An improved community structure method for catchment classification

Author:
Tumiran, Siti Aisyah

Publication Date:
2018

DOI:
https://doi.org/10.26190/unsworks/3711

An Improved Community Structure Method for Catchment Classification

Siti Aisyah Binti Tumiran

A thesis in fulfilment of the requirements for the degree of

Doctor of Philosophy

School of Civil and Environmental Engineering

Faculty of Engineering

September 2018

PLEASE TYPE

THE UNIVERSITY OF NEW SOUTH WALES
Thesis/Dissertation Sheet

Surname or Family name: **Siti Aisyah Binti**

First name: **Tumiran** Other name/s:

Abbreviation for degree as given in the University calendar: **PhD**

School: **School of Civil and Environmental Engineering** Faculty: **Faculty of Engineering**

Title: **An improved community structure method for catchment classification**

Abstract 350 words maximum: (PLEASE TYPE)

Catchment classification plays an important role in many hydrologic applications. There are numerous approaches for catchment classification, with different bases, assumptions, and methods. Applications of the concepts of complex networks, specifically community structure, for catchment classification are just emerging. Among the many community structure methods, the edge betweenness (EB) algorithm, which applies hierarchical clustering and *modularity function*, is widely used for community detection. However, the EB method is susceptible to network resolution problem, which is caused by the modularity function, in identifying the best split. Thus, it is important to address the resolution problem to obtain reliable classification. Motivated by this, the present study proposes a Modularity Density-based EB (MDEB) algorithm by considering the *modularity density function*, instead of the *modularity function*. The performance of the MDEB method and the EB method is first compared for the widely-studied Zachary's Karate Club network. Both methods are then tested for classification of 218 catchments in Australia and 639 catchments in the United States. For each region, three different types of networks are studied, to take into account the network size and regional similarity: (1) the entire network; (2) smaller network sizes based on random selection; and (3) smaller network sizes based on drainage divisions or hydrologic units. Considering streamflow data in a single-variable sense, the MDEB method is found to perform better than the EB method in catchment classification for both the study areas at all of the above three scenarios. The MDEB method is then also attempted for classification in a multi-variable sense for the 218 catchments in Australia, by considering rainfall and potential evapotranspiration (PET) in addition to streamflow. Four different combinations of variables are considered in the multi-variable approach. The results suggest that the multi-variable classification is nearly similar to that based on the single-variable classification using streamflow, but at different correlation thresholds. The present study is a significant advancement in the application of complex networks for catchment classification, as it offers an improved community structure methodology (EB) and an approach–based on multiple variables. Such advancement is certainly promising for the development of a generic catchment classification framework.

FOR OFFICE USE ONLY Date of completion of requirements for Award:

INCLUSION OF PUBLICATIONS STATEMENT

UNSW is supportive of candidates publishing their research results during their candidature as detailed in the UNSW Thesis Examination Procedure.

Publications can be used in their thesis in lieu of a Chapter if:

- The student contributed greater than 50% of the content in the publication and is the "primary author", ie. the student was responsible primarily for the planning, execution and preparation of the work for publication
- The student has approval to include the publication in their thesis in lieu of a Chapter from their supervisor and Postgraduate Coordinator.
- The publication is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in the thesis

Please indicate whether this thesis contains published material or not.

☒ *This thesis contains no publications, either published or submitted for publication (if this box is checked, you may delete all the material on page 2)*

☐ *Some of the work described in this thesis has been published and it has been documented in the relevant Chapters with acknowledgement (if this box is checked, you may delete all the material on page 2)*

☐ *This thesis has publications (either published or submitted for publication) incorporated into it in lieu of a chapter and the details are presented below*

CANDIDATE'S DECLARATION

I declare that:

- I have complied with the Thesis Examination Procedure
- where I have used a publication in lieu of a Chapter, the listed publication(s) below meet(s) the requirements to be included in the thesis.

Name	Signature	Date (dd/mm/yy)

Postgraduate Coordinator's Declaration (to be filled in where publications are used in lieu of Chapters)

I declare that:

- the information below is accurate
- where listed publication(s) have been used in lieu of Chapter(s), their use complies with the Thesis Examination Procedure
- the minimum requirements for the format of the thesis have been met.

PGC's Name	PGC's Signature	Date (dd/mm/yy)

Originality Statement

‘I hereby declare that this thesis is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in this thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project’s design and conception or in style, presentation and linguistic expression is acknowledged. ‘

Signed: _____

Date: _____

Copyright Statement

Signed: _____

Date: _____

Authenticity Statement

‘I certify that the Library deposit digital copy is a direct equivalent of the final officially approved version of my thesis. No emendation of content has occurred and if there are any minor variations in formatting, they are the result of the conversion to digital format.’

Signed: _____

Date: _____

Acknowledgements

First and foremost, I would like to express my gratitude to my supervisors Assoc. Professor Sivakumar Bellie and Assoc. Professor Samsung Lim. Your constant encouragement, enthusiasm and wisdom have guided me be my best. Thank you for everything.

I gratefully acknowledge the support of the Ministry of Higher Education Malaysia and the Universiti Malaysia Sabah for funding this research. I am also thankful to the School of Civil and Environmental Engineering, UNSW for the opportunity to conduct this research and to all the helpful, supportive administration staff especially to Pattie (I will miss you!).

To all my friends in Water Research Center that already finished their PhD and to who still fighting, thanks for the enjoyable moments and it is such a pleasure to share this wonderful journey with amazing bunch of people. Thanks in particular to Ha Nguyen and Nazly Yasmin with your help, to Shakera for your endless support of motivations, to Nur Hidayaty and Nur Fadhilah for being my instant close friends which I never thought of any Malaysian that I would work closely in Down Under. To Clare Stephens, Ademir, Suresh, Hung Pham, Yating, Ruth Fisher, Hasan, Rounak, Xia, James, Xudong, Rebecca, Thou, Philipa and everyone. You guys will be remembered!

I also feel grateful of having emotionally support throughout my PhD journey specifically my parents, Tumiran Wasidin and Rumsinah Baba@Kasdan, my parents in-law, Harun

Saidi and Jawarsir Su, for their endless prayers and blessing. I am truly grateful having them to be part in my PhD journey. My dear brothers and sister, my brothers and sisters in-law, thank you for your best wishes for all this while. Not to be forgotten, my instant sisters the 'Powerpuff Moms' group who lift each other up whenever one was felt down with their research works through lunch breaks, coffee breaks, potlucks, kids' birthday party and more! I am so blessed knowing all of you.

To my not-so-little baby, Mounissa my darling and my newborn, Medina, you both are my biggest achievement. My deepest apology my dear Mounissa that you have to sacrifice the most, although you are too young to understand it. For that reason, I am sure you are growing up to be an independent strong girl and you girls will always be my PhD babies.

Saving the best for last, my dearest husband/ my best friend/ my partner in crime, Ahmad Syahrul Anuar Saidi, you are the one who always encourages me to do things beyond my expectation. You have faith in me more than I believe in myself. This is our journey as a team work. Thank you for being my faithful teammate and I love you!

Glory to God, all praise is due to God alone, Allah the Almighty!

Expanded Abstract

In recent decades, there has been significant interest in the development of a catchment classification framework, especially due to the growing need of a common modeling framework. There exist numerous approaches for classification, with different bases, assumptions and methods, which have been applied for catchment classification. The concepts of complex networks, and particularly community structure, have emerged as important tools for classification, and are currently gaining attention in catchment classification. Among the many community structure-based methods, the edge betweenness (EB) algorithm, which applies a hierarchical clustering concept and *modularity function*, is one of the most basic methods for identification of communities (groups) in large dynamically-evolving networks, such as catchment systems. The method's signature steps include: (1) an iterative removal of edges (i.e. links) by calculation of edge betweenness values that pass through the shortest paths between vertices (i.e. nodes); (2) recalculation of the betweenness values after each iterative removal of edges; and (3) formation of communities using a modularity measure, as the maximum value of modularity representing the best partition of the network. Although the EB method has been effectively applied for classification in many different fields, including in hydrology, the modularity measure that is used to form the best partition of community structure is susceptible to network (or data) resolution or scale problem. As a consequence, communities may change when the size of the network changes. Since the

size of a network can change in many situations, as is the case with hydrologic monitoring stations (with the removal/addition of stations), it is important to address the resolution problem to obtain reliable classification outcomes.

Motivated by this, the present study proposes a modified EB (MDEB) algorithm by considering the *modularity density function*, instead of the *modularity function*, for catchment classification. The superior performance of the MDEB method over the EB method is first demonstrated on a real-world network, the Zachary's Karate Club network. The performance of both the EB and MDEB methods are then tested and compared by applying them for classification of a large number of catchments independently in two different countries: (1) 218 catchments in Australia; and (2) 639 catchments in the United States. For each study region, three different scenarios of network sizes are studied: (1) the entire network, as above; (2) 100 and 300 randomly selected stations (with 100 different realizations) from these 218 and 639 streamflow stations, respectively – purely to address the network size; and (3) stations in each of 9 different drainage divisions in Australia and 14 different hydrologic units in the US, respectively – to address the regional similarity and influence. The analysis is mainly performed using streamflow data, in a single-variable sense. For both study areas, the results indicate that the MDEB method performs better than the EB method in catchment classification, for all of the above three scenarios.

With the better performance of the MDEB method, classification is also attempted in a multi-variable sense for the 218 catchments in Australia, by considering, in addition to streamflow, also rainfall and potential evapotranspiration (PET). In addition to the single-variable cases (streamflow, rainfall, and PET independently), four different combinations are considered: streamflow and rainfall; streamflow and PET; rainfall and PET; and streamflow, rainfall, and PET. For each case of multi-variable classification, a count of number of stations within the identified communities and count of the connection links that occur within the network at different threshold values are interpreted. The results suggest that the classification based on the multi-variable approach is nearly similar to that based on the single-variable approach, especially streamflow, but at different correlation thresholds.

The present study is a significant advancement in the application of the concepts of complex networks, especially community structure, for catchment classification, as it offered an improved community structure methodology (edge betweenness) as well as an approach based on multiple variables. Such advancement is certainly promising for the development of a generic catchment classification framework.

List of Publications

Conference Presentation

S.A. Tumiran and B. Sivakumar. Catchment Classification using Complex Networks. 14th Annual Meeting Asia Oceania Geosciences Society. August 6 – 11. Singapore. 2017. *(Oral Presentation)*

S.A. Tumiran and B. Sivakumar. An improved community detection algorithm based on edge betweenness and modularity density for catchment classification. 22nd International Congress on Modelling and Simulation. December 3 – 8. Tasmania, Australia. 2017. *(Oral Presentation)*

S.A. Tumiran and B. Sivakumar. An improved community detection algorithm for classification of catchments in a large region. Japan Geoscience Union Meeting 2018. May 20 – 24. Makuhari Messe, Japan. *(Oral Presentation)*

Table of Contents

Originality Statement i

Copyright Statement ii

Authenticity Statement iii

Acknowledgements iv

Expanded Abstract vi

List of Publications ix

Table of Contents x

List of Figures xiv

List of Tables xx

List of Symbols xxii

1. Introduction 1

1.1 Background 1

1.2 A Brief Review of Catchment Classification 2

1.3 Complex Networks-based Methods for Catchment Classification 5

1.4 Statement of the Research Problem 6

1.5 Objectives of the Study 8

1.6 Outline of the Thesis 10

2. Catchment Classification – Literature Review 12

2.1 Introduction 12

2.2 Catchment Classification Framework Development 14

2.2.1 Catchment Classification Methods 14

2.2.2 Network Concept for Catchment Classification 18

2.2.3 Community Structure-based Methods for Catchment Classification 20

2.3 Research Gaps and Goals 22

3. Methodology 25

3.1 Introduction ……..……..25

3.2 Network: Concepts and Measures ……..……....27

3.3 Community Structure Methods ……..……..…..30

3.4 Edge Betweenness Method ……..……..33

3.4.1 Procedure ……..…..….33

3.4.2 Limitation: Issue of Resolution …….……..….38

3.5 Improvement to the Edge Betweenness Method …….39

3.5.1 Modularity Density Function ……..…….39

3.5.2 Modularity Density-based Edge Betweenness (MDEB) Method …..…..40

3.6 Zachary's Karate Club Network Benchmark: The EB and the MDEB methods …….……..…….41

3.7 Multi-variable Approach for Catchment Classification ….…….……..…….44

3.8 Summary…….……..……..….46

4. Study Area and Data….……48

4.1 Introduction ……..……..48

4.2 Australia ……...……..49

4.2.1 Streamflow ……..….54

4.2.2 Rainfall..…….……..….54

4.2.3 Potential Evapotranspiration...……..55

4.3 The United States ……...…….……..…….56

5. Catchment Classification using Edge Betweenness Method……60

5.1 Introduction ……..……..60

5.2 Classification of Australian catchments.....……..61

5.3 Classification of the catchments in the United States …….72

5.4 Summary …….……...80

6. Modularity Density- based Edge Betweenness (MDEB) Method for Catchment Classification…….....82

6.1 Introduction ……..……..82

6.2 Australian Streamflow ……..…….83

6.2.1 Entire Network (218 Stations)……..…..……..83

6.2.2 Network of 100 Stations through Random Realizations ……..….86

6.2.3 Networks of 9 Drainage Divisions ……..……..…..….93

6.2.4 Comparison between EB and MDEB methods for Catchment Classification ……...……..…..….99

6.3 The United States Streamflow ……106

6.3.1 Entire Network (639 stations)……...….106

6.3.2 Network of 300 Stations through Random Realizations ……..…..….108

6.3.3 Networks of 18 Hydrologic Unit Code (HUC) Regions ……...….114

6.3.4 Comparison between EB and MDEB methods ……..…..…..….123

6.4 Summary …….…….……129

7. Catchment Classification based on Multiple Variables…….…….....131

7.1 Introduction ……..……..…..131

7.2 Single-variable correlation analysis ……..……...133

7.2.1 Streamflow……..…..……..133

7.2.2 Rainfall ……..……134

7.2.3 Potential Evapotranspiration (PET)…….....…..….136

7.3 Multi-variable correlation analysis …….….137

7.3.1 Streamflow and Rainfall ……...….137

7.3.2 Streamflow and Potential Evapotranspiration ……..……..….139

7.3.3 Rainfall and Potential Evapotranspiration …….....….140

7.3.4 Streamflow, Rainfall and Potential Evapotranspiration ……...….141

7.4 Single-variable vs. Multi-variable Classification ……143

7.4.1 Station Count with the Influence of Variables and Threshold ……...143

7.4.2 Connection Link Count ……..……...…153

7.5 Summary ……..……....161

8. Conclusions….……...……163

8.1 Edge Betweenness method for catchment classification …….……164

8.2 Improved EB (MDEB) method for catchment classification ……165

8.3 Multi-variable approach for catchment classification ……...……...166

8.4 Limitations and future work ……...……..……167

REFERENCES...……...……169

APPENDICES...……...……182

Appendix A..……...……182

Appendix B..……...……187

List of Figures

Figure 3.1: Concept of a network……..……...……...…..…28

Figure 3.2: Example of a community structure in a small network. Three communities of densely-connected vertices (in the circles), with a lower density of connections (gray lines) between them …….…….…….…..30

Figure 3.3: The calculation of edge betweenness from (a) node 1, (b) node 2, (c) node 3, and (d) node 4, to every other node in sequence……37

Figure 3.4: (a) The summation of edge betweenness values is represented in each edge; (b) The first edge that has been removed; (c) The edge betweenness values after recalculation; (d) A dendogram is formed based on iterative removal of the highest value of edges, with the red dashed-horizontal line representing the cut for division of communities ……38

Figure 3.5: Zachary's Karate Club Network……42

Figure 3.6: Community structure in the Zachary's Karate Club network with application of the EB method……...……...……...……..43

Figure 3.7: Community structure in the Karate club network by applying the MDEB method. The red circle shows the only node that is incorrectly assigned by this method…….……..….44

Figure 4.1: Locations of 218 hydrologic monitoring stations in Australia……...50

Figure 4.2: Statistical characteristics of streamflow from 218 stations in Australia: (a) mean; (b) standard deviation; and (c) coefficient of variation…..…...……..51

Figure 4.3: Statistical characteristics of rainfall from 218 stations in Australia: (a) mean; (b) standard deviation; and (c) coefficient of variation……...……..…….52

Figure 4.4: Statistical characteristics of PET from 218 stations in Australia: (a) mean; (b) standard deviation; and (c) coefficient of variation……..……...…..52

Figure 4.5: Locations of 639 streamflow stations in the US…….57

Figure 4.6: Statistical characteristics of streamflow from 639 stations in the US: (a) mean; (b) standard deviation; and (c) coefficient of variation……..……...….58

Figure 5.1: Communities identified from the EB method at four different correlation thresholds for streamflow from Australia: (a) T = 0.65; (b) T = 0.7; (c) T = 0.75; and (d) T = 0.8. Each colour represents a community with at least 6 stations, while the open circles represent all communities with less than 6 stations. The different colours are used only to distinguish the communities and hold no meaning when comparing across thresholds.....…62

Figure 5.2: Communities identified from the EB method for correlation threshold T = 0.8 for Australia. Each colour represents a community with at least 6 stations, while the open circles represent all communities with less than 6 stations……..……64

Figure 5.3: Relationship between station drainage area (a), stream length (b) and elevation mean (c) against flow mean for 11 largest communities (150 stations) in Australia. Stations in 11 communities are plotted in colour, corresponding to the legend.…….……..67

Figure 5.4: Relationship between station drainage area (a), stream length (b) and elevation mean (c) against flow CV for 11 largest communities (150 stations) in Australia. Stations in 11 communities are plotted in colour, corresponding to the legend…… ……68

Figure 5.5: Distance-correlation relationship for 11 largest communities in Australia, corresponding to the colouring scheme in Figure 5.2; see text for additional details……..……..…72

Figure 5.6: Communities identified from the EB method at four different correlation thresholds for 639 catchments in the US: (a) T = 0.7; (b) T = 0.75; (c) T = 0.8; and (d) T = 0.85. Each colour represents a community with at least 20 stations, while the open circles represent all communities with less than 20 stations. The different colours are used only to distinguish the communities and hold no meaning when comparing across thresholds.…….73

Figure 5.7: Communities identified from the EB method for 639 stations in the US at the correlation threshold, T = 0.75. Each colour represents a community with at least 20 stations, while the open circles represent all communities with less than 20 stations……..……...76

Figure 5.8: Relationship between station drainage area (a) and elevation (b) against flow mean, as well as station drainage area (c) and elevation (d) against the flow

CV for ten largest communities (479 stations) in the US. Plots (a) and (b) are in log-log scale, while (c) and (d) are in semi-log scale. Stations in ten communities are plotted in colour, corresponding to the legend……...….78

Figure 5.9: Distance-correlation scatterplots for two catchment communities in the US:(a) community 23 (light purple); and (b) community 53 (yellow). The communities correspond to the legend in Figure 5.8(d)……..…..….79

Figure 6.1: Communities identified using the MDEB method for 218 streamflow stations in Australia, with threshold value T = 0.8. Different colours are used only to distinguish the communities……..…….……...84

Figure 6.2: Classification of 10 randomly selected streamflow networks of 100 catchments from Australia using the MDEB method. Each colour represents a different community……..…….……...89

Figure 6.3: Communities identified using the MDEB method for two different sizes of networks: (a) 218 stations and (b) 100 stations. Each colour represents a community with at least 10 stations and 4 stations, respectively. The open circles represent all communities with less than these numbers, respectively……...……...….89

Figure 6.4: Distance- correlation scatterplots for the selected communities from six regions in Australia by the MDEB method (see Table 6.1), ((a)1- 6) base classification and ((b)1- 6) 100 randomly selected stations …...……...93

Figure 6.5: Regions according to drainage divisions and river regions (source of the map: Website of Commonwealth of Australia (Bureau of Meteorology), 2016)...94

Figure 6.6: Circles in colours represent the streamflow locations for each drainage division in Australia……...……..……95

Figure 6.7: Distance-correlation relationship for nine drainage division regions (see Figure 6.6) in Australia……..…...……98

Figure 6.8: Number of communities identified for all 100 random realizations of 100 randomly selected stations for Australia using (a) EB method and (b) MDEB method. The red horizontal lines represent the number of communities with base classification (218 stations)…...…… ……100

Figure 6.9: Difference in the number of communities identified for all 100 random realizations for 100 randomly selected stations for Australia using (a) EB method and (b) MDEB method……..……..……...101

Figure 6.10: Bar plot to compare the number and the percentage of stations changed for 100 random realizations between the EB and MDEB methods……..….103

Figure 6.11: Communities identified using the MDEB method for 639 stations in the United States……..……...……..107

Figure 6.12: Communities identified with 300 randomly selected stations in the US using the MDEB method. Ten out of 100 random realizations are presented, as examples. Each colour represents a community, but the communities are not the same across all the 10 realizations…….…...…….111

Figure 6.13: Communities identified using the MDEB for two different sizes of networks in the US: (a) 639 stations and (b) 300 stations. Each colour represents a community with at least 20 stations (a) and 10 stations (b), while the open circles represent all communities with less than these……..……...111

Figure 6.14: Distance- correlation relationship for communities from seven selected regions in the US by the MDEB method (see Table 6.4, ((a)(1-7)) base classification and ((b)(1-7)) 300 randomly selected stations ……..……114

Figure 6.15: Regions according to hydrologic unit code (HUC) in the United States (source of the map: (Kiang et al., 2013))……...…..……115

Figure 6.16: Figure 6.16: Circles in colours represent the streamflow locations for each region according to the HUC in the US ……...……..…….118

Figure 6.17: Figure 6.17: Distance–correlation relationship for stations within the 18 HUC regions in the US. The colours correspond to those in Figure 6.16)...123

Figure 6.18: Number of communities identified for all 100 random realizations using (a) the EB method and (b) the MDEB method for the US. The red horizontal lines represent the number of communities with the base classification…..124

Figure 6.19: Difference in the number of communities identified for all 100 random realizations for the catchments in the US: (a) EB method; and (b) MDEB method……..……..……125

Figure 6.20: Bar plot to compare the number of stations changed for 100 random realizations between the EB and MDEB methods for the US. Horizontal lines

represent the average number of stations changed using 100 random realizations……..……126

Figure 6.21: Bar plot to compare the percentage of stations changed for 100 random realizations between the EB and MDEB methods for the US……..127

Figure 7.1: Correlation analysis for streamflow from 218 stations in Australia: (a) Correlation of each station to other stations; and (b) distance-correlation relationship…...……...……...134

Figure 7.2: Correlation analysis for rainfall from 218 stations in Australia: (a) Correlation of each station to other stations; and (b) distance-correlation relationship..135

Figure 7.3: Correlation analysis for potential evapotranspiration from 218 stations in Australia: (a) Correlation of each station to other stations; and (b) distance-correlation relationship……...…..……137

Figure 7.4: Correlation analysis for streamflow-rainfall combination for 218 stations in Australia: (a) Correlation of each station to other stations; and (b) distance-correlation relationship……..……...……138

Figure 7.5: Correlation analysis for streamflow-PET combination for 218 stations in Australia: (a) Correlation of each station to other stations; and (b) distance-correlation relationship……...……..140

Figure 7.6: Correlation analysis for rainfall-PET combination for 218 stations in Australia: (a) Correlation of each station to other stations; and (b) distance-correlation relationship…….……141

Figure 7.7: Correlation analysis for streamflow-rainfall-PET combination for 218 stations in Australia: (a) Correlation of each station to other stations; and (b) distance-correlation relationship……...……..……142

Figure 7.8: Number of communities identified for six selected threshold values (T = 0.65, 0.70, 0.75, 0.80, 0.85, and 0.90) for all seven single-variable and multi-variable cases. Each case is indicated with a different colour…..…..153

Figure 7.9: Line graph of the number of connection links based on six selected threshold values for all seven single-variable and multi-variable cases. Each case is indicated by a colour corresponding to the legend in Figure 7.8…..……155

Figure 7.10: Communities identified by: (a, c, and e) streamflow; and (b, d, and f) streamflow and PET with respect to different correlation threshold values using the MDEB method. Each colour represents a community and different

colours are used only to distinguish the communities and hold no meaning when comparing across plots……..157

Figure 7.11: Communities identified by: (a, c, and e) streamflow and rainfall; and (b, d, and f) streamflow, rainfall and PET with respect to different correlation threshold values using the MDEB method. Each colour represents a community, and different colours are used only to distinguish the communities and hold no meaning when comparing across plot……..…159

Figure 7.12: Communities identified by: (a, c, and e) rainfall and PET; and (b, d, and f) streamflow, rainfall and PET with respect to different correlation threshold values using the MDEB method. Each colour represents a community, and different colours are used only to distinguish the communities and hold no meaning when comparing across plot……..…...……161

List of Tables

Table 4.1: Characteristics of 218 catchments and monthly data in Australia ……..….53

Table 4.2: Characteristics of 639 catchments and monthly data in the US……..59

Table 5.1: Sizes of the identified catchment communities in Australia using the EB method at four different correlation thresholds (T = 0.65, 0.7, 0.75 and 0.8). (NSC is the number of stations in the identified communities, NC is the number of communities, and NS is the number of stations)…….……...…63

Table 5.2: Sizes of the identified communities in the US using the EB method at T = 0.7, 0.75, 0.8 and 0.85. (NSC is the number of stations in the identified communities, NC is the number of communities and NS is the number of stations) ……..74

Table 6.1: Sizes of the identified communities using the MDEB method for Australia...85

Table 6.2: Number of random realizations and the average number of stations for the EB and MDEB methods based on the stations changed in classification for Australia……103

Table 6.3: Number of stations changed using EB and MDEB methods according to the drainage divisions in Australia……..105

Table 6.4: Sizes of the identified communities using the MDEB method for 639 catchments in the US……...……..107

Table 6.5: Number of random realizations and the average number of stations for EB and MDEB methods based on the stations changed in classification for the US.…….……127

Table 6.6: Number of stations changed using EB and MDEB methods according to HUC regions in the US……...……..…..128

Table 7.1: Sizes of communities identified using the MDEB method for all single- and multi-variable cases for Australian catchments at threshold value, T = 0.65. (NSC is the number of stations in community, NC is the number of communities and NS is the number of stations)…….....……146

Table 7.2: Sizes of communities identified using the MDEB method for all single- and multi-variable cases for Australian catchments at threshold value, T = 0.7.

(NSC is the number of stations in community, NC is the number of communities and NS is the number of stations)…….……...147

Table 7.3: Sizes of communities identified using the MDEB method for all single- and multi-variable cases for Australian catchments at threshold value, T = 0.75.(NSC is the number of stations in community, NC is the number of communities and NS is the number of stations)…….…...……...……..148

Table 7.4: Sizes of communities identified using the MDEB method for all single- and multi-variable cases for Australian catchments at threshold value, T = 0.8. (NSC is the number of stations in community, NC is the number of communities and NS is the number of stations)……...…….……149

Table 7.5: Sizes of communities identified using the MDEB method for all single- and multi-variable cases for Australian catchments at threshold value, T = 0.85. (NSC is the number of stations in community, NC is the number of communities and NS is the number of stations)…….150

Table 7.6: Sizes of communities identified using the MDEB method for all single- and multi-variable cases for Australian catchments at threshold value, T = 0.9. (NSC is the number of stations in community, NC is the number of communities and NS is the number of stations)……...…….…..…151

Table 7.7: Number of connection links for seven different single-variable and multi-variable cases at six selected threshold values……...……..152

List of Symbols

Symbol	Meaning
N	Number of nodes
V	Vertex or node
E	Link
T	Threshold value
Q	Modularity function
W_{ij}	Weight of edge
m	Number of edges
A_{ij}	Adjacency matrix
$k_i \ and \ k_j$	Degree of node i and node j
$c_i \ and \ c_j$	type (or group) of node i and node j
$\delta(c_i, c_j)$	Kronecker delta
D	Modularity density function
n_i	Number of nodes of subgraph i
l_i	Number of internal links in subgraph i
l_i^{ext}	Number of external links of subgraph i
X	Correlation values of variable

Chapter 1
Introduction

1.1 Background

Hydrology has seen rapid growth during the last century, particularly facilitated by technological and methodological advances, including powerful computers, geographic information system (GIS), measurement devices, digital elevation models (DEMs), remote sensors, scientific theories and mathematical techniques, and networking facilities. Despite these advances, there remain a number of major concerns in hydrologic modelling and forecasting, especially in regards to the ever-increasing complexity of models (e.g., Jakeman and Hornberger, 1993; Perrin et al., 2001; Beven, 2002) and the disparate nature of hydrologic research without a generally acceptable framework for all (e.g., Sivakumar, 2008a, b). These concerns have led, among others, to an increasing realization on the need for a common modelling framework (e.g.,

Woods, 2002; Sivapalan et al., 2003; Gupta, 2004; McDonnell and Woods, 2004; Sivakumar, 2004, 2008a; Sivapalan, 2005; Dawdy, 2007; Sivakumar et al., 2007; Winsemius et al., 2009; Clark et al., 2011; McDonnell and Beven, 2014; Song et al., 2015). While there may be many different ways to achieve such a framework, catchment classification has gained significant attention in recent years (e.g., Olden and Poff, 2003; Snelder et al., 2005; Isik and Singh, 2008; Moliere et al., 2009; Kennard et al., 2010b; Sivakumar and Singh, 2012; Nguyen et al., 2015).

