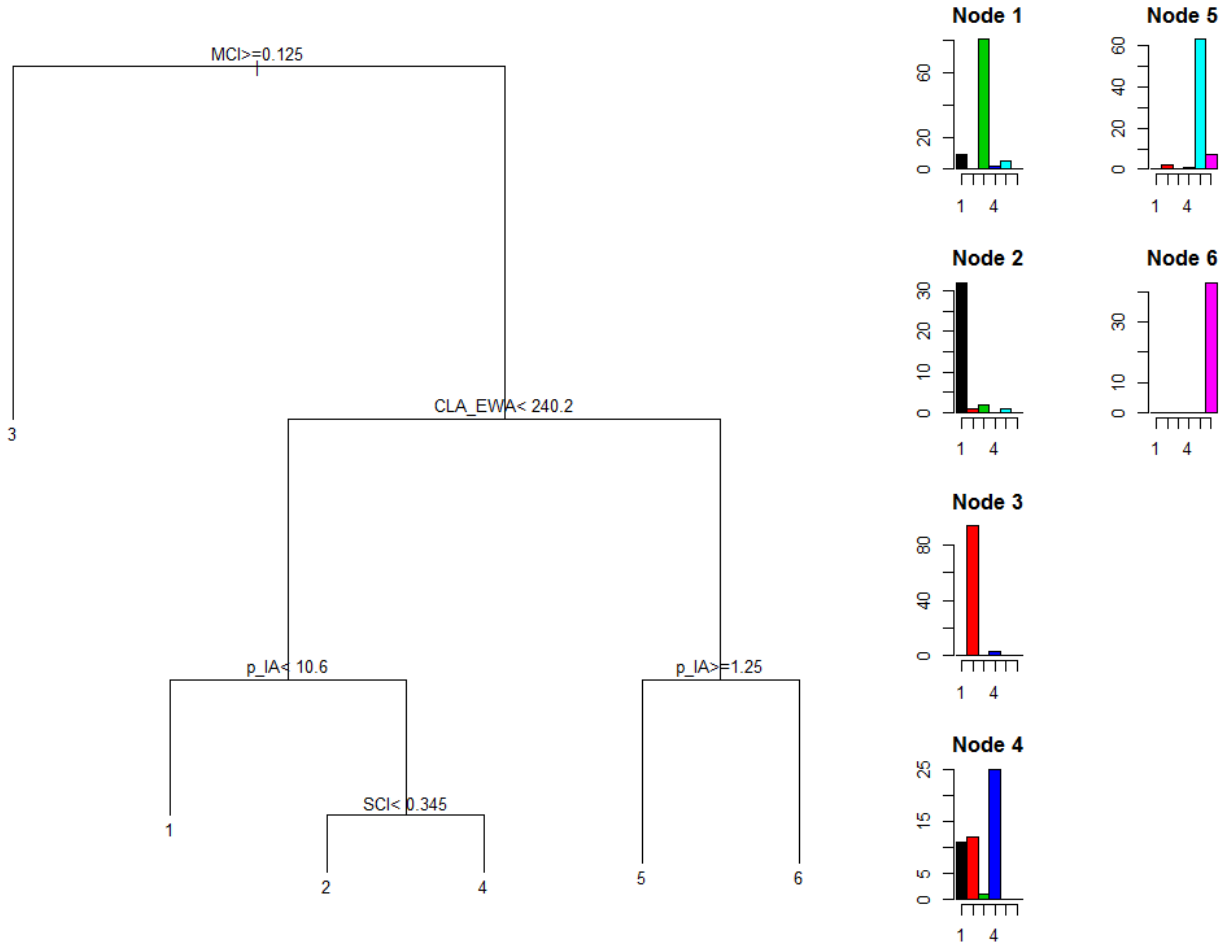

Figure 38. Classification tree for the eight group scheme identified by Euclidean hierarchical clustering. Groups 2 and 8 have been removed from the analysis and labels 2-6 refer to groups 3-7, respectively. Histograms to the right of the classification tree indicate the class make-up of each terminal node, where node numbering proceeds from left to right along the base of the classification tree.

Table 34. Classification accuracy of the classification tree developed for the eight group scheme identified by Euclidean hierarchical clustering.

Group	Classification accuracy (%)	No. in Group
1	61.5	52
2	100	37
3	86.2	109
4	96.4	84
5	80.6	31
6	91.3	69
7	86	50
8	100	11
Total	87.1	443

The resulting rule set achieves a good classification rate, with an overall success rate of 87.1%, and groupwise classification rates above 80% for all but group 1 (Table 34). Classification tree partitions, in addition to the rules for groups 2 and 9, are formed along STP/TEV (1), mean depth (1), MCI (1), % IA (2), CLA/EWA (1) and SCI (1) axes (Figure 38). Groups are therefore classified by:

(1) STP/TEV ≥ 0.0067, mean depth < 44, MCI < 0.125, CLA/EWA < 240.2, % IA < 10.6
(2) STP/TEV < 0.0067
(3) STP/TEV ≥ 0.0067, mean depth < 44, MCI < 0.125, CLA/EWA < 240.2, % IA ≥ 10.6, SCI < 0.345
(4) STP/TEV ≥ 0.0067, mean depth < 44, MCI ≥ 0.125
(5) STP/TEV ≥ 0.0067, mean depth < 44, MCI < 0.125, CLA/EWA < 240.2, % IA ≥ 10.6, SCI ≥ 0.345
(6) STP/TEV ≥ 0.0067, mean depth < 44, MCI < 0.125, CLA/EWA ≥ 240.2, % IA ≥ 1.25
(7) STP/TEV ≥ 0.0067, mean depth < 44, MCI < 0.125, CLA/EWA ≥ 240.2, % IA < 1.25
(8) STP/TEV ≥ 0.0067, mean depth ≥ 44

Nine groups

Similar to the eight group analysis, groups 2 and 9 are distinct from the remaining groups based on individual variable ranges (Table 35). Therefore a suitable rule set would begin with:

- STP < 0.00667
    - Y – Group 2
    - N – Mean Depth > 44
        - Y – Group 9
        - N – Decision tree (Figure 39)

Table 35. Physical variable ranges for each group based on the eight group scheme identified by Euclidean hierarchical clustering.

Group	R12/TEV	STP/TEV	SCI	MCI	Mean Depth	% IA	CLA/EWA
1	0, 0.099	0.047, 1.182	0.12, 0.7	0.02, 0.23	1.5, 31.4	0, 80.5	0.6, 223.2
2	0, 8.968	0, 0	0.13, 0.92	0, 0.03	0.1, 1	0, 95	0, 3305.8
3	0.001, 0.357	0.29, 13.135	0.08, 0.53	0, 0.07	0.2, 6.1	9.2, 98.2	1, 52.3
4	0, 0.09	0.039, 0.917	0.4, 0.84	0.1, 0.35	2.1, 38.3	0, 30.4	0, 170.2
5	0.028, 0.327	0.35, 1.155	0.1, 0.39	0, 0.11	1.1, 3.9	10.8, 90.8	28.2, 345
6	0.003, 0.288	0.662, 1.558	0.31, 0.77	0.01, 0.13	1, 3.9	46.6, 100	2.1, 296.6
7	0.167, 7.655	0.327, 1.921	0.08, 0.52	0, 0.18	1.1, 6.4	0.6, 100	116.1, 5309.9
8	0.307, 27.13	0.422, 1.357	0.1, 0.41	0, 0.1	1.5, 5.1	0, 11.7	492, 39156.9
9	0, 0.003	0.013, 0.036	0.1, 0.34	0, 0.07	50.2, 141.1	0, 3.7	3.7, 19

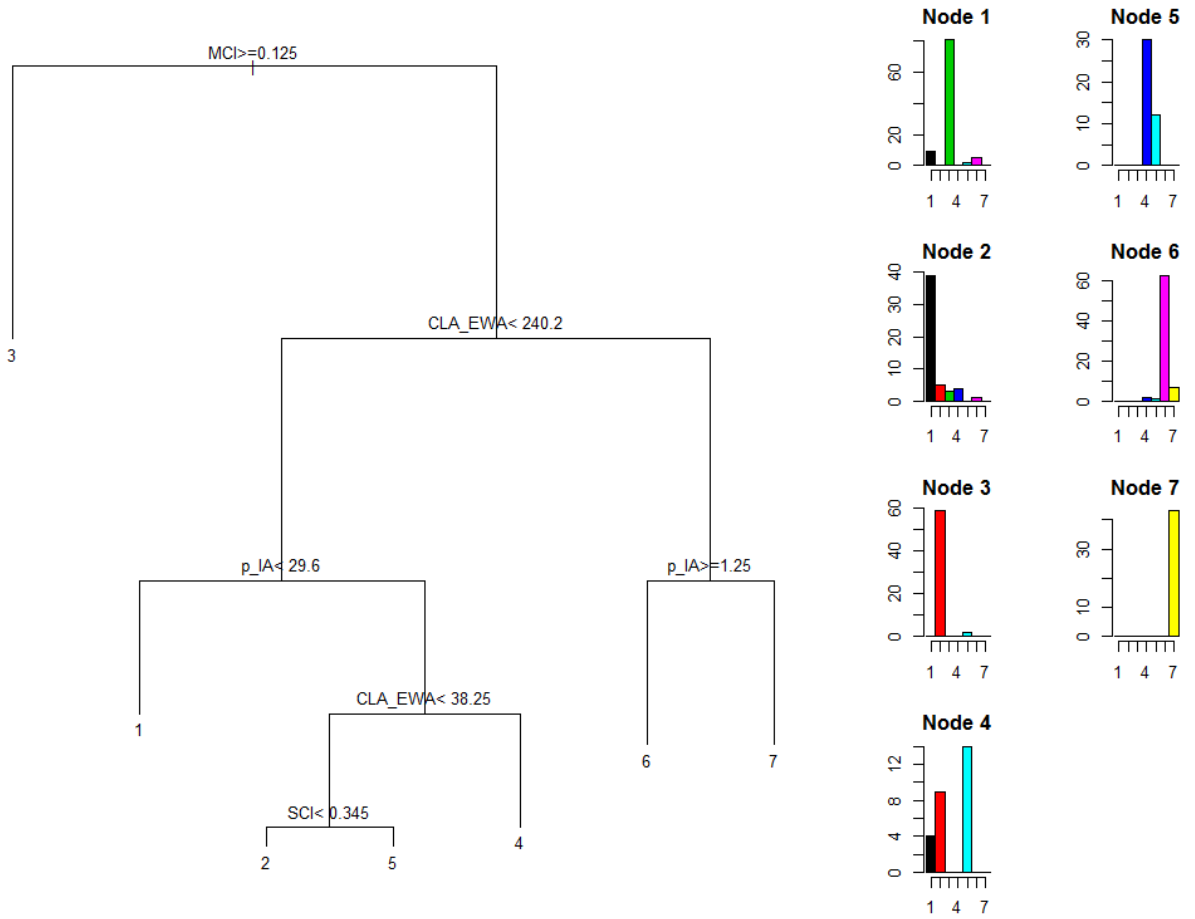

Figure 39. Classification tree for the nine group scheme identified by Euclidean hierarchical clustering. Groups 2 and 9 have been removed from the analysis and labels 2-7 refer to groups 3-8, respectively. Histograms to the right of the classification tree indicate the class make-up of each terminal node, where node numbering proceeds from left to right along the base of the classification tree.

Table 36. Classification accuracy of the classification tree developed for the nine group scheme identified by Euclidean hierarchical clustering.

Group	Classification accuracy (%)	No. in Group
1	75	52
2	100	37
3	80.8	73
4	96.4	84
5	83.3	36
6	45.2	31
7	91.3	69
8	86	50
9	100	11
Total	85.1	443

The resulting rule set achieves a good classification rate, with an overall success rate of 85.1%, and groupwise classification rates above 80% for all but group 1 and group 6 (Table 36). Classification tree partitions, in addition to the rules for groups 2 and 9, are formed along STP/TEV (1), mean depth (1), MCI (1), % IA (2), CLA/EWA (2) and SCI (1) axes (Figure 39). Groups are therefore classified by:

(1) STP/TEV $\geq$ 0.0067, mean depth < 44, MCI < 0.125, CLA/EWA < 240.2, % IA < 29.6
(2) STP/TEV < 0.0067
(3) STP/TEV $\geq$ 0.0067, mean depth < 44, MCI < 0.125, CLA/EWA < 240.2, % IA $\geq$ 29.6, CLA/EWA < 38.25, SCI < 0.345
(4) STP/TEV $\geq$ 0.0067, mean depth < 44, MCI $\geq$ 0.125
(5) STP/TEV $\geq$ 0.0067, mean depth < 44, MCI < 0.125, CLA/EWA < 240.2, % IA $\geq$ 29.6, CLA/EWA $\geq$ 38.25
(6) STP/TEV $\geq$ 0.0067, mean depth < 44, MCI < 0.125, CLA/EWA < 240.2, % IA $\geq$ 29.6, CLA/EWA < 38.25, SCI $\geq$ 0.345
(7) STP/TEV $\geq$ 0.0067, mean depth < 44, MCI < 0.125, CLA/EWA $\geq$ 240.2, % IA $\geq$ 1.25
(8) STP/TEV $\geq$ 0.0067, mean depth < 44, MCI < 0.125, CLA/EWA $\geq$ 240.2, % IA < 1.25
(9) STP/TEV $\geq$ 0.0067, mean depth $\geq$ 44

3.2.6 – Summary

The groupings resulting from this analysis method successfully distinguished among systems of different types, and also created a system that is readily interpretable in terms of large scale differences among systems. The major distinction between the groups corresponded to differences in closure, morphology (depth, structural complexity and intertidal coverage played a role in distinguishing a number of groups) and prevailing tidal or river input regimes. Five distinct super-groups could be interpreted from the data, which describe large differences in typology related to tidal and river forcings, with sub-groups describing differences within these super-groups that were related to openness and morphology (depth and % intertidal area).

3.3 – Method 2: Hierarchical clustering based on Random Forests distance measures

In this section hierarchical clustering, based on the random forests distance measure and Ward's linkage criterion, is applied to the dataset of R12/TEV, STP/TEV, SCI, MCI, % IA, mean depth and CLA/EWA for the 443 NZ coastal hydrosystems and the resulting grouping schemes are discussed and analysed further.

3.3.1 – Initial treatment of data

No pre-treatment, transformation or scaling of the data was undertaken (for an explanation of why no pre-treatment is required see Section 2.3.1).

3.3.2 – Deciding on the "best" number of groups

The method described in section 2.3.2 was similarly applied to this dataset to examine how OOB error rates vary with the number of groups in order to identify a suitable number of groups (Figure 40).

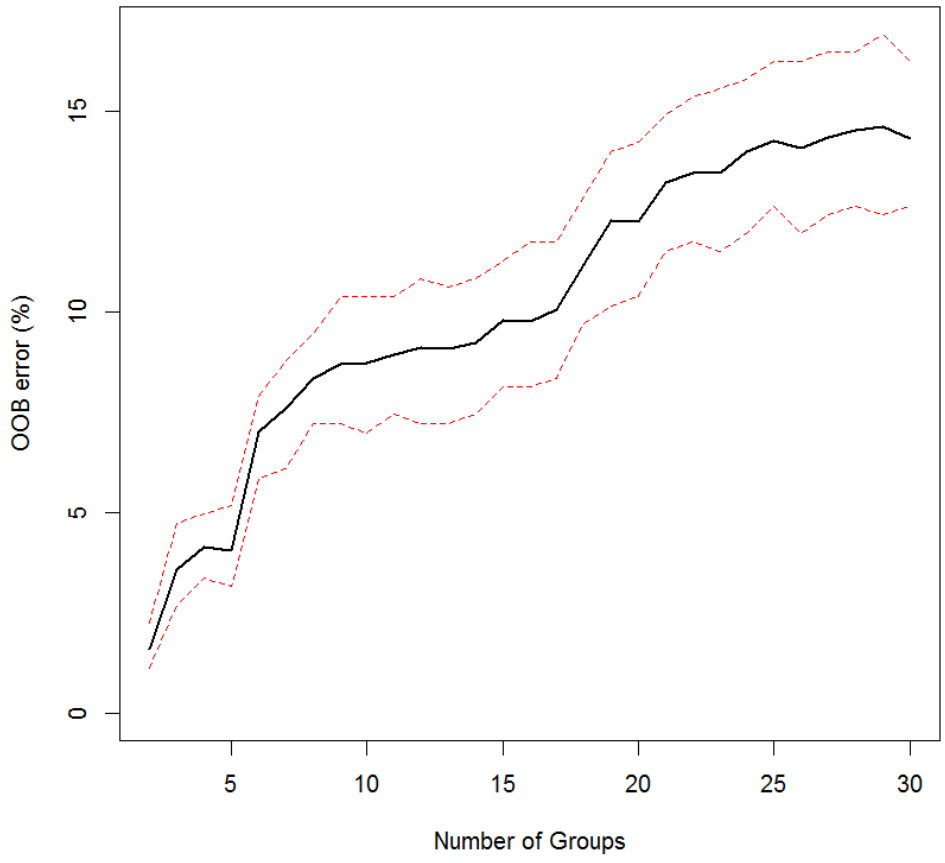

Figure 40. Out of bag error rates for random forests hierarchical clustering as a function of the number of groups. The thick black line is the mean OOB error, whilst red dotted lines are the 95% confidence interval.

Error rates increase sharply between five and six groups, and increase at a fast rate up to nine groups, at which point the increase in error rate levels off before rising sharply again for $k > 17$ (Figure 40). A 5% error rate corresponds to five or fewer groups, 7.5% corresponds to seven or fewer groups and 10% corresponds to 16 or fewer groups (Figure 40). This information is not easily interpreted to provide a robust number of groups to investigate. In the absence of further information, cluster analyses with seven and nine groups were investigated.