The basic idea in catchment classification is to streamline catchments into different groups and sub-groups based on their salient characteristics (e.g. system, process, scale, and data properties). Due to the various degrees of complexity exhibited by different types of catchments, grouping of catchments based on their salient characteristics is particularly useful for the identification of appropriate complexity of models and for the interpolation/extrapolation of data (including predictions in ungauged basins) that can aid in the planning and management of water, environmental, and ecologic systems; see, for example, Ley et al. (2011), Patil and Stieglitz (2011), Sawicz et al. (2011), Ali et al. (2012), Vignesh et al. (2015), Fang et al. (2017), and Tongal and Sivakumar (2017) for some recent studies for details.

1.2 A Brief Review of Catchment Classification

The idea of catchment classification had been addressed as early as in the 1930s (Pardé, 1933), and received further attention since the 1960s (e.g., Beckinsale, 1969; Budyko, 1974; Gottschalk et al., 1979; L'vovich, 1979; Haines et al., 1988; Nathan and

McMahon, 1990; Rosgen, 1994; Krasovskaia, 1997; Hall and Minns, 1999; Krasovskaia et al., 1999). However, studies during the last two decades or so have provided a catalyst to this issue. Part of this increased interest has been due to the availability of more sophisticated mathematical techniques and better-quality data for analysis as well as the need to study and obtain hydrologic data for ungauged basins (e.g., Sivapalan et al., 2003; Schröder, 2006; Kim and Kaluarachchi, 2008; Oudin et al., 2008; Reichl et al., 2009; Seibert and Beven, 2009; Zhang and Chiew, 2009; Sauquet and Catalogne, 2011; Sivakumar and Singh, 2012; Ali et al., 2012; Nguyen et al., 2015; Sivakumar et al., 2015; Fang et al., 2017; Tongal and Sivakumar, 2017). The initiative by the International Association of Hydrological Sciences (IAHS) on "Predictions in Ungauged Basins" (PUB), has also contributed to this (e.g., Sivapalan et al., 2003; Hrachowitz et al., 2013).

Studies on catchment classification have used different assumptions, bases, and approaches, both to take into account the properties of hydrologic systems (e.g., complexity, scale, nonlinear interdependence, hidden order and determinism, sensitivity to initial conditions, and nature and strength of connections within and among the components) and to take advantage of the mathematical techniques at our disposal. These include river/flow regimes (e.g., Moliere et al., 2009), hydroclimatic factors (e.g., Budyko, 1974; L'vovich, 1979), river morphology (e.g., Poff et al., 2006), hydrologic similarity indexes and signatures, including for regionalization (e.g., Patil and Stieglitz, 2011; Sawicz et al., 2011; Ali et al., 2012; Casper et al., 2012), landscape and land use parameters (e.g., Merz and Blöschl, 2004; Wardrop et al., 2005), eco-hydrologic and geomorphic factors (e.g., Kennard et al., 2010a; Olden et al., 2012), hydropedological factors (e.g., Bouma et al., 2011), geostatistical properties (e.g., Vormoor et al., 2011),

entropy (e.g., Krasovskaia, 1995, 1997), symbolic dynamics into biological signatures (e.g., Hauhs and Lange, 2008), nonlinear dynamic properties (e.g., Sivakumar and Singh, 2012), data-based mechanistic strategies (e.g., Wagener and McIntyre, 2012), data-driven approaches (e.g., Di Prinzio et al., 2011; Ley et al., 2011), and other relevant characteristics/methods (e.g., Isik and Singh, 2008; Wagener et al., 2008). The aforementioned studies have resulted in encouraging outcomes for the provision of a generalisation framework; however, they have also led to additional difficulties. These difficulties are caused by a combination of external factors, internal catchment/process properties, scientific concepts, and mathematical techniques that are driven to realize their importance to specifically outline the processes that are involved to form a classification. However, these have not been sufficiently considered in the past studies and, thus, such studies have limited our ability to accomplish a proper basis of classification (e.g., Olden et al., 2012). With regard to the importance of the basis of classification, some of the crucial processes involved in classification are: consideration of criteria used and the associated data (in terms of selection, treatment, and assessment) (e.g., Olden et al., 2012), dealing with anthropogenic effects (i.e., uncertainties associated with data on land cover, water quality, and climate projections) (e.g., Carillo et al., 2011; Bocchiola et al., 2011; Casper et al., 2012), and the suitability of mathematical techniques (i.e., the advantages and limitations) to obtain a reliable classification of catchments (Sivakumar et al., 2015). In the context of identifying ideal mathematical techniques for catchment classification, assessing the suitability of methods for broader sets of catchment conditions and process properties required by models could be useful. Therefore, there is a need for a coherent and more general approach that can consider these in one way or another.

1.3 Complex Networks-based Methods for Catchment Classification

Considering that hydrologic systems exhibit different levels of complexity and that hydrologic models need to be developed to represent such complexity, there are arguments in favor of using hydrologic 'system complexity' as an appropriate basis for catchment classification (e.g., Sivakumar and Singh, 2012). In this context of system complexity and connections, the concepts and methods developed in the field of *complex systems science*, especially recent developments under the umbrella of *complex networks* (e.g., Watts and Strogatz, 1998; Barabási and Albert, 1999; Girvan and Newman, 2002) can be particularly useful. A network (or graph) is a set of points connected together by a set of lines, where the points are known as *nodes* or *vertices* and the lines are called as *links* or *edges*. The concepts of complex networks, and particularly *community structure*, have emerged as important tools for studying the dynamic connections in complex systems and for their classification and, hence, are currently gaining attention in hydrology, including for catchment classification (e.g., Sivakumar and Woldemeskel, 2014; Halverson and Fleming, 2015; Braga et al., 2016; Serinaldi and Kilsby, 2016; Fang et al., 2017; Han et al., 2018; Yasmin and Sivakumar, 2018).

Community structure is defined as a network structure where distinct groups (i.e., communities) are formed by a cluster of nodes (e.g., catchments), where each of them is more densely linked together when compared to the rest of the network. To our knowledge, only two studies (Halverson and Fleming, 2015; Fang et al., 2017) have,

thus far, applied the concepts of complex networks, especially community structure, for catchment classification. Halverson and Fleming (2015) applied eight different community structure methods (walktrap, fast greedy, leading eigenvector, edge betweenness, multi-level, label propagation, info map, and optimal) to daily streamflow data from 127 catchments in the Coast Mountains of British Columbia and Yukon in Canada for their classification. Fang et al. (2017) applied six community structure methods (edge betweenness, greedy, multilevel modularity optimization, leading eigenvector, label propagation, and walktrap) to daily streamflow data from 1663 stations in the Mississippi River basin, USA (as a representative large-scale basin) for catchment classification.

The outcomes of these studies on the suitability and effectiveness of community structure methods, and complex networks concepts more broadly, are certainly encouraging. They are particularly promising, as the communities are found to offer useful catchment system/process interpretations, including in terms of catchment properties (e.g., drainage area, elevation), flow properties (e.g., mean, coefficient of variation, correlation-distance, unit hydrograph), and others (e.g., river network formation), as appropriate.

1.4 Statement of the Research Problem

Despite their encouraging outcomes, the above community structure-based studies are still insufficient in assessing the general suitability of community structure methods for catchment classification. For instance, the study by Halverson and Fleming (2015) only

examined a relatively small number of stations (127) from a relatively small region (west coast of Canada), despite the differences in topographic/catchment properties in the region. Similarly, although the study by Fang et al. (2017) examined a large number of catchments (1663) from a large-scale river basin (the Mississippi River basin), covering a wide range of hydroclimatic, topographic, and land use properties, it cannot account for catchments that are spread across large regions and/or different river basins and, thus, cannot offer reliable and convincing information as to the suitability and effectiveness of community structure methods for catchment classification, in a general sense. Therefore, it is important to apply the community structure methods to catchments across large regions and different river basins, which, in all likelihood, cover a much wider range of possibilities in terms of hydroclimatic, topographic, geomorphic, land use, and other relevant properties.

At the same time, although community structure methods have been shown to be useful and effective for classification of many different systems, they also often have limitations when applied to real dynamically-evolving systems (Fortunato and Barthélemy, 2007). Therefore, finding the limitations of any community structure method is important to more reliably assess its usefulness and effectiveness of classification of real systems. Consequently, it is also important to modify the existing algorithms to overcome such limitations or even develop new methods for more reliable outcomes.

Finally, in catchment classification studies, it is a common practice to use only a single variable representing the catchments. This is especially the case in studies that have applied the concepts of community structure (and complex networks, more

broadly). Since streamflow is the central and representative component of catchments, most studies have essentially used the streamflow data for catchment classification, in the sense of single-variable analysis. However, inclusion of additional variables that govern the catchment dynamics in one way or another (e.g., rainfall and potential evapotranspiration) can often lead to more reliable classification, since their inclusion will bring more stringent conditions for classification. The fact that almost all the variables influencing the catchment dynamics are also often interconnected and interact in a nonlinear manner provides additional stimulus on their inclusion for catchment classification. Therefore, it is important to perform a multi-variable analysis, with as many variables influencing the catchment dynamics as possible, by including variables, in addition to streamflow.

1.5 Objectives of the Study

The observations made above reveal that a proper assessment on the general suitability of the community structure methods for catchment classification is still lacking. Finding the limitation(s) of any community structure method is certainly useful to enhance the performance and assess its usefulness for classification. More stringent conditions also need to be imposed to examine the usefulness and efficacy of a particular community structure method. To achieve these, a coherent effort is clearly needed. This provides the motivation for the present thesis.

The overall aim of this thesis is to propose a better community structure-based approach for catchment classification. This is proposed to be achieved through the following specific objectives:

1. Application of a community structure method, specifically the edge betweenness (EB) method, to classify numerous catchments in two large regions (with different climatic conditions, topographic characteristics, land uses, and other relevant properties) to assess the general suitability and effectiveness of the method. To this end, 218 catchments across Australia and 639 catchments across the United States are studied. Streamflow data are used, in a single-variable sense, for identifying the connections between catchments and for classifying the catchments;
2. Development of an improved edge betweenness method for catchment classification. This is proposed to be achieved by addressing the issue of resolution limit (i.e., network size), an important limitation of the *modularity function* used in the EB method to split the entire network into different communities. To this end, instead of the modularity function, a *modularity density function* is used in the EB method. The improved method is termed as the *modularity density-based edge betweenness* (MDEB) method. The method is implemented for the above catchments in Australia and in the United States, using streamflow data in a single-variable sense.
3. Proposal of a multi-variable approach for catchment classification, by involving multiple variables influencing the catchments. This is achieved by including rainfall and potential evapotranspiration (PET), in addition to streamflow. Four

different combinations of these variables are considered, and the approach is implemented for classification of catchments in Australia.

As for the methodology development, the performance of the MDEB method in classifying catchments is compared against that of the EB method, to assess its superiority, if any. To assess the influence of catchment variables on classification, the classification results from the multi-variable approach are compared with those from the single-variable approach. These comparisons and the associated interpretations and conclusions are expected to shed some light on the usefulness and effectiveness of the community structure concepts for catchment classification, and complex networks more broadly for hydrologic modelling and forecasting.

1.6 Outline of the Thesis

The rest of this thesis is organized as follows. Chapter 2 presents a review of the literature on catchment classification studies. Particular emphasis is given to discussing the role of the concepts of complex networks for hydrologic studies, especially community structure concepts for catchment classification. Chapter 3 provides a detailed description of the network methodology used in this thesis. The concepts of a network, complex networks, and community structure are reviewed. After a detailed description of the edge betweenness method using an example, the issue of the resolution limit is highlighted using a widely-used real network, the Zachary Karate Club network. This leads to the proposal of an improved edge betweenness method, the modularity density-based edge betweeness (MDEB) method. Finally, a multi-variable

approach for catchment classification is also presented. Chapter 4 describes the two study areas (218 catchments in Australia and 639 catchments in the United States) and the data considered in this study.

Chapter 5 presents the application of the edge betweenness method for classification of catchments in Australia and in the United States, using only streamflow data in a single-variable sense. The catchment communities are interpreted in terms of catchment/flow properties, among others. Chapter 6 presents the application of the MDEB method for classification of catchments in Australia and in the United States, using only streamflow data in a single-variable sense. A comparison between the performance of the EB method and the MDEB method for catchment classification is also made.

In Chapter 7, the application of the multi-variable approach in the MDEB method for classification of catchments in Australia is presented, with the inclusion of rainfall and PET, in addition to streamflow. The classification results from the multi-variable-based MDEB analysis are also compared with those from the single-variable-based MDEB analysis. Interpretations of the classification results are also made in terms of correlations between stations, distance-correlation relationship, and accurate station count for the communities identified, among others. Finally, Chapter 8 draws some key conclusions from the present study and offers potential directions for further research.

Chapter 2

Catchment Classification – Literature Review

2.1 Introduction

There have been tremendous advances in hydrology and water resources, facilitated by the invention of powerful computers, scientific theories and mathematical techniques, measurement devices, geographic information system (GIS), digital elevation models (DEMs), and networking facilities. Nevertheless, there remain many big challenges in teaching, research, and practice in hydrology and water resources. It has been increasingly realized, in recent years, that there is a need for simplification in hydrologic models and a common framework for hydrologic modelling (e.g., Grayson and Blöschl, 2000; Woods, 2002; Sivapalan et al., 2003; McDonnell and Woods, 2004; Sivakumar, 2004, 2008b; Wagener et al., 2007; Young and Ratto, 2009; Olden et al., 2012). In the context of a general framework for hydrology, catchment classification has been proposed as one possible means,

and has gained significant attention (e.g., Harris et al., 2000; Olden and Poff, 2003; McDonnell and Woods, 2004; Snelder et al., 2005; Poff et al., 2006; Sivakumar et al., 2007; Wagener et al., 2007; Isik and Singh, 2008; Moliere et al., 2009; Kennard et al., 2010a; Sivakumar and Singh, 2012; Nguyen et al., 2015).

The basic idea of catchment classification is to streamline catchments into different groups and sub-groups based on their salient characteristics (e.g., system, process, scale, and data properties). As Sivakumar et al. (2015) pointed out, most of the existing approaches to catchment modelling adopt either of the following extreme views: (1) regardless of the differences among all the catchments, all catchments are treated in the same way; and (2) regardless of the similarities among all the catchments, each catchment is treated in its own way. Therefore, any attempt to offer a reliable approach for catchment classification (and modelling more broadly) should adopt a middle-ground approach. In the end, a catchment classification framework should have at least three main expected outcomes: (1) it should be designed to provide an ideal way of studying catchments, taking into account both the minimization of the costs and the maximization of the benefits; (2) it should be able to accommodate important general characteristics as well as specific ones for catchments and the associated processes; and (3) it must also be simple and able to provide a common language for communication and discussion among academics, researchers, and practitioners in the fields of hydrology, water resources, ecology, geography, geomorphology, and beyond. By achieving these to form such a framework, catchment classification can be useful for the identification of the appropriate complexity of models for different types of catchments and for the interpolation/extrapolation of data, including predictions in ungauged basins (e.g.,

McDonnell and Woods, 2004; Wagener et al., 2007; Hauhs and Lange, 2008; Sivakumar, 2008a; Wagener et al., 2008; Bocchiola et al., 2011; Carrillo et al., 2011; Di Prinzio et al., 2011; Ley et al., 2011; Patil and Stieglitz, 2011; Sawicz et al., 2011; Ali et al., 2012; Sawicz, 2013; Sivakumar et al., 2015; Fang et al., 2017; Tongal and Sivakumar, 2017).

2.2 Catchment Classification Framework Development

An effective and reliable classification scheme must be able to provide names or types of catchments, hydrologic information transfer (i.e., regionalization of information), development in generalization (i.e., able to develop new theories), and also to provide a first-order environmental change impact assessment (i.e., the hydrologic implications of climate and land use change) (e.g., Grigg, 1965, 1967; Milly et al., 2008). There exist many different methods for catchment classification. In what follows, a brief review of such methods is presented. Particular emphasis is given to the complex networks-based methods, as such is the focus of this thesis.

2.2.1 Catchment Classification Methods

As early as during the 1930s and, especially since the 1960s, numerous attempts have been made to advance the idea for a catchment classification framework. Consequently, different theoretical bases and a variety of mathematical techniques have been used to classify numerous catchments in different geomorphologic and

climatic settings. Based on such studies, various important implications have been offered for hydrology, ecohydrology, environment, and water resources, including for under the conditions of climate change.

Until the end of the twentieth century, attempts on catchment classification were mainly focused on river flow regimes, hydroclimatic factors, and hydrologic similarity, which were perhaps mainly driven towards aiding hydrologic modeling in regionalization analyses (e.g., Pardé, 1933; Beckinsale, 1969; Budyko, 1974; Gottschalk et al., 1979; L'vovich, 1979; Tasker, 1982; Haines et al., 1988; Chapman, 1989; Nathan and McMahon, 1990; McMahon and Finlayson, 1992; Lins, 1997; Krasovskaia et al., 1999). However, the realization of the impacts that river flows have (floods, low flows, and droughts) on water resources, environment, and ecosystems has led researchers to view the catchment classification problem in such contexts as well. Consequently, many other studies have attempted environmental and ecosystem classification, particularly facilitated by the aforementioned advances in studying flow regimes, river geomorphology, and hydrologic similarity and signatures (e.g., Hughes and James, 1989; Claussen and Biggs, 2000; Detenbeck et al., 2000; Harris et al., 2000; Snelder and Biggs, 2002; Loveland and Merchant, 2004; Snelder et al., 2004, 2005; Snelder and Hughey, 2005; Kampichler et al., 2010; Kennard et al., 2010a, b; Zhang et al., 2012). Since the beginning of this century, many other bases and approaches have been used for catchment classification, with emphasis on model complexity and predictions in ungaged basins, among others (e.g., Krasovskaia and Gottschalk, 2002; Sivakumar, 2003; Sivapalan et al., 2003a; Merz and Blöschl, 2004; Poff et al., 2006; Rao and Srinivas, 2006a, b; Schröder, 2006; McMahon et al., 2007a, b; Wagener et al., 2007; Isik and Singh, 2008; Kim

and Kaluarachchi, 2008; Oudin et al., 2008; Hrachowitz et al., 2009; Moliere et al., 2009; Reichl et al., 2009; Seibert and Beven, 2009; Zhang and Chiew, 2009; Patil and Stieglitz, 2011; Vormoor et al., 2011; Sims et al., 2012; Sivakumar and Singh, 2012; Ali et al., 2012; Nguyen et al., 2015; Sivakumar et al., 2015; Fang et al., 2017; Tongal and Sivakumar, 2017).

The various methods for studying the classification of catchments (and others) may generally be grouped into deductive approaches and inductive approaches (Olden et al., 2012). As emphasized by Olden et al. (2012), it is essential to explicitly describe the steps taken in the development of a classification system, including standards used for data choice, data handling and assessment, metric choice and basis, and classification method, including the explicit basis for derivation of the final group number. The need to address the aforementioned issues is crucial for such studies and the methods to be useful in wider practice.

During the past century, studies on the influences of anthropogenic activities and their impacts for catchment classification have become extremely important. Many attempts have been made to address the uncertainty in catchment classification associated with land use, land cover, and climate (e.g., Krasovskaia and Gottschalk, 2002; Bower et al., 2004; Snelder et al., 2005; Carillo et al., 2011; Bocchiola et al., 2011; Casper et al., 2012). However, the outcomes of such studies need to be carefully interpreted, especially with the uncertainties associated with data on land cover, water quality, and climate projections.

There exists a plethora of alternate avenues, mathematical techniques, and scientific concepts that are beneficial to attain catchment classification, whether

classification is based on hydrology or ecohydrology or other. For instance, the current approaches for classification include regression-based methods (e.g., Kennard et al., 2010); cluster analysis, including fuzzy clustering and partitioning (e.g., Rao and Srinivas, 2006a, b; Moliere et al., 2009; Kennard et al., 2010; Sawicz et al., 2011; Ali et al., 2012); principal component analysis (e.g., Snelder et al., 2005); entropy-based methods (e.g., Krasovskaia, 1995, 1997); symbolic dynamic and nonlinear dynamic concepts (e.g., Krasovskaia and Gottschalk, 2002; Sivakumar, 2003; Sivakumar et al., 2007; Hauhs and Lange, 2008; Sivakumar and Singh, 2012); and other methods, such as data-driven, data-based mechanistic, and geostatistical (e.g., Castiglioni et al., 2011; Di Prinzio et al., 2011; Ley et al., 2011; Vormoor et al., 2011; Wagener and McIntyre, 2012).

In the context of finding an appropriate approach for catchment classification, as reported by Sivakumar et al. (2015), concepts of nonlinear dynamics and networks seem to offer a practical methodology for identification of the catchment complexity and classification of catchments. Such concepts have been shown to be useful in the study of a wide range of systems, processes, and problems encountered in diverse fields, including hydrology, ecology, atmospheric sciences, physics, chemistry, biology, engineering, technology, economics, medicine, psychology, politics, and social sciences (e.g., Tsonis, 1992; Goerner, 1994; Strogatz, 1994; Abarbanel, 1996; Kantz and Schreiber, 1997; Phillips, 1999; Watts, 1999; Sivakumar, 2000; Liljeros et al., 2001; Barabási, 2002; Newman et al., 2003; Tsonis and Roebber, 2004; Barrat et al., 2008). In addition, studying catchment classification essentially involves their grouping or regionalization based on one or more criteria as a network, which could differ in terms of the purpose, data accessibility and application, metric considered,

methodology, and others. The size and nature of such a network depend on the geographic extent, number of catchments, type and resolution (spatial and temporal) of data, and other factors.

The basic idea in catchment classification is to examine if some connections exist (e.g., correlation, similarity) between/among catchments and to then use the strength of such connections for grouping (with due consideration to the possible spuriousness of connections), regardless of any approach. In the system of catchments, the connections may be developed in the form of geographic proximity, hydrologic similarity and signatures, hydroclimatic factors, landscape and land use parameters, and others. These observations make it clear that a methodology that can reliably represent the network in its entirety is needed for an effective and efficient formulation of a catchment classification framework. Recent developments in the field of complex systems, especially *complex networks science*, seem to offer such a methodology.

2.2.2 Network Concept for Catchment Classification

The catchment classification problem basically consists of studying a network of catchments and dividing them based on one or more criteria. The number of catchments, type and resolution (spatial or temporal) of data, geographic, and other factors are the common variabilities involved in the size and nature of a network. The purpose, data availability and use, metric considered, methodology employed, and others may lead to fluctuations on the criteria for grouping. The basic idea in

catchment classification is to examine the existence of some connections (e.g., correlation, similarity) between/among catchments and then use the strength of such connections for grouping (despite the consideration in a possible fallacy of connections) regardless of the approach. As reviewed previously, the connections may be in the form of geographic proximity, hydrologic similarity and signatures, hydroclimatic factors, landscape and land use parameters, and others. In the context of system complexity, the suitability of concepts and methods developed in the field of *complex systems science*, especially recent developments under the umbrella of *complex networks* (e.g., Watts and Strogatz, 1998; Barabási and Albert, 1999; Girvan and Newman, 2002). A network (or graph) is a set of points connected together by a set of lines, where the points are known as *nodes* or *vertices* and the lines are called as *links* or *edges*.

A large number of measures have been developed to study the properties of complex networks, including clustering coefficient, degree distribution, shortest (or geodesic) path, and community structure. The concepts of complex networks and the associated measures have been widely applied in numerous research fields for almost two decades now. However, their applications in hydrology and closely-related fields are fairly new, and applications have started to emerge only recently, including for rainfall networks (e.g., Malik et al., 2012; Boers et al., 2013; Scarsoglio et al., 2013; Sivakumar and Woldemeskel, 2015), river networks (e.g., Rinaldo et al., 2006; Zaliapin et al, 2010), streamflow networks (e.g., Sivakumar and Woldemeskel, 2014; Braga et al., 2016; Serinaldi and Kiksby, 2016; Han et al., 2018; Yasmin and Sivakumar, 2018), and virtual water networks (e.g., Suweis et al, 2011; Konar et al., 2011; Carr et al., 2012; Dalin et al., 2012; D'Odorico et al., 2012; Tamea et al.,

2013). Application of the concepts of complex networks, and particularly community structure, for catchment classification is currently gaining attention (e.g., Halverson and Fleming, 2015; Fang et al., 2017).

2.2.3 Community Structure-based Methods for Catchment Classification

In many complex networks, nodes cluster together into distinct groups, with each group more densely linked together when compared to the rest of the network. The properties of these groups are generally independent of the properties of the individual nodes and of the network as a whole. These groups are known as *communities*, and this kind of network structure is known as *community structure*. Intuitively, community is deemed as a set of entities in the network sense, where each entity is closer to the other entities within the community than to the entities outside it (Coscia et al., 2012). To our knowledge, there have, thus far, been only two studies that have specifically applied the community structure methods for catchment classification (Halverson and Fleming, 2015; Fang et al., 2017). Halverson and Fleming (2015) studied 127 stations from a network of streamflow gauging stations along the west coast of Canada. In addition to the investigation of whether regional streamflow hydrology might be quantitatively represented as a formal network, their study was aimed at assessing whether the results from the network-based methods might offer important information as to the optimal design of streamflow monitoring systems. They employed a host of network-based methods, including clustering coefficient, degree distribution, average shortest path length, betweenness, and

community structure. They identified ten groups by applying eight different community structure methods and also presented the representative unit hydrographs for the ten groups and discussed the classification of stations in terms of elevation and drainage area. They proposed that an idealized sampling network should sample high-betweenness stations as well as small-membership communities which are, by definition, a rare or undersampled relative to other communities, while at the same time retaining some degree of redundancy to maintain network robustness.

Fang et al. (2017) attempted catchment classification using six community structure methods on daily streamflow data from a network of 1663 stations in the Mississippi River Basin. The study was the first ever to apply the community structure methods for a large river basin. The consistency of the identified communities from each method was assessed using the Normalized Mutual Information (NMI) value. The results indicated that, in addition to geographic proximity, organization of the river network (e.g., main stem, river branching) also plays an important role in the formation of different communities of catchments. An examination of the identified communities against some important catchment/flow properties (altitude, drainage area, flow mean, and flow coefficient of variation) offered some interesting observations, as did the distance-correlation relationship.

The outcomes of the studies by Halverson and Fleming (2015) and Fang et al. (2017) are certainly promising, regarding the suitability and effectiveness of the community structure methods, and complex networks concepts more broadly, for catchment classification. This is particularly so, as the communities offer useful catchment system/process interpretations in terms of catchment properties (e.g.,

drainage area, elevation), flow properties (e.g., mean, coefficient of variation, correlation-distance, unit hydrograph), and others (e.g., river network formation). Nevertheless, these studies are largely insufficient in assessing the general suitability of community structure methods for catchment classification. For instance, the study by Halverson and Fleming (2015) only examined a relatively small number of stations (127) from a relatively small region (west coast of Canada), despite the differences in topographic/catchment properties in the region. Even though the study by Fang et al. (2017) examined a large number of stations (1663) from a large-scale river basin (the Mississippi River basin), covering a wide range of hydroclimatic, topographic, and land use properties, there are also possible strong inherent connections between the catchments, since all the stations essentially belong to only one large river basin. As a result, these studies cannot account for catchments that are spread across large regions and/or different river basins and, thus, cannot offer reliable and convincing information as to the general suitability and effectiveness of community structure methods for catchment classification.

2.3 Research Gaps and Goals

The issues mentioned above can only be addressed by studying catchments across large regions and different river basins, which, in all likelihood, cover a wide range of possibilities in terms of hydroclimatic, topographic, geomorphic, land use, and other relevant properties. Apart from that, despite their effectiveness for catchment classification, each of the community structure methods has its own limitations. For instance, the edge betweenness (EB) method is susceptible to the network resolution

(size) problem and, thus, can influence the classification outcomes when the network size (e.g., number of catchments used for classification) changes. Therefore, it is extremely important to carefully examine the limitations of the community structure methods, so that appropriate modifications or new developments can be undertaken for a more reliable classification framework. These provide the motivation for the present study.

In light of the above, the present study attempts to improve the existing traditional community structure-based methods for applications in catchment classification. In particular, the study focuses on the edge betweenness (EB) algorithm (Girvan and Newman, 2002). The EB method, which applies a hierarchical clustering concept and modularity function, is one of the widely-used methods for identification of communities (groups) in large dynamically-evolving networks, such as catchment systems. Although the EB method has been effectively applied for classification in many different fields, including in hydrology, it is also susceptible to network (or data) resolution or scale problem caused by the modularity function. Since the size of a network can change in many situations, it is important to address the resolution problem to obtain reliable classification outcomes. To overcome this resolution or scale problem for catchment classification, an improved EB algorithm is proposed by replacing the modularity function with the modularity density function to decide on the best division of communities in the network. Both the EB method and the improved EB method (the latter is termed as the *modularity density-based edge betweenness* (MDEB)) are first applied to one of the most-studied real networks, the Zachary's Karate Club network, to assess their performance. To assess the general suitability of the methods for catchment classification, the methods are

then applied independently for classification of a large number of catchments in two different countries: (1) 218 catchments in Australia; and (2) 639 catchments in the United States. In each case, three different network scenarios that take into account the influence of network size and regional similarity are considered. The analysis is first performed in a single-variable sense, with use of only streamflow data from the respective catchments.

It is important to recognize the role of interactions between surface water and climatic inputs (e.g., rainfall and potential evapotranspiration (PET)) and, hence, on the influence of climate inputs on catchment functions and classification. Therefore, studying streamflow data alone for catchment classification is often not sufficient, and there is indeed a need to also include other variables that may have notable influence on catchment functions and, thus, classification. In light of this, the present study considers, in addition to streamflow, data of two climate variables: rainfall and PET. With this, the study makes the first ever attempt on a multi-variable approach in the application of community structure method for catchment classification. In the multi-variable-based analysis, different combinations of multiple variables are considered, by combining any two and all three variables. The results obtained from the multi-variable approach are also compared with those from the single-variable classification, especially streamflow.

Chapter 3

METHODOLOGY

3.1 Introduction

An effective and reliable catchment classification framework requires an approach that is able to address all the key aspects associated with hydrologic system dynamics, especially their complexity, nonlinearity, and dynamically-evolving properties. In this regard, modern developments in complex systems science have been found to be very useful to represent many of the key properties of hydrologic systems and for their classification (e.g., Regonda et al., 2004; Salas et al., 2005; Sivakumar et al., 2007; Mishra et al., 2009; Li et al., 2010; Tongal et al., 2013); see also Sivakumar and Berndtsson (2010) and Sivakumar (2017) for comprehensive accounts. Among such developments, complex networks-based methods are particularly useful for their ability to study system connections and dynamics in a holistic manner and, hence, their applications in hydrology and water resources are of great interest at the current time (e.g., Rinaldo et al., 2006; Scarsoglio et al., 2013; Sivakumar, 2015; Sivakumar and Woldemeskel, 2014, 2015; Halverson and Fleming, 2015; Braga et al., 2016; Serinaldi

and Kilsby, 2016; Fang et al., 2017; Jha and Sivakumar, 2018; Han et al., 2018; Yasmin and Sivakumar, 2018). In the specific context of classification, the concept of community structure is useful. Community structure is defined as a network structure where distinct groups (i.e., communities) are formed by a cluster of nodes (e.g., catchments), where each of them is more densely linked together when compared to the rest of the network. Application of the community structure concept, and complex networks more broadly, for catchment classification is fairly new. To our knowledge, the studies by Halverson and Fleming (2015) and Fang et al. (2017) have been the only ones to specifically apply the concepts of community structure for catchment classification. The outcomes are certainly encouraging, as they indicate the suitability and effectiveness of community structure methods, and complex networks concepts more broadly.

While there is no question that community structure algorithms are useful for catchment classification, one must be careful in implementing them, especially when it comes to real dynamically-evolving systems. This is because, community structure algorithms often have limitations, and the limitations may also be different for the different methods. Therefore, finding the limitations of any community structure method is necessary to assess the potential errors in the resulting classification and, consequently, to modify and improve the method further for a more reliable classification. The present study addresses this issue, with particular reference to the edge betweenness algorithm (Girvan and Newman, 2002). The edge betweenness method measures the shortest path of a particular link in the network and the communities are defined using a quality measure function called *modularity function* (or Q value) and the maximum value of modularity leads to the best community formation.

However, the method is susceptible to the issue of scale or resolution. This study addresses this issue in the implementation of the method for catchment classification and also proposes an improved edge betweenness algorithm.

3.2 Network: Concepts and Measures

A network (or graph) is a set of points called nodes (or vertices) connected together by links (or edges), as shown in Figure 3.1. Mathematically, a network can be represented as G = {V, E}, where V is a set of N nodes ($V_1, V_2, \ldots V_N$) and E is a set of n links. Figure 3.1 shows a network with N = 5 nodes and n = 6 links. Hence, in this network, V = {1,2,3,4,5} and the set of links is E = {[1,2], [1,4], [2,4], [3,4], [3,5], [4,5]}.

Figure 3.1 shows the simplest form of a network consisting of a set of identical nodes connected by identical links. There are many ways in which networks may be (made) more complex. For instance: (1) a network may have more than one different type of node and/or link; (2) the nodes and links may have a variety of properties associated with them, such as different weights for different nodes and links depending on the strength of nodes and connections; and (3) the links can be directed, pointing in only one direction. The directed networks can be either cyclic (i.e., containing closed loops of links) or acyclic. Further, networks can have multi-links (i.e., repeated links between the same pair of nodes), self-links (i.e., links connecting a node to itself), and hyperlinks (i.e., links connecting more than two nodes together). Networks can also be bipartite, i.e., containing nodes of two distinct types, with links running only between unlike types. Further details of the different forms of networks can be seen in Newman (2010) and Estrada (2012), among others.

In a network, the assignment of a link to a pair of nodes need not be a straightforward binary relationship but can be based on a measure that represents the strength of the link, such as the linear correlation coefficient. For instance, node pairs that have correlation coefficients exceeding a certain threshold value (T) may be assigned links. For example, the network shown in Figure 3.1 is based on a streamflow monitoring network, and the links are assigned based on calculation of correlations in streamflow data between the stations and a correlation threshold value of T = 0.75.

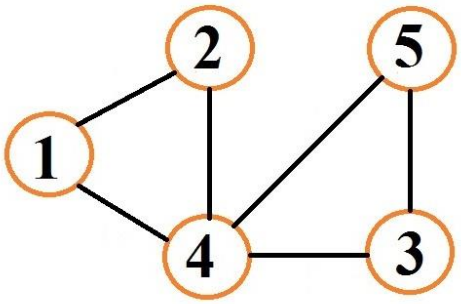


Figure 3.1: Concept of a network

Network theory or graph theory originated in the seventeenth century and developed into topology, trees, and random graph theory until around the mid-20th century (e.g., Listing, 1848; Cayley, 1857; Euler, 1741; Erdös and Rényi, 1960). However, such concepts, especially the random graph theory, were found to possess some major deficiencies. First, they largely focused on networks that are regular, simple, small, and static and are generally unsuitable for examining real networks, which are often highly irregular, complex, large, and dynamically-evolving in time. Second, the random graph theory assumed that complex and large-scale networks are

wired randomly together (Erdös and Rényi, 1960). However, for real networks, such an assumption is not necessarily valid, since order and determinism are inherent in real systems and networks. Real networks are neither completely ordered nor completely random, but generally exhibit important properties of both.

Motivated by the deficiencies of random graph theory, advances in network study, together with other complex systems science concepts, including nonlinear dynamics, chaos, self-organization, and scale-invariance (e.g., Lorenz, 1963; Mandelbrot, 1983; Bak, 1996), led to the new science of networks, called *Complex Networks* (e.g., Watts and Strogatz, 1998; Barabasi and Albert, 1999). Notable advances in such are the small-world networks (Watts and Strogatz, 1998), scale-free networks (Barabási and Albert, 1999), network motifs (Milo et al., 2002), and community structure (Girvan and Newman, 2002). There are also different measures to identify and quantify different properties of networks, and, for some, there are also different definitions, sub-measures, and corresponding methods, as appropriate. These network measures include centrality, clustering, adjacency, distance, community structure, bipartivity, fragments (or subgraphs), communicability, and global invariants (Estrada, 2012).

The discovery of complex networks and associated methods led to their applications to a wide range of natural and artificial networks, including the spread of diseases and sexual contact (e.g., Liljeros et al., 2001), World Wide Web (e.g., Albert et al., 1999), distribution of wealth (e.g., Bouchaud and Mezard, 2000), climate (e.g., Tsonis and Roebber, 2004), and virtual water trade (e.g., Suweis et al., 2011); also refer to Watts (1999), Barabási (2002), Newman et al. (2003), Barrat et al. (2008), and Estrada (2012) for comprehensive accounts of such applications.

3.3 Community Structure Methods

In many complex networks, nodes cluster together into distinct groups, with each group more densely linked together when compared to the rest of the network, as shown in Figure 3.2. The properties of these groups are generally independent of the properties of the individual nodes and of the network as a whole. These groups are known as "communities" and this kind of network structure is known as a "community structure." Intuitively, the community is deemed as a set of entities, in the network sense, where each entity is closer to the other entities within the community than to the entities outside it (Coscia et al., 2012).

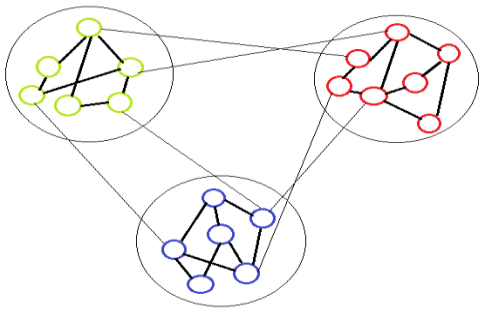


Figure 3.2: Example of a community structure in a small network. Three communities of densely-connected vertices (in the circles), with a lower density of connections (gray lines) between them.

Since nodes belonging to the same community are more likely to share network properties and dynamics (e.g., centrality, clustering, degree distribution, shortest path

length, communicability), detection of communities in networks is particularly useful. Further, the number and characteristics of the existing communities provide subsidies for identifying the type of a network and in understanding its dynamic evolution and organization. Community detection is highly relevant and useful in hydrology, as in the case of catchment classification, where the purpose is to identify a region(s) or a group(s) of monitoring stations that have similar properties. A number of methods have been developed for community detection in networks, including edge betweenness, greedy algorithm, multilevel optimisation, leading eigenvector, label propagation, and walktrap. Some of these methods rely on the modularity, Q, which quantifies the quality or strength of a community. A brief description of the above-mentioned community structure methods is as follows:

i. **Edge betweenness centrality:** Edge betweenness centrality (Newman and Girvan, 2004) is a measure to identify a particular link that has the most 'betweenness' in a network, by only considering the number of shortest paths that pass through the link.

ii. **Greedy algorithm:** The greedy algorithm is an attempt to optimise the modularity (Q) directly by self-organizing each node to its community, as proposed by Clauset et al. (2004).

iii. **Multilevel modularity optimisation**: This method, proposed by Blondel et al. (2008), is somewhat similar to the greedy algorithm. However, instead of self-organizing each node as a community, this method assigns the nodes and the nodes are also re-assigned one by one at a time into a community until the highest increase in modularity is reached.

iv. **Leading eigenvector method**: The leading eigenvector method also employs modularity optimisation aided by the algebraic properties to define the modularity matrix based on random networks (Newman, 2006).

v. **Label propagation method**: This method, proposed by Raghavan et al. (2007), identifies communities based on the nodes with the same label after the labels stopped propagating. The label is at first uniquely assigned to each node; then, with iterative process by adopting and propagating the label of its neighbours, densely connected nodes will comprise a unique label for the group; and, finally, each unique label represents as community.

vi. **Walktrap method**: Walktrap method is a measure of distance between nodes and communities based on short random walks. It forms the dendrogram and apply the modularity maximization to split the network (Pons and Latapy, 2005).

In the present study, the edge betweenness method is applied for catchment classification. While the edge betweenness method has been shown to be effective, the modularity function, Q, that is used to measure the strength of the community structure is susceptible to resolution (or scale) problem. To overcome this issue, an improvement is also proposed by considering the modularity density function, instead of the modularity function; the proposed method is called the modularity density-based edge betweenness (MDEB) method. Additional improvement to the MDEB method and, hence, catchment classification is also done through development of a multi-variable approach, instead of a single-variable approach that is widely used. The edge betweenness method and the proposed improvements are detailed next.

3.4 Edge Betweenness Method

3.4.1 Procedure

Edge betweenness centrality is a measure of how central a particular link is in a network, and it measures the number of shortest (also known as geodesic) paths which pass through the link. This means, the links with the highest centrality are considered to be acting as bridges connecting the communities together and, therefore, removing these central links will split the network into more densely connected communities (Newman and Girvan, 2004).

To implement the edge betweenness procedure, the correlations between nodes are first determined. For instance, in the present study, with streamflow stations as the nodes in the network, the cross-correlation values in streamflow values between the different streamflow stations are determined. The presence/absence of a link between any two stations is identified by considering a threshold value (T) and comparing the correlation between any two stations with such a threshold value, as mentioned earlier; see Sivakumar and Woldemeskel (2014) for further details on the selection of the threshold value, especially in regards to streamflow.

Let us consider, for simplicity, an unweighted and undirected network with five nodes (streamflow stations) and six links, as shown in Figure 3.1. In this case, the connection between node *i* to node *j* is similar to that between node *j* to node *i* (Newman and Girvan, 2004). For this network, the main steps to measure the edge betweenness to identify communities are explained as follows:

1. **Calculate the value of edge betweenness for each and every node in the network.**

a) Every node is given a weight, $w = 1$. In the case of multiple shortest paths between a pair of stations, they will be given equal weights, i.e., the summation of the weights of those multiple shortest paths will always be equal to 1. Figure 3.3 illustrates the calculation of the weight from each node (to score the edge betweenness) when connecting to other nodes in the network based on the shortest paths. For example, Figure 3.3(a) starts from node 1, and node 1 to node 2 has only one shortest path and given a weight of 1.

b) To calculate the edge of any node i to any other node j, the betweenness value is based on $w_{ij} = w_i + w_j$. For example, in Figure 3.3(a), node 1 has to pass node 4 in order to connect with node 3 via the shortest paths. Thus, node 1 to node 4 will be given a weight of 1 and node 4 to node 3 another weight of 1. The weight that is given at edge node 1 to node 4 will remain and be accumulated where this particular edge is also used to connect node 1 to node 4 as well as node 1 to node 5. Therefore, the edge of node 1 to node 4 has accumulated a score of 3 for one case of node.

c) Repeat steps (a) and (b) for other nodes (Figure 3.3(b to d)) until there are no remaining nodes.

d) Prior to each iteration of removal of edge, all the weights are accumulated, as given by $W_i = \sum_{i=1}^{n} w_{ij}$, with j representing nodes that are connected to node i, to measure the betweenness value, as seen in Figure 3.4(a) where each edge is denoted with the value of edge betweenness.

This whole process is considered as one iteration to remove the first edge, as shown in Figure 3.4(b).

2. Remove the edge with the highest value of edge betweenness in the network.

The removal of an edge is the crucial step in this method to split the network naturally according to the strength of the connection between communities (Newman, 2004). Thus, the first edge to be removed is found by following the order in numbering, i.e., from node 1 to node 4, as shown in Figure 3.4(b).

3. **Recalculate the edge betweenness for all remaining edges.**

Repeating step (1) means recalculating the remaining edges and, therefore, only the betweenness of other edges that are in the same community will be affected by the removal and only the betweenness in that community will be recalculated. Figure 3.4(c) shows the value of edge betweenness after the recalculation of the first edge removal.

4. **Repeat steps (1) to (3) until no remaining edges in the network.**

The iterative removal of edges will be stopped when there are no remaining edges in the network and each node in the network is formed as a community. This means, the number of communities will be equal to the number of nodes in the network.

5. **Form the dendrogram and network division by modularity function.**

A dendrogram is used to represent the whole process of the algorithm and is formed after there are no remaining edges available, based on the hierarchical clustering concept (Girvan and Newman, 2002). In Figure 3.4(d), each horizontal level in the dendrogram, from top to bottom, is represented as the first edge until the last edge is removed. The division of the network (i.e., formation of communities) by this algorithm is evaluated using a measure called “modularity function,” which is a numerical index for forming the best partition of the network.

The first division of the network is identified by the first level at which the dendrogram is cut. The division of the group should be stopped when the value of

modularity begins to decrease, as the modularity is calculated with respect to the given membership vector and value, i.e., it ranges within 0 to 1. The modularity function, that is applied in the present study, is given by:

$$Q = \frac{1}{2m}\sum_{i,j}(A_{ij} - \frac{k_i k_j}{2m})\delta(c_i, c_j) \qquad (3.1)$$

where m is the number of edges, A_{ij} is the adjacency matrix in row i and column j, k_i and k_j are degree of i and j , respectively, derived from the sum of all rows/columns of the adjacency matrix, c_i and c_j represent a type (or group) of i and j, and $\delta(c_i, c_j)$ is the Kronecker delta that will be denoted as 1 if $c_i = c_j$ (if nodes i and j are in the same community) or as 0 if otherwise. In this way, Q is only concerned with the links that are within the group and ignores links from the other groups. Further details on the calculation of the modularity are presented in Appendix A.1. From the first level of the horizontal cut in the dendrogram (Figure 3.4(d)), the calculated modularity value is 0.11, while, for the second horizontal cut, the modularity value decreases to 0 with communities of [1 2 3 3 3]. Since the Q value has decreased and reached 0, the calculation of modularity will be stopped. Thus, the first cut of the dendrogram with communities [1 1 2 2 2] is derived with the maximum Q and eventually represents the best split of this network.

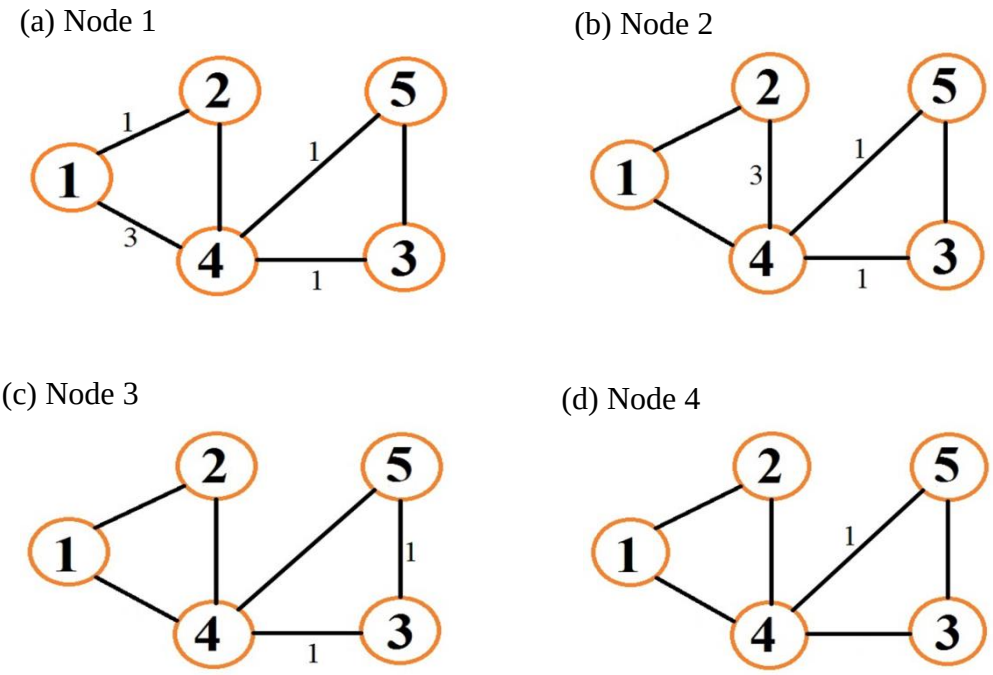


Figure 3.3**:** The calculation of edge betweenness from (a) node 1, (b) node 2, (c) node 3, and (d) node 4, to every other node in sequence.

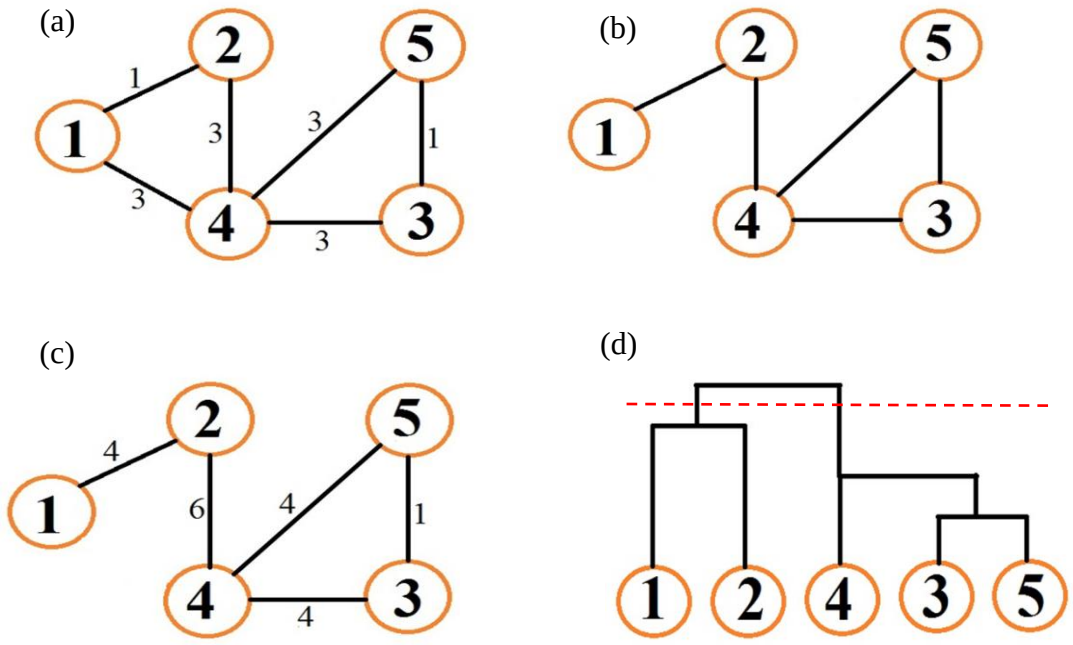


Figure 3.4**:** (a) The summation of edge betweenness values is represented in each edge; (b) The first edge that has been removed; (c) The edge betweenness values after recalculation; (d) A dendogram is formed based on iterative removal of the highest value of edges, with the red dashed-horizontal line representing the cut for division of communities.

3.4.2 Limitation: Issue of Resolution

Although the edge betweenness (EB) algorithm has been effective in identifying communities in many synthetic and real-world networks, the modularity function that is used to measure the strength of the community structure is susceptible to a scale problem. In other words, the function is incapable of classifying a network correctly. This deficiency is mainly due to the fact that modularity does not consider the size of the community, and the measure is mainly dependent on the size of the total links in the network for community formation (Rosvall and Bergstrom, 2007). As a consequence, communities can change when the size of the network changes. In the context of the catchment classification framework, in order to gain an accurate framework for classification, it is important for an approach to be able to consistently classify catchments regardless of the number of catchments (or size of network in this context) available. Regarding consistency as one of the efficacies of the method for catchment classification framework, it is important to note that the number of stream-gauges may change in the future either due to the removal of some existing ones or due to newly installed ones. This may lead to unreliable classification outcomes, when the above modularity function is used specifically for the development of a generic catchment

classification framework. Therefore, there is a need to improve the edge betweenneess method for a more consistent and reliable classification.

3.5 Improvement to the Edge Betweenness Method

The above-mentioned limitation of the edge betweenness method motivated Li et al. (2008) to propose a new quantitative measure to evaluate the partition of the network to communities based on the density of links by subgraphs. The new measure is called modularity density function (or known as the D value). In the present study, we propose an improvement to the edge betweenness algorithm using the modularity density function. The improved method is called as the Modularity Density-based Edge Betweenness method (MDEB). The modularity function and the procedure for the improved method are described next.

3.5.1 Modularity Density Function

The modularity density function (D) mainly focuses on the density of links and the number of nodes within communities regardless of the size of network, and form the partition naturally on the local properties of the network. Thus, it can resolve more precise communities and is able to alleviate the scale problem. This measure is mainly to address the uncertainty of a modularity function that is dependent on the total number of nodes regardless of taking into consideration a count of the number of links. Several studies have applied the modularity density function (D value) as an optimisation function for community detection and also for other applications (e.g., Zhang et al., 2009; Chen et al., 2014; Botta and Del Genio, 2016).

In the proposed improvement to the EB method, the main four steps of the traditional method, i.e. steps 1 to 4, as described above in Section 3.4.1, will remain. However, for the fifth step, the maximum modularity density function or D value is obtained instead of the Q value, to determine which level of the dendrogram will be cut to represent the best split of the network. Hence, in the present study, the modularity density function (D value) is applied as follows:

$$D = \sum_i \left(\frac{2l_i}{n_i} - \frac{l_i^{ext}}{n_i}\right) \qquad (2)$$

where n_i is the number of nodes of subgraph i, l_i is the number of internal links in subgraph i, and l_i^{ext} is the number of external links of subgraph i.

3.5.2 Modularity Density-based Edge Betweenness (MDEB) Method

The procedure of the MDEB method is briefly explained in this section. As the EB is still the base for the MDEB method, some of the main steps from Section 3.4.1 will remain, including: (1) calculation of the edge betweenness from each node to every other node in the network; (2) removal of the edge with the highest value of betweenness; (3) recalculation of betweenness of the remaining edges; and (4) repeat of the steps (1) – (3) until no remaining edges in the network. For the fifth step, in which the improvement is made, the modularity density is applied to quantify the network partition by maximization of the D value. For better understanding, the same example of the network presented in Figure 3.1 is used to illustrate the differences between the two community detection methods (i.e. EB and MDEB). Since the iterative processes in MDEB are similar to that in the EB method, the dendrogram is also similar, as in Figure

3.4(d). From the dendrogram, as in Figure 3.4(d), the calculation of the D value should be stopped when the number of communities formed has the same count as the size of the network (i.e. this means until the final bottom level of dendrogram), and the final division of the network based on which level is derived with the maximum D value. Thus, from the first level of horizontal cut in the dendrogram (Figure 3.4(d)), the calculated D value is 1.33 with communities of [1 1 2 2 2], while for the second horizontal cut, the modularity density (D value) decreases to -2.67, -8 and -12, with communities of [1 2 3 3 3], [1 2 3 4 3], and [1 2 3 4 5], respectively. The details to calculate the D value are presented in Appendix A.2. As explained earlier, when the EB method is applied to the network in Figure 3.1, the first cut of the dendrogram with communities [1 1 2 2 2] has the maximum D representing the best split of this example network. Based on the example, the best split shown by both methods are the same. Thus, to prove the superiority of the MDEB method, further benchmarking of these methods is performed by applying the methods to a real-world network, the Zachary's Karate Club network, as presented in the next section.

3.6 Zachary's Karate Club Network Benchmark: The EB and the MDEB methods

The Zachary's Karate Club (Zachary, 1977) network is one of the most commonly used benchmark networks in community mining and, hence, is considered here to assess the effectiveness of the EB and MDEB methods for classification. This real-world network is purposely used to find out whether groupings can be successfully uncovered, without prior knowledge of the potential behaviour of the network, and also to be able to

determine which nodes belong to which community. The Zachary's Karate Club network considered here consists of 34 club members (i.e., nodes) and 78 edges (i.e., links) that represent the friendship between members of the club that has been observed over a period of two years, as shown in Figure 3.5. However, the administrator (node 1) and the instructor (node 33) have decided not to cooperate with each other, which has led to the members splitting into two groups; either joining with the administrator or joining with the instructor. The application of the EB method, which is based on the quality measure of maximum modularity (Q value), divides the network into five communities, separated in rectangles with coloured borders, as shown in Figure 3.6. As seen, the EB method is unable to split the club network correctly according to the real communities of this network as a result.

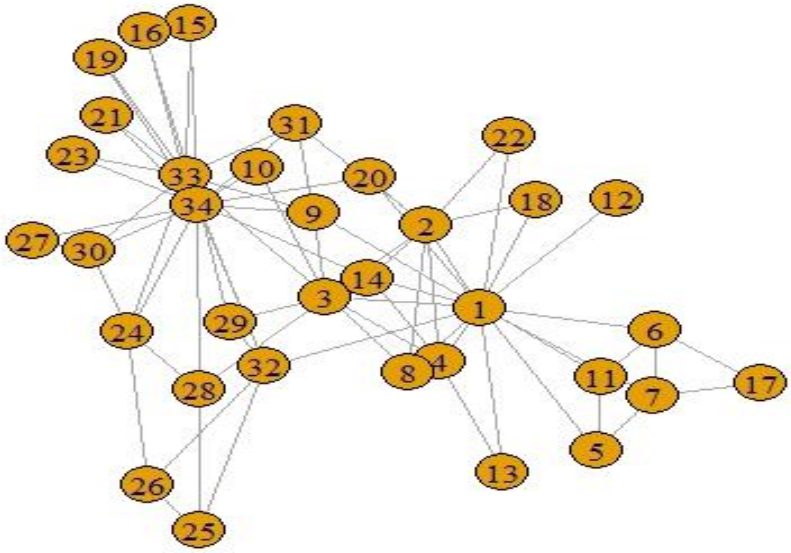


Figure 3.5: Zachary's Karate Club Network

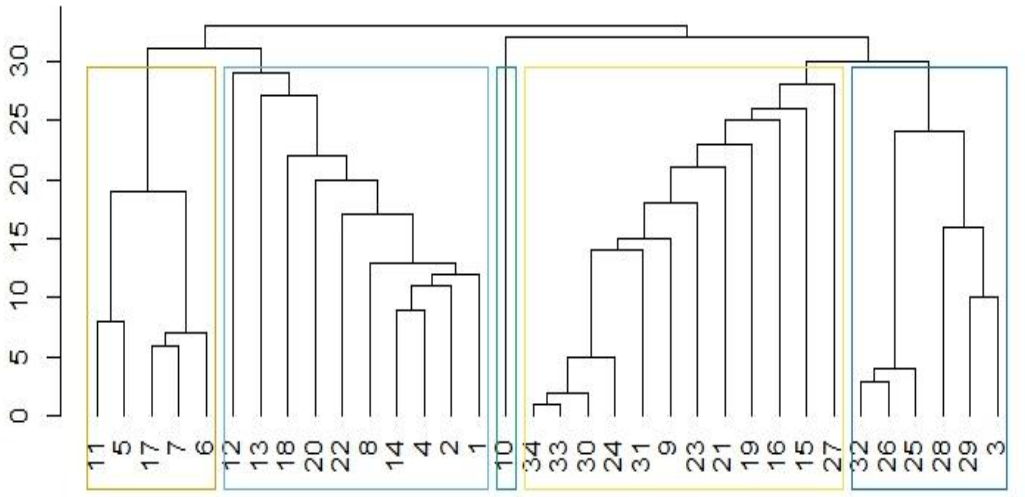


Figure 3.6: Community structure in the Zachary's Karate Club network with application of the EB method.