3.3.3 – Grouping scheme for seven groups

Performing the hierarchical cluster analysis with seven groups resulted in the group structure illustrated in Figure 41. Several groups are distinct along particular variable axes. Group 2 is visually separate along the STP/TEV axis, group 5 and 6 along the STP/TEV, MCI and depth axes, and group 7 along the R12/TEV and CLA/EWA axes (Figure 41). Examining group attributes (Figure 42) and systems contained in each group, groups can be characterised as:

- 1 – moderate R12/TEV and % IA, but considerable overlap with other groups, primarily river mouths and estuaries
- 2 – zero STP/TEV, lakes and lagoons
- 3 – high % IA, primarily inlets, harbours and harbour systems,
- 4 – high SCI, moderate depth and STP/TEV, primarily bays and harbours
- 5 – high SCI, high STP/TEV, bays
- 6 – moderate-high depth and moderate STP/TEV, primarily sounds, fiords and harbour systems
- 7 – high R12/TEV, river mouths

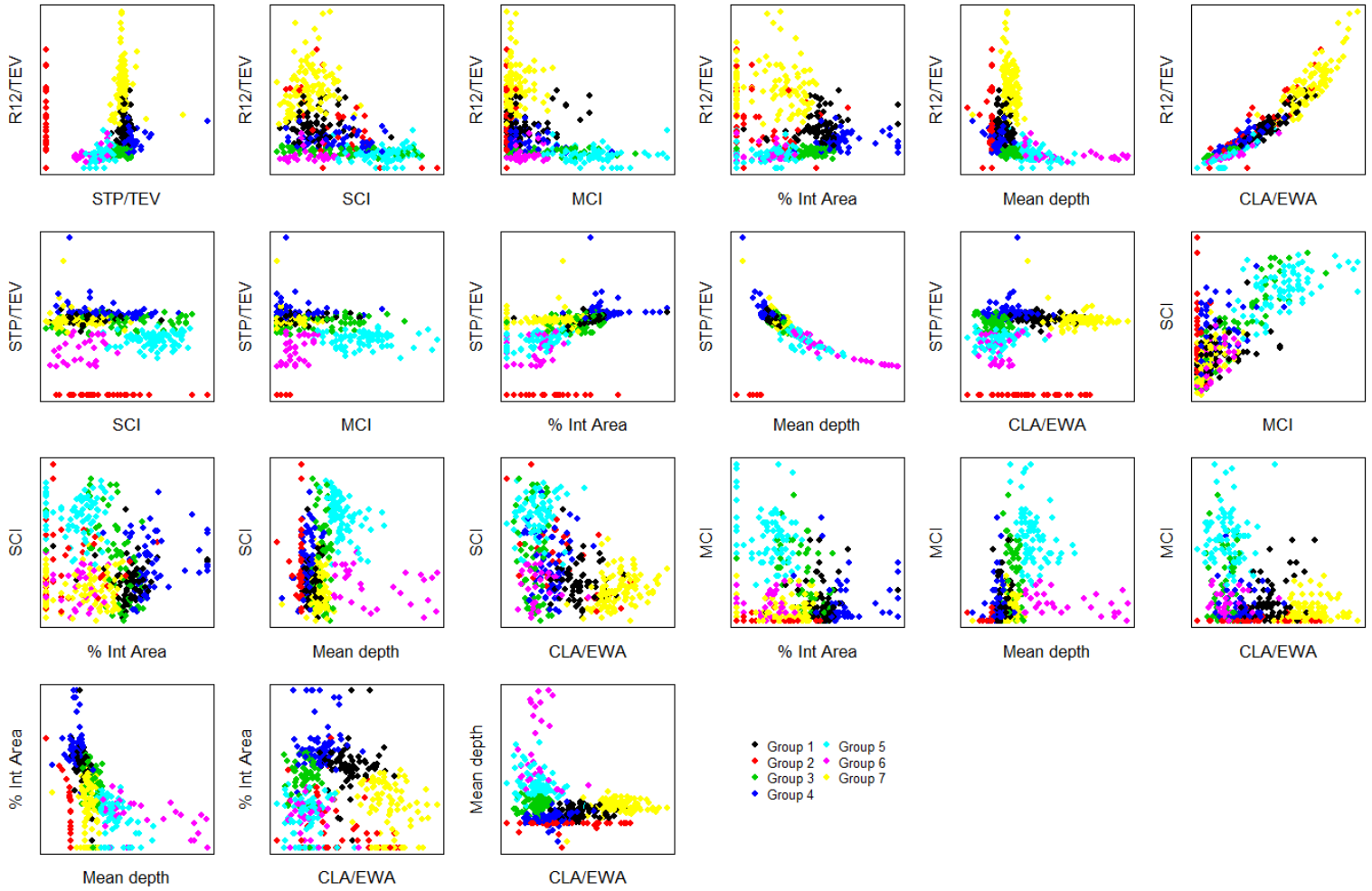

Figure 41. Biplots of all possible combinations of the seven physical variables, R12/TEV, STP/TEV, SCI, MCI, % IA, mean depth and CLA/EWA colour coded by group label identified by random forests hierarchical clustering with $k = 7$. All variables are transformed and scaled as detailed in section 3.2.1.

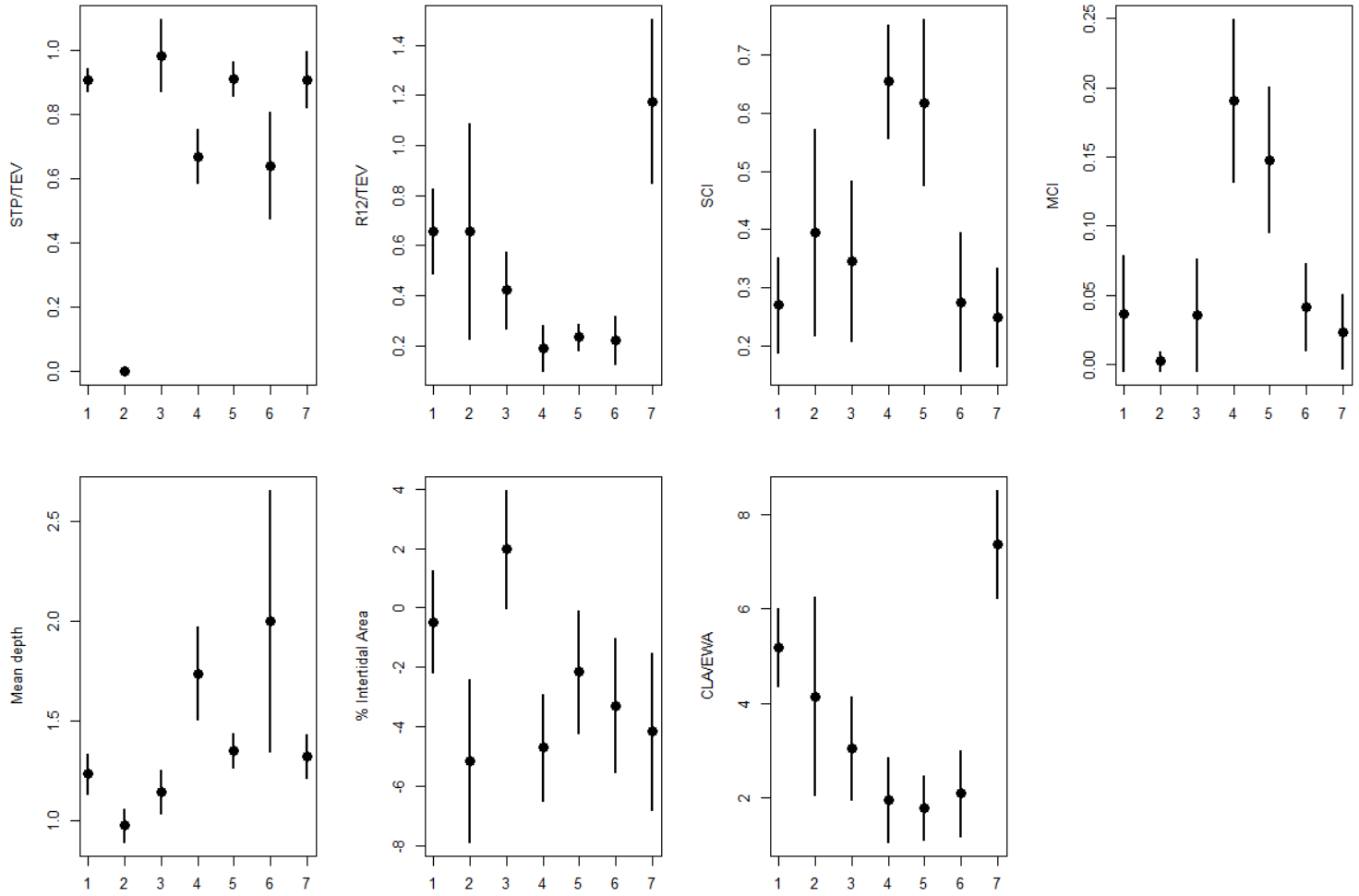

Figure 42. Means (±1 SD) of each of the seven physical variables for each group, based on the group structure identified by random forests hierarchical clustering with $k = 7$. All representations are based on the transformed variables as detailed in section 3.2.1.

Comparing these labels to the Hume classification (Table 37):

- 1 – corresponds to a mix of B-F classes, but predominantly B and F class systems
- 2 – corresponds exactly to class A
- 3 – corresponds to a mix of B-F classes, but predominantly E and F class systems
- 4 – corresponds to the majority of class D, and half of class H
- 5 – corresponds to a portion of class D systems
- 6 – corresponds to a mix of D-H classes, containing the majority of G class systems
- 7 – corresponds strongly to class B

Table 37. Table illustrating the number of systems in each group that were of the different hydro classes outlined in the Hume et al. (2007) classification. This is based on random forests hierarchical clustering, with $k = 7$.

Hume Hydro Class	Group						
	1	2	3	4	5	6	7
A	0	37	0	0	0	0	0
B	13	0	5	0	0	0	103
D	3	0	9	70	24	7	0
E	4	0	49	0	3	12	0
F	21	0	43	0	0	17	3
G	0	0	0	1	0	10	0
H	0	0	0	5	0	4	0

There is considerable agreement between this grouping scheme and the Hume classification and is particularly strong for classes A, B, D (split into two groups), but the majority of E, F, G and H are split among three groups. In contrast to the Euclidean analysis class D is classified into completely separate groups to those containing the majority of E and F, which are not separated in this classification and are placed into groups with other classes, primarily B (Group 1), G (Group 6) and D (Group 3). Class D is split into two major groups, with group 4 corresponding to deeper Bays with lower intertidal area, and group 5 corresponding to shallower bays with greater tidal influence (STP/TEV and % IA). Groups 1 – 3 – 6 form a gradient in R12/TEV (decreasing), and also differ based on % IA, which is highest for 3 and lowest for 6, and depth, which is highest for 6 and lowest for 3. All three groups have similar SCI and MCI, which indicate partially closed systems with high structural complexity. Therefore these groups represent harbours or estuarine systems that are partially closed and can be further defined as high river input, which may include complex barrier enclosed river mouths (Group 1), tidal flat dominated systems (Group 3) and predominantly subtidal systems,

such as large harbour systems, sounds and fiords (Group 6). As with the Euclidean analysis, this seven group system can be simplified to a system containing four super-groups with one group containing three sub-divisions and another containing two.

1. Fully closed, little to no tidal influence – Group 2
2. Open river mouths – Group 7
3. Bays
    a. Deep, lower tidal influence relative to volume – Group 4
    b. Shallow, higher tidal influence relative to volume – Group 5
4. Partially closed, structurally complex
    a. High river input – Group 1
    b. Tidal flat dominated – Group 3
    c. Predominantly subtidal – Group 6

3.3.4 – Grouping scheme for nine groups

The nine group analysis builds upon the seven group analysis by dividing the previous group 3 into three separate groups, group 3, 4 and 6, primarily separated along the SCI axis (Figure 43). Examining group attributes (Figure 44) and systems contained in each group, groups can be characterised as:

- 1 – moderate R12/TEV and % IA, but considerable overlap with other groups, primarily river mouths and estuaries
- 2 – zero STP/TEV, lakes and lagoons
- 3 – low SCI, moderate-high % IA, primarily harbour systems
- 4 – high SCI, high % IA, primarily inlets and estuaries
- 5 – high SCI, moderate depth and STP/TEV, primarily bays
- 6 – highest STP/TEV, primarily harbours and inlets
- 7 – high SCI, high STP/TEV, bays and harbours
- 8 – moderate-high depth and moderate STP/TEV, primarily sounds, fiords and harbour systems
- 9 – high R12/TEV, river mouths

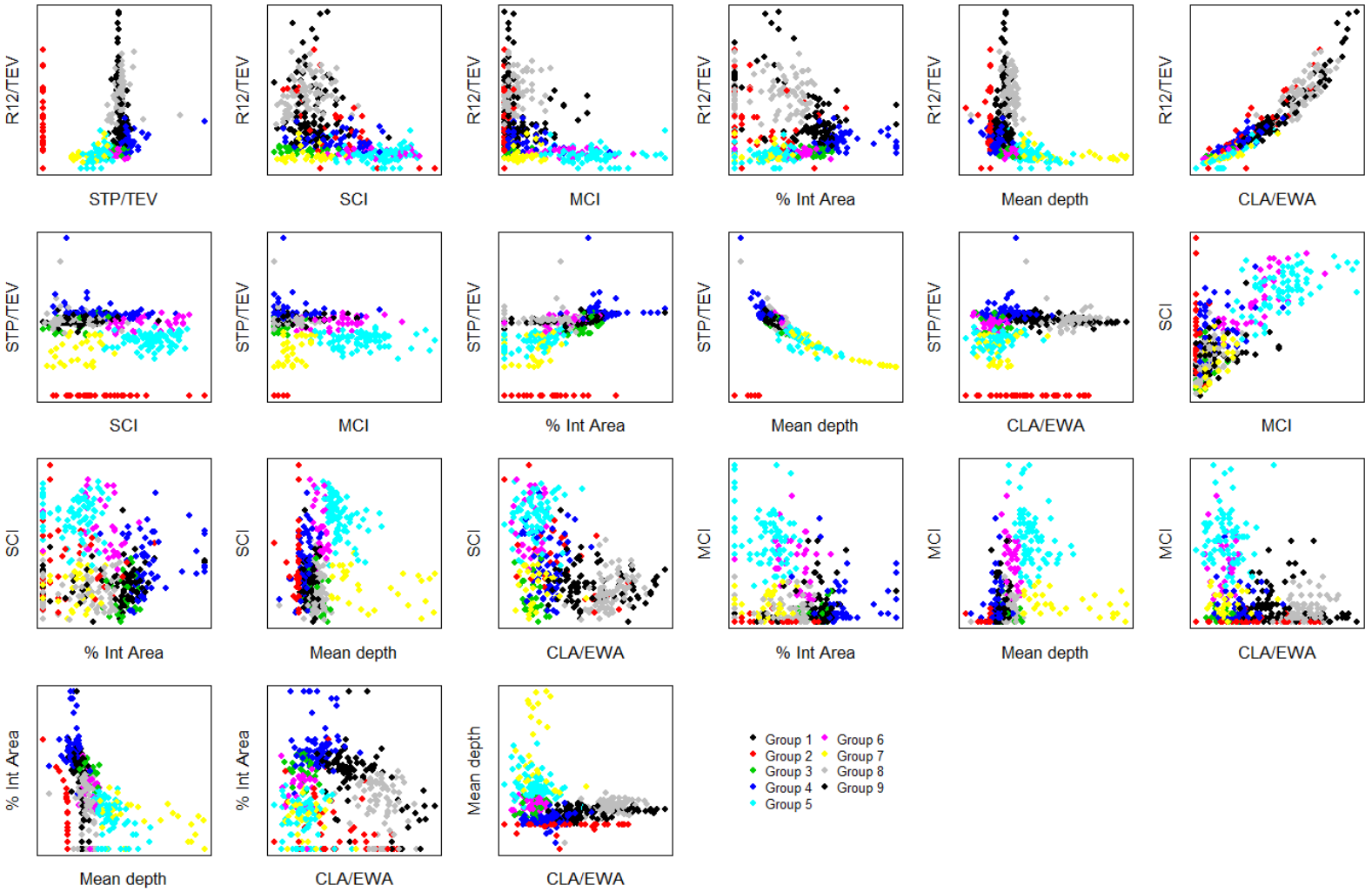

Figure 43. Biplots of all possible combinations of the seven physical variables, R12/TEV, STP/TEV, SCI, MCI, % IA, mean depth and CLA/EWA colour coded by group label as identified by random forests hierarchical clustering with $k = 9$. All variables are transformed and scaled as detailed in section 3.2.1.

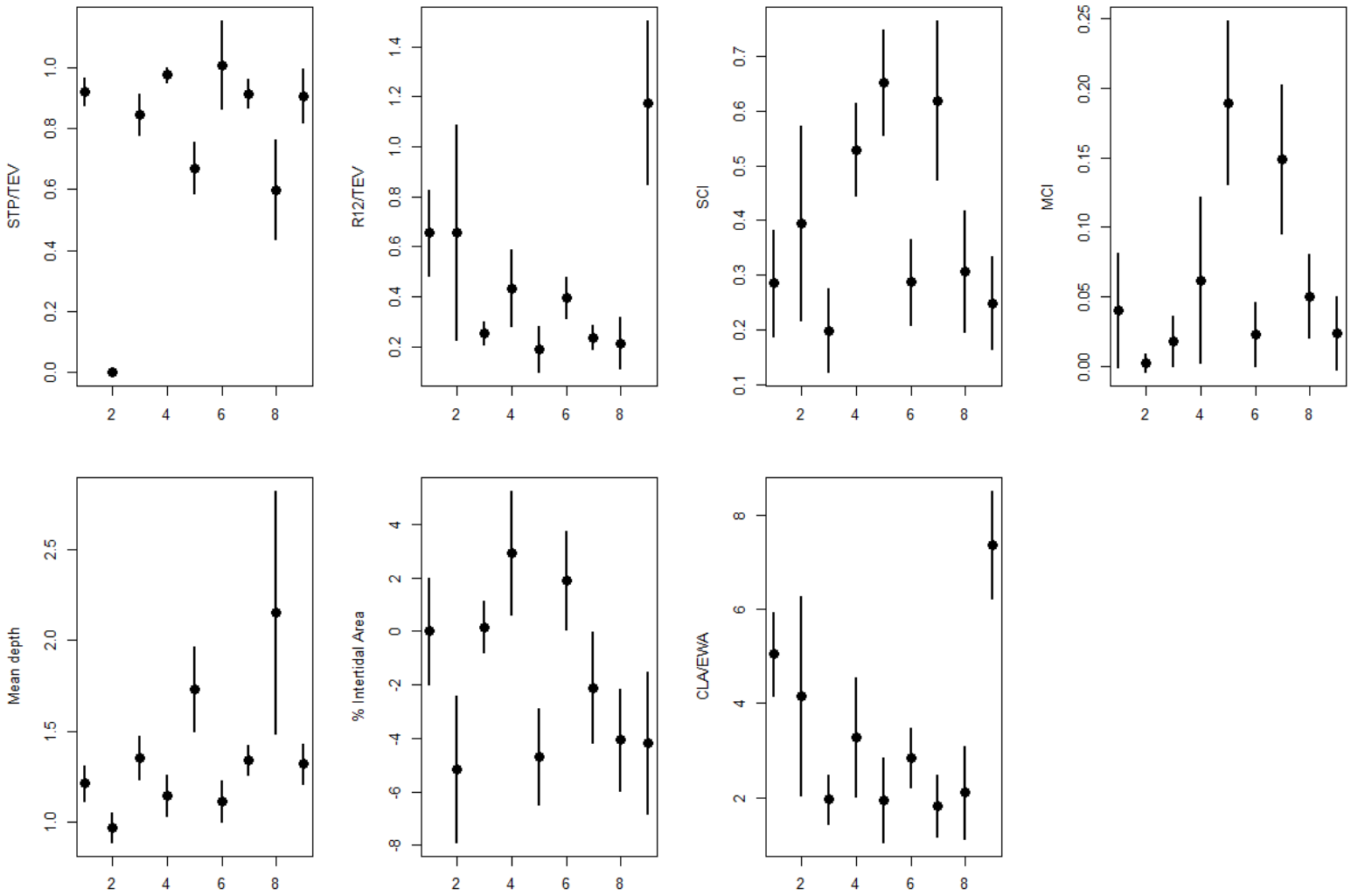

Figure 44. Means (±1 SD) of each of the seven physical variables for each group, based on the group structure identified by random forests hierarchical clustering, with $k = 9$. All representations are based on the transformed variables as detailed in section 3.2.1.

Comparing these labels to the Hume classification (Table 38):

- 1 – corresponds to a mix of B-F classes, but predominantly B and F class systems
- 2 – corresponds exactly to class A
- 3 – corresponds to a selection of class F systems
- 4 – corresponds primarily to a selection of class E, with some D class systems
- 5 – corresponds to the majority of class D, but also contains half of class H
- 6 – corresponds to a mix of E and F class systems
- 7 – corresponds to a selection of class D systems
- 8 – corresponds to a mix of D-H systems
- 9 – corresponds to the majority of class B systems

Table 38. Table illustrating the number of systems in each group that were of the different hydro classes outlined in the Hume et al. (2007) classification. This is based on random forests hierarchical clustering, with $k = 9$.