Figure 3.7 shows the dendrogram as the result to the classification of the Zachary Karate Club network when the MDEB method is applied, where each community is also represented by coloured rectangular boxes. As may be seen, the MDEB method results in classification outcomes that correspond almost perfectly to the actual community formation in the club, with the exception of only node 10 (circled in red) that is misclassified. In comparison to the outcomes from the traditional EB method, this improved method certainly progresses in two ways: (1) links (i.e., internal and external) to enhance for a smaller community's detection in a large network; and (2) more accurate measures in defining community structure in a particular network. Thus, the MDEB method performs better than the EB method in addressing and overcoming the issue of the resolution limit (i.e. network size). As the size of the networks associated with hydrologic systems can change, such as removal/addition of streamflow and raingauge monitoring stations, a method that adequately addresses and overcomes the

issue of resolution limit is clearly needed for studying catchments, including for the purpose of catchment classification. Therefore, in the present study, additional emphasis is given in the application of the MDEB method for catchment classification.

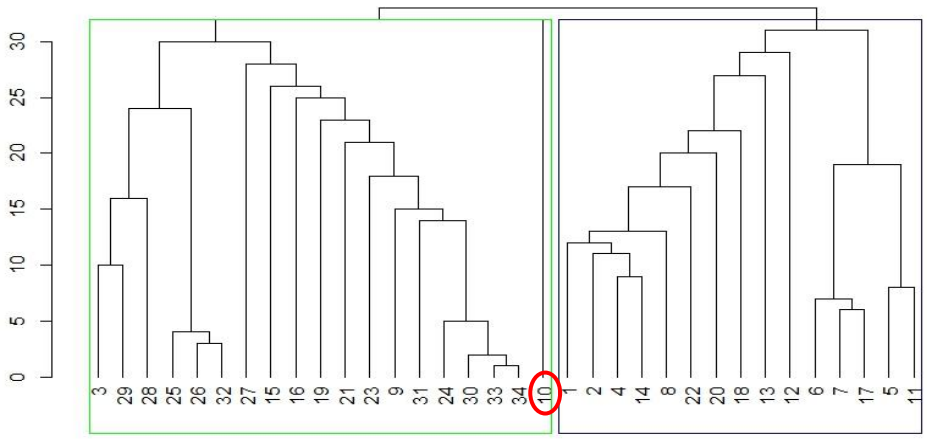


Figure 3.7: Community structure in the Karate club network by applying the MDEB method. The red circle shows the only node that is incorrectly assigned by this method.

3.7 Multi-variable Approach for Catchment Classification

An extension to examine the effectiveness of the MDEB method, in the form of a multi-variable analysis for catchment classification, is proposed to compare with the classification outcomes from a single-variable streamflow analysis, to assess the influence of other hydroclimatic inputs as well. For the multi-variable approach, inclusion of additional variables that govern the catchment dynamics in one way or another (e.g., rainfall and potential evapotranspiration) could form more stringent conditions for classification. Since streamflow is the central and representative component of catchments, most studies have essentially used the streamflow data for catchment classification, in the sense of a single-variable analysis. By taking into

account that almost all the variables influencing the catchment dynamics are also often interconnected and interact in a nonlinear manner, it would be useful to include all the available influencing variables, in addition to streamflow, for catchment classification. Furthermore, exploring the potential of each of the other hydroclimatic inputs as a single variable and their combinations as multiple variables may give insights into which variable(s) could be more important for a catchment classification framework.

In the present study, the basic idea for multi-variable analysis lies in creating a new correlation value (for identifying connections), X, such that the new value is obtained by averaging the correlation values of variables, say X^1, X^2,. . . , X^N where N= number of variables. Consider combining two discrete variables as an example of correlation values from two hydroclimatic inputs. If it is assumed that X^1 is [0.9, 0.5, 0.5, 0.8, 0.85, 0.7, 0.6] and X^2 is [0.8, 0.4, 0.45, 0.7, 0.75, 0.8, 0.85], then the new variable X can be obtained by the average of pairwise summation, i.e., X is [0.85, 0.45, 0.48, 0.75, 0.8, 0.75, 0.73]. Similar to a single-variable case, the new correlation value is then assigned with the assumed threshold value, i.e., if the new value is more than or equal to the assumed threshold value, then the pair of stations of interest is considered as connected; otherwise, there is no link between the pair. In the multi-variable approach, the classification is carried out using only the MDEB method and several threshold values are selected: T = 0.65, 0.7, 0.75, 0.8, 0.85, and 0.9. Each correlation threshold is assigned to each single-variable case (i.e., streamflow, rainfall, and PET, separately) and multi-variable cases (i.e., any and all of the combinations of the three variables) to examine the connections and, thus, classification.

3.8 Summary

In order to obtain reliable catchment classification outcomes, it is essential to decide which particular approach should be taken, and how that particular approach could be useful in identifying the groups of catchments. As the edge betweenness method has been widely used for classification in many different fields, it may be appropriate to use the method for catchment classification as well. However, the method is susceptible to the issue of resolution (or scale) limit, mainly because of the use of the modularity function. To overcome this issue, an improvement to the EB method is proposed in this study, by replacing the modularity function with the modularity density-based function, and the new method is termed as the MDEB method. The modularity density-based function considers the density of links and the number of nodes within communities regardless of the size of network (i.e., scale or resolution). This allows the modified edge betweenness method to perform better in classification when compared to the traditional edge betweenness method, which uses only the modularity function that is depended on the size of the network. The MDEB method, therefore, has a particular advantage in catchment classification, since the size of the streamflow (or other hydrologic) network (i.e., number of stations) can change either due to removal of one or more of the existing stations or due to addition of one or more new stations. Since removal or addition of streamflow (or other hydrologic) stations may be commonplace, due to a variety of reasons, the modularity density function (and, hence, the MDEB method) has great conceptual and practical significance in studying hydrologic networks and their classification. Further, to take into account the influence of other catchment/climate variables, in addition to streamflow, on catchment classification, it is appropriate to use a multi-variable approach, by including additional variables. The rest

of this thesis presents the details of such, with Chapter 4 presenting the details of the study area and data, Chapter 5 and 6 presenting the application of the EB and the MDEB methods in a single-variable sense, and Chapter 7 presenting the multi-variable analysis (only with the MDEB method).

Chapter 4

Study Area and Data

4.1 Introduction

In the present study, to investigate the utility and effectiveness of the concepts of community structure, specifically using the edge betweenness (EB) method and the improved EB (MDEB) method, for catchment classification, numerous catchments from two different countries, Australia and the United States, are studied. Each of these two large regions cover a wide range of hydroclimatic, topographic, land use properties, and other relevant properties and, thus, offer proper test beds for catchment classification and to generalize the classification outcomes. For each region, first, the catchment classification is performed based on the streamflow data in a single-variable sense using the EB method and the MDEB method. Further, for the Australian catchments, a multi-variable approach for classification is also performed by using rainfall and potential evapotranspiration (PET), in addition to streamflow. For this, only the MDEB method is applied, due to its superior performance than the EB method, as presented in Chapter 3.

Details of the two study areas and the associated data considered for the present analysis are presented next.

4.2 Australia

Australia, including Tasmania, has almost the same size of the United States of America (excluding Alaska) with a land-mass of 7,682,300 km^2 (Ghassemi and White, 2007). Over the continent, there is a wide range of climatic zones, varying from tropical to sub-Arctic, with large parts of the central and western Australia influenced by arid and semi-arid climatic conditions. The north is more influenced by a tropical climate, and the south-east as well as the south-west parts have a moderately temperate climate affected by the oceans. In the entire continent, about 87 percent of the area has an altitude less than 500 m, and 99.5 percent is less than 1000 m with a mean altitude of 300 m above mean sea level. The average annual air temperature ranges from 28 °C along the extreme north of western Australia to 4 °C in the Alpine regions of south-eastern Australia.

For the present study, a total of 218 catchments across Australia are considered for catchment classification. The locations of these stations are shown in Figure 4.1. These 218 stations are from the Hydrological Reference Stations (HRS) database, maintained by the Australian Bureau of Meteorology (BoM). The HRS database has been developed based on several factors, including the period of observation and anthropogenic influences. Extensive details associated with the selection of HRS database by the BoM are available in Zhang et al. (2016). A few stations from the HRS database are not considered in this study, because of some missing data. For these 218

catchments, streamflow data are mainly considered in a single-variable sense, while streamflow, rainfall, and potential evapotranspiration (PET) are considered in a multi-variable sense. In the multi-variable approach, combinations of any two and all three variables are considered. The data considered in this study covers a 26-year period, i.e., from 01 January, 1981 to 31 December, 2006. The data are average monthly values. Table 4.1 presents a summary of the minimum and maximum values of some important characteristics of the stations/data, including the corresponding station numbers. Figures 4.2 to 4.4 present the statistical characteristics of streamflow, rainfall, and PET: (a) mean, (b) standard deviation, and (c) coefficient of variation in all the 218 stations. Catchment and flow characteristics play important roles in the nature and extent of connections in each variable between the different stations. However, the focus of the present study is in identifying the extent of connections among the stations based on each variable as single-variable and by combinations (two out of three and all three of them) as multiple variables.

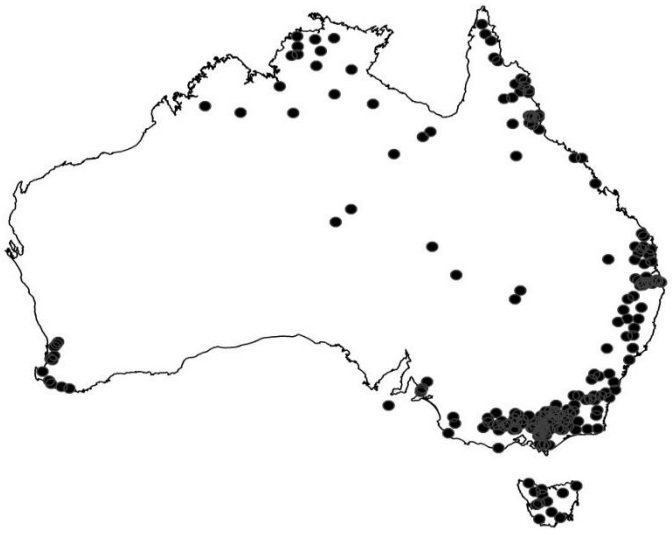


Figure 4.1: Locations of 218 hydrologic monitoring stations in Australia

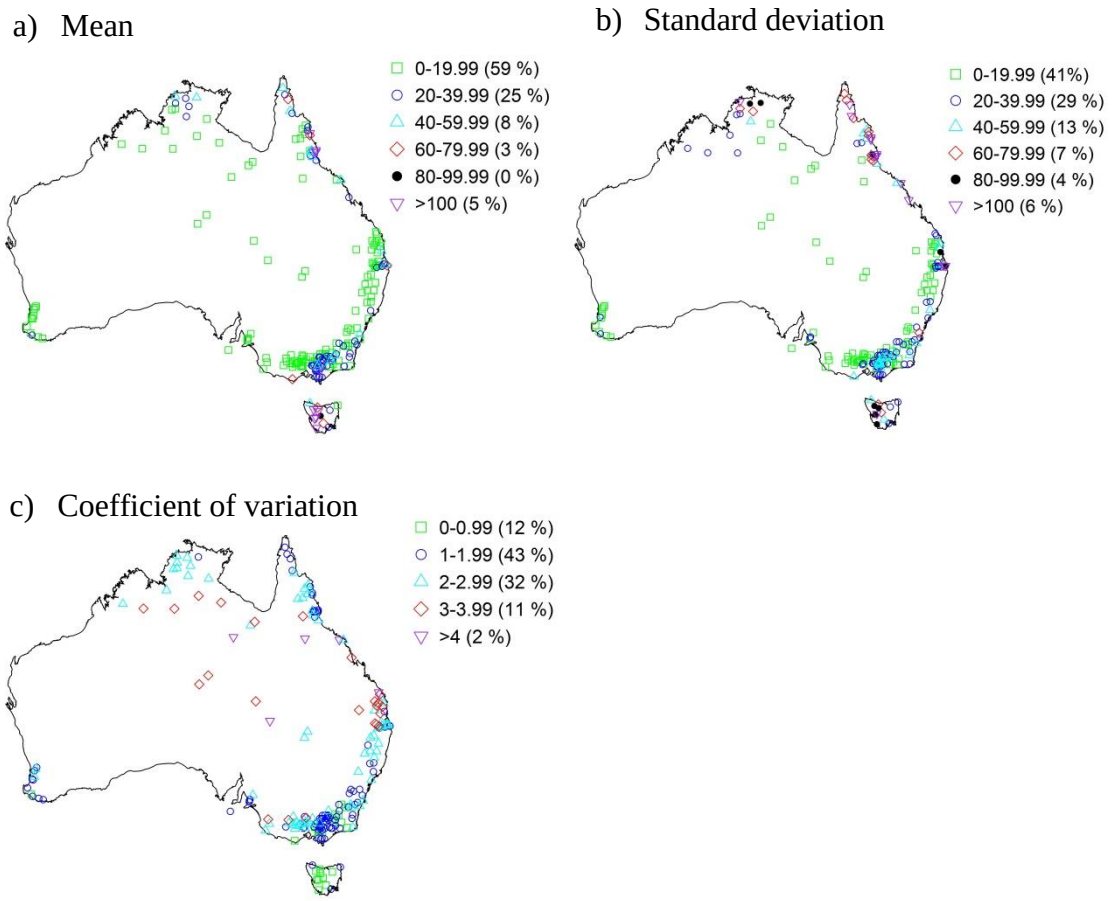
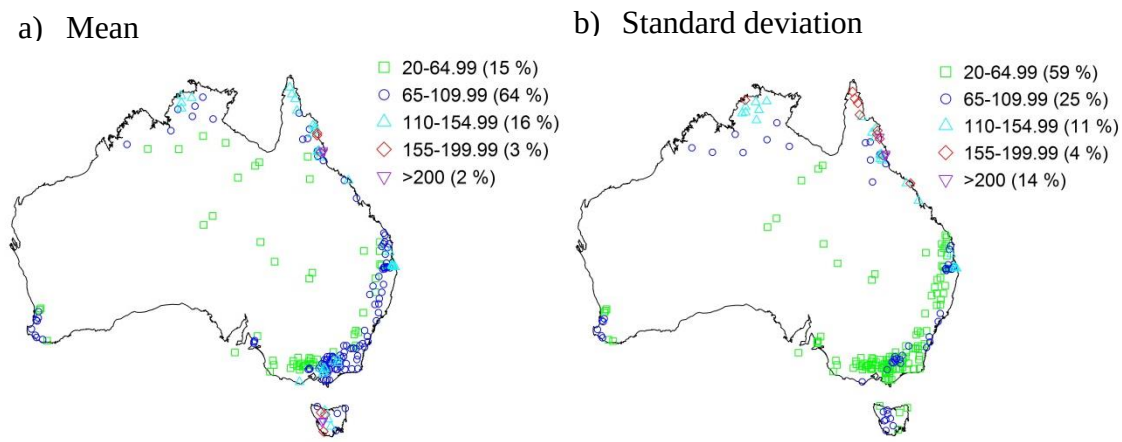


Figure 4.2: Statistical characteristics of streamflow from 218 stations in Australia: (a) mean; (b) standard deviation; and (c) coefficient of variation.



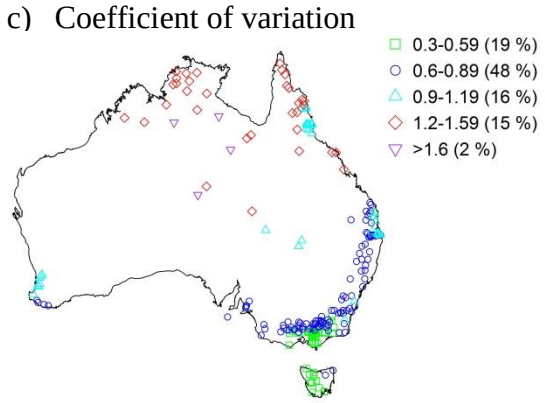


Figure 4.3: Statistical characteristics of rainfall from 218 stations in Australia: (a) mean; (b) standard deviation; and (c) coefficient of variation.

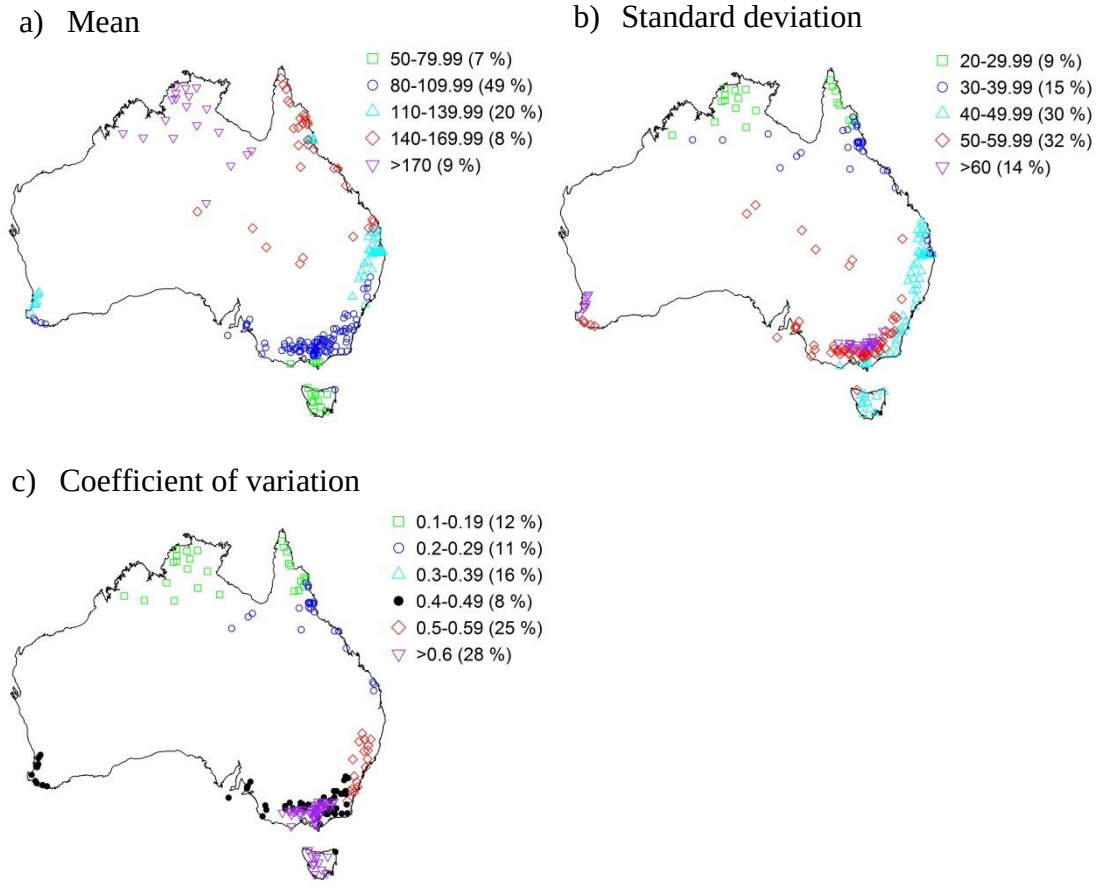


Figure 4.4: Statistical characteristics of PET from 218 stations in Australia: (a) mean; (b) standard deviation; and (c) coefficient of variation.

Table 4.1: Characteristics of 218 catchments and monthly data in Australia.

	Minimum	Maximum	Station (State)
Latitude	-43.14°	-11.83°	Minimum: #473 (TAS)[a] Maximum:#926002A (QLD)
Longitude	115.44°	153.42°	Minimum: #610008 (WA) Maximum:# 146012A (QLD)
Drainage area (km^2)	11.65 (4.5 mi^2)	603069.15 (232846.3 mi^2)	Minimum: #235205 (VIC) Maximum:#A0030501 (SA)
Elevation (m)	5 (16.37 ft)	2181.55 (7157.32 ft)	Minimum: #G8140040 (NT) Maximum:#401012 (NSW)
Flow mean (m^3/s)	0.36 (12.83 ft^3/s)	182.42 (6442 ft^3/s)	Minimum: #A0030501 (SA) Maximum: #112002A (QLD)
Flow standard deviation m^3/s)	0.944 (33.337 ft^3/s)	233.9082 (8260.39 ft^3/s)	Minimum: #616013 (WA) Maximum: #112002A (QLD)
Flow CV	0.471	6.12	Minimum: #226222 (VIC) Maximum: #G0010005 (NT)
Rainfall Mean (mm)	22.68	282.31	Minimum: #A0020101 (SA) Maximum: #112002A (QLD)
Rainfall standard deviation (mm)	28.96	290.5	Minimum: #407253 (VIC) Maximum: #112002A (QLD)
Rainfall CV	0.468	1.658	Minimum: #473 (TAS) Maximum: #G0050115 (NT)
PET Mean (mm)	61.17	192	Minimum: #473 (TAS) Maximum: #G9030124 (NT)

PET standard deviation (mm)	21.51	71.9	Minimum: #G8170002 (NT) Maximum: #616002 (WA)
PET CV	0.118	0.706	Minimum: #G8170002 (NT) Maximum: #473 (TAS)

[a] NT - Northern Territory; SA - South Australia; TAS - Tasmania; QLD - Queensland; VIC - Victoria; WA - Western Australia.

4.2.1 Streamflow

In the context of rivers and streamflow, there are 12 regions based on drainage and river divisions of Australia. Streamflow data are measured at numerous gaging stations across the country. In the present study, monthly streamflow data from the above 218 stations across entire Australia are considered. The 218 streamflow stations and their observed streamflow data show enormous variations in their characteristics, as presented in Table 4.1. For instance: (1) basin drainage area ranges from 11.65 km^2 (4.5 mi^2) to 603,069.15 km^2 (232,846.3 mi^2); (2) station elevation ranges from 5 m (16.37 ft) to 2181.55 m (7157.32 ft); and (3) mean flow ranges from 0.36 m^3/s (12.83 ft^3/s) to 182.42 m^3/s (6,442 ft^3/s).

4.2.2 Rainfall

The rainfall variability is significant in Australia due to different climates in different regions of the country. The annual rainfall in the far north exceeds 4000 mm, while more than 80% of the country receives an annual rainfall of less than 600 mm (many

parts of the interior gets less than 200 mm per year). The mean annual rainfall across the country is only about 465 mm. Influenced by seasons, much of the rainfall occurs during winter (June–August) in parts of the south (especially to eastern New South Wales, Victoria, Tasmania, and southwest of Western Australia), while the main rainfall season in the north is during summer (December–February). More than 17,000 raingage stations are archived by the Australian Bureau of Meteorology (BoM). More details about the raingages for the entire Australia are available in Lavery et al. (1992, 1997). For the present study, similar to the streamflow data, monthly rainfall data from a network of 218 stations (from the HRS database) across Australia are considered. Table 4.1 presents some important station/rainfall characteristics. As seen, the mean rainfall ranges from 22.68 mm to 282.31 mm, and the standard deviation from 28.96 mm to 290.5 mm.

4.2.3 Potential Evapotranspiration

Evapotranspiration is a collective term for the transfer of water, as water vapour, to the atmosphere from both vegetated and unvegetated surfaces greatly affected by climate, availability of water and vegetation (Chiew et al., 2002). The potential evapotranspiration (PET) values for Australia can be found from the Evapotranspiration Maps in the Climatic Atlas of Australia (http://www.bom.gov.au/climate/averages; Australian Bureau of Meteorology, 2001). In the present study, the monthly PET data from 218 catchments across Australia are studied (Figure 4.1), the same stations used for streamflow and rainfall. The PET data covers a period of 26 years (January 1981 to December 2006), similar to the streamflow and rainfall data periods. The mean PET

ranges from 61.17 mm to 192 mm, and the standard deviation from 21.51 mm to 71.9 mm, as presented in Table 4.1.

4.3 The United States

In the analysis for catchment classification in the US in the present study, only streamflow data are studied in a single-variable sense. Monthly streamflow data from an extensive network of 639 streamflow gaging stations in the contiguous US are studied. The locations of these 639 stations are shown in Figure 4.5. The streamflow data are obtained from the US Geological Survey (USGS) database, in particular from the Hydro-Climatic Data Network (HCDN), originally developed by Slack and Landwehr (1992) and subsequently updated at different times, with the last update in 2009; see Lins (2012) for details (http://water.usgs.gov/osw/hcdn-2009/).

Streamflow data in the US are commonly expressed in water years, which commence in October. The data used in this study are those observed over a period of 53 years (1950–2002) from each of the above 639 stations. During the past few decades, many studies have used the streamflow data set from HCDN or a sub-set for numerous purposes by applying a variety of methodologies (e.g., Slack and Landwehr, 1992; Kahya and Dracup, 1993; Vogel and Sankarasubramanian, 2000; Sivakumar, 2003; Tootle and Piechota, 2006; Patil and Stieglitz, 2012; Kiang et al., 2013; Yasmin and Sivakumar, 2018). However, the studies by Sivakumar and Singh (2012), Sivakumar and Woldemeskel (2014), Vignesh et al. (2015), and Yasmin and Sivakumar (2018) are worth mentioning, as they have addressed the connections in streamflow among the stations in the contexts of nonlinear dynamics and complex networks, including for

catchment classification framework, which are of particular interest in the present study. The 639 streamflow stations and the streamflow data exhibit enormous variations in their characteristics, even up to four orders of magnitude, as presented in Table 4.2. For instance: (1) basin drainage area ranges from 10.62 km^2 (4.1 mi^2) to 35,224 km^2 (13,600 mi^2); (2) station elevation ranges from 0 m to 2996 m (9830 ft); and (3) mean flow ranges from 0.0549 m^3/s (1.94 ft^3/s) to 381.59 m^3/s (13,476 ft^3/s). Figure 4.6, for instance, presents the variations in the mean (Figure 4.6(a)), standard deviation (Figure 4.6(b)), and coefficient of variation (Figure 4.6(c)) of flow values in all the 639 stations. More details about this streamflow data set are available in Vignesh et al. (2015), including for a summary of maximum and minimum values of some important characteristics of the stations and flows.

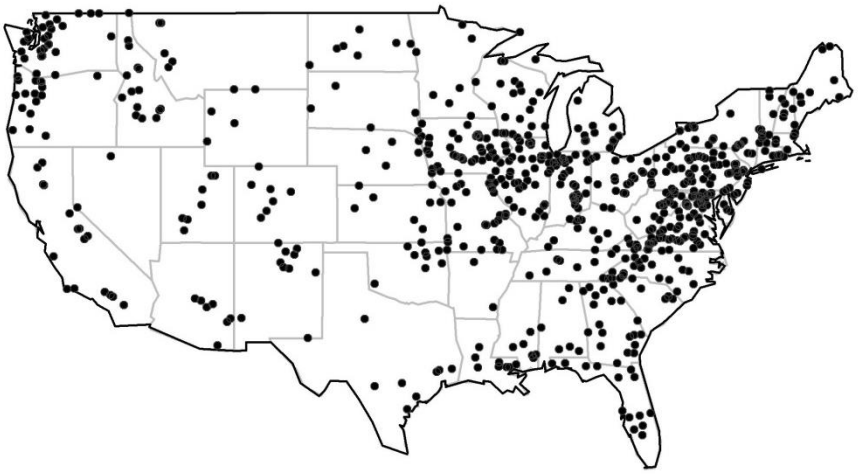


Figure 4.5: Locations of 639 streamflow stations in the US.

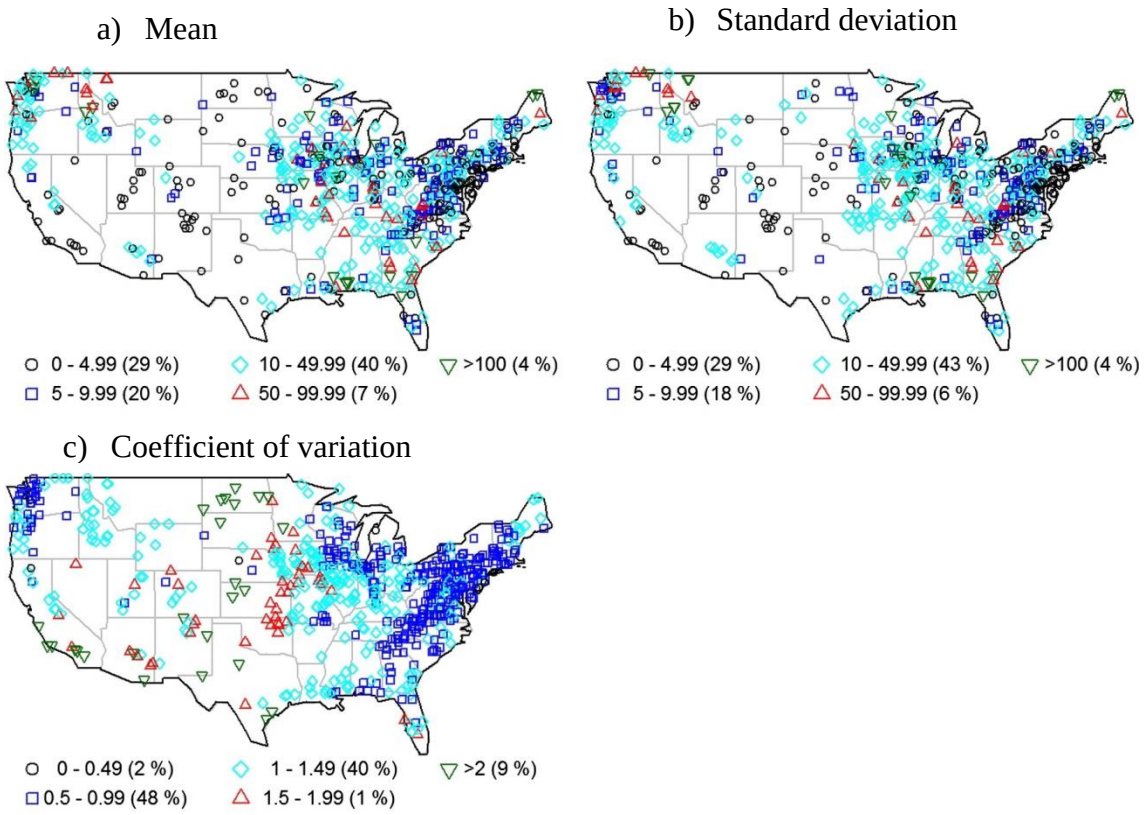


Figure 4.6: Statistical characteristics of streamflow from 639 stations in the US: (a) mean; (b) standard deviation; and (c) coefficient of variation.

Table 4.2: Characteristics of 639 catchments and monthly data in the US.

	Minimum	Maximum	Station (State)
Latitude	26.93°	49°	Minimum: #02256500 (FL)[a] Maximum:#12306500 (ID)
Longitude	- 124.07°	- 67.94°	Minimum: #14325000 (OR) Maximum:#01022500 (MA)
Drainage area (km^2)	10.62 (4.1 mi^2)	35,224 (13,600 mi^2)	Minimum: #1188000 (CT) Maximum:#2226000 (GA)
Elevation (m)	0	2996 (9830 ft)	Minimum: #2310000 (FL) Maximum:#7083000 (CO)
Flow mean (m^3/s)	0.0549 (1.94ft^3/s)	381.589 (13475.70 ft^3/s)	Minimum: #11063500 (CA) Maximum: #2226000 (GA)
Flow standard deviation m^3/s)	0.1101 (3.89 ft^3/s)	373.77 (13199.64 ft^3/s)	Minimum: #11063500 (CA) Maximum: #13317000 (ID)
Flow CV	0.11	5.56	Minimum: #6775500 (NE) Maximum: #6860000 (KS)

[a] CA - California; CO - Colorado; CT - Connecticut; DE - Delaware; FL - Florida; GA - Georgia; ID - Idaho; KS - Kansas; MO - Missouri; MA- Massachusetts; NE - Nebraska; OR - Oregon.

Chapter 5

Catchment classification using edge betweenness method

5.1 Introduction

In the present study, for the application of community structure methods for catchment classification, the edge betweenness (EB) method (Girvan and Newman, 2002) is used as a representative method. The method is applied for catchment classification in the two study areas discussed in Chapter 4: Australia and the United States. For implementation, monthly hydrologic data from a network of 218 catchments across Australia and from a network of 639 catchments across the United States are studied. This chapter presents the analysis of streamflow data for classification in a purely single-variable sense. For the analysis, different correlation thresholds are selected (i.e., spatial correlation in streamflow between stations) to examine the sensitivity of classification to threshold. Results of catchment classification are interpreted in terms of catchment properties (stream length, elevation, drainage area) and flow properties (mean, coefficient of variation, correlation-distance).

5.2 Classification of Australian catchments

In the implementation of the EB method to monthly streamflow data from 218 catchments in Australia for their classification, links between node pairs (i.e., stations) are assigned based upon the Pearson correlation coefficient of the streamflow data. The range of correlation threshold is chosen to better reflect the influence of the threshold, and is also based on the analysis of streamflow (and other hydrologic) data using networks-based methods (Sivakumar and Woldemeskel 2014; Jha et al., 2015). The threshold levels considered are: T = 0.65, 0.7, 0.75, and 0.8. For a given threshold, any node pair with a correlation coefficient above that threshold value is assigned a link.

Figure 5.1(a) to (d) presents the communities for the four different threshold values: 0.65, 0.70, 0.75, and 0.80. In this figure, the communities are represented with different colors. In general, when the threshold is too low, there will be a very large number of links identified that form large-size communities and will cover very large portion of the study area, which will not be useful for studying the variability of catchment's properties. On the other hand, an extremely high threshold will lead to less connected links that eventually break down the network into very close and smaller neighbours and form more isolated communities, which again will not be meaningful for catchment classification studies. Thus, choosing an ideal threshold value is crucial in order to examine the catchment characteristics/properties to be able to assess each community that is formed in the network and could cover the regions as widely as possible.

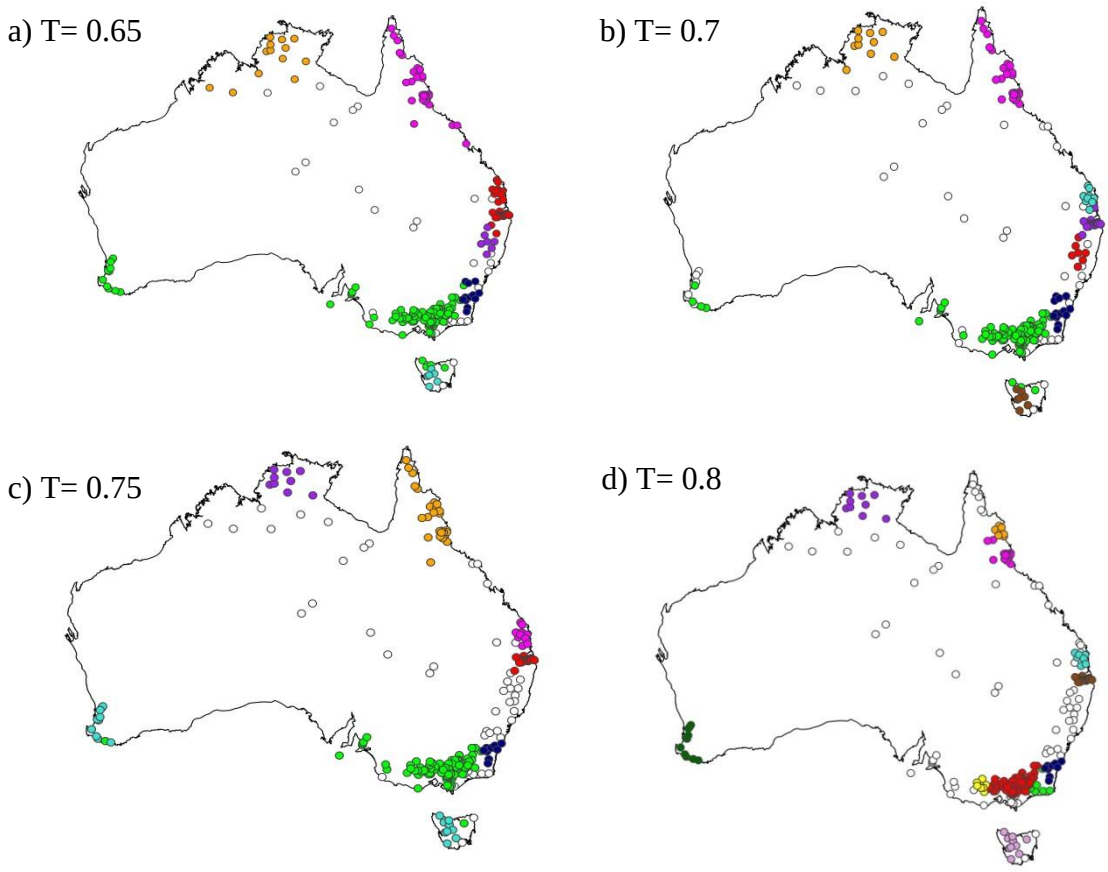


Figure 5.1: Communities identified from the EB method at four different correlation thresholds for streamflow from Australia: (a) T = 0.65; (b) T = 0.7; (c) T = 0.75; and (d) T = 0.8. Each color represents a community with at least 6 stations, while the open circles represent all communities with less than 6 stations. The different colors are used only to distinguish the communities and hold no meaning when comparing across thresholds.

To better understand the hydrologic similarities and better physical explanation and interpretations of the catchment classification, an accurate count of the number of communities with number of stations is obtained, as presented in Table 5.1. As seen from Table 5.1, the number of stations for the largest community decreases when the

threshold value increases, while the number of communities with only a very few catchments within them (one and two catchments) increases when the threshold value increases. Nevertheless, the total number of identified communities varies with the increase in the threshold values.

Table 5.1: Sizes of the identified catchment communities in Australia using the EB method at four different correlation thresholds (T = 0.65, 0.7, 0.75 and 0.8). (NSC is the number of stations in the identified communities, NC is the number of communities, and NS is the number of stations)

T = 0.65			T = 0.7			T = 0.75			T = 0.8		
NSC	NC	NS	NSC	NC	NS	NSC	NC	NS	NSC	NC	NS
1	11	11	1	23	23	1	24	24	1	41	41
2	3	6	2	7	14	2	3	6	2	3	6
3	2	6	3	1	3	3	1	3	3	4	12
5	1	5	5	1	5	4	2	8	4	1	4
6	1	6	7	2	14	5	1	5	5	1	5
7	1	7	9	1	9	9	1	9	7	4	28
13	2	26	10	1	10	10	2	20	9	1	9
28	1	28	14	1	14	17	1	17	10	2	20
29	1	29	18	1	18	19	1	19	11	1	11
94	1	94	25	1	25	26	1	26	13	1	13
Total	**24**	**218**	83	1	83	81	1	81	16	1	16
			Total	**40**	**218**	**Total**	**38**	**218**	53	1	53
									Total	**61**	**218**

In view of the above observations, the communities identified for the threshold value T = 0.8 (Figure 5.1(d)) are chosen, based on their boundaries and regions, for better interpretation of the catchment characteristics. Figure 5.2 shows these communities (an enlarged version of Figure 5.1(d)), merely for better visualization. As seen from Figure 5.2, a total of 11 communities that have at least 6 stations (indicated with different colors) are studied.

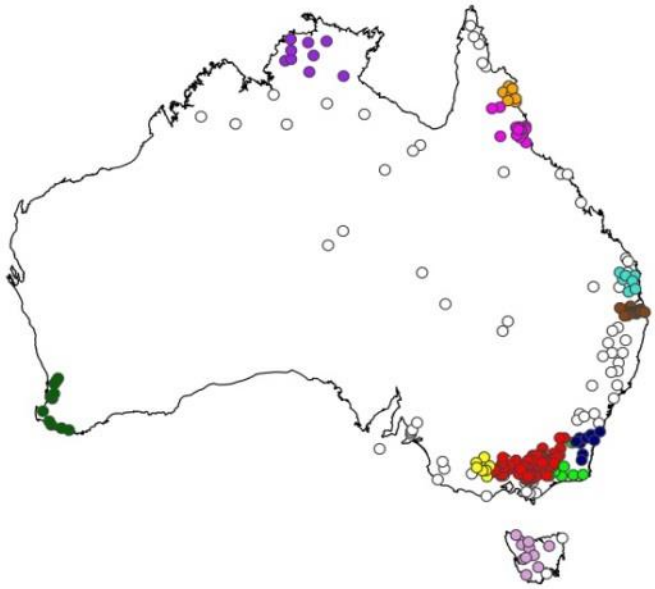


Figure 5.2: Communities identified from the EB method for correlation threshold T = 0.8 for Australia. Each colour represents a community with at least 6 stations, while the open circles represent all communities with less than 6 stations.

As seen from Table 5.1, for T = 0.8, four communities have more than 10 stations (with 53, 16, 13 and 11 stations), combining to form about 43% of the total

number of stations (93 out of 218), and only the 11 largest communities (out of the total 61) combine to have almost 70% of the total number of stations (150 out of 218). This seems to suggest that each catchment within a large community has strong connections with the rest of the catchments in that community, regardless of the distance between them or whether they are located in different basins/regions. Communities with only one catchment (41 communities) form almost 70% of the total number of communities identified (61), but combine to have only about 20% of the total number of stations (41 out of 218). This seems to indicate that each catchment in these small communities has no connection or only very little connection with the other catchments, regardless of their presence within the same basin/region. These observations suggest the important role of catchment and flow properties (which, in turn, are influenced by geographic and climatic characteristics), among others, in community formation, i.e., for catchment classification. This can be examined further, by linking the identified communities to catchment/flow properties. In what follows, this is done with respect to the following: station drainage area, station stream length, and station elevation (as the catchment characteristics) as well as station flow mean and station flow coefficient of variation (CV) (as the flow characteristics).

Figure 5.3 presents the relationship of the three catchment characteristics (i.e., drainage area, stream length, and elevation mean) with the flow mean for the 11 largest communities (150 stations), while Figure 5.4 presents their relationship with the flow CV. The colors used in these figures for the selected communities correspond to the colors of the communities in Figure 5.2, purposely for visualization of the communities' locations. By referring to Figure 5.2, communities that are located in the southeast (coloured in green – community 1, blue – community 2, red – community 13, and

yellow – community 54), north (coloured in purple – community 23), northeast (coloured in orange – community 26 and pink – community 27), east (coloured in cyan – community 29 and brown – community 32), Tasmania (coloured in light purple – community 43), southwest (coloured in dark green – community 57) are studied.

As seen from Figure 5.3(a) and (b), the relationships (drainage area vs. flow and stream length vs. flow) are not significantly different for most communities, since only a few stations are out of the cluster in terms of stream length as compared to the drainage area, and so may be neglected. Similarly, the relationship between drainage area and flow CV (Figure 5.4(a)) and the stream length and flow CV (Figure 5.4(b)) have sparse distribution. However, as seen from Figure 5.3(c) and Figure 5.4(c), the relationship of the elevation mean with the flow mean (Figure 5.3(c)) and the flow CV (Figure 5.4(c)) is more clustered and almost a straight line, especially for the four communities that are mostly located in the south-eastern part, i.e., Communities numbered as 1 (green), 2 (blue), 13 (red), 54 (yellow) (Figure 5.2). This seems to suggest that the straight line relationship is most likely due to the closeness factors (geographic proximity) within the catchments in the community.

Overall, the EB method and its ability to classify catchments according to connectivity as its basis, without prior information about the catchment physics but solely relying on connection measurement, has proven to be useful. Furthermore, the communities that are in the northern part (community 23 (purple)) and in the south-western part (community 57 (dark green)) vary in their elevation mean, suggesting that the catchments have strong connections among themselves, regardless of the difference in their elevations.

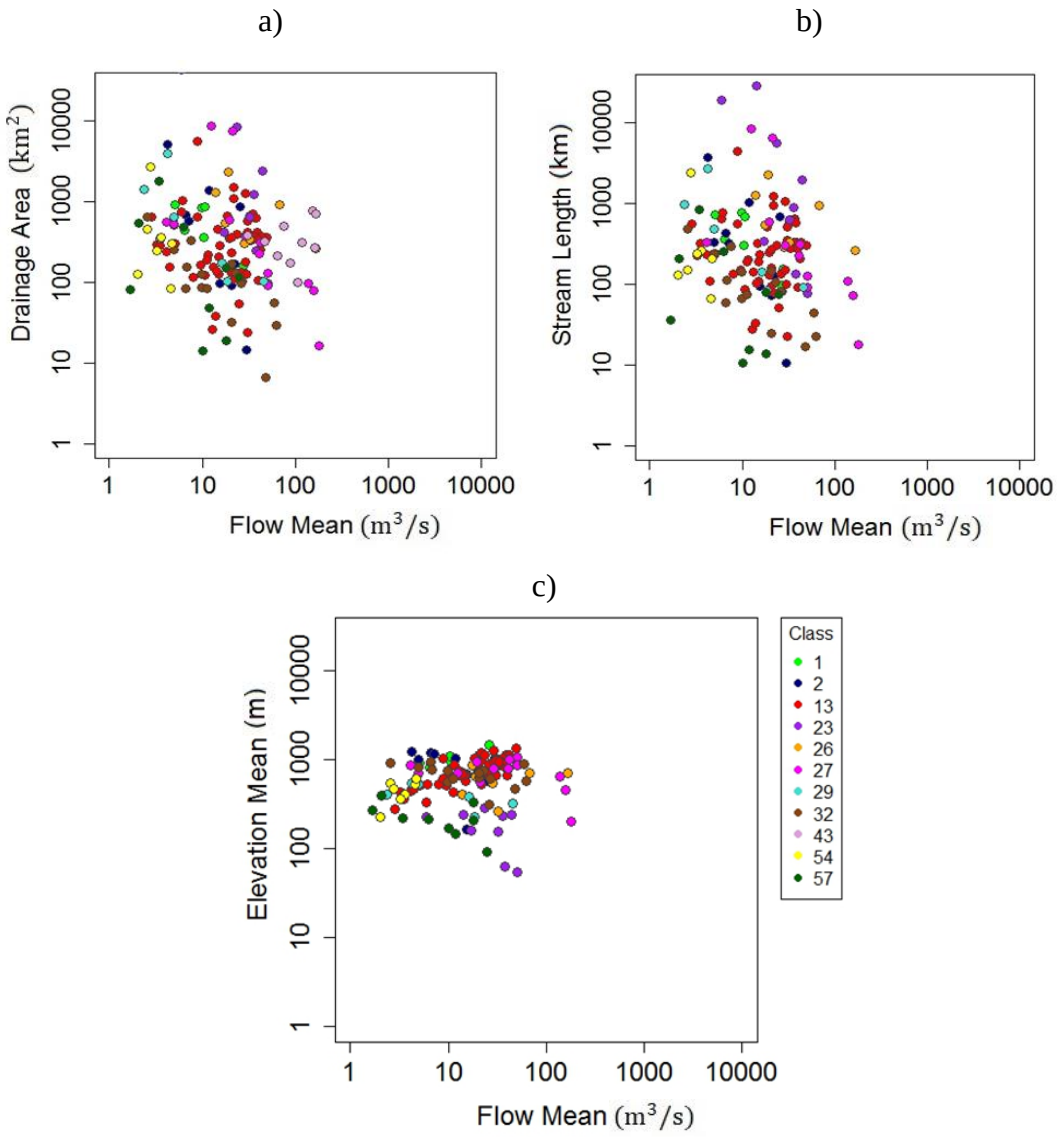


Figure 5.3: Relationship between station drainage area (a), stream length (b), and elevation mean (c) against flow mean for 11 largest communities (150 stations) in Australia. Stations in 11 communities are plotted in colour, corresponding to the legend.

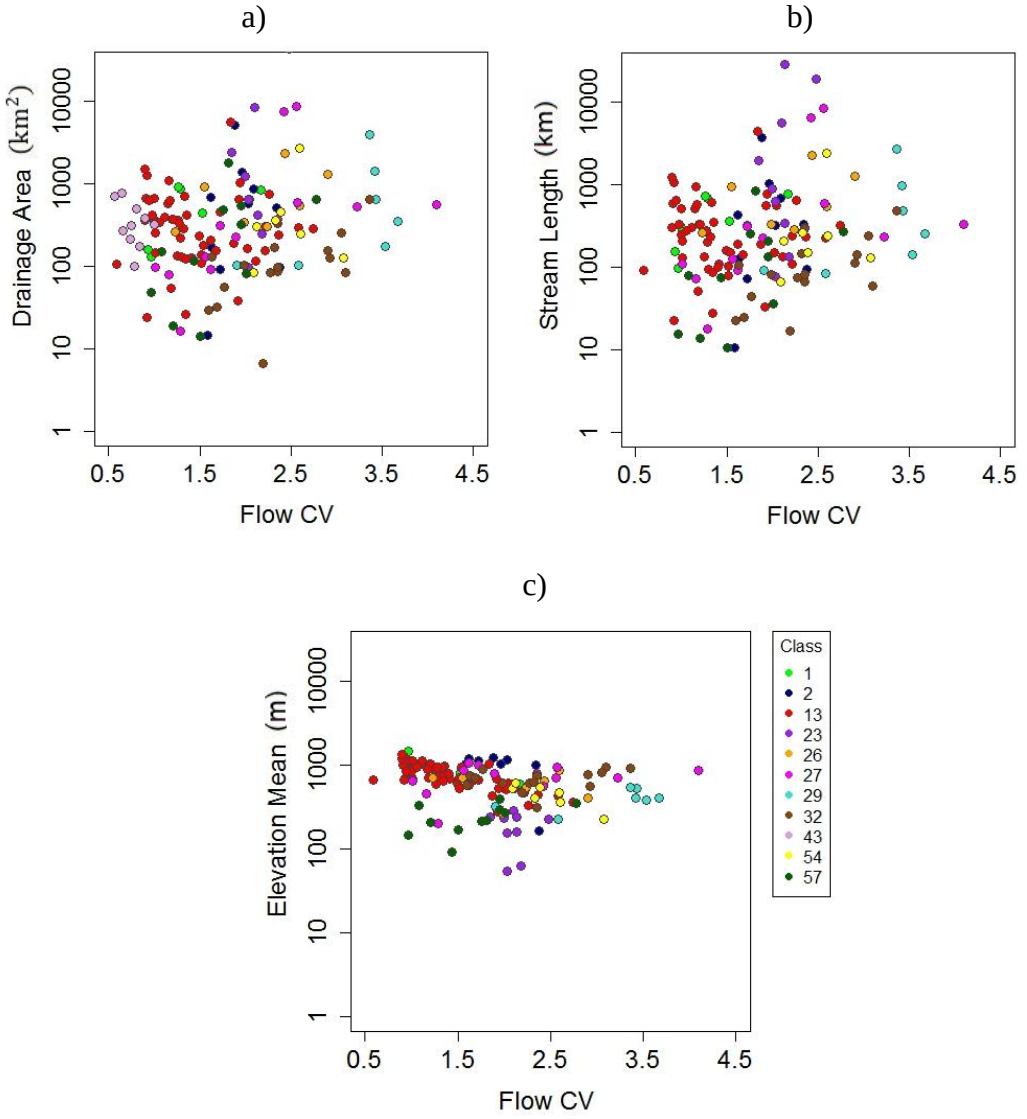
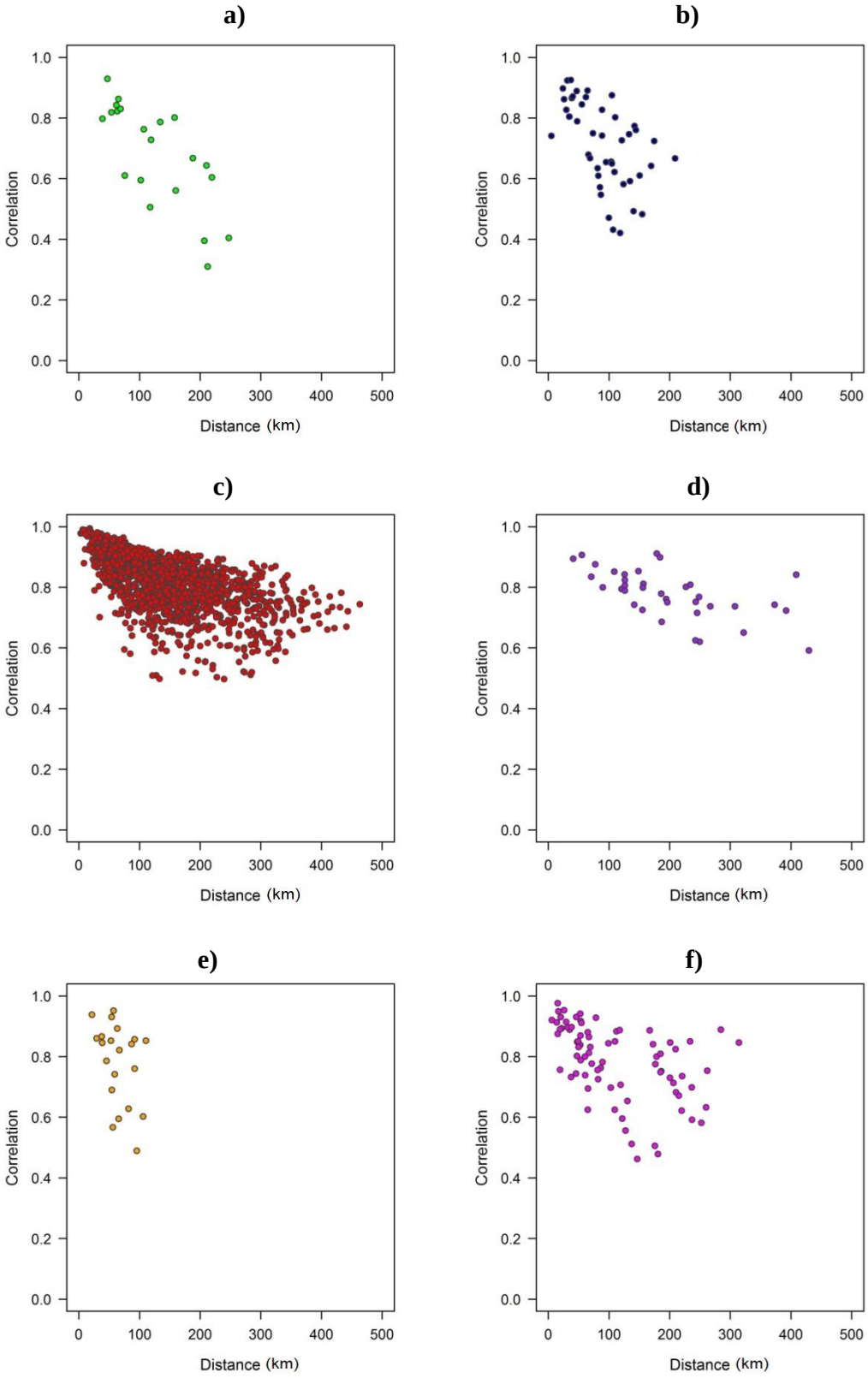


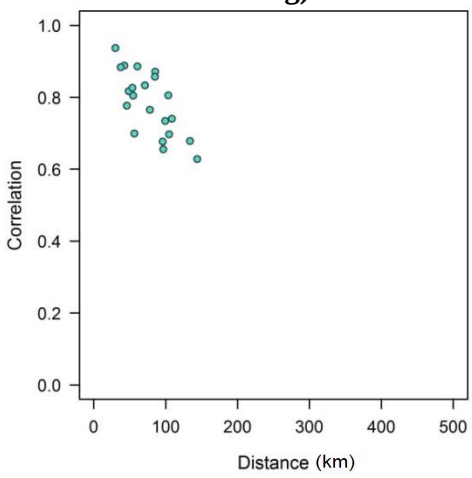
Figure 5.4: Relationship between station drainage area (a), stream length (b), and elevation mean (c) against flow CV for 11 largest communities (150 stations) in Australia. Stations in 11 communities are plotted in colour, corresponding to the legend.