Hume Hydro class	Group								
	1	2	3	4	5	6	7	8	9
A	0	37	0	0	0	0	0	0	0
B	16	0	0	2	0	0	0	0	103
D	4	0	0	6	71	2	23	7	0
E	8	0	3	18	0	25	3	11	0
F	26	0	20	0	0	28	0	7	3
G	0	0	0	0	1	0	0	10	0
H	0	0	0	0	5	0	0	4	0

This grouping scheme builds upon the seven group system in that it successfully differentiates the majority of classes A, B and D into one or more almost pure groups, with group 2 corresponding to closed lakes and lagoons, group 5 corresponding to deep bays, group 7 corresponding to shallow bays and group 9 corresponding to open river mouths. As stated for the seven group analysis, the remaining groups (1, 3, 4, 6, 8) all correspond to partially closed (low MCI) systems of varying typology. Group 1 corresponds to partially closed systems with high river input, potentially barrier or partially enclosed river mouths. Group 3 has the lowest SCI and moderate intertidal coverage and therefore corresponds to structurally complex systems with extensive tidal flats, such as harbour systems with multiple arms. Group 4 has the highest intertidal coverage and high SCI, and is therefore less structurally complex and may resemble a single enclosed basin dominated by intertidal flats. Group 6 overlaps with many of the other partially enclosed groups along any particular axis, but contains systems with the

highest tidal influx relative to their volume, high structural complexity, moderate river input, shallow depth and extensive tidal flats. Therefore this group is perhaps characterised as more dynamic than group 3, which is marginally deeper, and has lower STP/TEV and R12/TEV than group 6. Group 8 is characterised by depth and corresponds to systems that are predominantly subtidal. Therefore this group scheme can be summarised as four super-groups with one group containing five sub-divisions and another containing two.

1. Fully closed, little to no tidal influence – Group 2
2. Open river mouths – Group 9
3. Bays
    a. Deep, lower tidal influence relative to volume – Group 5
    b. Shallow, higher tidal influence relative to volume – Group 7
4. Partially closed
    a. High river input – Group 1
    b. Tidal flat dominated, structurally simple – Group 4
    c. Dynamic shallow systems with extensive tidal flats, large tidal influence and high structural complex – Group 6
    d. Less dynamic systems with moderate tidal flats and high structural complexity – Group 3
    e. Predominantly subtidal – Group 8

3.3.5 – Rule-based summary of grouping schemes

Seven groups

Based on variable ranges, group 2 is distinct from the remaining groups in that it has STP/TEV=0, whereas the remaining groups have a minimum STP/TEV of 0.01336 (Table 39). Therefore a suitable initial rule would be

- STP < 0.0067
    - Y – Group 2
    - N – decision tree (Figure 45)

Table 39. Physical variable ranges for each group based on the seven group scheme identified by random forests hierarchical clustering.

Group	R12/TEV	STP/TEV	SCI	MCI	Mean Depth	% IA	CLA/EWA
1	0.028, 1.097	0.35, 0.819	0.1, 0.48	0, 0.18	1.1, 3.9	0.1, 78.7	28.2, 859.6
2	0, 8.968	0, 0	0.13, 0.92	0, 0.03	0.1, 1	0, 95	0, 3305.8
3	0.001, 1.613	0.492, 13.135	0.13, 0.77	0, 0.23	0.2, 3.9	40, 100	1, 1186.6
4	0, 0.09	0.039, 0.459	0.4, 0.83	0.08, 0.35	3.2, 38.3	0, 18.9	0, 170.2
5	0, 0.009	0.378, 1	0.36, 0.84	0.05, 0.28	1.9, 5	0, 80.5	0.6, 14.7
6	0, 0.065	0.013, 0.518	0.08, 0.49	0, 0.15	3.8, 141.1	0, 61.8	1.1, 223.2
7	0.167, 27.13	0.327, 6.862	0.08, 0.52	0, 0.18	0.2, 6.4	0, 47.3	52.3, 39156.9

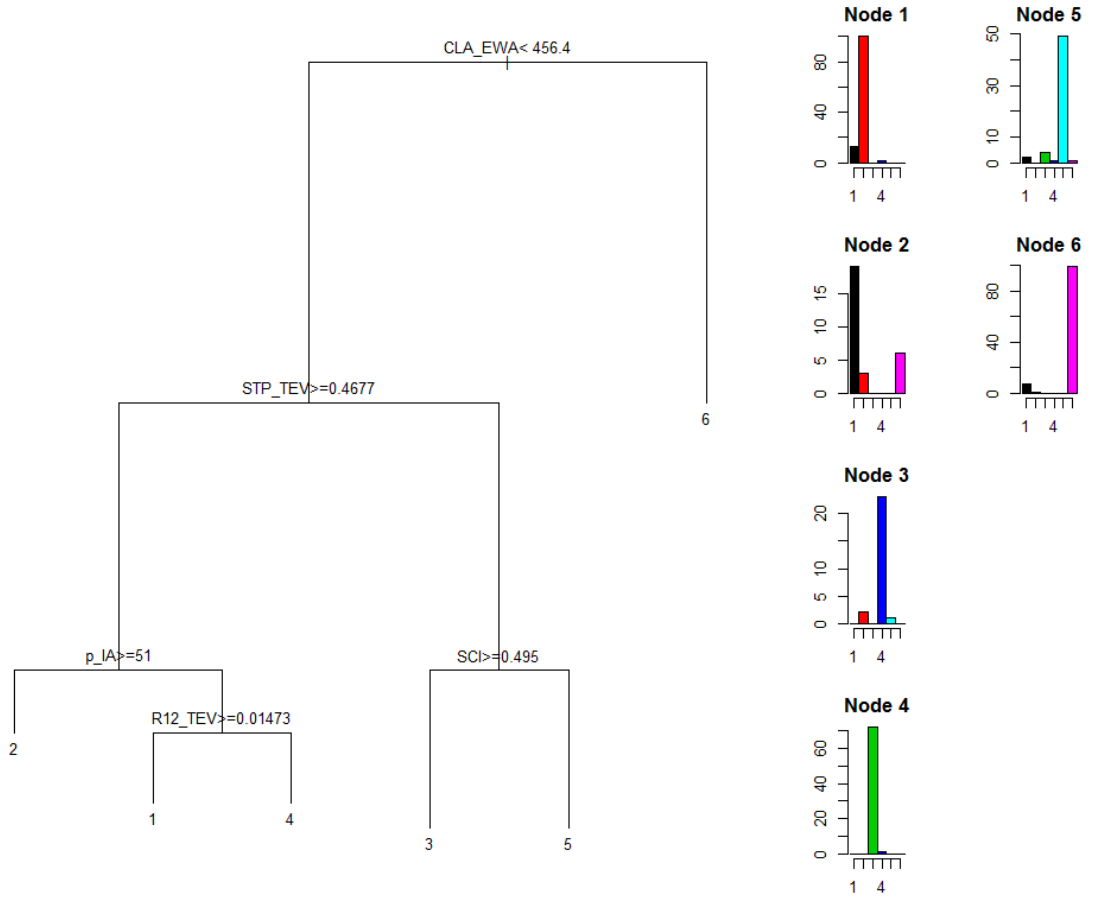

Figure 45. Classification tree for the seven group scheme identified by random forest hierarchical clustering. Group 2 has been removed from the analysis and labels 2-6 refer to groups 3-7, respectively. Histograms to the right of the classification tree indicate the class make-up of each terminal node, where node numbering proceeds from left to right along the base of the classification tree.

Table 40. Classification accuracy of the classification tree developed for the seven group scheme identified by random forests hierarchical clustering.

Group	Classification accuracy (%)	No. in Group
1	46.3	41
2	100	37
3	94.3	106
4	94.7	76
5	85.2	27
6	98	50
7	93.4	106
Total	90.1	406

The overall classification accuracy of the identified rule-set was high, with 90.1% of all observations assigned to the correct class, but classification accuracy was lower for group 1 (Table 40). Classification tree partitions, in addition to the rule separating group 2, are formed along STP/TEV (2), SCI (1), % IA (1), CLA/EWA (1) and R12/TEV (1) axes (Figure 45). The resulting rule set is:

(1) STP/TEV ≥ 0.0067, CLA/EWA < 456.4, STP/TEV ≥ 0.4677, % IA < 51, R12/TEV ≥ 0.01473
(2) STP/TEV < 0.0067
(3) STP/TEV ≥ 0.0067, CLA/EWA < 456.4, STP/TEV ≥ 0.4677, % IA ≥ 51
(4) STP/TEV ≥ 0.0067, CLA/EWA < 456.4, STP/TEV < 0.4677, SCI ≥ 0.495
(5) STP/TEV ≥ 0.0067, CLA/EWA < 456.4, STP/TEV ≥ 0.4677, % IA < 51, R12/TEV < 0.01473
(6) STP/TEV ≥ 0.0067, CLA/EWA < 456.4, STP/TEV < 0.4677, SCI < 0.495
(7) STP/TEV ≥ 0.0067, CLA/EWA ≥ 456.4

Nine groups

Based on variable ranges, group 2 is distinct from the remaining groups in that it has STP/TEV=0, whereas the remaining groups have a minimum STP/TEV of 0.01336 (Table 41). Therefore a suitable initial rule would be

- STP < 0.0067
    - Y – Group 2
    - N – decision tree (Figure 46)

Table 41. Physical variable ranges for each group based on the nine group scheme identified by random forests hierarchical clustering.

Group	R12/TEV	STP/TEV	SCI	MCI	Mean Depth	% IA	CLA/EWA
1	0.028, 1.613	0.35, 1.155	0.1, 0.68	0, 0.18	1.1, 3.9	0.1, 100	28.2, 1186.6
2	0, 8.968	0, 0	0.13, 0.92	0, 0.03	0.1, 1	0, 95	0, 3305.8
3	0.001, 0.012	0.308, 0.779	0.08, 0.34	0, 0.07	1.7, 6.1	11.4, 82	1.8, 18.3
4	0.003, 0.288	0.724, 1.182	0.41, 0.77	0.01, 0.23	1, 3.9	46.6, 100	2.1, 296.6
5	0, 0.09	0.039, 0.459	0.4, 0.83	0.08, 0.35	3.2, 38.3	0, 18.9	0, 170.2
6	0.001, 0.23	0.601, 13.135	0.13, 0.43	0, 0.11	0.2, 3.4	46.5, 100	1, 42.3
7	0, 0.009	0.431, 1	0.36, 0.84	0.05, 0.28	1.9, 4.8	0, 80.5	0.6, 14.7
8	0, 0.065	0.013, 0.45	0.1, 0.49	0, 0.15	4.2, 141.1	0, 37	1.1, 223.2
9	0.167, 27.13	0.327, 6.862	0.08, 0.52	0, 0.18	0.2, 6.4	0, 47.3	52.3, 39156.9

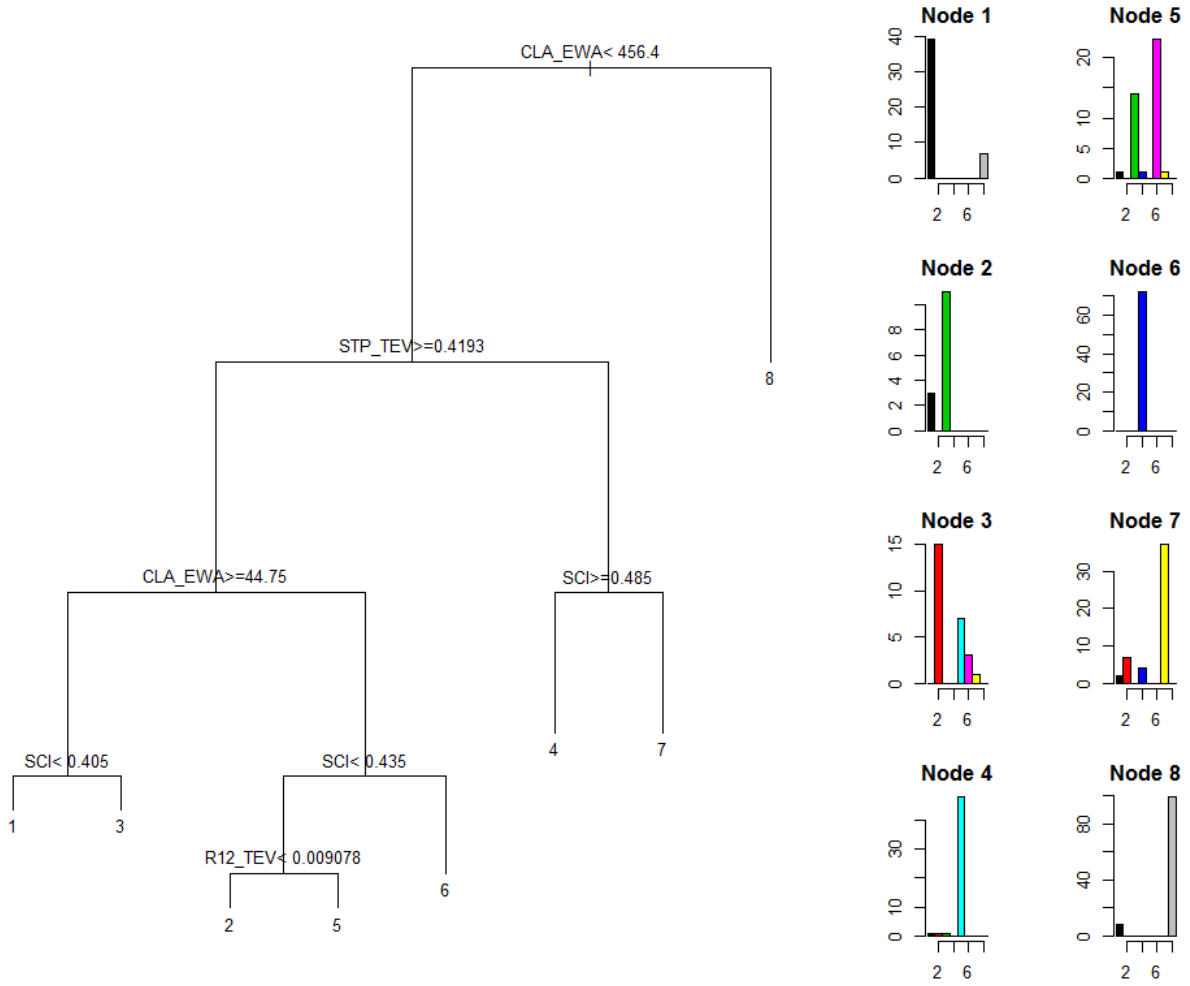

Figure 46. Classification tree for the nine group scheme identified by random forest hierarchical clustering. Group 2 has been removed from the analysis and labels 2-8 refer to groups 3-9, respectively. Histograms to the right of the classification tree indicate the class make-up of each terminal node, where node numbering proceeds from left to right along the base of the classification tree.

Table 42. Classification accuracy of the classification tree developed for the nine group scheme identified by random forests hierarchical clustering.

Group	Classification accuracy (%)	No. in Group
1	72.2	54
2	100.0	37
3	65.2	23
4	42.3	26
5	93.5	77
6	87.3	55
7	88.5	26
8	94.9	39
9	93.4	106
Total	86.0	443

The classification rule set achieved reasonably high classification overall, but was less good at classifying groups 3 and 4 (Table 42). Classification tree partitions, in addition to the rule separating group 2, are formed along STP/TEV (2), SCI (3), CLA/EWA (2) and R12/TEV (1) axes (Figure 46). The resulting rules are:

(1) STP/TEV ≥ 0.0067, CLA/EWA < 456.4, STP/TEV ≥ 0.4193, CLA/EWA ≥ 44.75, SCI < 0.405

(2) STP/TEV < 0.0067

(3) STP/TEV ≥ 0.0067, CLA/EWA < 456.4, STP/TEV ≥ 0.4193, CLA/EWA < 44.75, SCI < 0.435, R12/TEV < 0.00908

(4) STP/TEV ≥ 0.0067, CLA/EWA < 456.4, STP/TEV ≥ 0.4193, CLA/EWA ≥ 44.75, SCI ≥ 0.405

(5) STP/TEV ≥ 0.0067, CLA/EWA < 456.4, STP/TEV < 0.4193, SCI ≥ 0.485

(6) STP/TEV ≥ 0.0067, CLA/EWA < 456.4, STP/TEV ≥ 0.4193, CLA/EWA < 44.75, SCI < 0.435, R12/TEV ≥ 0.00908

(7) STP/TEV ≥ 0.0067, CLA/EWA < 456.4, STP/TEV ≥ 0.4193, CLA/EWA < 44.75, SCI ≥ 0.435

(8) STP/TEV ≥ 0.0067, CLA/EWA < 456.4, STP/TEV < 0.4193, SCI < 0.485

(9) STP/TEV ≥ 0.0067, CLA/EWA ≥ 456.4

3.3.6 – Summary

In summary the groups identified by the random forests hierarchical clustering method were meaningful in terms of environmental and morphological differences among systems.

The seven group analysis successfully distinguished among completely closed systems (lakes and lagoons), open systems, including river mouths and bays, which were separated into deep and shallow types, and partially closed systems, which were separated into high river input, tidal flat dominated and predominantly subtidal systems. The nine group system expanded upon this by splitting the tidal flat dominated group into three separate groups. These three groups were characterised according to structural complexity, with one group containing those systems with relatively simple outlines, and a mixture of inflow/outflow regime (R12/TEV and STP/TEV), depth and intertidal coverage separating the remaining two structurally complex partially closed groups. These two groups could be considered as dynamic (larger tidal influence in terms of STP/TEV and % IA and higher R12/TEV) and relatively static (lower tidal influence and R12/TEV) system types. Finally, identifying a suitable number of groups to investigate using this method involved a large measure of guesswork and trialling a large range of group numbers (up to 11 were investigated but not reported here, see Classification's Appendix). As a result identifying the "best" number of groups for this method may require the analysis of multiple group numbers followed by expert analysis of group attributes to identify what number of groups distinguishes among the systems without adding unnecessary groupings.

3.4 – Method 3: Model-based clustering

In this section model-based clustering, is applied to the dataset of R12/TEV, STP/TEV, SCI, MCI, % IA, mean depth and CLA/EWA for the 443 NZ coastal hydrosystems and the resulting grouping schemes are discussed and analysed further.

3.4.1 – Initial treatment of data

Analyses are based on the transformed and scaled dataset, with transformations and scaling as detailed in section 3.2.1.

3.4.2 – Deciding on the "best" number of groups

All possible model types (see section 2.4 for a description of the different model types) were investigated for 2-20 groups and their BIC statistics were recorded (Table 43).

Table 43. BIC statistics for all multivariate model-based clustering types for 2-20 groups. Numbers in bold indicate those models with ΔBIC < 10 relative to the model with minimal BIC.

No. Groups	Model Type									
	EII	VII	EEI	VEI	EVI	VVI	EEE	EEV	VEV	VVV
2	-8195.1	-8097.7	-8063.0	-8084.2	-7603.7	-7211.8	-6939.2	-5806.6	-5654.1	-6124.8
3	-7684.3	-7229.1	-7391.2	-7192.5	-6429.4	-6603.8	-6610.5	-4734.0	-5241.2	-5224.3
4	-7254.1	-6742.4	-7032.2	-6603.1	-5942.0	-5842.6	-5647.9	-4597.7	-4476.1	-4495.9
5	-6536.0	-6510.3	-6363.0	-6321.4	NA	NA	-5268.6	-4345.6	-4206.1	NA
6	-6400.4	-6351.8	-6175.6	-6162.2	NA	NA	-5207.9	-4366.9	-4252.2	NA
7	-6337.7	-6194.6	-6087.1	-5992.4	NA	NA	-5170.6	-4369.2	-4207.9	NA
8	-6233.3	-6039.8	-6026.3	-5858.6	NA	NA	-5145.5	-4341.4	-4225.9	NA
9	-6183.8	-5930.6	-5983.3	-5773.2	NA	NA	-5065.1	-4258.7	-4094.7	NA
10	-6192.9	-5840.1	-6003.4	-5688.6	NA	NA	-5045.9	-4235.3	**-4065.8**	NA
11	-6106.7	-5792.9	-5879.6	-5597.9	NA	NA	-5061.4	-4228.0	-4096.6	NA
12	-6100.6	-5733.9	-5860.4	-5467.4	NA	NA	-4972.9	-4357.9	-4151.7	NA
13	-6076.2	-5694.9	-5842.3	-5452.1	NA	NA	-5009.9	-4457.8	-4240.6	NA
14	-6081.4	-5678.7	-5856.4	-5427.0	NA	NA	-5028.7	-4624.6	-4339.1	NA
15	-5849.3	-5614.6	-5632.6	-5339.0	NA	NA	-4679.9	-4761.3	-4403.9	NA
16	-5827.2	-5604.6	-5627.4	-5313.1	NA	NA	-4672.9	-4791.9	-4801.6	NA
17	-5769.3	-5588.2	-5645.6	-5314.2	NA	NA	-4846.3	-5032.8	-4564.6	NA
18	-5741.0	-5572.8	-5308.4	-5258.7	NA	NA	-4868.8	-5067.3	-5020.3	NA
19	-5738.8	-5574.8	-5311.0	-5304.8	NA	NA	-4634.4	-5207.3	-4905.7	NA
20	-5755.9	-5525.9	-5304.0	-5256.4	NA	NA	-4953.7	-4927.2	-5364.2	NA

The best model (VEV, k = 10) was significantly better than all other models, with a difference in BIC between it and the second best model (VEV, k = 9, ΔBIC=29) indicating very strong (ΔBIC > 10) support. Therefore only the ten group model will be investigated further here.