The usefulness of the EB method for classification is also examined by comparing the distance and correlation between the stations, for the respective communities. Figure 5.5 shows the distance-correlation comparison for the above 11 communities. It can be seen that communities 13 (red), 23 (purple), 27 (pink), 43 (light

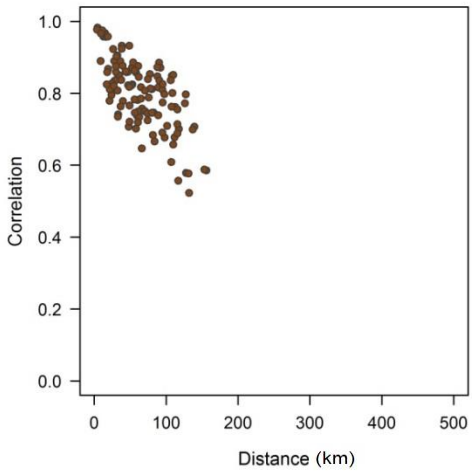
purple), and 57 (dark green) (Figure 5.5(c), (d), (f), (i) and (k)) retain relatively higher correlations as the distance increases. For community 13 (red), it is not surprising for a large community to span large distances, since very strong correlations could help the connection links to be connected with stations that are located over long distances. Nonetheless, communities 1 (green), 2 (blue), and 26 (orange) (Figure 5.5(a), (b), and (e)) with low correlations are found to have connections, perhaps due to the short distances. Thus, the geographic proximity and river network could also be important factors in catchment classification. For the communities 23 (purple), 27 (pink), 43 (light purple), and 57 (dark green), the number of stations in each community is relatively smaller and the distributions are sparser (Figure 5.5(d), (f), (i) and (k)). These communities are able to be formed regardless of the (long) distance. This seems to suggest that the strong correlations lead to the stations spanning over long distances.


a)
b)
c)
d)
e)
f)
Correlation
Distance (km)
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
400
500

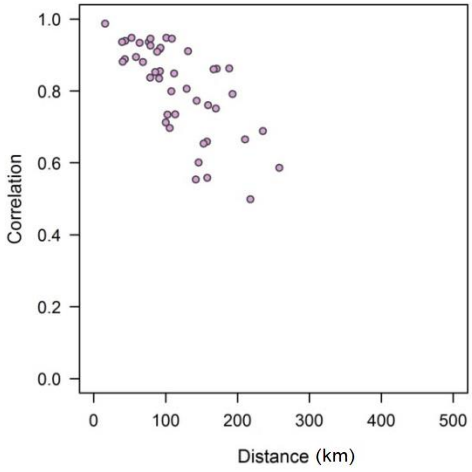
g)


Correlation
1.0
0.8
0.6
0.4
0.2
0.0
0
100
200
300
400
500
Distance (km)

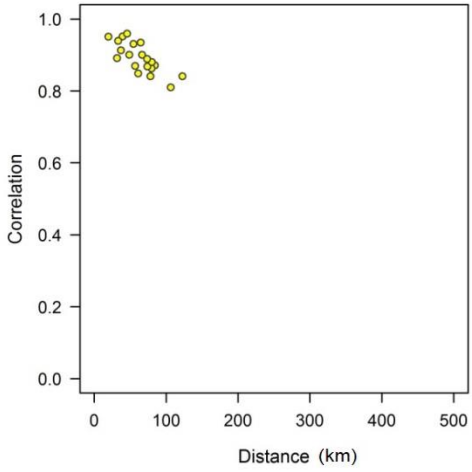
h)


Correlation
1.0
0.8
0.6
0.4
0.2
0.0
0
100
200
300
400
500
Distance (km)

i)


Correlation
1.0
0.8
0.6
0.4
0.2
0.0
0
100
200
300
400
500
Distance (km)

j)


Correlation
1.0
0.8
0.6
0.4
0.2
0.0
0
100
200
300
400
500
Distance (km)

k)

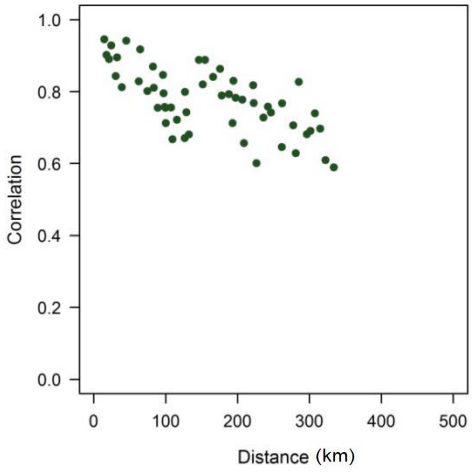

Correlation
1.0
0.8
0.6
0.4
0.2
0.0
0
100
200
300
400
500
Distance (km)

Figure 5.5 Distance-correlation relationship for 11 largest communities in Australia, corresponding to the colouring scheme in Figure 5.2; see text for additional details.

5.3 Classification of the catchments in the United States

In this section, the EB method is implemented to a large network of catchments in the United States to further examine the usefulness and effectiveness of the method to a large region with physically different climates and catchment conditions. Monthly streamflow data from a network of 639 stations across the contiguous US are analyzed. As with the analysis for the Australian catchments above, four different threshold values (T = 0.7, 0.75, 0.8, 0.85) are considered to assess the influence of the correlation threshold on the classification of the catchments. Figure 5.6(a) to (d) presents the catchment communities identified for these four threshold values.

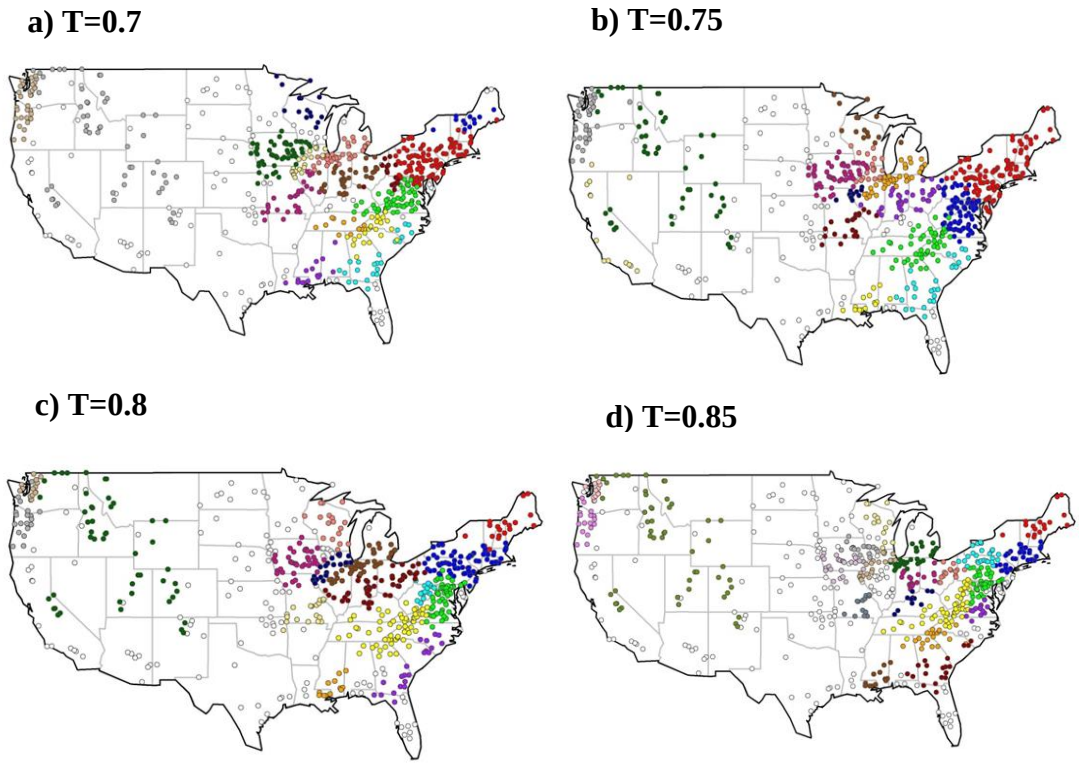


Figure 5.6: Communities identified from the EB method at four different correlation thresholds for 639 catchments in the US: (a) T = 0.7; (b) T = 0.75; (c) T = 0.8; and (d) T = 0.85. Each colour represents a community with at least 20 stations, while the open circles represent all communities with less than 20 stations. The different colours are used only to distinguish the communities and hold no meaning when comparing across thresholds.

Similar to the classification for Australian catchments above, an accurate count of the number of communities with the number of stations for the selected threshold values is obtained, as shown in Table 5.2. As seen, the number of stations of the largest community decreases when the threshold value increases. It seems that some of the stations form/merge with other communities, as shown by the increases in the total number of communities with the increase in the threshold value. However, this

interpretation is not necessarily valid for every situation, since the total number of communities identified for T = 0.75 (61) is also less than the number of communities based on T = 0.7 (69).

Table 5.2: Sizes of the identified communities in the US using the EB method at *T* = 0.7, 0.75, 0.8 and 0.85. (NSC is the number of stations in the identified communities, NC is the number of communities and NS is the number of stations)

T = 0.7			T = 0.75			T = 0.8			T = 0.85		
NSC	NC	NS	NSC	NC	NS	NSC	NC	NS	NSC	NC	NS
1	39	39	1	33	33	1	59	59	1	83	83
2	3	6	2	5	10	2	8	16	2	11	22
3	1	3	3	2	6	3	4	12	3	4	12
4	1	4	4	1	4	4	1	4	4	2	8
5	2	10	6	1	6	5	1	5	5	1	5
6	3	18	7	3	21	6	2	12	7	1	7
7	2	14	8	1	8	7	2	14	8	2	16
8	1	8	12	1	12	12	1	12	11	2	22
10	1	10	13	1	13	14	1	14	12	1	12
11	1	11	14	1	14	16	1	16	13	1	13
12	1	12	15	1	15	17	1	17	14	2	28
15	2	30	18	1	18	18	1	18	15	1	15
17	1	17	21	1	21	19	1	19	17	1	17
21	1	21	27	1	27	21	2	42	18	1	18
23	1	23	35	1	35	23	1	23	19	3	57

25	1	25
27	1	27
31	1	31
37	2	74
50	1	50
52	1	52
54	1	54
100	1	100
Total	**69**	**639**

36	1	36
40	1	40
48	1	48
49	1	49
50	1	50
76	1	76
97	1	97
Total	**61**	**639**

39	2	78
48	1	48
49	2	98
61	1	61
71	1	71
Total	**93**	**639**

20	1	20
22	1	22
24	1	24
26	1	26
30	1	30
37	1	37
40	1	40
51	1	51
54	1	54
Total	**125**	**639**

Figure 5.7 presents an enlarged version of the communities identified with a correlation threshold value of T = 0.75, i.e., an enlarged version of Figure 5.6(b), for further discussion of the classification results. The threshold value of T = 0.75 is chosen based on the division of the communities that seem to be following the river basins. The ten largest communities indicated with different colours (each with at least 20 stations) are considered for better visualization of community formation and discussion.

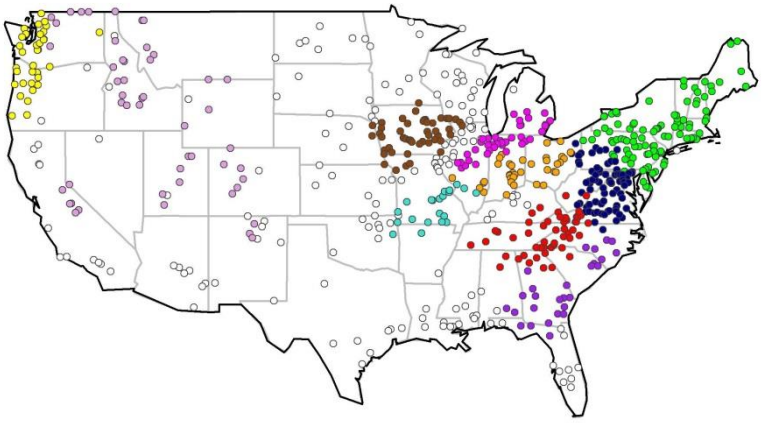


Figure 5.7: Communities identified from the EB method for 639 stations in the US at the correlation threshold, T = 0.75. Each colour represents a community with at least 20 stations, while the open circles represent all communities with less than 20 stations.

As seen from Table 5.2, based on T = 0.75, the ten largest communities only cover only a small percentage of the total number of communities (i.e., 16%), but combine to have almost three-quarters (75%) of the total number of stations (479 out of 639). Additionally, communities with only a few catchments, such as one catchment (33 communities) and two catchments (5 communities), combine to form over 60% of the total number of communities identified (61), although they only cover 7% of the total number of stations (43 out of 639). Hence, to explore further the classification results, including to compare against the outcomes for the Australian catchments, the ten largest identified communities are also interpreted in terms of the following four properties: drainage area, elevation, flow mean, and flow coefficient of variation (CV).

Figure 5.8(a) and (b) presents the relationship of the two catchment characteristics (i.e., drainage area and elevation) with the flow mean for the 10 largest

communities (479 stations), while Figure 5.8(c) to (d) presents their relationship with the flow CV. Again, different colours are used to represent the different communities. The results presented in Figure 5.8 offer some interesting observations, especially since they are different from those obtained for the Australian catchments. For instance, Figure 5.8(a) generally shows a linear relationship between the drainage area and flow mean, even though a few stations in community 1 (coloured in green) are out of the cluster (Even the anomaly presented by the community 1 can be neglected because, for a large community, it has only very little or almost no influence in the network). However, such a linear relationship is not observed for the Australian catchments, as discussed earlier. As for the relationship between elevation and flow mean for the US catchments (Figure 5.8(b)), it appears that communities 1 (green), 3 (blue), 5 (purple), and 53 (yellow) are scattered, while the other communities are more clustered.

The relationship between the drainage area and the flow CV (Figure 5.8(c)) does not seem to indicate a linear relationship. However, the relationship between the elevation and flow CV (Figure 5.8(d)) is mostly linear, except for communities 1 (green), 3 (blue), 5 (purple), and 53 (yellow).

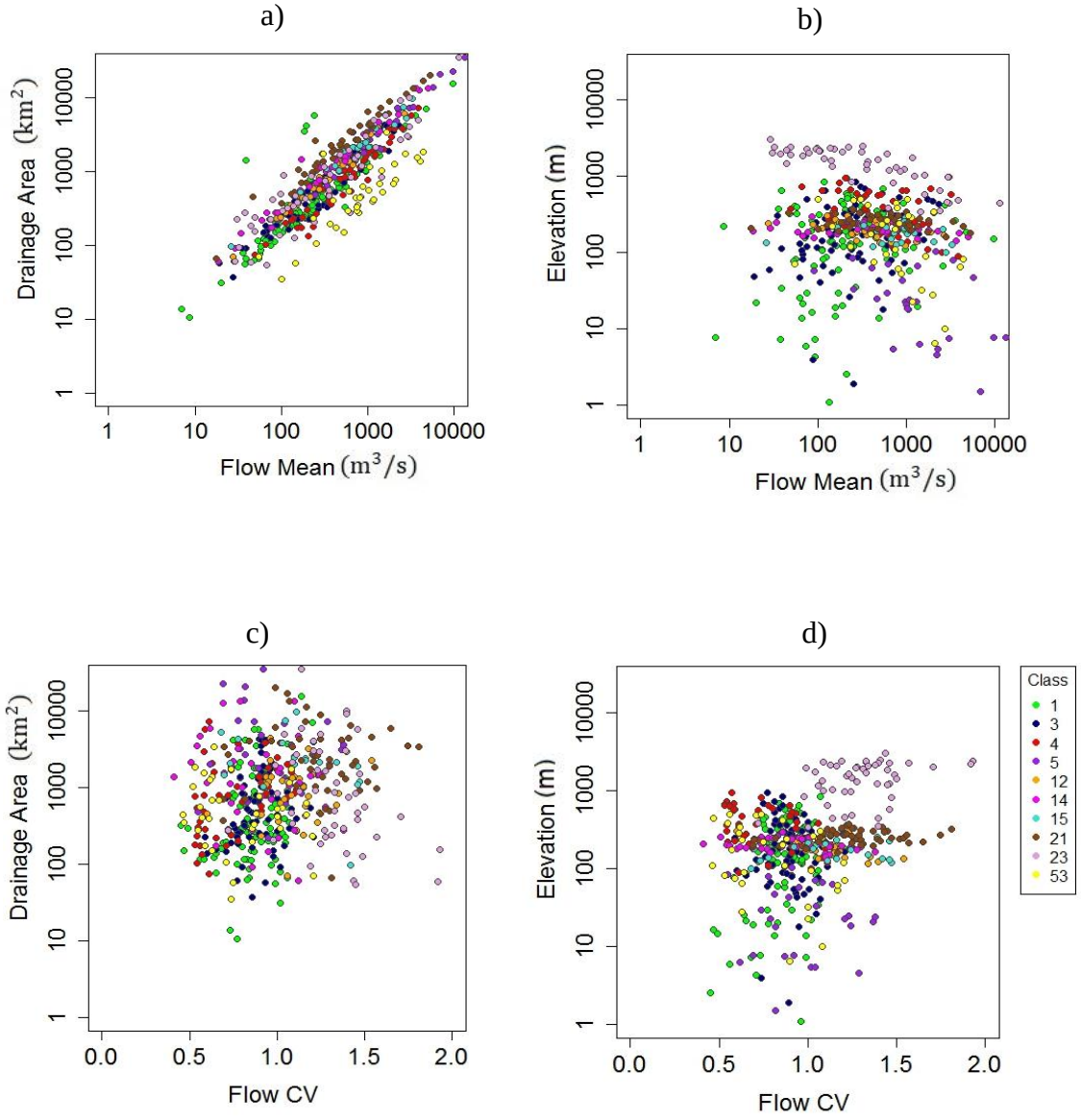


Figure 5.8: Relationship between station drainage area (a) and elevation (b) against flow mean, as well as station drainage area (c) and elevation (d) against the flow CV for ten largest communities (479 stations) in the US. Plots (a) and (b) are in log-log scale, while (c) and (d) are in semi-log scale. Stations in ten communities are plotted in colour, corresponding to the legend.

Figure 5.9 shows, for instance, the relationship between the distance and correlation of two of the communities identified using the EB method: communities 23

(light purple) and 53 (yellow). These two communities are chosen, as they are unique and offer interesting observations and, thus, deserve more discussion. Discussion on other communities is not made essentially to avoid redundancy in the presentation, since the other communities (can be found in the Appendix B.1) have quite similar patterns to the outcomes for the Australian catchments. The correlations of community 23 (light purple) (Figure 5.9(a)) retain relatively higher values as the distance increases. This kind of behavior is observed in at least one community from both the study areas (Australia and the US), as strong correlations tend to be connected with stations over long distances. Figure 5.9(b) shows that the communities have lower correlations as the distance increases. This indicates that connections between stations still exist over short distances (i.e., as neighbours) with a great number of stations in the community.

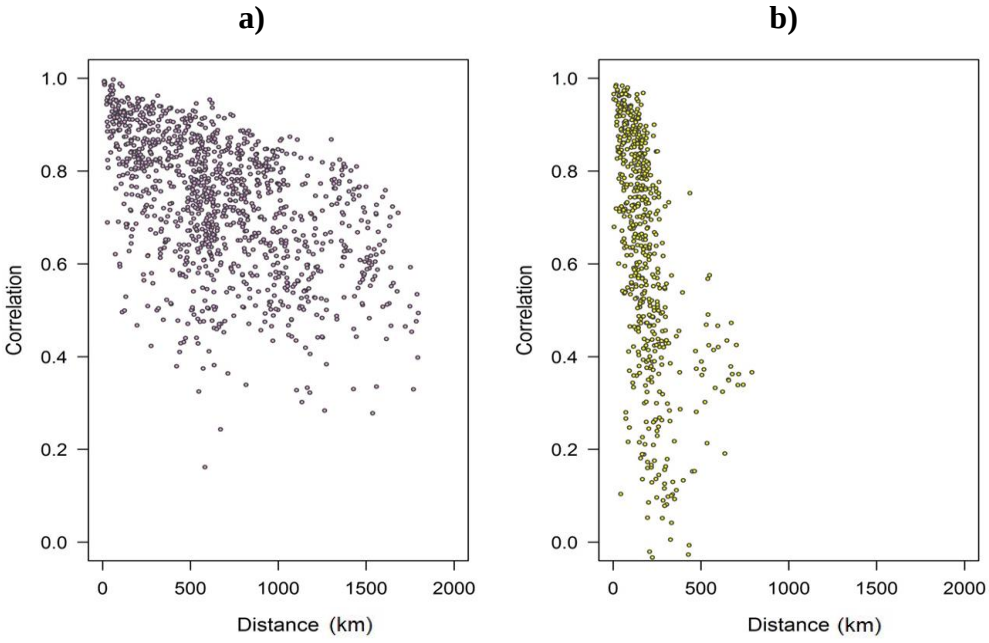


Figure 5.9: Distance-correlation scatterplots for two catchment communities in the US: (a) community 23 (light purple); and (b) community 53 (yellow). The communities correspond to the legend in Figure 5.8(d).

5.4 Summary

This chapter has presented the application of the edge betweenness method for classification of a large number of catchments from two large regions: Australia and the United States. Application of the method to monthly streamflow data from a network of 218 catchments across Australia and 639 catchments across the United States has provided promising results on catchment classification. Although the classification has been carried out to examine the two large regions with different climates and catchment characteristics, the results generally indicate the following; (1) for each study area, a very small number of communities have a large number of catchments within them – for instance, 11 of the largest communities from Australia and 10 of the largest communities from the US combine to represent as much as 70% of the catchments; and (2) a significantly large number of communities have only a very few catchments within them – for instance, almost 70% of the communities have only one or two stations within them, and thus represent only about 20% and 10% of the catchments in Australia and the US, respectively. An examination of the identified communities against some important catchment/flow properties (drainage area, stream length, elevation, flow mean, and flow CV) has offered some interesting observations, as has the distance-correlation relationship. The results also indicate that a similar correlation threshold can be used to study the monthly streamflow for classification in both regions, regardless of the differences in climatic factors and catchment characteristics.

The analysis presented in this chapter is the first ever attempt to apply the concept of community structure for classification of catchments in two different and large regions. The assessment of the EB method, in particular, for the two regions also

sheds some light on the general suitability of the method for catchment classification. Nevertheless, as explained in Chapter 3, the EB method has an important limitation in that it is susceptible to the issue of resolution limit. Therefore, it is necessary to explore possible improvements to the classification of catchments in Australia and in the US. To this end, application of an improved EB method, the Modularity-Density based EB (MDEB) method, proposed in Chapter 3, will be presented in Chapter 6.

Chapter 6

Modularity Density-based Edge Betweenness (MDEB) Method for Catchment Classification

6.1 Introduction

The efficiency of the proposed Modularity Density-based Edge Betweenness (MDEB) method, to overcome the resolution limit of the network, for catchment classification is evaluated using monthly streamflow data from Australia and the United States. For Australia, streamflow data observed over a period of 26 years (1981–2006) from each of 218 stations are considered. For the United States, data observed over a period of 53 years (1950–2002) from each of 639 stations are considered. Different correlation threshold (T) values are used to identify links between the stations, but T = 0.8 for the Australian catchments and T = 0.75 for the US catchments are given particular focus here, following up on the results obtained from the EB method (see Chapter 5). For each

region, the performance of the MDEB method is evaluated with three different scenarios of network sizes: (1) the entire network (i.e., 218 and 639 streamflow stations, respectively); (2) smaller network sizes based on random selection with 100 different realizations (i.e., 100 and 300 randomly selected stations, from the 218 and 639 streamflow stations, respectively – purely to address the network size); and (3) smaller network sizes based on drainage divisions or hydrologic regions (i.e., stations in each of 9 different drainage divisions in Australia and 18 different hydrologic units in the US – to address the network size and the influence of regional similarity). The results are interpreted based on the number of identified communities and the number of stations that change communities when different sizes of networks are used based on the above scenarios. The distance-correlation relationship is also examined for scenarios (2) and (3) (for the first scenario, it was already presented in Chapter 5) based on the selected communities and drainage divisions/hydrologic units.

6.2 Australian Streamflow

6.2.1 Entire Network (218 Stations)

Similar to the application of the EB method in Chapter 5, the MDEB method is applied to monthly streamflow data from the 218 stations across Australia. Figure 6.1, for instance, presents the communities identified using the MDEB method with a threshold value of T = 0.8. In this figure, different colours are used to distinguish the different communities. The results seem to suggest that the communities identified are largely divided geographically. The MDEB method forms large communities, especially in the area clustered with stream-gauges specifically near the coast in the south-eastern part.

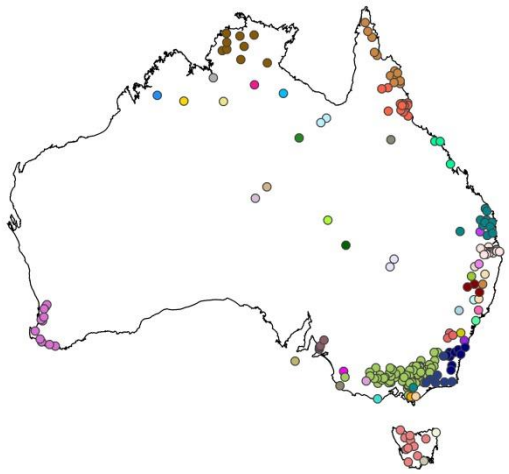


Figure 6.1: Communities identified using the MDEB method for 218 streamflow stations in Australia, with threshold value T = 0.8. Different colours are used only to distinguish the communities.

Table 6.1 presents the count of the number of stations within each community. This type of count can be helpful in identifying whether a given catchment stands on its own as a community or has a certain level of similarity with other catchments, and how. The results in Table 6.1 indicate the following, among others: (1) A total of 52 communities is formed using the MDEB method, for T = 0.8; (2) A significantly large number of communities have only very few catchments within them. For instance, communities with only one catchment and two catchments using the MDEB method are 35 and 9 communities, respectively, and combine to form almost 85% of the total number of communities identified (52). These results indicate that each of these catchments has no or only very little connection with the other catchments, regardless of their geographic proximity with other or their presence within the same basin,

specifically within the river/stream network; and (3) A very small number of communities have a large number of catchments within them. For instance, eight communities have at least 10 catchments within them and combine to form over 65% of the total number of catchments (148 out of 218). This means that each catchment within a given large community has strong connections with the rest of the catchments, regardless of their basin characteristics, specifically in terms of the river/stream network.

With these results, to account for possible changes in network size and to overcome the resolution problem, an attempt is made here to study smaller network sizes and assess the classification outcomes. As mentioned earlier, such smaller networks are produced either in terms of the sheer number of stations (through random realizations) or in terms of the drainage divisions. The results are presented next.

Table 6.1: Sizes of the identified communities using the MDEB method for Australia.

MDEB		
Number of stations in community	Number of Communities	Number of stations
1	35	35
2	3	6
3	3	9
4	1	4

7	1	7
9	1	9
10	3	30
11	1	11
12	1	12
13	1	13
16	1	16
66	1	66
Total	**52**	**218**

6.2.2 Network of 100 Stations through Random Realizations

To account for the change in network size, a network size of 100 streamflow stations is randomly selected from the entire network of 218 catchments in Australia. To account for different combinations of stations to be included in the 100 stations, 100 different realizations are carried out. This means, 100 networks of 100 streamflow stations are analyzed. It is decided to use a network of 100 stations, since this number is still large and roughly about half of the total number of stations (218) and, thus, is sufficient for classification and comparison of results. The random realizations also help not only in considering a specific number of stations but also can cover other aspects, including selection of stations from different geographic regions.

Figure 6.2, for instance, presents the classification results for the case of 100 catchments for 10 out of the 100 random realizations using the MDEB method. Each plot indicates the 100 stations that are selected for each random realization, and these plots are presented to indicate the variability of the locations of the monitoring stations. To examine the influence of network size on classification outcomes, Figure 6.3 shows the communities identified using the MDEB method of two different networks sizes that are studied; (1) network size of 218 stations, where communities with at least 10 stations are indicated with colours (Figure 6.3(a)); and (2) network size of 100 stations, i.e., one of the random realizations, where communities with at least 4 stations are indicated with colours (Figure 6.3(b)). The latter (i.e., just one realization of the 100 realizations) network is shown only for basic comparison purposes, and should not be construed to offer any broader interpretations. Communities coloured with purple in Figure 6.3(a) and coloured with green in Figure 6.3(b) are not examined due to the inconsistency of the stations in the communities.

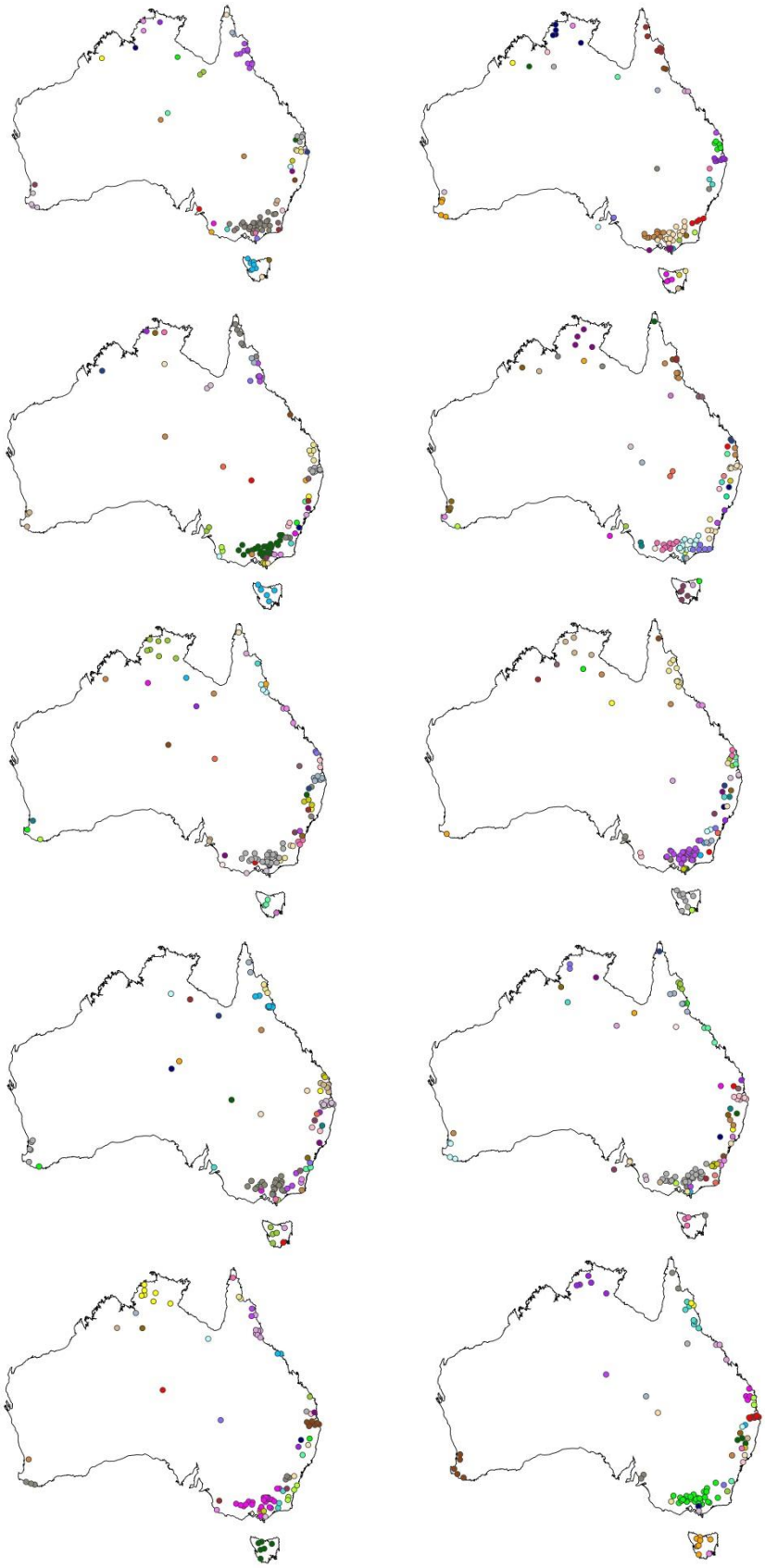


Figure 6.2: Classification of 10 randomly selected streamflow networks of 100 catchments from Australia using the MDEB method. Each colour represents a different community.

a) N = 218 stations b) N = 100 stations

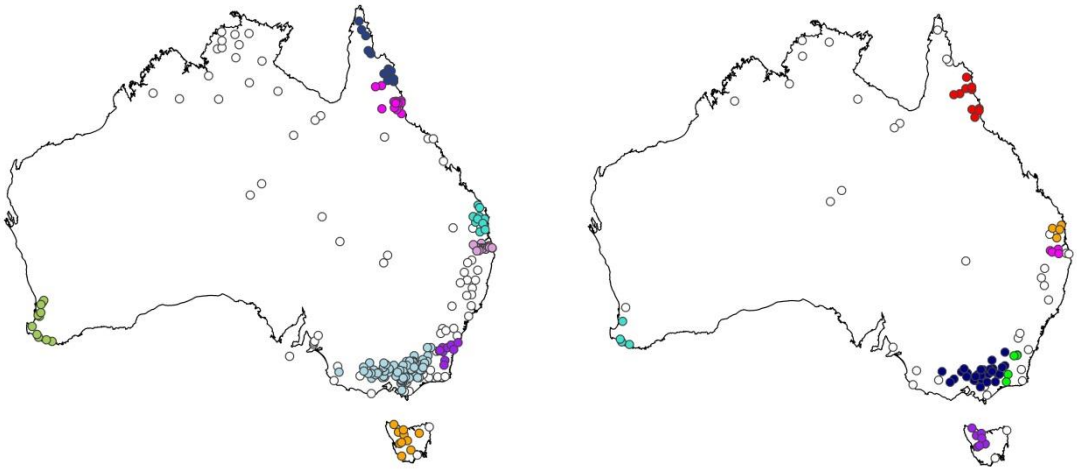
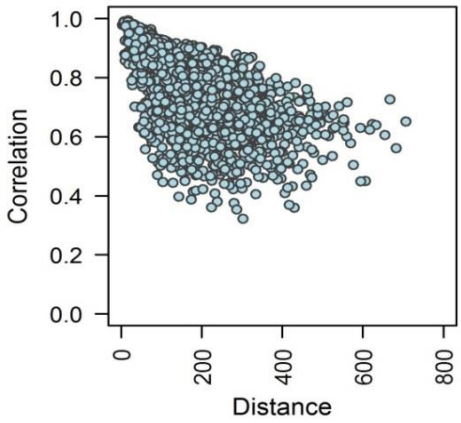
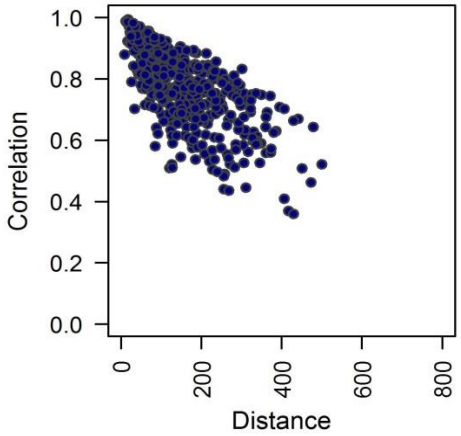
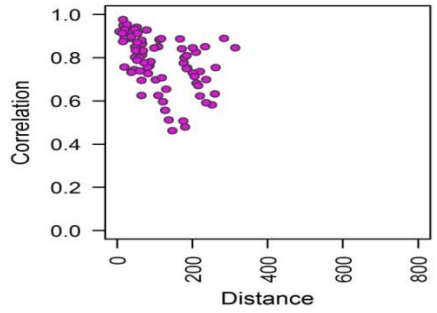
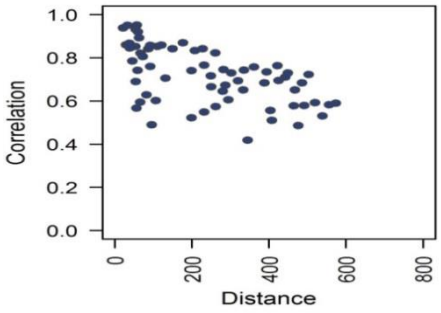
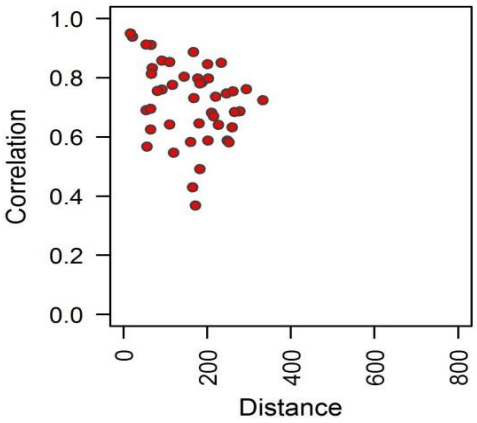
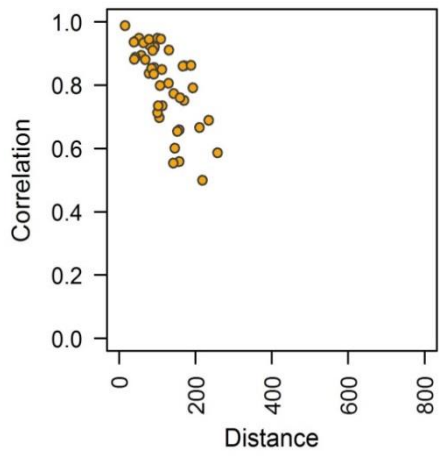
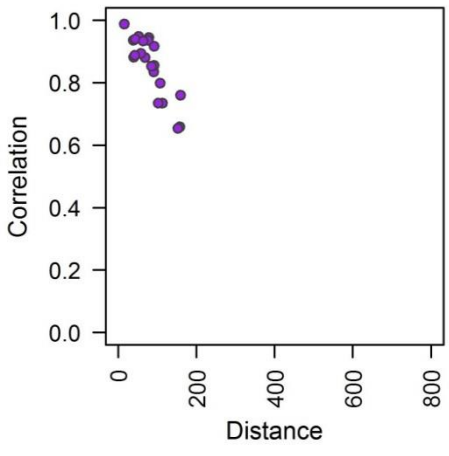
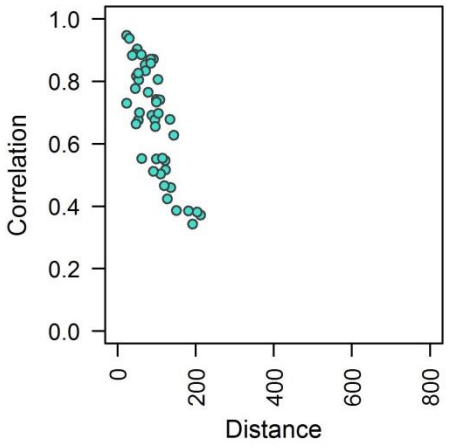
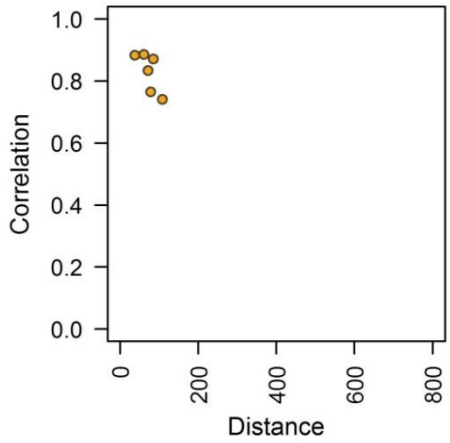
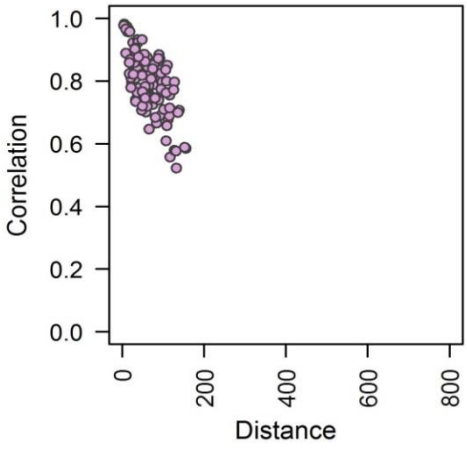
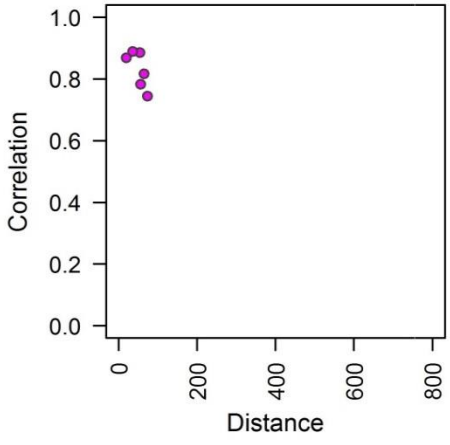


Figure 6.3: Communities identified using the MDEB method for two different sizes of networks: (a) 218 stations and (b) 100 stations. Each color represents a community with at least 10 stations and 4 stations, respectively. The open circles represent all communities with less than these numbers, respectively.

Figure 6.4 shows the scatterplots of the distance-correlation comparison for (a) seven selected communities from the classification based on 218 stations (Figure 6.3(a) except the community colored in purple) and (b) six communities are based on classification of one of the realizations of 100 randomly selected stations (Figure 6.3(b) except the community colored in green) using the MDEB method. These results can be

used to examine whether the communities from the six selected regions (including one each in the north, west, part of southeast, and Tasmania, and two in the east) either from 218 stations or from 100 randomly selected ones retaining their behaviour (when correlation against the distance) for a scale limit problem assessment. Overall, communities from Figure 6.4(a) of rows 1, 3, 4, 5, and 6 are seen to be very similar to the communities obtained from the random realization, although the communities from the 100 stations are mostly relatively very small (at most 4 stations due to the size of the network). Also, some communities tend to merge into different communities, as shown by some stations from communities that are coloured in purple and blue (Figure 6.4(a) at row 2) and are combined to form a community that is coloured in red (Figure 6.3(b) at row 2). These results might be due, as may be seen from Figure 6.4 at row 2, the community in red that has relatively high correlations in short distances. Although some stations from the community in blue (Figure 6.4(a)) are not selected in the realization, the remaining stations are merged with the community that is closer; i.e., community in red (Figure 6.4(b)). This kind of scenario is termed as the ‘stations that changed’ in further discussion in the following sections. Therefore, this also shows that there certainly exist a number of central links to accommodate the connections among the catchments and the communities may change if stations with high centrality are removed from or added to the network.

	(a) 218 stations	(b) 100 stations
1.		
2.	 	

3.	 Correlation 1.0 0.8 0.6 0.4 0.2 0.0 0 200 400 600 800 Distance	 Correlation 1.0 0.8 0.6 0.4 0.2 0.0 0 200 400 600 800 Distance
4.	 Correlation 1.0 0.8 0.6 0.4 0.2 0.0 0 200 400 600 800 Distance	 Correlation 1.0 0.8 0.6 0.4 0.2 0.0 0 200 400 600 800 Distance
5.	 Correlation 1.0 0.8 0.6 0.4 0.2 0.0 0 200 400 600 800 Distance	 Correlation 1.0 0.8 0.6 0.4 0.2 0.0 0 200 400 600 800 Distance

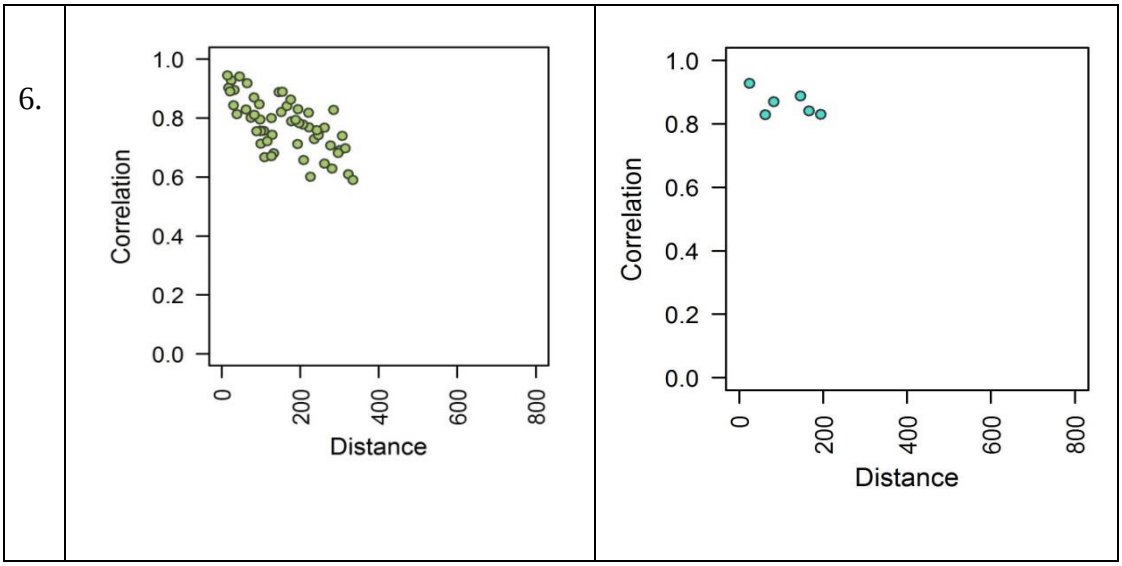


Figure 6.4: Distance-correlation scatterplots for the selected communities from six regions in Australia by the MDEB method (see Table 6.1), ((a)1–6) base classification and ((b)1–6) 100 randomly selected stations.

6.2.3 Networks of 9 Drainage Divisions

In addition to addressing the network size in terms of the sheer number of stations (through random selection), an attempt is also made to examine the influence of drainage division on the classification outcomes. To this end, the drainage divisions and river regions reported by the Bureau of Meteorology (BoM) of the Commonwealth of Australia are considered, as shown in Figure 6.5. There is a total of 13 drainage divisions in Australia. However, in this present study, only nine out of the 13 regions are used, to be consistent with the locations of the 218 streamflow stations considered in this study. The four regions that are not included in this study are the North Western Plateau (NWP), Pilbara-Gascoyne (PG), the South Australian Gulf (SAG), and the South Western Plateau (SWP). The locations of the stations in the nine regions are presented in Figure 6.6. The stations are indicated in different colours to represent each

particular region, and also for better visualization and interpretation for assessment of the network size. Figure 6.6 indicates, for instance: (1) regions such as SWC (stations colored with brown) (Figure 6.6(g)), TTS (stations colored with yellow) (Figure 6.6(j)) and Tasmania (stations colored with pink) (Figure 6.6(i)) have only a small number of stations, yet are compacted along the coast apart from the middle region (i.e., the LEB (stations colored with turquoise) (Figure 6.6(b)); (2) most areas in the north down to south-east, which includes the CC (stations colored with red) (Figure 6.6(a)), NEC (stations colored with blue) (Figure 6.6(d)), SEC for NSW (stations colored with purple) (Figure 6.6(e)), SEC for Victoria (stations colored with orange) (Figure 6.6(f)) and MDB (green) (Figure 6.6(c)), have a large number of stations and also are stretched in distances. These observations are important in offering further explanations related to distance-correlation relationships, discussed next (see Figure 6.7).

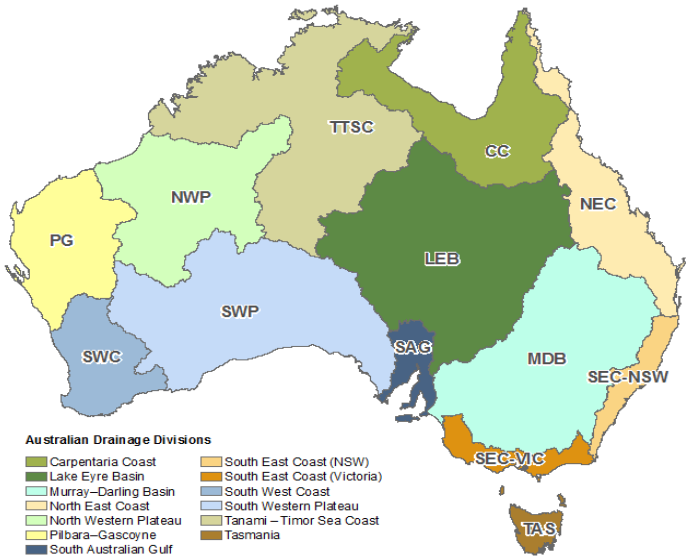


Figure 6.5: Regions according to drainage divisions and river regions (**source of the map**: Website of Commonwealth of Australia (Bureau of Meteorology), 2016).

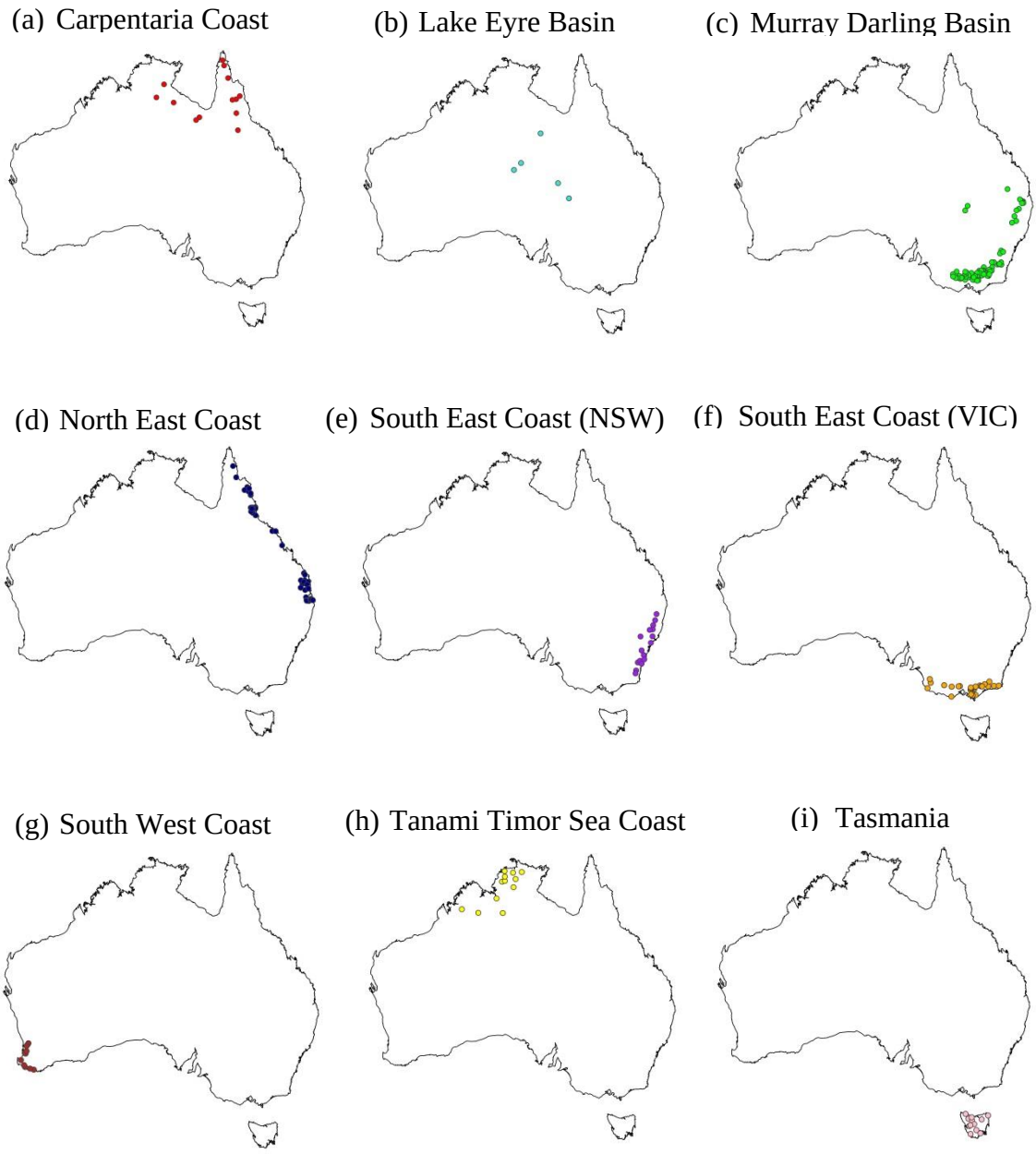
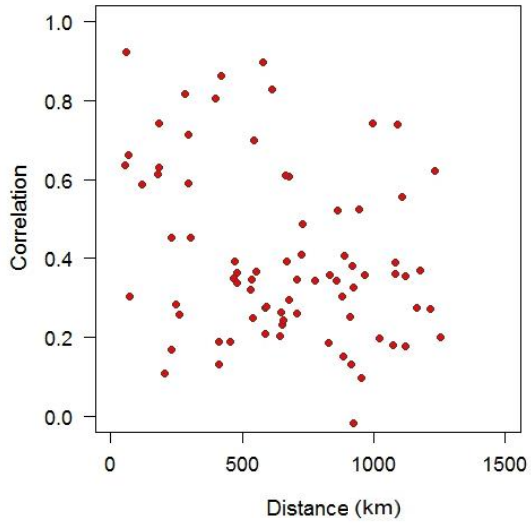


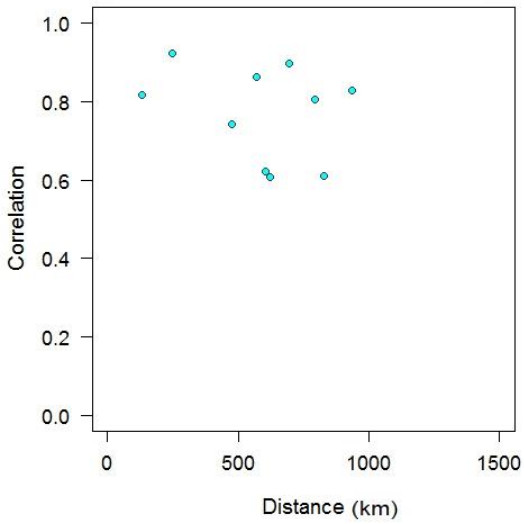
Figure 6.6: Streamflow station locations in nine drainage divisions in Australia considered in this study.

Figure 6.7(a-i) presents the distance-correlation relationship for each of the nine regions. It appears that, in general, different regions tend to have different relationships. For instance, catchments in regions like CC, LEB, and TTS (red, turquoise and yellow) (Figure 6.7(a, b, and h)) are relatively sparse in distribution compared to the rest of the regions, which means that these catchments have very low correlations as distances between themselves increases. This is not surprising when referring to Figure 6.6(a, b, and h), since the station locations are distributed over long distances. On the other hand, Figure 6.7(e), (f), and (i), coloured in purple, orange, and pink, respectively, shows a decrease in correlations when the distance decreases. It is important to mention that the regions from MDB and NEC (as coloured in green and blue) (Figure 6.7(c) and (d)), where the distributions are separated into two groups (which seems sensible according to their catchment locations), are stretched in distance leading to a greater variation of this kind of relationship.

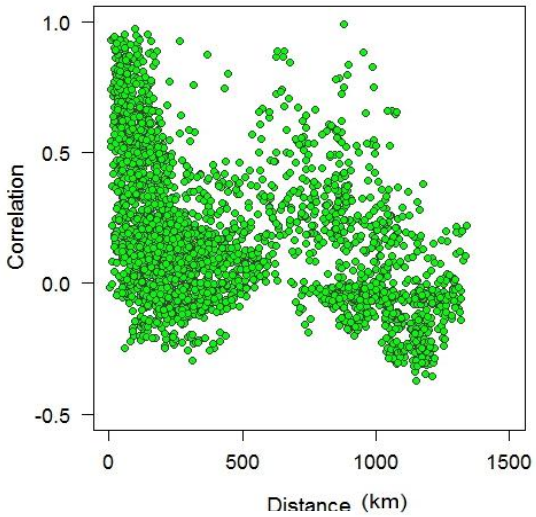
(a) Carpentaria Coast



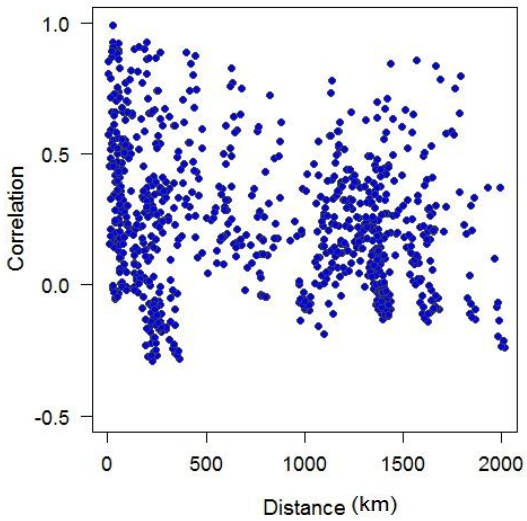
(b) Lake Eyre Basin



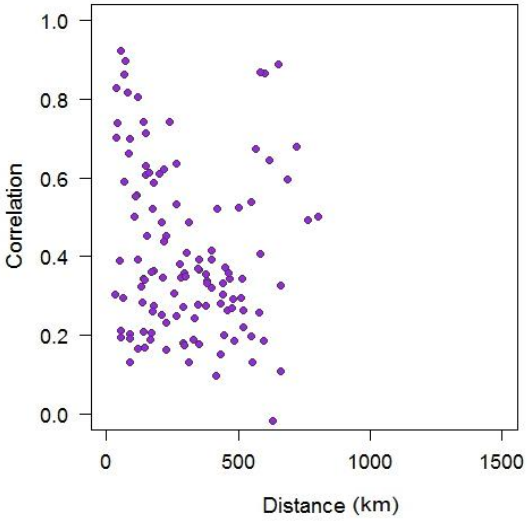
(c) Murray Darling Basin



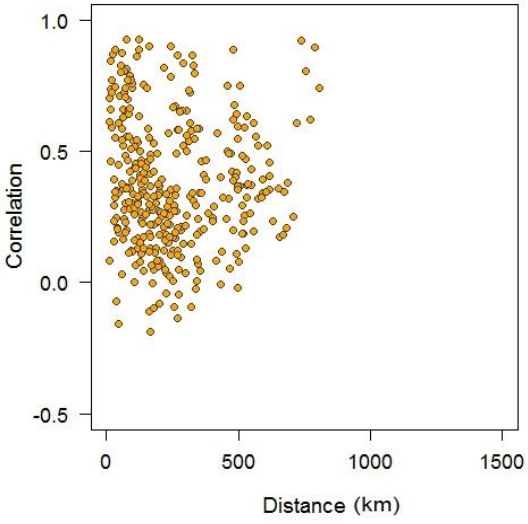
(d) North East Coast



(e) South East Coast (NSW)



(f) South East Coast (VIC)



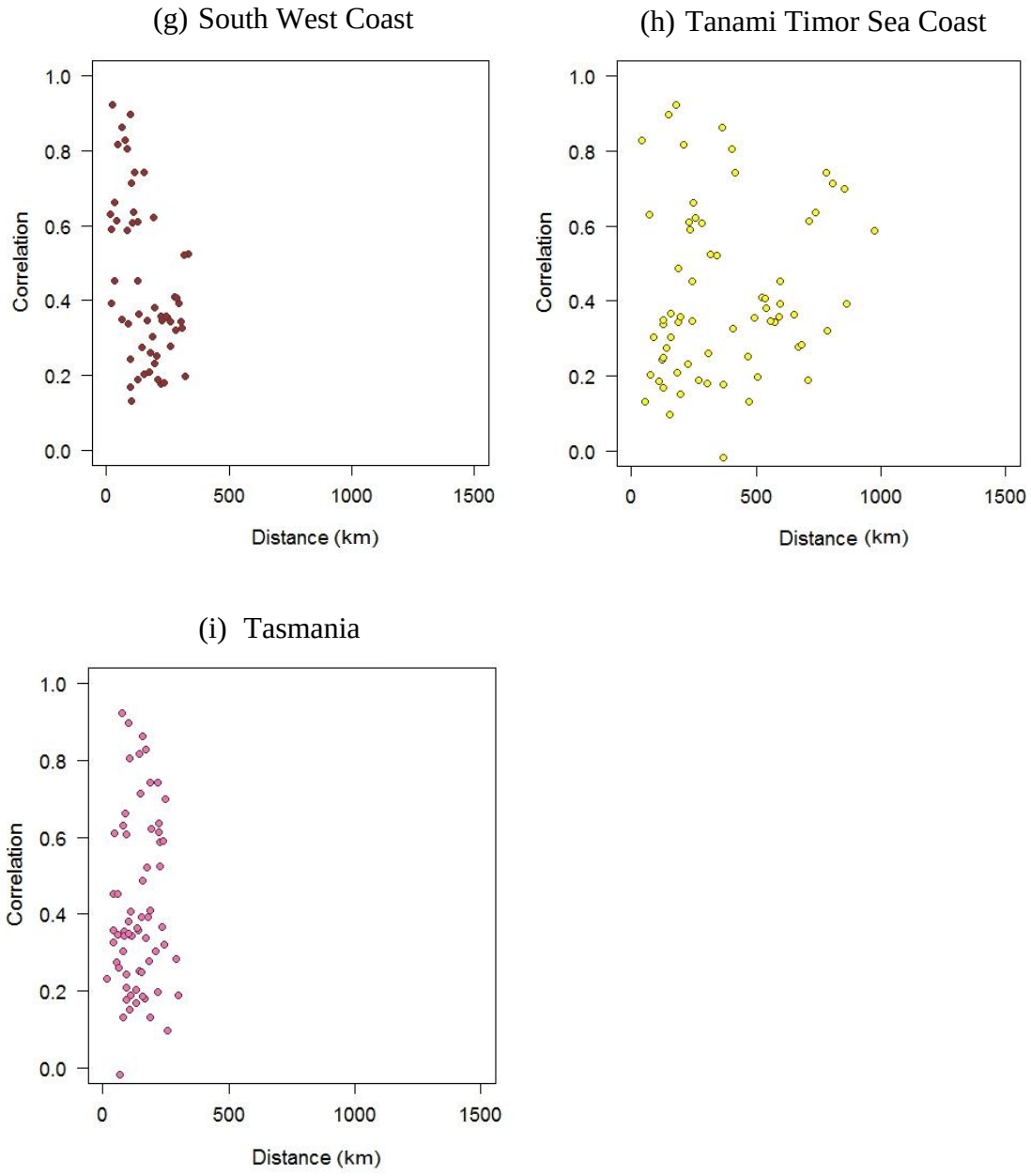


Figure 6.7: Distance-correlation relationship for nine drainage division regions (see Figure 6.6) in Australia.