3.4.3 – Grouping scheme for ten groups

The ten group scheme identifies groups that are visually distinctive along multiple axes (Figure 47). In particular group 2 is separate along the STP/TEV axis, group 5 along the SCI/MCI axes, group 9 along R12/TEV axes and group 10 along the mean depth axis (Figure 47).

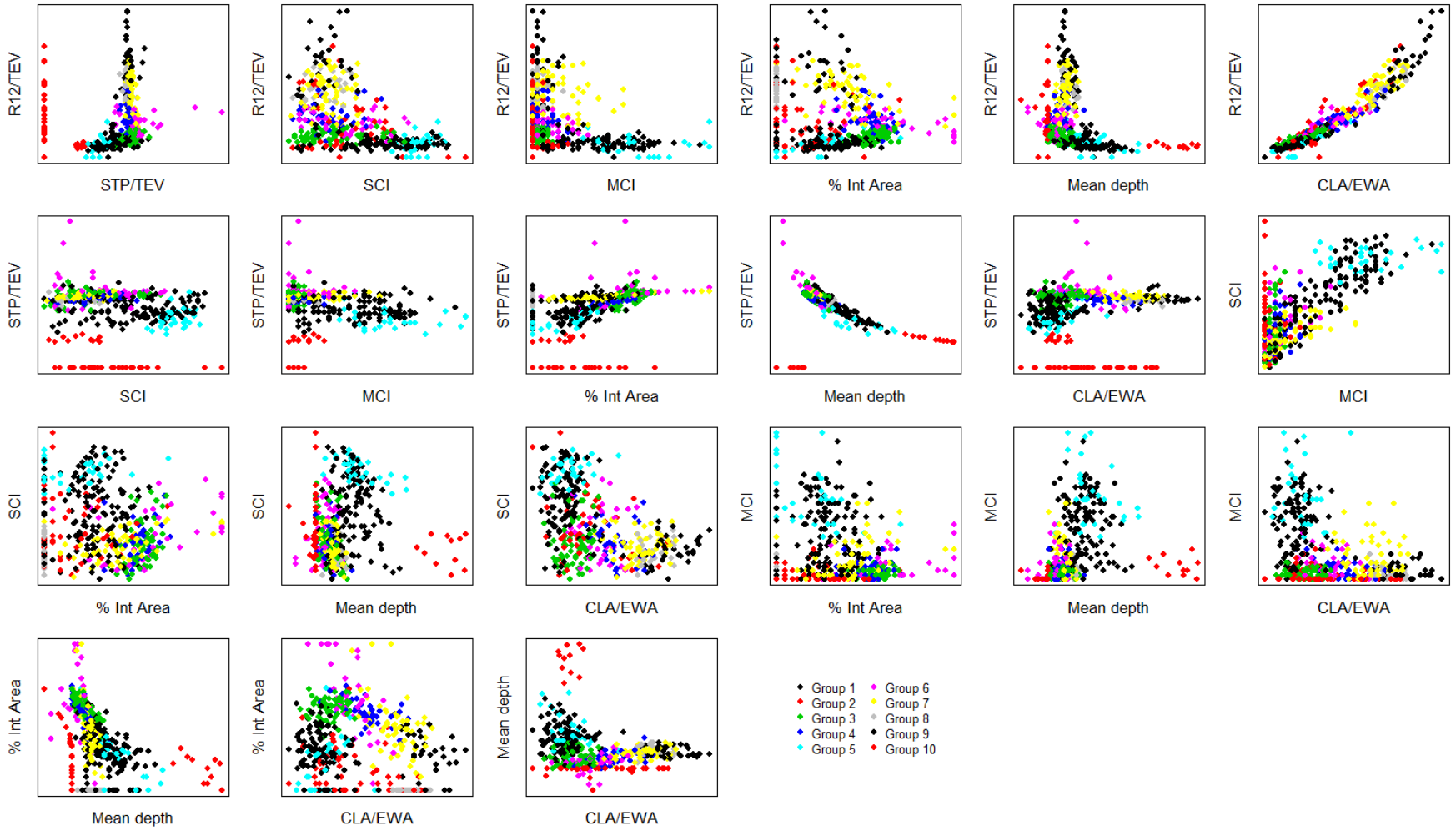

Figure 47. Biplots of all possible combinations of the seven physical variables, R12/TEV, STP/TEV, SCI, MCI, % IA, mean depth and CLA/EWA colour coded by group label identified by model-based clustering with $k = 10$. All variables are transformed and scaled as detailed in section 3.2.1.

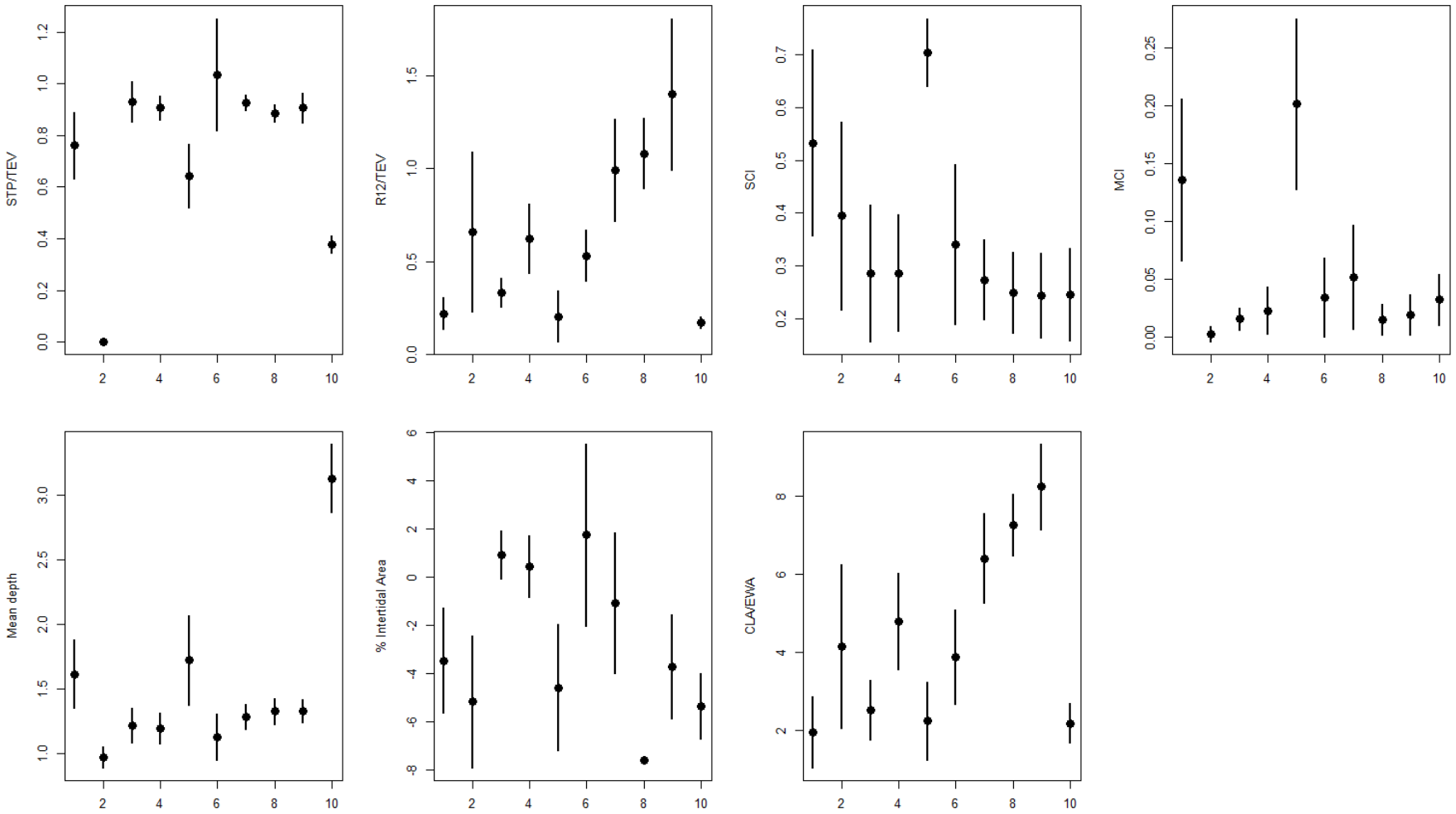

Figure 48. Means (±1 SD) of each of the seven physical variables for each group, based on the group structure identified by model-based clustering with k = 10. All representations are based on the transformed variables as detailed in section 3.2.1.

Examining group attributes (Figure 48) and systems contained in each group, groups can be characterised as:

- 1 – low R12/TEV and moderate-high SCI/MCI, primarily bays and harbours
- 2 – zero STP/TEV, lakes and lagoons
- 3 – high intertidal area, low R12/TEV, harbours, harbour systems and inlets
- 4 – high intertidal area, moderate R12/TEV, rivers and estuaries
- 5 – high SCI/MCI, bays
- 6 – high STP/TEV, high intertidal area, harbours and inlets
- 7 – moderate R12/TEV, moderate intertidal cover, river mouths
- 8 – moderate R12/TEV, zero intertidal cover, river mouths
- 9 – highest R12/TEV, low-moderate intertidal cover, river mouths
- 10 – high depth, fiords and sounds

Comparing the groups to the Hume classification (Table 44):

- 1 – Mix of predominantly D and E class systems
- 2 – corresponds exactly to class A
- 3 – consists of a mix of E and F class systems
- 4 – consists of a mix of B, E and F class systems
- 5 – corresponds to a portion of class D, with some H
- 6 – consists of a mix of B-F class systems, but primarily E and F class systems
- 7 – corresponds to a selection of B class systems, with some D-F class systems
- 8 – corresponds to a selection of B class systems
- 9 – corresponds to a selection of B class systems
- 10 – corresponds strongly to class G

Table 44. Table illustrating the number of systems in each group that were of the different hydro classes outlined in the Hume et al. (2007) classification. This is based on model-based clustering, with $k = 10$.

Hume Hydro class	Group									
	1	2	3	4	5	6	7	8	9	10
A	0	37	0	0	0	0	0	0	0	0
B	0	0	0	10	0	4	47	27	33	0
D	83	0	0	0	23	5	2	0	0	0
E	18	0	24	10	0	13	3	0	0	0
F	8	0	41	17	0	12	6	0	0	0
G	1	0	0	0	1	0	0	0	0	9
H	2	0	0	0	5	0	0	0	0	2

This grouping scheme successfully distinguishes classes A, B and G from the remaining classes, placing the majority of class B into four groups, of which three are primarily class B. However, as with the Euclidean analysis of the same data, classes D-F are placed primarily into five groups of mixed class. Groups 2 and 10 are easily identified as closed systems (lakes and lagoons) and systems characterised by their great depth (fiords and sounds). Groups containing systems of class B, 4, 7, 8 and 9 describe a gradient in types of river mouth from low river input relative to their volume (group 4) to high input relative to their volume (group 9). However, group 4 and 7 also exist along the continuum of the remaining groups which make up the majority of classes D-F. Group 5 has the highest SCI and MCI, indicating very simple structures and is strongly aligned with class D, indicating open coastal embayments. Group 1 is also characterised by high SCI and MCI, but slightly lower than that of group 5 indicating a relatively simple structure that are almost completely open, but perhaps more recessed than those of group 5. Groups 3, 4, 6 and 7 have similar SCI and MCI, depth and intertidal cover, indicating systems that are structurally complex, with a partial gradient in closure regime from 3 (almost fully closed) to 7 (partially closed). These groups mainly differ, however, based on R12/TEV which is highest for group 7, which is closely aligned with river mouths, and could therefore be considered as a partially closed, or barricaded river mouth, and lowest for group 3, which are characteristically large harbours and harbour systems. This grouping system can therefore be classified as five super-groups containing sub divisions:

(1) Completely closed systems – group 2
(2) Bays (low structural complexity)
    a. Completely open – group 5
    b. Slightly recessed with narrower mouth – group 1

(3) Structurally complex, partially closed systems
    a. Narrowest mouth, lowest river input relative to volume – group 3
    b. Moderate width mouth, moderate river input – groups 4 and 6
    c. Widest mouth, highest river input – group 7
(4) River Mouths
    a. Moderate river input relative to volume – group 8
    b. Highest river input relative to volume – group 9
(5) Depth dominated systems, such as fiords and sounds – group 10

3.4.4 – Rule-based summary of grouping schemes

Based on variable ranges group 2 and group 10 are distinct from the other groups based on STP/TEV and mean depth, respectively (Table 45). Group 2 has zero STP, and all other groups have a minimum STP of 0.01336, whilst group 10 ranges in depth from 50.2-141.1, whereas all other groups have a maximum depth of 38.3 (Table 45). Therefore a suitable rule set would begin with:

- STP < 0.0067
    - Y – Group 2
    - N – Mean Depth > 44
        - Y – Group 10
        - N – decision tree (Figure 49)

Table 45. Physical variable ranges for each group based on the ten group scheme identified by model-based clustering.

Group	R12/TEV	STP/TEV	SCI	MCI	Mean Depth	% IA	CLA/EWA
1	0, 0.065	0.047, 1.558	0.12, 0.84	0, 0.33	1.2, 31.4	0, 80.5	0, 223.2
2	0, 8.968	0, 0	0.13, 0.92	0, 0.03	0.1, 1	0, 95	0, 3305.8
3	0.001, 0.051	0.308, 1.591	0.08, 0.63	0, 0.04	1, 5.4	11.4, 94.8	1, 50.5
4	0.004, 1.184	0.426, 0.962	0.1, 0.56	0, 0.07	1, 6.4	7.7, 96.1	5.5, 1301.3
5	0, 0.09	0.039, 0.973	0.57, 0.82	0.1, 0.35	1.6, 38.3	0, 95.3	0.6, 170.2
6	0.003, 0.375	0.327, 13.135	0.13, 0.65	0, 0.13	0.2, 4.4	0.1, 100	2.1, 544.9
7	0.052, 5.059	0.58, 1	0.09, 0.42	0, 0.18	1.1, 5.1	0, 100	42.3, 5309.9
8	0.307, 3.81	0.422, 0.971	0.1, 0.38	0, 0.06	1.5, 5.1	0, 0	492, 4704.7
9	0.329, 27.13	0.341, 1.357	0.08, 0.41	0, 0.06	1.9, 5	0, 32.5	615.7, 39156.9
10	0, 0.003	0.013, 0.036	0.1, 0.34	0, 0.07	50.2, 141.1	0, 3.7	3.7, 19

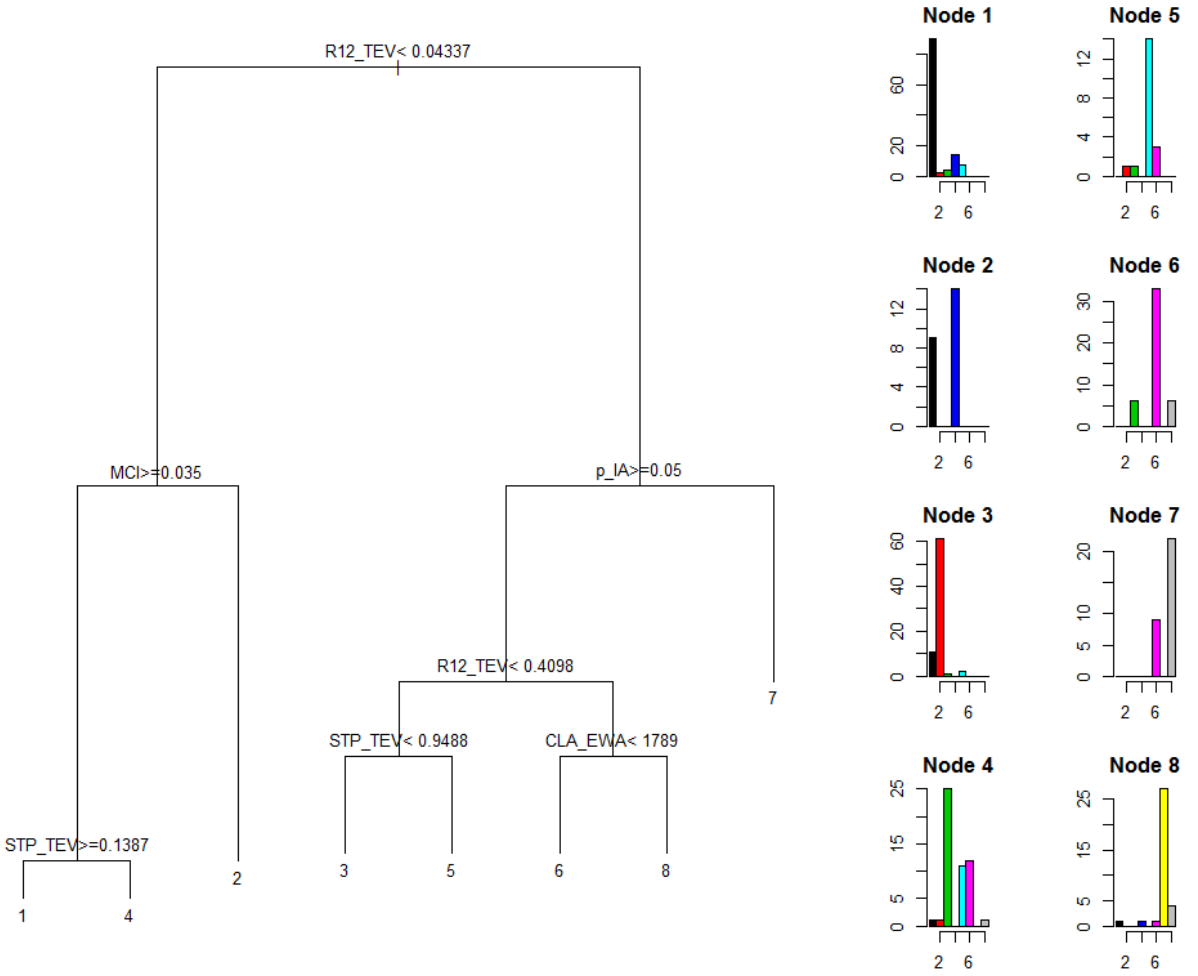

Figure 49. Classification tree for the ten group scheme identified by model-based clustering. Groups 2 and 10 have been removed from the analysis and labels 2-8 refer to groups 3-9, respectively. Histograms to the right of the classification tree indicate the class make-up of each terminal node, where node numbering proceeds from left to right along the base of the classification tree.

Table 46. Classification accuracy of the classification tree developed for the ten group scheme identified by model-based clustering.

Group	Classification accuracy (%)	No. in Group
1	80.4	112
2	100.0	37
3	93.8	65
4	67.6	37
5	48.3	29
6	41.2	34
7	56.9	58
8	100.0	27
9	66.7	33
10	100.0	11
Total	75.4	443

The classification rule set didn't perform as well as previous rule sets, and failed to achieve a reasonable classification accuracy for groups 5, 6 and 7 (Table 46). Classification tree partitions, in addition to the rule separating groups 2 and 10, are formed along STP/TEV (3), MCI (1), mean depth (1), % IA (1), CLA/EWA (1) and R12/TEV (2) axes (Figure 49). The resulting rules are:

(1) STP/TEV ≥ 0.0067, mean depth < 44, R12/TEV < 0.04337, MCI < 0.035, STP/TEV < 0.1387
(2) STP/TEV < 0.0067
(3) STP/TEV ≥ 0.0067, mean depth < 44, R12/TEV < 0.04337, MCI ≥ 0.035
(4) STP/TEV ≥ 0.0067, mean depth < 44, R12/TEV ≥ 0.04337, % IA ≥ 0.05, R12/TEV < 0.4098, STP/TEV < 0.9488
(5) STP/TEV ≥ 0.0067, mean depth < 44, R12/TEV < 0.04337, MCI < 0.035, STP/TEV ≥ 0.1387
(6) STP/TEV ≥ 0.0067, mean depth < 44, R12/TEV ≥ 0.04337, % IA ≥ 0.05, R12/TEV < 0.4098, STP/TEV ≥ 0.9488
(7) STP/TEV ≥ 0.0067, mean depth < 44, R12/TEV ≥ 0.04337, % IA ≥ 0.05, R12/TEV ≥ 0.4098, CLA/EWA < 1789
(8) STP/TEV ≥ 0.0067, mean depth < 44, R12/TEV ≥ 0.04337, % IA < 0.05
(9) STP/TEV ≥ 0.0067, mean depth < 44, R12/TEV ≥ 0.04337, % IA ≥ 0.05, R12/TEV ≥ 0.4098, CLA/EWA ≥ 1789
(10) STP/TEV ≥ 0.0067, mean depth > 44

3.4.5 – Summary

The grouping scheme identified using model-based clustering could be considered as five major groups distinguishing among; (1) closed systems, (2) bays, sub-divided into completely open and recessed groups, (3) open river mouths, which were sub divided into low and high river input groups, (4) structurally complex partially closed systems, which were sub-divided into four groups based on closure extent and river input and finally (5) systems characterised by great depth. Some of the groups were characterised by certain physical attributes, but those groups with high structural complexity were difficult to distinguish from one another as there was considerable overlap in physical characteristics among them. It seems that these groups were actually modelling a gradient in R12/TEV, which arguably could be modelled using fewer groups, which would simplify interpretation considerably.

3.5 – Conclusion

Overall the group structures identified by incorporating the additional variables of mean depth and % intertidal area lead to the identification of groups that differed based on morphology, whereas CLA/EWA, because of its strong correlation with R12/TEV, emphasised the differences in river input between systems. All three classification methods were similar in that closed systems were distinguished from open river mouths and bays, with additional groups accounting for systems that were partially closed. However, the exact breakdown of groups differed between classification methods as summarised in Table 47. The greatest similarity between methods was between Euclidean and model based analyses that grouped bays by SCI and MCI, differentiated fiords and sounds from complex harbour systems, and grouped partially closed systems by structural complexity and river input (Table 47). In contrast the random forests method split bays into shallow and deep variants, and grouped fiords and sounds with harbour systems. This is likely associated with the fact that the same datasets are used for the Euclidean and model-based analyses, and hence the distance between or dissimilarity of points is the same for these analyses, whereas for the random forests method the distance measure is determined in a completely different manner, and provides an alternative way to distinguish among systems.

Table 47. Summary of grouping schemes developed from Euclidean hierarchical clustering (k = 9), random forests hierarchical clustering (k = 9) and model-based clustering (k = 10).

Method	Closed	Open		Partially Closed	
		River Mouths	Bays	Harbours, Inlets, Estuaries	Fiords and Sounds
Euclidean	1 Group	1 – Low % IA 2 - High % IA	1 - Open and simple 2 - Recessed and complex	1 – Low structural complexity 2 – High structural complexity, low R12/TEV 3 – High structural complexity, high R12/TEV	1 Group
Random Forests	1 Group	1 Group	1 - Deep 2 - Shallow	1 – High R12/TEV 2 – Low structural complexity, high % IA 3 – High structural complexity, moderate % IA 4 – High structural complexity, high % IA 5 – High structural complexity, predominantly subtidal	NA
Model Based	1 Group	1 - Low R12/TEV 2 - High R12/TEV	1 - Open and simple 2 - Recessed and complex	1 – High structural complexity, low R12/TEV 2 – High structural complexity, moderate R12/TEV 3 – High structural complexity, high R12/TEV	1 Group

As highlighted in Table 47, the additional variables of mean depth and % IA in combination with SCI and MCI enabled systems of different morphologies to be distinguished,

with many of the sub-divisions characterised either by depth, SCI or % IA. In contrast the large scale groupings (i.e. the super groups in Table 47) predominantly refer to prevailing inflow/outflow regime (STP/TEV and R12/TEV) and so the grouping schemes highlighted by these analyses could be considered as primarily driven by inflow/outflow regimes, with morphology as a secondary consideration. The additional variable of CLA/EWA largely replicates R12/TEV and therefore differences in CLA/EWA can equally be interpreted as differences in R12/TEV. This also likely places greater emphasis on river input as a discriminatory variable for Euclidean and model-based analyses as it has the effect of essentially doubling the weight of the R12/TEV variable in the calculation of Euclidean distances. Removing this variable may therefore be advisable, as no additional information in terms of group structure is gained by using it.

Examining the rule-based systems that were developed to approximate each grouping scheme, the splits occur at different locations in multivariate space for different analysis methods, with the only similarities across methods being 0.0067 for STP/TEV and 44 for mean depth (Table 48). For the remaining variables no partitions were shared among methods.

Table 48. Locations of splits for each of the analysis methods and group numbers investigated along each of the seven physical variable axes. Values in parentheses indicate the location in the classification tree where the rule was implemented.

Cluster analysis type	Split location						
	STP/TEV	R12/TEV	SCI	MCI	% IA	Mean Depth	CLA/EWA
Euclidean (k=9)	0.0067 (1)		0.345 (7)	0.125 (3)	29.6 (5) 1.25 (5)	44 (2)	240 (4) 38.25 (6)
Random Forests (k=9)	0.0067 (1) 0.4193 (3)	0.009 (6)	0.485 (4) 0.405 (5) 0.435 (5)				456 (2) 44.8 (4)
Model-based (k=10)	0.0067 (1) 0.139 (5) 0.949 (6)	0.043 (3) 0.410 (5)		0.035 (4)	0.05 (4)	44 (2)	1789 (6)

4.0 – Conclusion

The analyses performed herein provide an overview of the methods currently available for clustering data, and highlight the similarities and differences among methods with regards to clustering NZ coastal hydrosystems. Each method highlighted alternative ways to group the data, and in most cases the groupings reflected genuine differences among systems that were easily identifiable. In addition, in both the initial (four variables) and the latter (seven variables) analyses the grouping schemes displayed some similarity with the Hume et al. (2007) classification, in particular for open (bays and river mouths) and closed (lakes and lagoons) systems. Partially closed systems were often classified in a different manner to that in the Hume classification, and the statistical grouping schemes for these systems often varied considerably among methods. However, in all cases the partially closed groups were characterised by one or more of the physical variables, and therefore the differences among methods simply reflect alternate ways to cluster the data.

In contrast to other statistical analyses there is no single "Gold-standard" analysis method for clustering data, and all three methods used in this report provide a valid assessment of the grouping structure of the data. In some cases specific methods may provide a greater advantage given the types of data being analysed and/or the objectives of the cluster analysis. However, in most cases an extensive exploration of the methods, and the additional analysis decisions regarding data transformation, linkage criterion and number of groups, will be required to identify the classification scheme that best meets the objectives of the clustering exercise and/or reflects current expert opinion. Given that each method is subject to a number of pre-analysis decisions it could be argued that the analysis type that removes the most of these decisions (i.e. makes it most independent of human input) is the most advantageous (or defensible). Both model-based and random forests hierarchical clustering methods remove one or more of the analysis decisions, which the Euclidean hierarchical analysis was sensitive to. As the random forests method is not governed by the scale of the variables (i.e. splits are chosen based on the relative differences among points, rather than the absolute differences) it is perhaps the most appropriate for this dataset given the large disparity in the scales and variances of the physical variables. This leaves only linkage type and number of groups as variables to be investigated. The model-based analysis is advantageous in that there are no linkage types, and the number of groups is decided based on BIC. However, given the necessity of transforming each variable prior to the cluster analysis, uncertainty over transformation type can lead to many more alternate transformed datasets on which to perform the analysis. For example suppose there are

two potential transformations for each variable as was the case for many of the variables in this dataset. For a four variable dataset the number of possible combinations of transformed variables, and hence the number of potential transformed datasets and grouping schemes is 2^4=16. Comparatively for the random forests analysis given three linkage types (average, complete and Ward's) and a choice of five group numbers there is a total of 15 potential grouping schemes and so for smaller datasets (fewer physical variables), the two are approximately equal. However, when the number of variables rises from four to seven, the number of combinations for model-based rises from 16 to 2^7=128, whereas for random forests the number of possible sets remains approximately the same. Therefore the random forests method provides the best possible analysis method, or one with the fewest potential sources of uncertainty, given the types of data in the NZ coastal hydrosystem dataset.

5.0 – Recommendations

- The random forests hierarchical method is the most advantageous given the dataset.
- A minimum group number should be chosen that meets the objectives of the clustering exercise, and the maximum group number should be set once the grouping structures become overly complicated relative to the data.
- The range of group numbers should be investigated, and grouping structures judged based on specialist knowledge to identify the number of groups that provides the most distinction between systems without adding unnecessary complication
- Ward's linkage criterion appeared to produce groups with minimal overlap, but average and complete linkage methods should also be investigated.
- Given the dataset the single linkage criterion does not appear appropriate due to the creation of one large group, with very few observations (~1-2) in remaining groups.
- Once a number of groups and linkage criterion is chosen, multiple runs of the random forests procedure should be performed to identify the group membership (i.e. uncertainty thereof) of each system across groups.
- An absolute system label can be identified as the modal group for each system, or a primary/secondary group labelling scheme can be utilised.
- Group characteristics should be identified via the mean and ranges of physical variables within each group.
- An approximate rule-based system can be identified using classification trees, and/or examination of physical variable ranges

- For management purposes, grouping structures could be compared against an independent criterion (e.g. presence/absence of some biological entity such as seagrass or Inanga spawning sites) as a measure of how well this group structure based on physical variables reflects biological traits within each system group.

6.0 – References

Breiman, L. (2001). Random forests. *Machine learning*, **45**: 5-32.

Cohen, J. (1960). "A coefficient of agreement for nominal scales". *Educational and Psychological Measurement*, **20**: 37–46.

Cutler, D.R., Edwards Jr, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., & Lawler, J.J. (2007). Random forests for classification in ecology. *Ecology*, **88**: 2783-2792.

De'ath, G. & Fabricius, K.E. (2000). Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, **81**: 3178–3192.

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, ***97***: 611-631.

Fraley, C., Raftery, A.E., Murphy, T.B. & Scrucca L. (2012). mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation Technical Report No. 597, Department of Statistics, University of Washington.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). The Elements of Statistical Learning (2nd ed.); *Section 14.3.12 Hierarchical Clustering*. New York: Springer. pp. 520–528. Retrieved 28/07/2014.

Hume, T.M., Snelder, T., Weatherhead, M. & Liefting, R. (2007). A controlling factor approach to estuary classification. *Ocean and Coastal Management*, **50**: 905-929.

Landis, J.R. & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**: 159-174.

Liaw, A. & Wiener M. (2002). Classification and Regression by randomForest. *R News*, **2**: 18-22.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Shi, T. & Horvath, S. (2006). Unsupervised Learning with Random Forest Predictors. *Journal of Computational and Graphical Statistics*, **15**: 118-138.

Therneau, T., Atkinson, B., & Ripley B. (2014). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-8. http://CRAN.R-project.org/package=rpart

7.0 – Appendices

7.1 – Appendix A: Examination of data transformations

Initial examination of the data revealed that the variables considered in Section 2, R12/TEV, STP/TEV would potentially require transformation to reduce the influence of extreme values, and to centre the distributions of the data. Square-root, fourth-root and log (X+1) were trialled on each variable and their histograms examined.

For STP/TEV the raw data is heavily left-skewed, with the majority of datapoints less than 1, but with several extreme points (Figure A1). The fourth-root transformation produced the distribution with the least skew, whereas square-root and log(X+1) produce very similar distributions, but still displayed an element of left-skew (Figure A1).

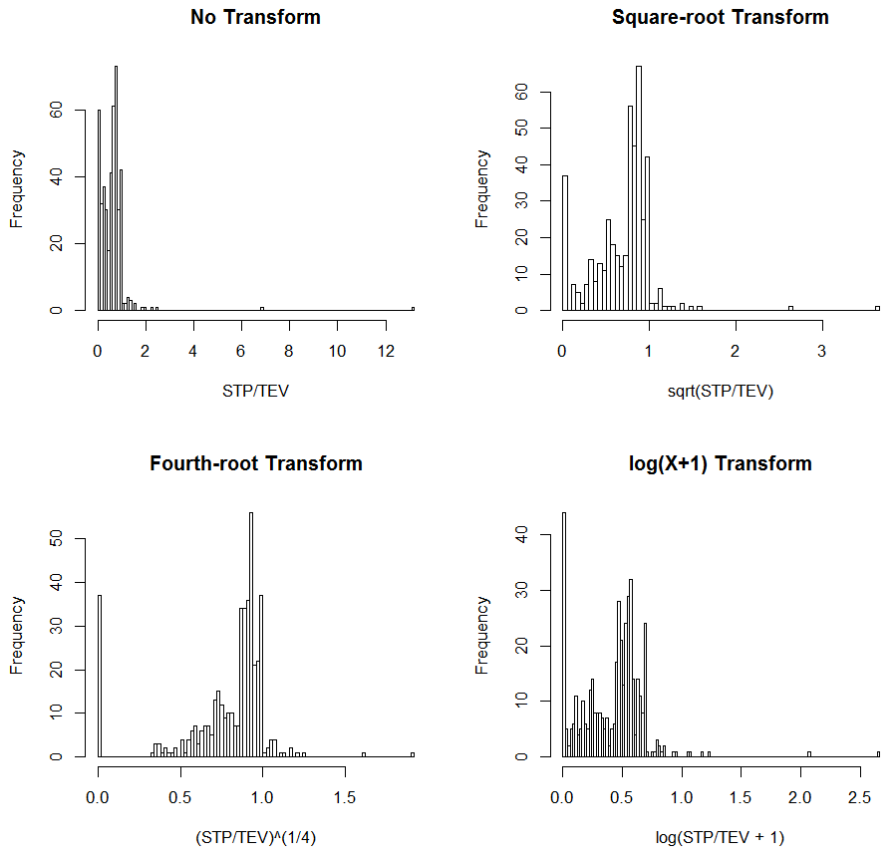

Figure A1. Histograms illustrating the distribution of raw and transformed STP/TEV data under three transformations: square-root, fourth-root and log(X+1)

For R12/TEV the raw data is heavily left-skewed, with the majority of datapoints less than 5, but with several extreme points (Figure A2). The fourth-root transformation produced the distribution with the least skew, but still wasn't central, but best controlled for the extreme datapoints in the tail of the distribution (Figure A2). Square-root and log(X+1) transformations

produce very similar distributions to each other, but neither were capable of reducing the amount of skew as much as the fourth-root transform (Figure A2).

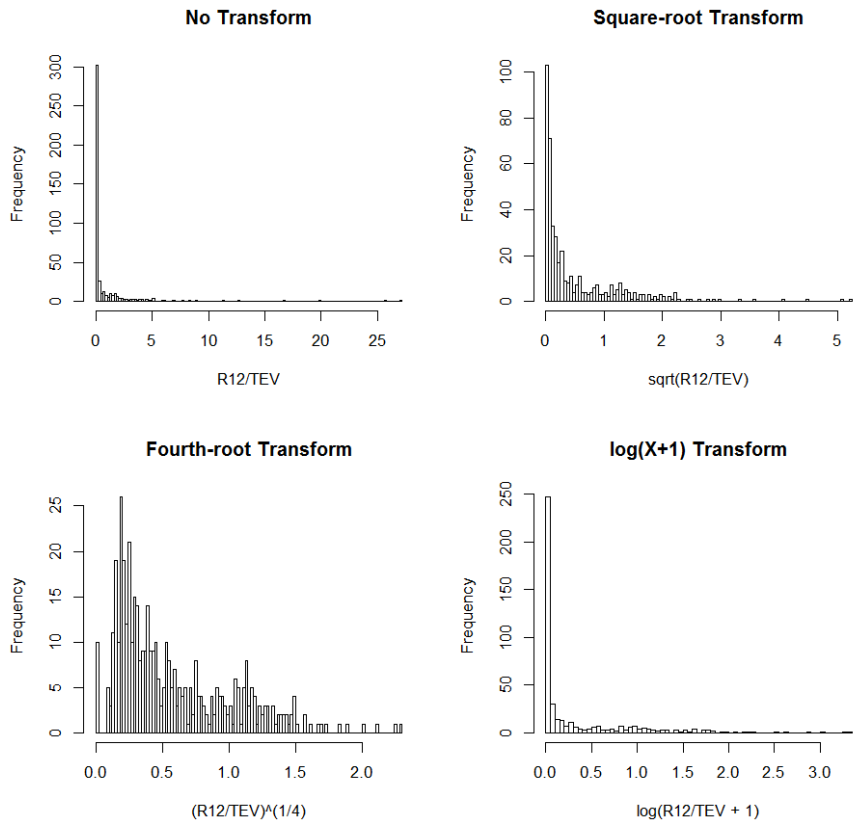

Figure A2. Histograms illustrating the distribution of raw and transformed R12/TEV data under three transformations: square-root, fourth-root and log(X+1)

The distributions of the additional variables added in section 3, mean depth, % IA and CLA/EWA, were examined for a range of appropriate transformations (Figures A3 – A5). For mean depth, the fourth-root and log(X) transformations successfully reduce the range of the data, and approximately centre the distribution (Figure A3). The fourth-root transform was chosen as the log transform creates outliers at the low end (i.e. very shallow observations) of the depth range, but both transforms seem appropriate for the data. For % IA no particular transform seemed to be absolutely necessary, but the fourth root and logistic transforms seem to highlight multiple peaks in the data distribution, perhaps suggestive of a group structure, and the logistic transform was chosen of the two (Figure A4). The CLA/EWA data is heavily left skewed, and the log(X+1) transformation was chosen as it achieved the most centrally distributed data, and appeared to illustrate multiple peaks, suggestive of a group structure (Figure A5).

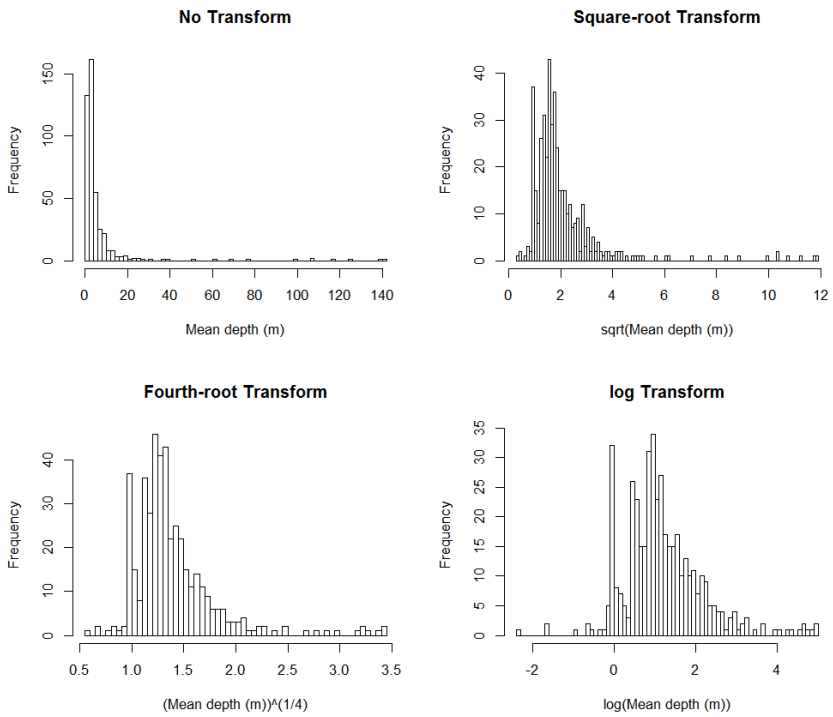

Figure A3. Histograms illustrating the distribution of raw and transformed mean depth data under three transformations: square-root, fourth-root and log.

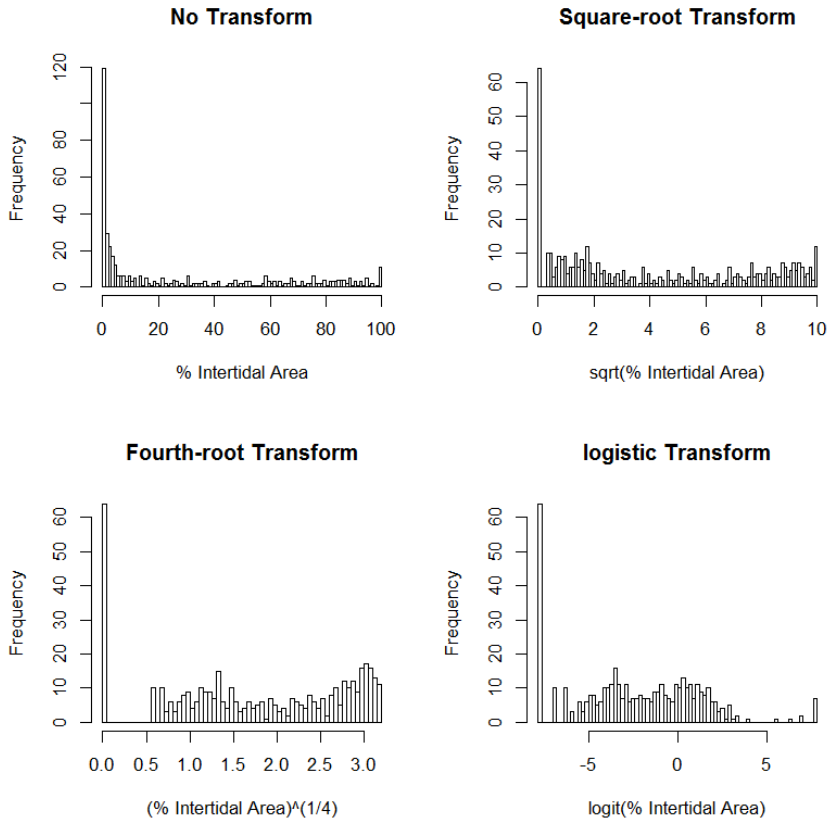

Figure A4. Histograms illustrating the distribution of raw and transformed % intertidal area data under three transformations: square-root, fourth-root and logistic.

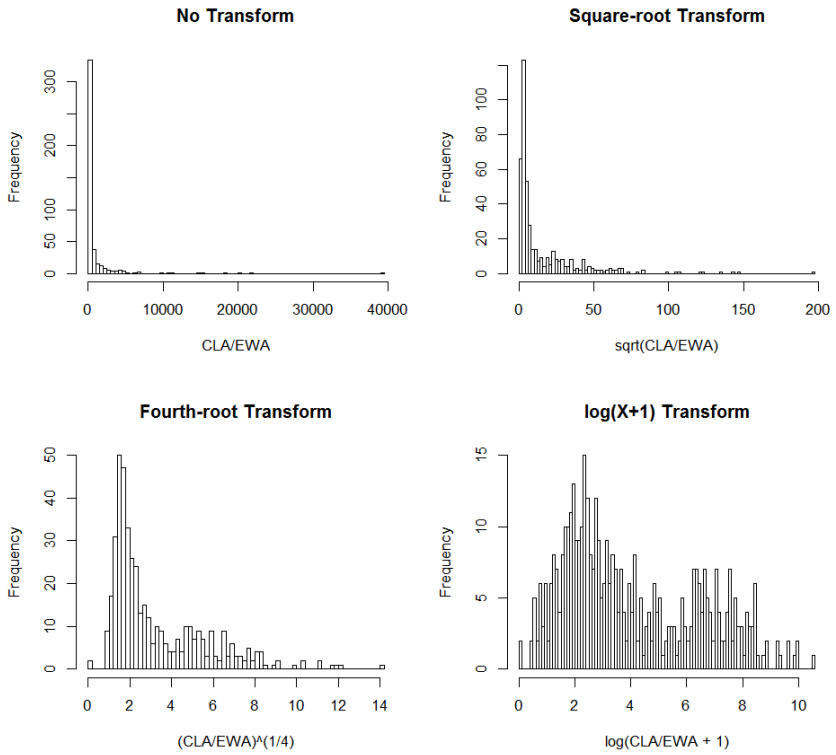

Figure A5. Histograms illustrating the distribution of raw and transformed CLA/EWA data under three transformations: square-root, fourth-root and log(X+1).

7.2 – Appendix B: The influence of alternate transformations and linkage on group structure

To examine how data transformation can alter grouping structure four datasets were examined; (1) all variables untransformed, (2) R12/TEV and STP/TEV were square-root transformed, whilst SCI and MCI were untransformed, (3) R12/TEV is fourth-root transformed whilst STP/TEV is square-root transformed and (4) R12/TEV and STP/TEV were fourth-root transformed. After transformation each variable was scaled and each dataset subjected to a Euclidean hierarchical cluster analysis with k=5. Groups were visualised by plotting the principal component axes of the 4th dataset, with points colour coded according to group label and vectors overlaid to illustrate the alignment of major physical variables.

There are noticeable differences among the resulting grouping structures, with the untransformed dataset showing only splits along the R12/TEV axis and the SCI/MCI axes (Figure B1). When R12/TEV is transformed, groups are identified as separate along the STP/TEV axes, as the importance of R12/TEV is reduced (Figure B1). The group identified with high SCI, MCI (green in panel A, C and D and blue in panel B, Figure B1) is almost the same across all datasets, and is therefore likely driven by SCI and MCI, which remains the

same across the four datasets. However, the remaining groups change considerably among the datasets (Figure B1).

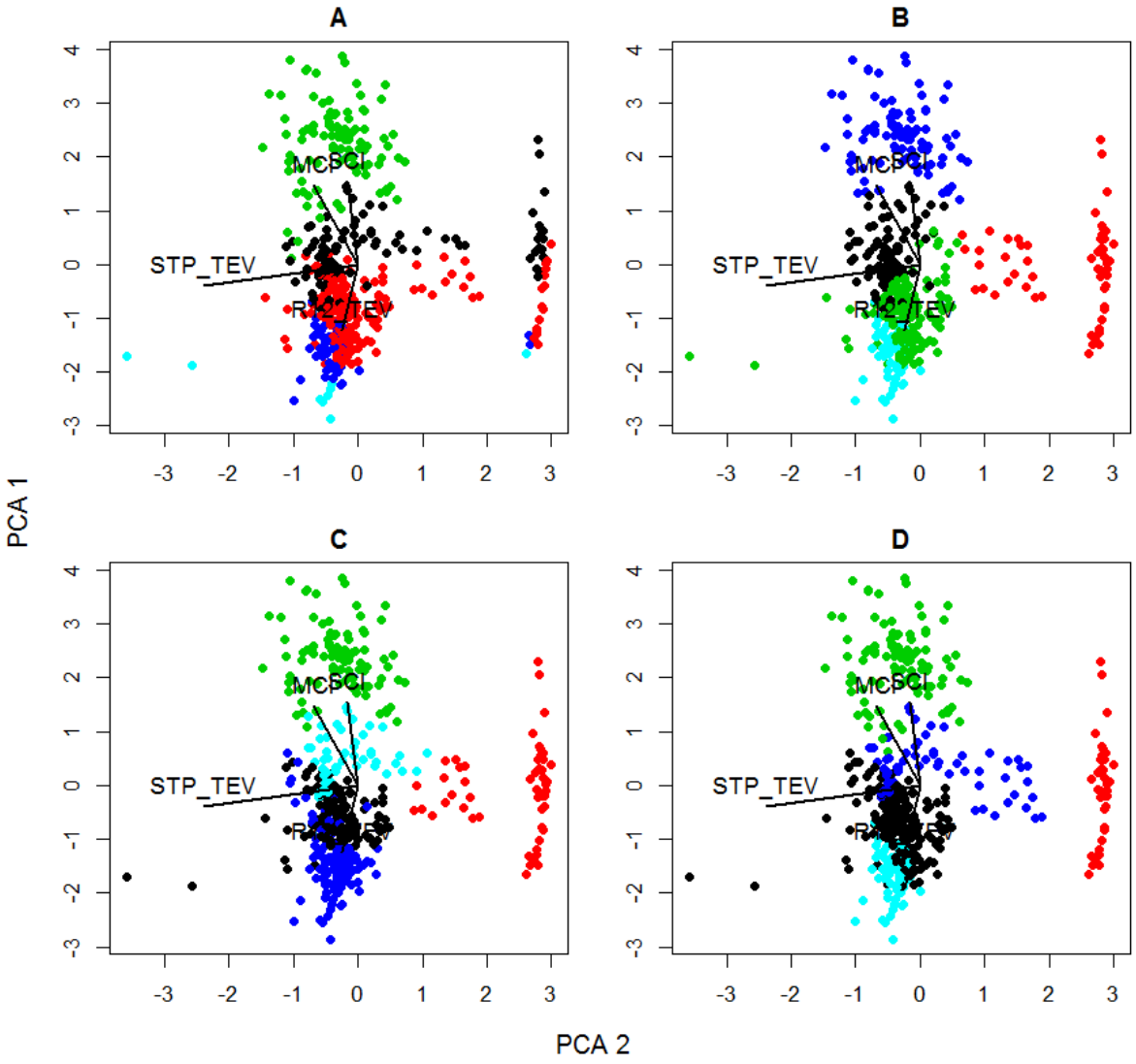

Figure B1. Principal component representations of the hydrosystem dataset based on scaled R12/TEV (fourth-root transformed), STP/TEV (fourth-root transformed), SCI and MCI physical parameters. Each panel is colour coded to illustrate a grouping structure for 5 groups based on Euclidean hierarchical clustering of (A) untransformed data, (B) square-root transformed R12/TEV and STP/TEV, (C) square-root transformed STP/TEV and fourth-root transformed R12/TEV and (D) fourth-root transformed STP/TEV and R12/TEV. In all tested datasets SCI and MCI were untransformed and all variables were scaled.

To identify how much group structure is likely to vary among linkage methods for hierarchical clustering four different linkage methods were investigated. The methods investigated were:

- Complete linkage – at each stage groups are combined by finding the two groups that have the minimum of the maximum distance between any two members of the

constituent groups (i.e. minimises the furthest "neighbour" within each combined group)

- Single linkage – at each stage groups are combined by finding the two groups that have the minimum distance between any two observations that are in different groups, (i.e. minimises the nearest "neighbour" within each combined group)
- Average linkage – aims for a compromise between single and complete linkage methods
- Linkage based on Ward's criterion – at each stage groups are combined that minimise the sum of the within-cluster variances, identifying compact clusters

Using the random forests distance measure each of these linkage criteria were trialled for a k=7 grouping scenario using R12/TEV, STP/TEV, SCI and MCI as the physical variable dataset. Groups were visualised by plotting the principal component axes of this data, with points colour coded according to group label (PCA analysis was performed on transformed and scaled data).

The PCA visualisation indicates that the single linkage method is not useful in this case as the vast majority of points are considered to be in a single group, with a few observations classed separately (Figure B2). This is likely due to an effect known as chaining that can adversely affect nearest neighbour linkage methods. Ward's linkage produces the most compact clusters (i.e. they aren't spread too far over multivariate space, judged from the PCA) and will therefore correspond to clusters with broadly similar characteristics (Figure B2). The Average linkage method is prone to placing outlier data into their own group and has several groups that contain only a few observations (Figure B2). The ward and complete linkage methods both produce reasonable clusters that are similar in some clusters, but different for others, and seem to highlight the clusters that are visible from the PCA visualisation. However, Ward's method seems to produce clusters with less overlap, but this is a marginal as both seem reasonable (Figure B2). There is approximately a 68% similarity in grouping structure among Ward's and the Complete linkage methods.

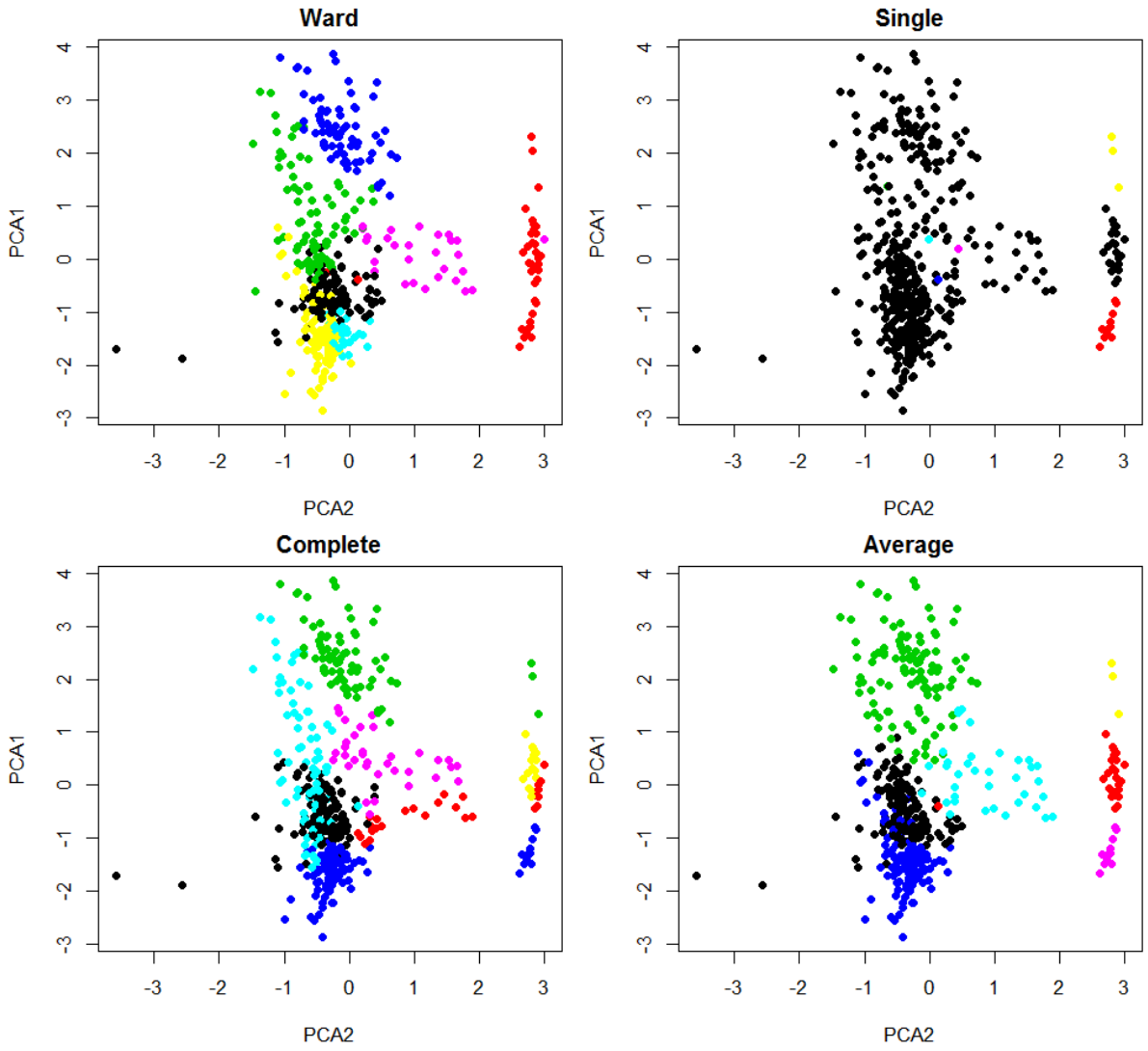

Figure B2. Principal component representations of the hydrosystem dataset based on R12/TEV, STP/TEV, SCI and MCI physical parameters which were transformed and scaled. Each panel is colour coded to illustrate a grouping structure for 7 groups based on random forests distances using one of four hierarchical linkage methods; Ward's (top left), Single Linkage (top right), Complete linkage (bottom left) and Average linkage (bottom right).

7.3 – Appendix C: Using stochasticity of random forests to determine group membership

The random forests method of producing distances is stochastic in that it varies between alternate runs of the distance calculations. This can lead to differences in group structure. This aspect can be used to identify each systems membership to any particular group, and whether it is in the core of a group, or intermediate between two or more groups. This is highlighted here by performing 1000 runs of the same random forests distance calculation and hierarchical cluster analysis, with k=4, to produce 1000 sets of group labels. Group labels are aligned across

all sets (i.e. to make sure 1 matches 1, as label 1 in one classification could contain exactly the same systems as another classifications label 2) and the proportion of times that the group labels matched 1, 2, 3 and 4 were calculated. The modal group labels were taken to be the final label for each observation. These group memberships are evaluated for each classification scheme presented in the main text and can be found in the separate Classifications Appendix. The resulting groups are plotted along SCI vs R12/TEV (fourth-root transformed) axes and SCI vs STP/TEV (fourth-root transformed) axes to visualise the model results and particular cases are highlighted.

Figure C1 illustrates the resulting group structure along with some particular examples of systems that show varying membership to their modal group. Particular systems such as Blind/Big Bay and Taramakau River show very strong membership to their modal group, and these systems occur in the centre of each of their respective clusters (Figure C1). Other systems such as Waikopua Creek, Kaiteretere Estuary and Opua Inlet System show a gradient of membership strength and all exist closer to the boundaries of groups 1 and 2, but show a particular affinity to one or the other group (Figure C1). Finally systems such as Bland Bay are on the boundary between group 2 and 3, and its group membership reflects this, with nearly half of all classifications placing it in group 3 rather than 2 (Figure C1). This can therefore be used to acknowledge the uncertainty associated with placing any particular observation within a group, as well as the overall uncertainty of any particular group, or grouping structure.

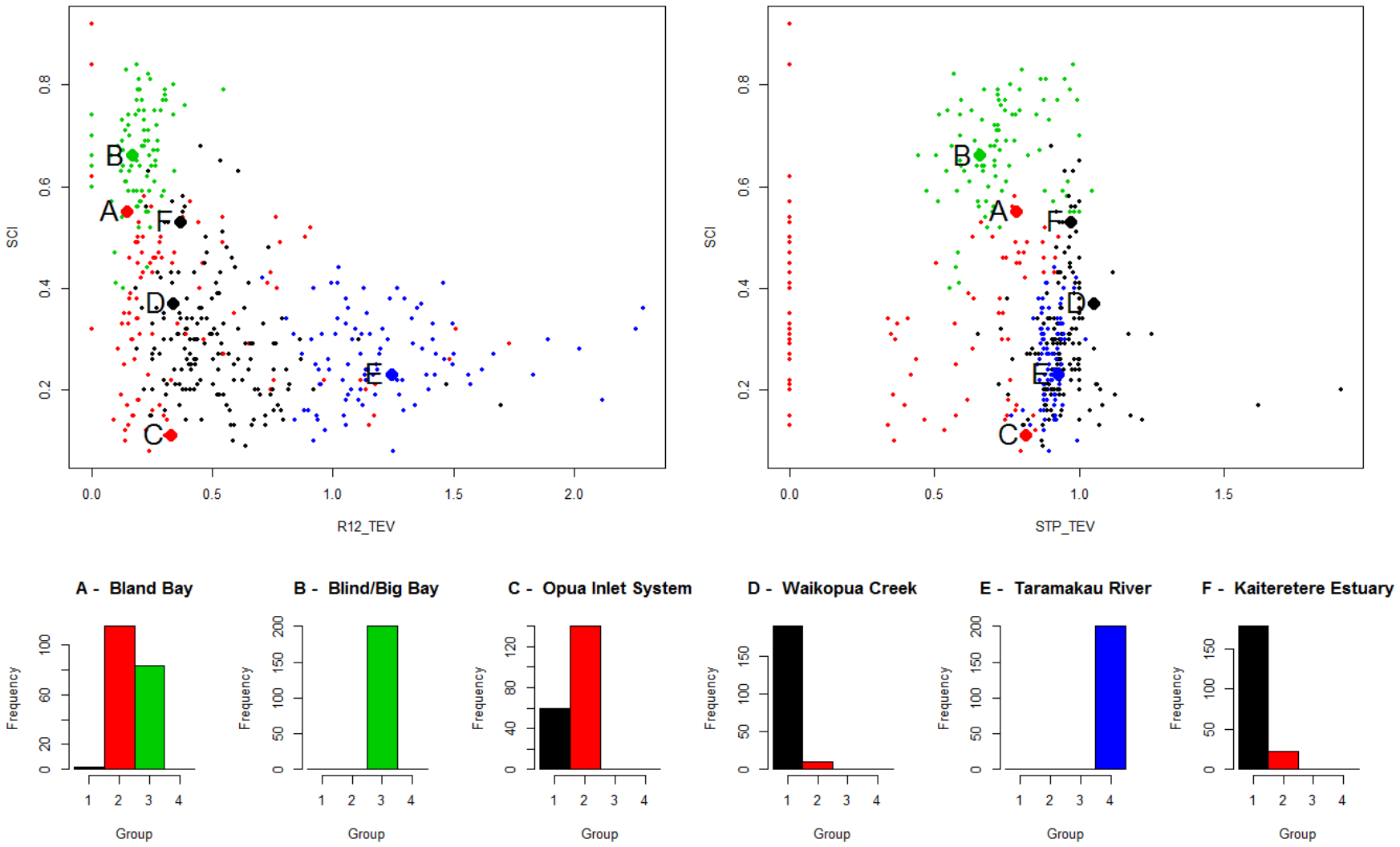

Figure C1. Plots of SCI against R12/TEV and STP/TEV colour coded by group membership according to a random forests hierarchical cluster analysis (k=4). Enlarged and labelled points refer to selected systems whose group memberships are plotted as histograms on the bottom row.

7.4 – Appendix D: Model-based clustering of unscaled data

In section 2.4, the model-based clustering method is applied to transformed and scaled data, but as a comparison an equivalent analysis was performed on unscaled data. The best model (as selected by BIC) consisted of 12 clusters, but a model with nine groups had ΔBIC=3.5 (Table D1). In comparison the best fitting models for scaled analyses were those with six to seven groups, suggesting additional groups are required to model the unscaled variables.

Table D1. BIC statistics of model-based clustering types for group numbers ranging from 2-20. NA indicates models failed to fit the data due to singularities in variance-covariance matrices. BIC statistics in **bold** indicate the two best fitting models based on BIC.

No. Groups	Model Type									
	EII	VII	EEI	VEI	EVI	VVI	EEE	EEV	VEV	VVV
2	69.2	66.1	734.5	1139.7	1492.3	1516.4	1143.5	1458.3	1607.3	1543.5
3	230.9	423.0	1316.0	1334.8	2015.4	2045.3	1370.7	2068.0	1633.9	2173.8
4	636.6	826.3	1647.5	1677.3	NA	NA	1734.1	2167.4	2183.2	NA
5	809.9	1052.6	1655.9	1799.1	NA	2220.3	1752.9	2125.7	2264.6	NA
6	1011.9	1199.9	1753.8	1864.8	NA	2324.1	1842.6	2150.6	2181.9	NA
7	1095.8	1298.3	1796.2	1905.3	NA	NA	1905.5	2124.2	2217.5	NA
8	1259.8	1397.0	1869.0	1978.6	NA	NA	1904.6	2223.0	2286.4	NA
9	1273.0	1425.3	1937.6	2009.4	NA	NA	2045.9	2299.3	**2329.0**	NA
10	1375.2	1469.7	1981.4	1963.5	NA	NA	1958.5	2308.3	2281.6	NA
11	1439.6	1469.7	2019.7	1959.9	NA	NA	1939.3	2324.5	2248.3	NA
12	1452.7	1495.4	2048.8	2062.9	NA	NA	2124.3	**2332.5**	2268.2	NA
13	1460.5	1525.8	2109.6	1977.0	NA	NA	2159.0	2249.8	2262.0	NA
14	1457.4	1604.6	2097.4	2061.2	NA	NA	2145.2	2261.9	2219.8	NA
15	1464.0	1605.3	2103.5	2065.1	NA	NA	2141.4	2239.7	2237.3	NA
16	1463.7	1617.7	2096.4	2089.7	NA	NA	2153.6	2157.5	2183.2	NA
17	1614.2	1615.5	2108.9	2130.4	NA	NA	2167.1	2150.8	2143.7	NA
18	1601.7	1620.3	2078.7	2115.3	NA	NA	2138.0	2088.3	2112.3	NA
19	1602.9	1638.0	2125.3	2126.2	NA	NA	2182.6	2008.1	2086.2	NA
20	1577.0	1617.2	2120.5	2110.7	NA	NA	2195.6	1959.4	2021.1	NA

Twelve groups

Examining the characteristics of the 12 group scheme (Figure D1) reveals that:

- Groups 2 and 12 have low STP/TEV, 4,8-11 have high STP/TEV and the remainder have moderate values
- Groups 1-7 have low R12/TEV, 8-9 have moderate R12/TEV and 10-12 have high R12/TEV
- Groups 1,8-12 have low SCI, 2-4,7 have moderate SCI and 5-6 have high SCI
- Groups 1,8,10-12 have low MCI, 7,9 have moderate MCI and 3,5-6 have high MCI

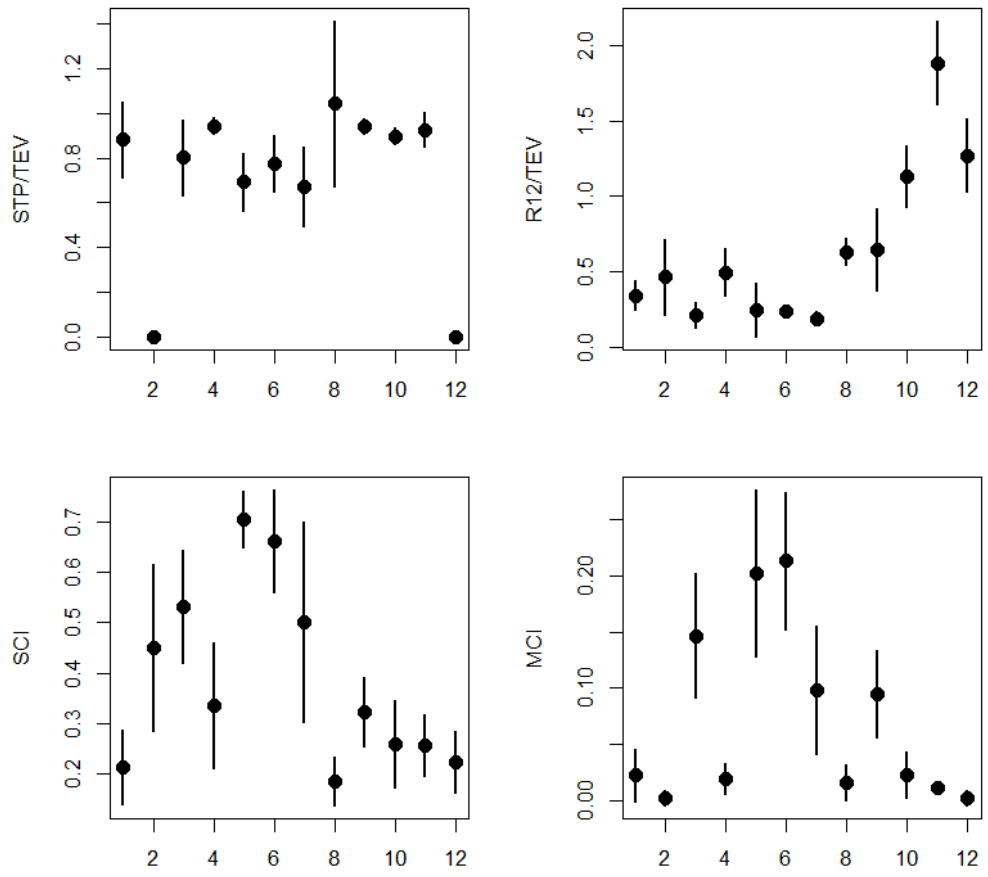

Figure D1. Group means (±1 SD) of the four physical variables with group labels identified by model-based clustering, with k=12. STP/TEV and R12/TEV are fourth-root transformed, whilst SCI and MCI are untransformed.

Table D2. Table illustrating the number of systems in each group that were of the different hydro classes outlined in the Hume et al. (2007) classification. This is based on model-based clustering of the unscaled data, with k=12.

Hume Hydro Class	Group											
	1	2	3	4	5	6	7	8	9	10	11	12
A	0	28	0	0	0	0	0	0	0	0	0	9
B	0	0	0	9	0	0	0	3	9	90	10	0
D	3	0	25	0	21	24	36	0	4	0	0	0
E	4	0	6	41	2	0	11	0	4	0	0	0
F	40	0	0	27	0	0	2	9	6	0	0	0
G	1	0	0	0	0	0	10	0	0	0	0	0
H	1	0	4	0	1	0	3	0	0	0	0	0

Comparing the grouping structure to the Hume classification (Table D2) reveals that

- Groups 2 and 12 correspond to variants of class A
- Group 10 strongly corresponds to class B
- Group 1 corresponds fairly to class F

- Group 3,5-7 correspond to class D, but group 7 has the highest mix of other classes
- Group 4 is a mix of E and F classes
- Group 8 is a small portion of classes B and F

7.5 – Appendix E: Quantifying uncertainty in group membership using model-based clustering

The model-based clustering analysis involves the construction of a statistical model to describe the groups in multivariate space. As a result the probability of any observation belonging to a particular group can be found by calculating the probability of that observation arising from the multivariate normal distribution that is assigned to that group. Calculating these across all groups can give an indication of how strongly an observation belongs to a particular group, and what objects are intermediate. These relative probabilities were extracted for the seven group (VEV) analysis of the dataset detailed in section 2.4. Plotting the resultant probabilities by group reveals that although group 1 only has a few actual members, many other points have at least partial membership due to the size and orientation of group 1 (Figure E1). Group 2 is completely distinct from the remaining groups, with all members having a high probability of membership to group 2, and points of other classes have zero probability of belonging to group 2 (Figure E1). The remaining groups all have a core membership zone, which tails off toward the edges or boundaries of the cluster with another (Figure E1). This information can therefore be used to identify core and intermediate systems for each class. These group membership probabilities are evaluated for each classification scheme presented in the main text and can be found in the separate Classifications Appendix.

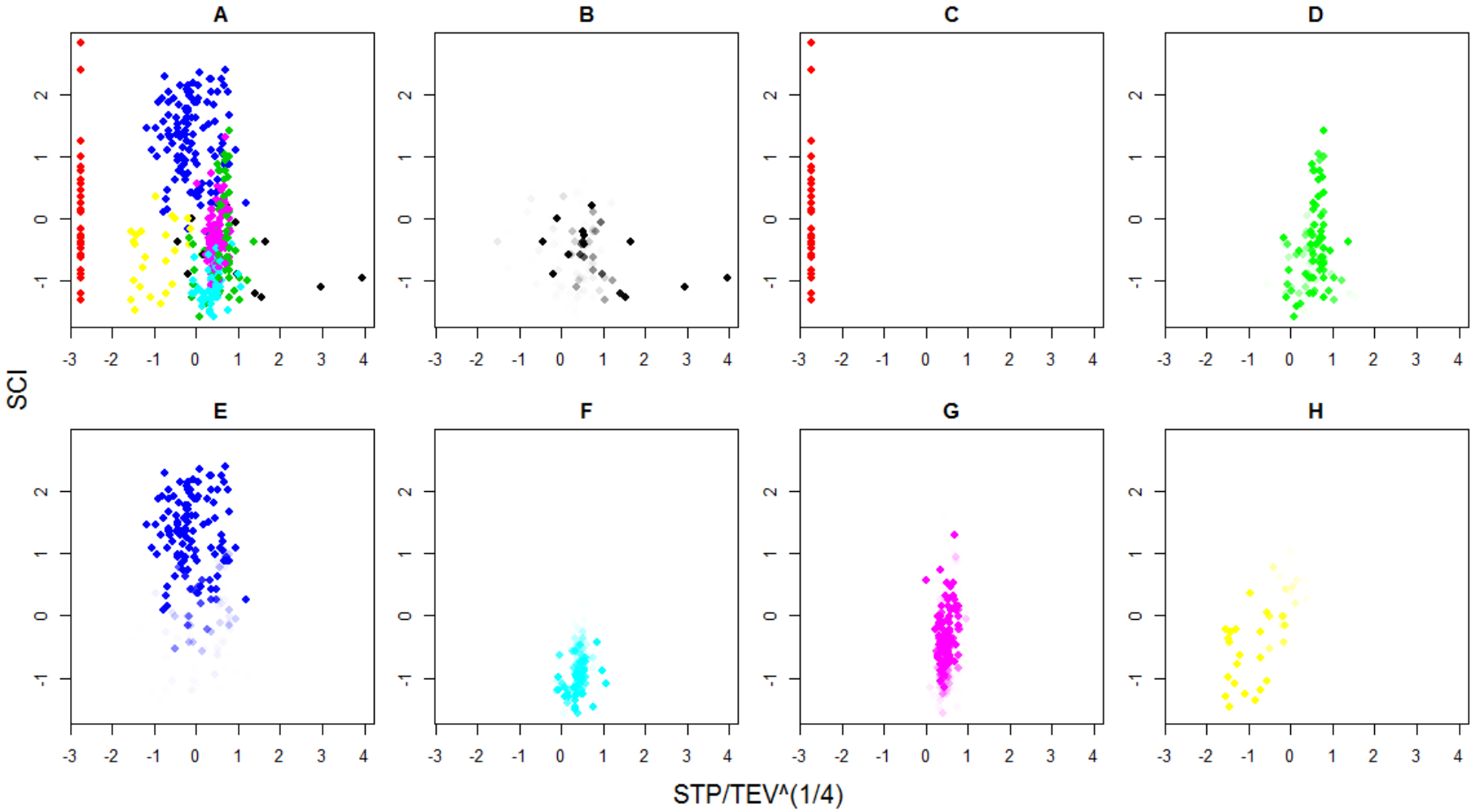

Figure E1. SCI plotted against STP/TEV (fourth-root transformed). Panel A illustrates the grouping scheme identified by model-based clustering with k = 7 and in each of the subsequent panels the transparency of the points are set to reflect the probability of group membership from 0 (transparent) to 1 (not transparent) for (B) group 1, (C) group 2, (D) group 3, (E) group 4, (F) group 5, (G) group 6 and (H) group 7.

7.6 – Appendix F: Alternate variable set

In addition to the variable sets investigated in sections 2 and 3, a third variable set consisting of all variables was examined. This dataset consisted of R12, STP, SCI, MCI, CLA, EWA, % IA, TEV and mean depth. The results presented in this appendix provide a simple overview of the results obtained when applying each of the methods to this dataset.

All variable distributions were initially examined and variables were suitably transformed and scaled (R12 – fourth root, STP – fourth root, SCI - none, MCI - none, CLA - log, EWA - log, % IA - logistic, TEV - log, mean depth – fourth root). Biplots of physical variables revealed that R12 and CLA were positively correlated, as was EWA and TEV. The correlation coefficients of all possible variable pairs was also examined (Table F1). CLA and R12 essentially provide the same information as they are 93% similar, and so CLA was removed from the analysis and R12 was retained (Table F1). Similarly EWA and TEV provide the same information on the size of the system and EWA was removed from the analysis and TEV retained (Table F1). Thus the remaining dataset consisted of seven variables: % IA, mean depth, STP, R12, TEV, MCI and SCI.

Table F1. Matrix of correlation coefficients between every possible pairing of physical variables. Values above 0.9 are in bold.

	EWA	CLA	% IA	Mean depth	STP	R12	TEV	MCI	SCI
EWA	1.00								
CLA	0.28	1.00							
% IA	0.18	-0.03	1.00						
Mean depth	0.34	-0.02	-0.38	1.00					
STP	0.84	0.31	0.15	0.47	1.00				
R12	0.25	**0.93**	-0.07	0.12	0.33	1.00			
TEV	**0.93**	0.22	0.00	0.66	0.86	0.24	1.00		
MCI	-0.09	-0.58	-0.18	0.32	-0.03	-0.48	0.08	1.00	
SCI	-0.17	-0.72	-0.13	0.14	-0.20	-0.63	-0.06	0.77	1.00

Euclidean hierarchical clustering

The cross-validation routine detailed in section 2.2.2 was applied to this dataset to determine the most appropriate number of groups (Figure F1).

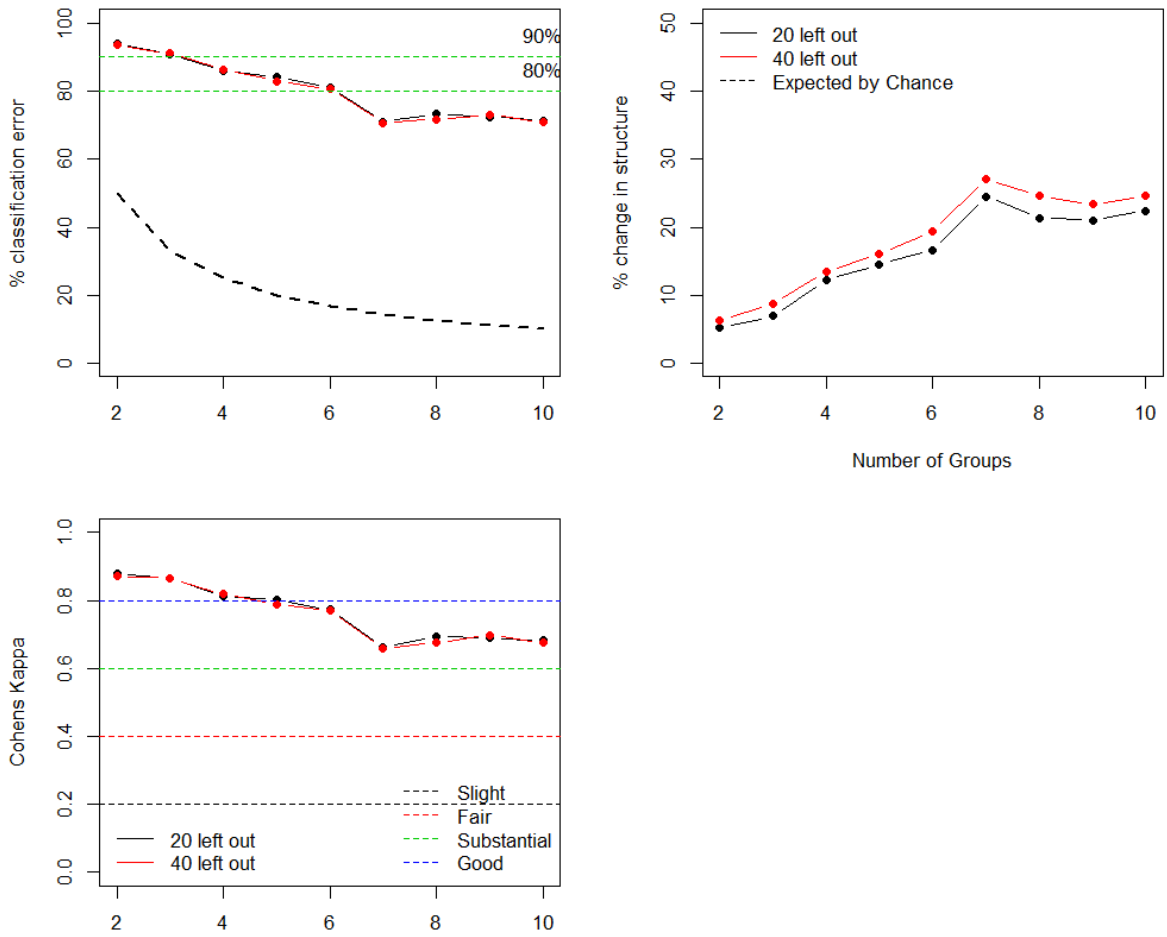

Figure F1. Cross validation metrics of % classification accuracy, % change in group structure and Cohens-κ plotted against number of groups.

The classification accuracy drops sharply between six and seven groups, and also drops below 80%. Cohens-κ drops below 0.8 between five and six groups (Figure F1). Therefore five to six groups are supported and both k = 5 and 6 scenarios were investigated but only the six group scheme will be presented here (see Classifications Appendix for other classifications). Examining the groups attributes (Figure F2) and the systems contained within each group it can be surmised that

- 1 - Lakes/Lagoons/Streams/Rivers, characterised by low STP and TEV
- 2 - Harbours/Harbour Systems/Inlets, characterised by high STP, moderate R12
- 3 - River/estuary/inlet/creek, characterised by high % IA
- 4 – Bays, characterised by high SCI
- 5 – River mouths characterised by high R12 and low STP
- 6 – Fiords/sounds characterised by high STP and TEV

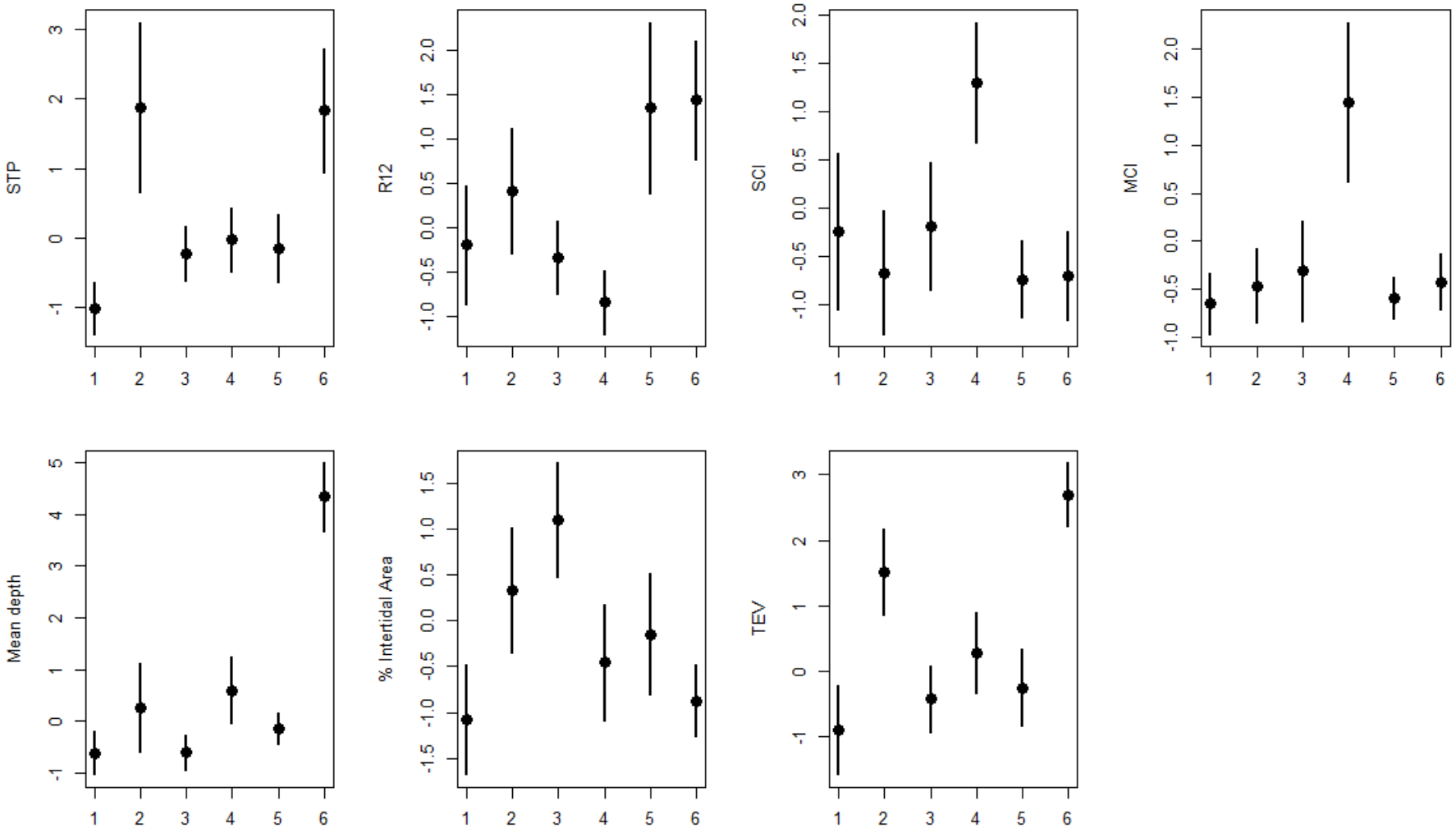

Figure F2. Means of the seven physical variables, STP, R12, SCI, MCI, mean depth, % IA and TEV for each group based on the six group scheme identified by Euclidean hierarchical clustering. Variable are transformed and scaled as detailed in section 7.6.

This grouping scheme is compared to the Hume classification in Table F2.

Table F2. Table illustrating the number of systems in each group that were of the different hydro classes outlined in the Hume et al. (2007) classification. This is based on Euclidean hierarchical clustering, with $k = 6$.

Hume Hydro class	Group					
	1	2	3	4	5	6
A	35	0	2	0	0	0
B	32	0	20	0	69	0
D	2	4	10	97	0	0
E	2	13	45	7	1	0
F	0	30	43	0	11	0
G	0	1	0	1	0	9
H	0	2	0	5	0	2

Random forests hierarchical clustering

Grouping schemes with $k = 5$, 7 and 9 were analysed using random forests hierarchichal clustering, but only the analysis with $k = 7$ is presented here (for other classifications see the Classifications Appendix). Examining the groups attributes (Figure F3) and the systems contained within each group it can be surmised that:

- 1 – River mouths, characterised by low TEV and low STP
- 2 – Lakes and lagoons, characterised by zero STP
- 3 – Harbours, harbour systems and inlets, characterised by high % IA and high TEV
- 4 – Bays, sounds and harbours, characterised by high STP, TEV and depth
- 5 – River mouths and harbour systems, characterised by high R12
- 6 – Bays, characterised by high SCI and MCI
- 7 – River mouths and estuaries, characterised by high % IA and low TEV

This group scheme is compared to the Hume classification in Table F3.

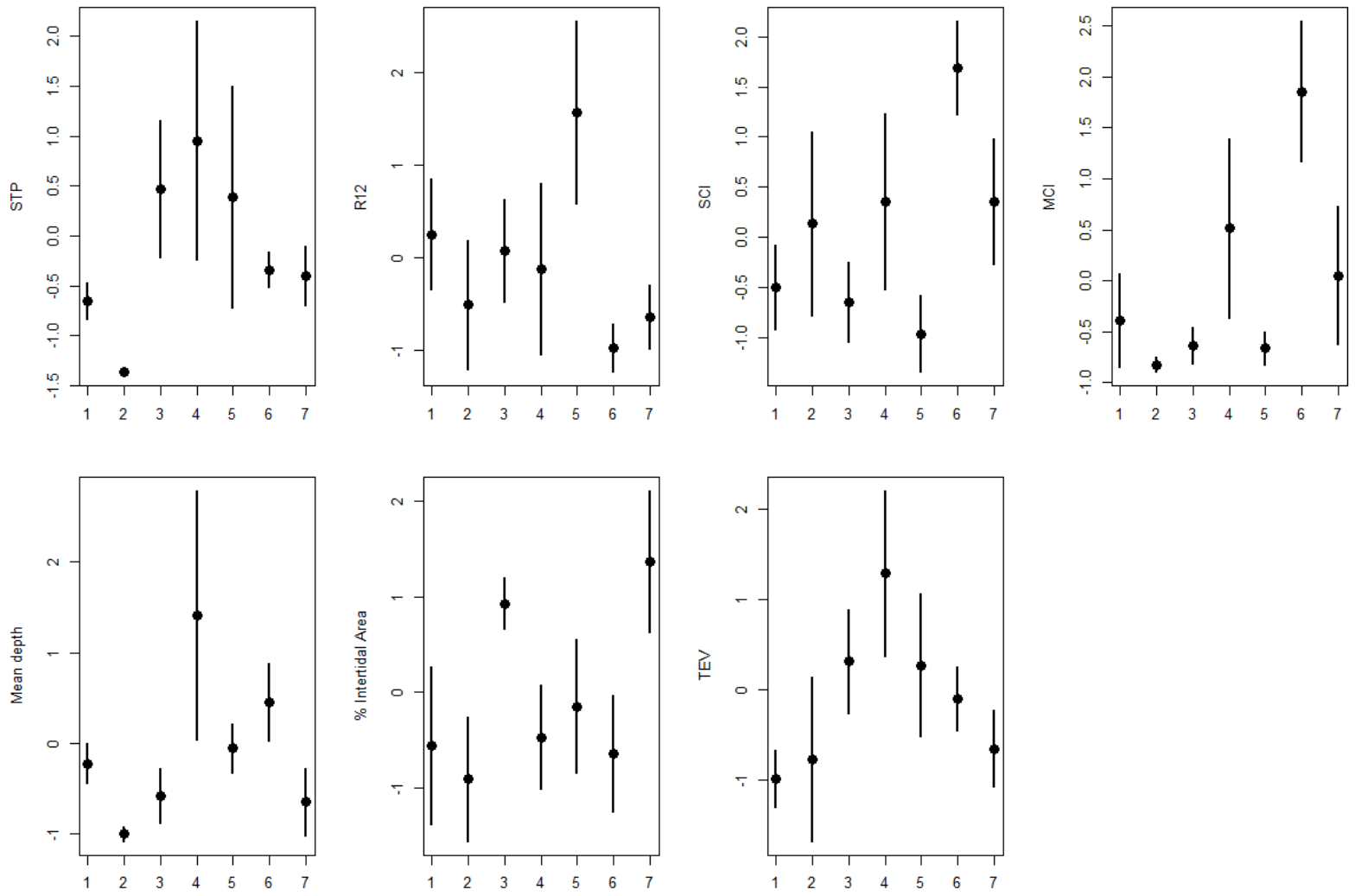

Figure F3. Means of the seven physical variables, STP, R12, SCI, MCI, mean depth, % IA and TEV for each group based on the seven group scheme identified by random forests hierarchical clustering. Variable are transformed and scaled as detailed in section 7.6.

Table F3. Table illustrating the number of systems in each group that were of the different hydro classes outlined in the Hume et al. (2007) classification. This is based on random forests hierarchical clustering, with $k = 7$.

Hume Hydro class	Group						
	1	2	3	4	5	6	7
A	2	33	0	0	0	0	2
B	64	0	8	0	43	0	6
D	2	0	0	41	0	55	15
E	3	0	20	14	0	0	31
F	5	0	52	6	17	0	4
G	0	0	0	11	0	0	0
H	0	0	0	8	0	1	0

Model-based clustering

All multivariate model types for k =2:20 were investigated and a model with $k = 7$ was the best-fitting based on BIC (Table F4).

Table F4. BIC statistics of model-based clustering types for group numbers ranging from 2-20. NA indicates models failed to fit the data due to singularities in variance-covariance matrices. BIC statistics in **bold** indicate those models with ΔBIC < 10 relative to the best model.

No. Groups	Model Type									
	EII	VII	EEI	VEI	EVI	VVI	EEE	EEV	VEV	VVV
2	-8402.5	-8234.2	-8157.6	-8105.4	-7800.3	-7801.8	-6967.0	-6275.9	-6106.5	-6082.2
3	-7733.0	-7553.0	-7555.3	-7382.9	-7110.6	-7038.5	-6825.8	-5735.8	-5622.9	-5530.3
4	-7477.4	-7212.8	-7405.8	-7180.0	-6806.6	-6767.3	-6716.9	-5239.8	-5158.0	NA
5	-7236.4	-7026.6	-7193.3	-7005.1	-6650.8	-6512.6	-6492.1	-5006.6	-4902.7	NA
6	-6976.0	-6814.9	-6978.7	-6783.4	NA	NA	-6278.5	-4994.1	-5067.1	-4997.0
7	-6904.3	-6639.3	-6917.8	-6597.5	NA	NA	-6233.8	-5062.4	**-4692.9**	-4998.1
8	-6855.5	-6560.4	-6867.4	-6505.8	NA	NA	-6174.9	-5037.7	-4951.4	-4996.0
9	-6704.3	-6492.3	-6734.2	-6415.2	NA	NA	-6114.6	-4905.8	-4745.6	-5035.7
10	-6702.6	-6463.5	-6674.5	-6396.7	NA	NA	-6089.7	-4974.9	-4936.9	-5161.6
11	-6647.2	-6381.5	-6664.1	-6341.5	NA	NA	-6068.6	-5105.9	-4955.4	NA
12	-6511.3	-6352.8	-6463.5	-6270.5	NA	NA	-5717.9	-5127.1	-5137.1	NA
13	-6471.6	-6347.7	-6455.4	-6266.2	NA	NA	-5715.4	-5229.7	-5358.4	NA
14	-6438.2	-6312.9	-6427.5	-6225.2	NA	NA	-6121.0	-5081.5	-5390.0	NA
15	-6448.0	-6316.3	-6434.5	-6221.7	NA	NA	-6121.5	-5384.4	-5427.0	NA
16	-6455.8	-6278.1	-6427.7	-6174.0	NA	NA	-5722.5	-5450.8	-5599.6	NA
17	-6435.9	-6255.9	-6332.9	-6112.5	NA	NA	-5664.9	-5375.7	-5583.2	NA
18	-6410.7	-6245.9	-6340.7	-6079.4	NA	NA	-5686.0	-5519.1	-5704.1	NA
19	-6396.9	-6243.5	-6374.5	-6077.5	NA	NA	-5666.2	-5496.9	-5788.2	NA
20	-6373.7	-6252.3	-6347.8	-6080.5	NA	NA	-5665.7	-5661.0	-5825.2	NA

Examining the groups attributes (Figure F4) and the systems contained within each group it can be surmised that:

- 1 – Lakes and lagoons, characterised by zero STP
- 2 – Harbours, harbour systems and inlets, characterised by moderate % IA, STP, R12 and TEV
- 3 – Bays and harbours, characterised by high SCI and MCI
- 4 – River mouths and inlets, characterised by high % IA, low MCI
- 5 – River, creek and estuary, characterised by high % IA, low TEV
- 6 – Sounds and harbour systems, characterised by high TEV and mean depth
- 7 – River mouths, characterised by high R12, low MCI

This group scheme is compared to the Hume classification in Table F5.

Table F5. Table illustrating the number of systems in each group that were of the different hydro classes outlined in the Hume et al. (2007) classification. This is based on model-based clustering, with $k = 7$.

Hume Hydro class	Group						
	1	2	3	4	5	6	7
A	36	0	1	0	0	0	0
B	0	18	0	7	16	0	80
D	0	0	89	0	10	14	0
E	0	9	15	23	19	1	1
F	0	26	6	31	10	8	3
G	0	1	0	0	0	10	0
H	0	0	1	0	0	8	0

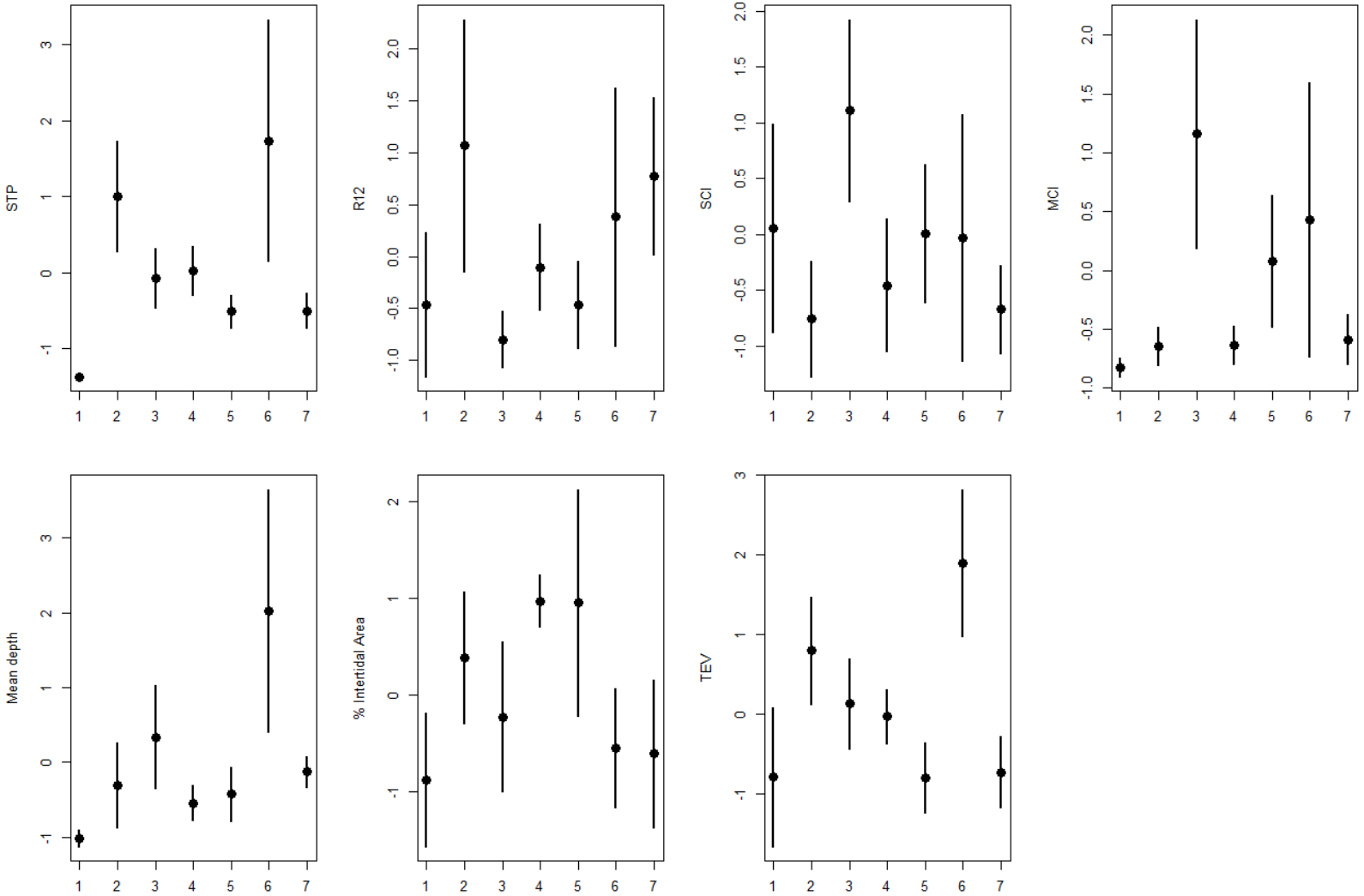

Figure F4. Means of the seven physical variables, STP, R12, SCI, MCI, mean depth, % IA and TEV for each group based on the seven group scheme identified by model-based clustering. Variable are transformed and scaled as detailed in section 7.6.