6.2.4 Comparison between EB and MDEB methods for Catchment Classification

In order to examine any improvement made by the MDEB method over the EB method, it will be helpful to have an accurate count of the communities identified for each realization by the methods. To this end, Figure 6.8 shows the number of communities identified for each realization using the EB method (Figure 6.8(a)) and the MDEB method (Figure 6.8(b)). In the plots, the horizontal red solid lines represent the number of communities identified from the base classification (i.e., 218 stations), which are 61 and 52, by the respective methods, and considered as a reference for comparison.

As seen, the gap between the number of communities identified for the 218 catchments and the 100 catchments is relatively larger for the EB method (Figure 6.8(a)) when compared to the MDEB method (Figure 6.8(b)). The results also show that the MDEB method tends to form small-size communities, since the number of communities identified for the 100 stations (with the different realizations) is very close to the base classification (Figure 6.8(b)). Moreover, the community formation is quite varied for the MDEB method when compared to the EB method. For instance, considering the 100 realizations, the range of the number of communities identified by the MDEB is mostly between 20 to 50, while the EB method has number of communities mostly ranging from 30 to 50. These results seem to suggest that the MDEB method is more natural in network partition when compared to the EB method. In the EB method, the partition is almost constant considering the sparseness of distribution of the monitoring stations. This is not surprising, as the EB algorithm is dependent on the size of the network, an important limitation, as discussed in Chapter 3.

The MDEB method mainly forms the communities by taking into account the robustness of connectivity by the station pair, regardless of the size of the network. It is important to note that the number of communities identified for 100 catchments will generally be smaller due to the smaller number of total catchments when compared to the communities identified with 218 catchments, regardless of whether the EB method is used or the MDEB method is used. What is also important to note from the above results is that the range of communities identified for 100 catchments using the MDEB method is closer to that identified for the 218 catchments when compared to that using the EB method. This can be considered to reflect the superiority of the MDEB over the EB method.

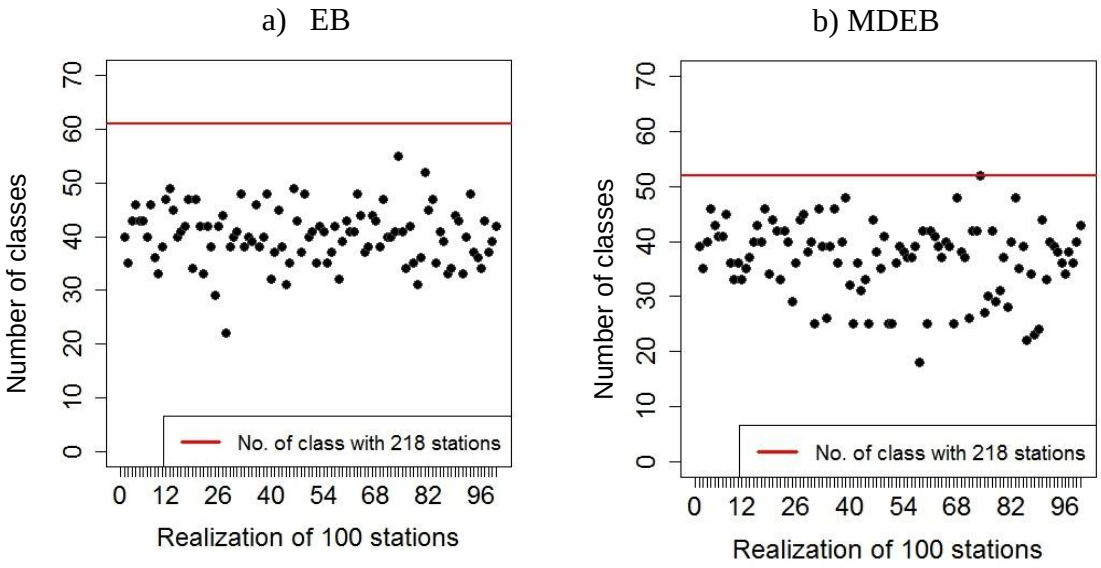


Figure 6.8: Number of communities identified for all 100 realizations of 100 randomly selected stations for Australia using (a) EB method and (b) MDEB method. The red horizontal lines represent the number of communities with the base classification (218 stations).

Figure 6.9 presents the difference in the number of communities identified between the case of 218 catchments and the case of 100 catchments for the EB method (Figure 6.9(a)) and for the MDEB method (Figure 6.9(b)). The plots indicate a distribution similar to that in Figure 6.8. However, the EB method (Figure 6.9(a)) results in a greater difference, even up to almost 40 communities, when compared to the MDEB method (Figure 6.9(b)) that shows a maximum difference of only slightly over 30 communities. The MDEB method also has more than 10 realizations where the difference in the number of communities is below 10 (Figure 6.9(b)). Therefore, as far as the classification of catchments in Australia is concerned, the results support the proposition that the MDEB method is generally superior and that it tends to form smaller-size communities with the smaller size of the network.

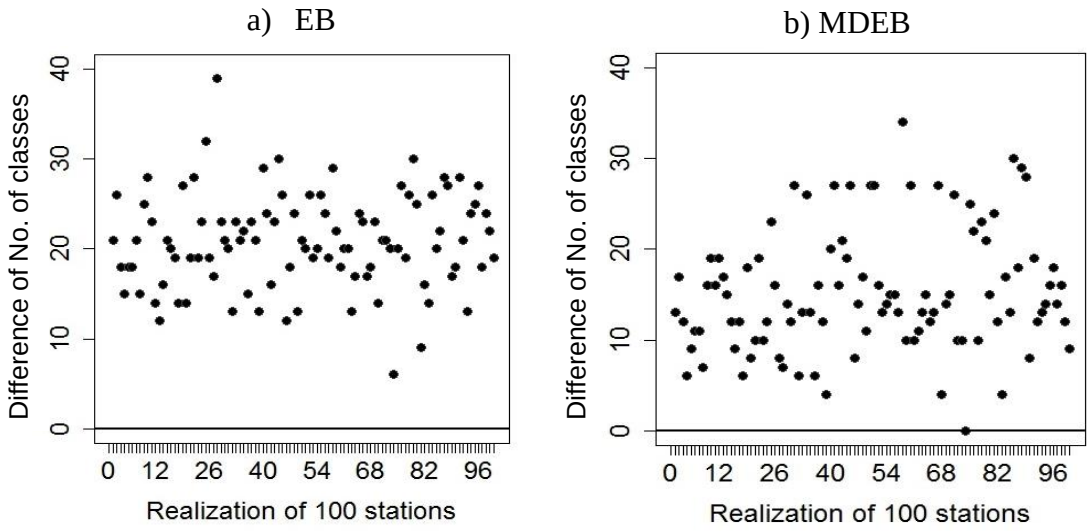


Figure 6.9: Difference in the number of communities identified for all 100 realizations for 100 randomly selected stations for Australia using (a) EB method and (b) MDEB method.

In addition to examining the performance of the methods for community detection, in light of the possible changes in the number of communities/stations at each random realization, it is also essential to determine the number of stations that change at each realization. To this end, the percentage number of stations that change is examined here. The reason for determining the percentage is due to the importance of identifying the rate of change formed by a particular method, specifically in catchment classification. Figure 6.10 presents a bar plot that compares the change in the number of stations at each iteration of random realization by the EB and the MDEB methods, illustrated by blue and red boxes, respectively. It is important to note that change in the number of stations is equal to the change in the percentage of stations, since the number of stations considered for classification is 100. The number of stations that change is obtained by comparing stations at each iteration in the random realizations (either merging to other existing communities or forming other different communities) with the base classification outcomes (i.e. when the number of stations is 218). The number and percentage of stations changed in Figure 6.10 indicates the number of iterations that have a lesser amount of stations changed between the methods. It can be seen that the EB method produces a fewer number of stations changed, as the blue boxes appear mostly lower than the red boxes (which represent the MDEB method). These differences are further observed by having an accurate count of the number of stations changed and the average of the number of stations changed based on the 100 random realizations, as shown in Table 6.2.

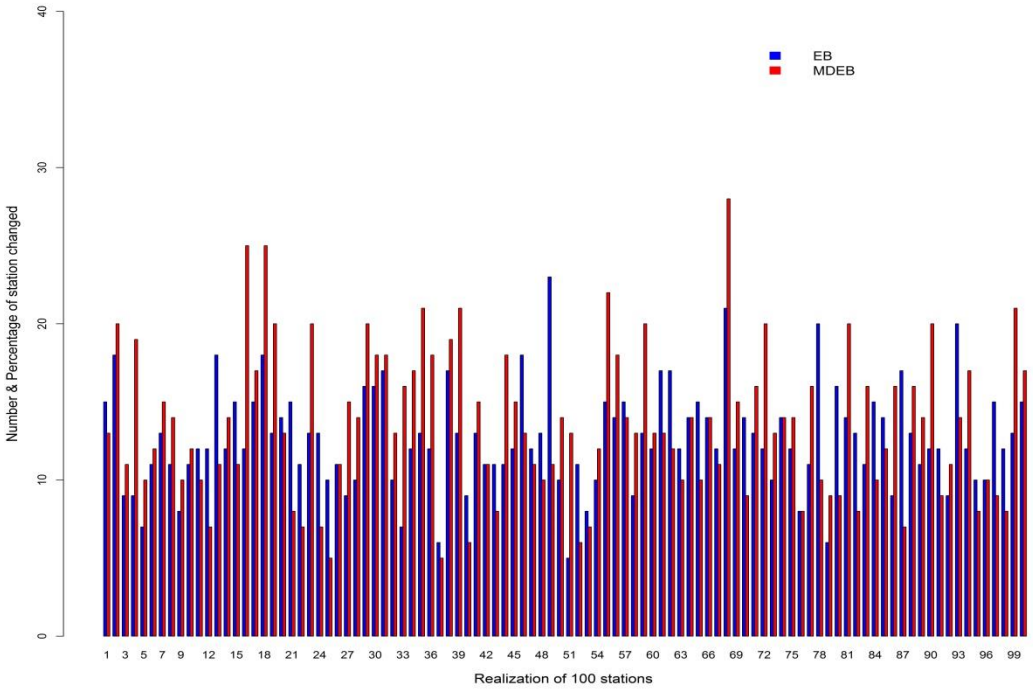


Figure 6.10: Bar plot to compare the number and the percentage of stations changed for 100 random realizations between the EB and MDEB methods.

Table 6.2: Number of random realizations and the average number of stations for the EB and MDEB methods based on the stations changed in classification for Australia.

	Method	
	EB	MDEB
Number of random realizations with lesser number of stations changed (refer Figure 6.10)	**56**	**37**
Average number of stations changed by 100 realizations	**13**	**14**

As presented in Table 6.2, the EB method is found to have a fewer number of iterations with changed stations as counted at 56 times compared to the MDEB method where the number of iterations of changed stations is 37 times. There might be some significant differences regarding which method has fewer number of stations changed at each iteration. However, in terms of the average number of stations changed, the counts are much closer being 13 and 14, respectively, for the EB and the MDEB methods. It might be suggested, from these results, that the EB method tends to form a significantly larger amount of changes (measured by stations that are merged to/formed as another community) at several iterations, e.g., at iteration 49, 78, and 87 (refer to Figure 6.10). However, the MDEB method might form more iterations with the number of stations changed but are not significantly very different (or perhaps similar) when compared to the EB method. Statistically, the percentage of the number of stations changed by the EB method is 13% at each iteration and for the MDEB method is 14%, when the size of the network (or N = number of stations) is considered half of the actual (i.e. base classification or 218 stations) network.

In addition to the change in network size through the random realization analysis for classification, an attempt is also made to examine the influence of network size on classification dividing the base network (218 stations) based on drainage divisions and river region boundaries. The communities identified based on such regional classification are also compared with those for the base classification. Table 6.3 presents the count of the number of stations changed (i.e., how many stations from the identified communities based on the regionalization are different to the communities from the base classification) for both the EB and the MDEB methods.

Table 6.3: Number of stations changed using the EB and MDEB methods according to the drainage divisions in Australia.

	Number of stations changed	
Regions	**EB**	**MDEB**
Carpentaria Coast (CC) -13 stations	1	3
Lake Eyre Basin (LEB)- 5 stations	0	0
Murray Darling Basin (MDB) -75 stations	11	1
North East Coast (NEC)- 42 stations	5	7
South East Coast (NSW) (SEN)- 16 stations	3	4
South East Coast (Victoria) (SEV)- 27 stations	6	8
South West Coast (SWC) – 11 stations	5	1
Tanami Timor Sea Coast (TTS)- 12 stations	3	1
Tasmania (TAS) – 12 stations	2	2
Total number of stations changed	**36**	**27**

According to Table 6.3, the MDEB method results in fewer stations changed, with a count of 27 stations from all the nine different regions considered in Australia, and the EB method results in a count of 36 stations in total. The MDEB method seems to perform better with larger networks compared to the EB method that works well with smaller networks. Having said that, the MDEB method still works better also with very small networks, i.e., from regions South West Coast (11 stations) and Tanami Timor Sea Coast (12 stations). Both methods have similar results with regions from Lake Eyre Basin (5 stations) and Tasmania (12 stations). This seems to suggest that small networks

that are relatively very sparse will have few connections or no connection at all and tend to have no changes either, as in the case of those that are very compact (i.e., short distances and strong correlations among the catchments).

6.3 The United States Streamflow

6.3.1 Entire Network (639 stations)

The MDEB method is applied to monthly streamflow data from 639 stations across the United States for classification. Figure 6.11 presents the communities identified for the 639 stations in the US using the MDEB method. Different colours are used to distinguish the different communities. Table 6.4 presents the number of communities and the number of stations, arranged according to the number of stations in each community, for the MDEB method. The results indicate that: (1) the MDEB method forms 76 communities for the 639 catchments; (2) a significantly large number of communities have only a very few catchments within them. For instance, communities with only one catchment and two catchments from the MDEB method are 34 and 9, respectively, forming over 50% of the total number of communities identified (76); (3) a very small number of communities have a large number of catchments within them. For instance, there are 11 communities that have more than 20 catchments in each, making up about 65% of the total number of catchments (416 out of 639).

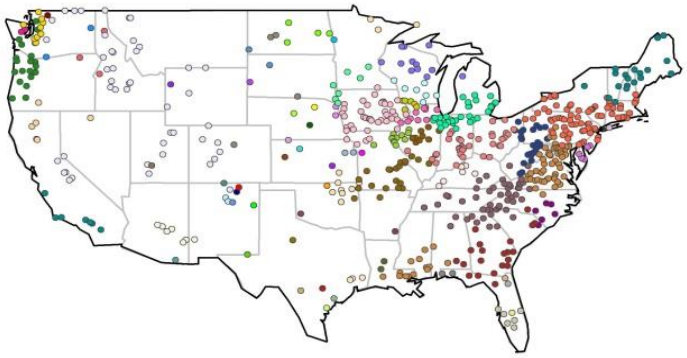


Figure 6.11: Communities identified using the MDEB method for 639 stations in the United States.

Table 6.4: Sizes of the identified communities using the MDEB method for 639 catchments in the US.

MDEB		
Number of stations in community	Number of Communities	Number of stations
1	34	34
2	9	18
3	3	9
4	1	4
5	1	5
6	4	24
7	5	35
8	1	8

9	2	18
11	1	11
12	2	24
14	1	14
19	1	19
21	2	42
22	1	22
25	1	25
31	1	31
36	1	36
40	1	40
46	1	46
50	1	50
51	1	51
73	1	73
Total	**76**	**639**

6.3.2 Network of 300 Stations through Random Realizations

To account for the influence of network size on catchment classification, a network size of 300 randomly selected catchments (i.e., almost half out of 639 stations) with 100 sets of random realizations are analyzed. As an example of classification of the 300 randomly selected catchments using the MDEB method, Figures 6.12 presents 10 out of

100 sets of random realizations. In these plots, the different colours represent different communities but hold no meaning when comparing across other plots.

For better visualization of the communities and classification outcomes, Figure 6.13 shows the communities identified using the MDEB method for two scenarios: (1) a network size of 639 stations where communities with at least 20 stations are indicated with colours (Figure 6.13(a)); and (2) a network size of 300 stations, i.e., one of the random realizations, where communities with at least 10 stations are indicated with colours (Figure 6.13(b)). Communities that are coloured in brown, green, and dark blue in Figure 6.13(a) and coloured in green and dark blue in Figure 6.13(b) are not examined for comparison, due to lack of coverage of stations.

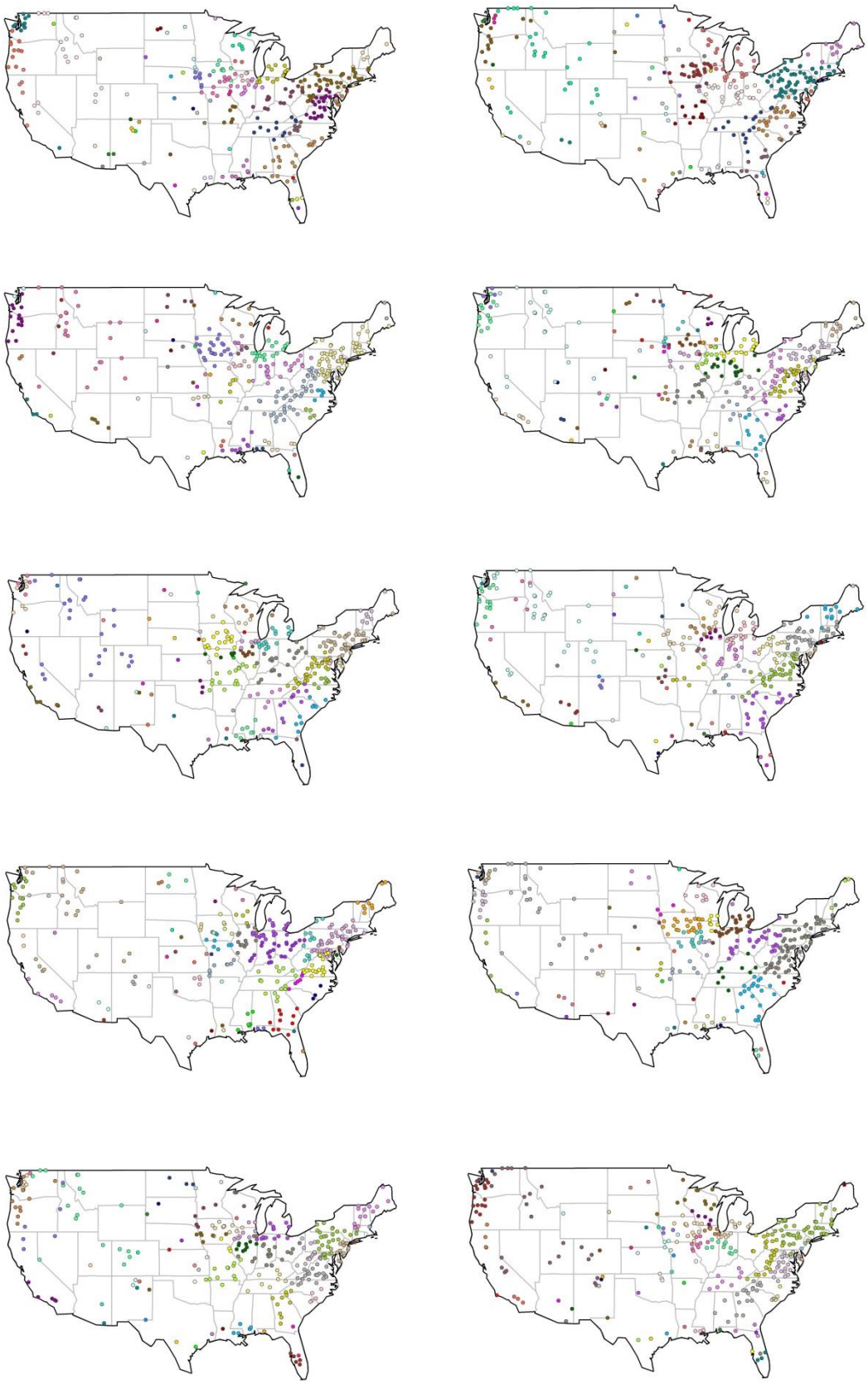


Figure 6.12: Communities identified with 300 randomly selected stations in the US using the MDEB method. Ten out of 100 random realizations are presented, as examples. Each colour represents a community, but the communities are not the same across all the 10 realizations.

a) N = 639 stations b) N = 300 stations

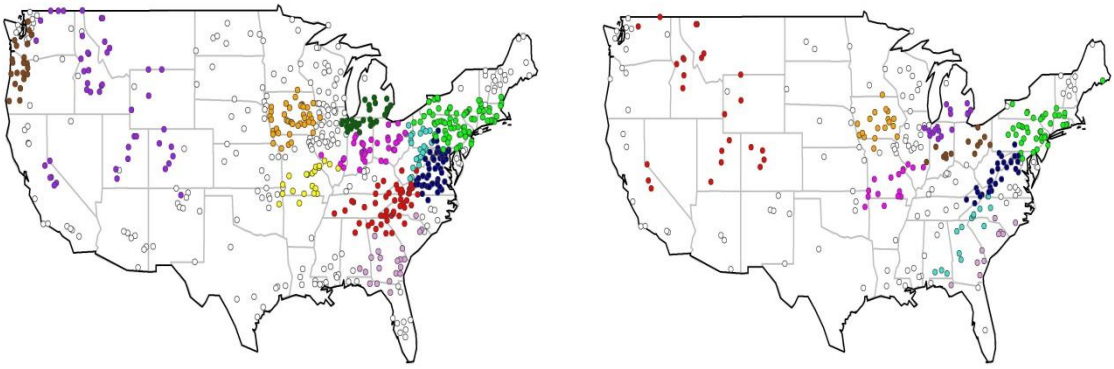
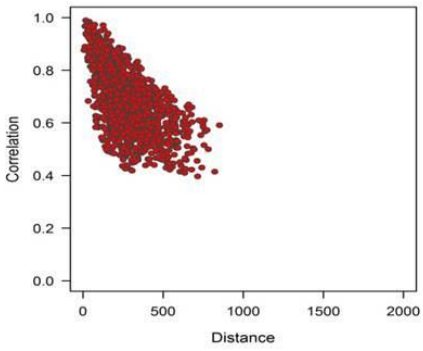
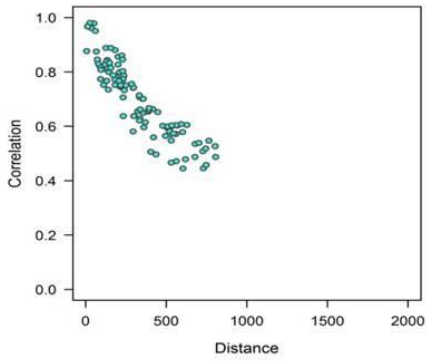
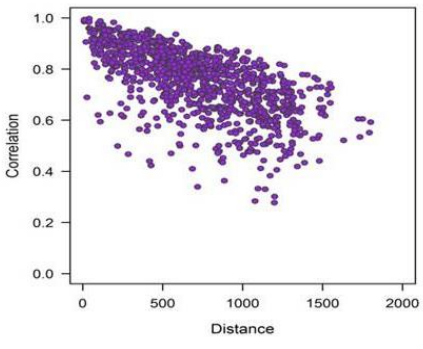
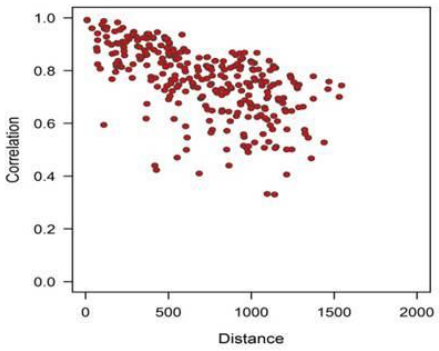
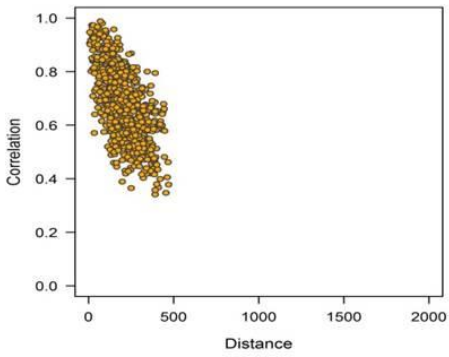
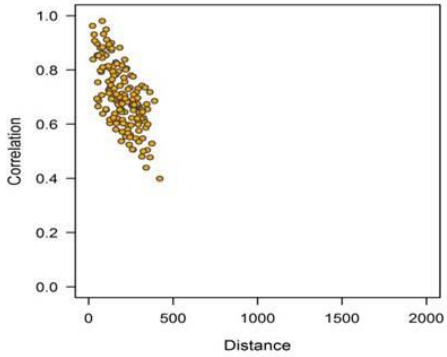
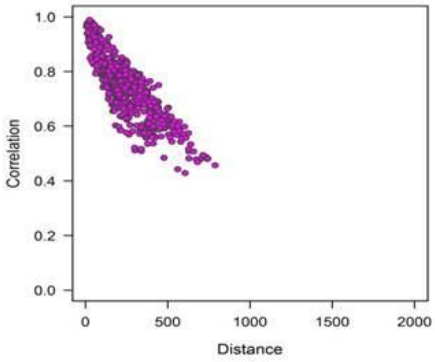
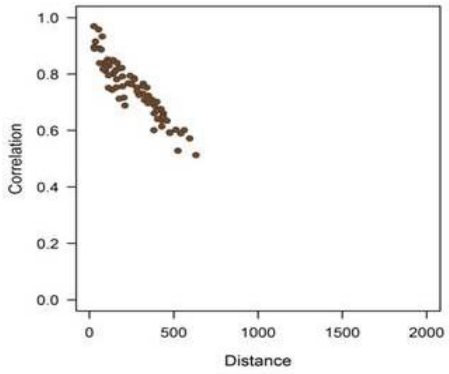
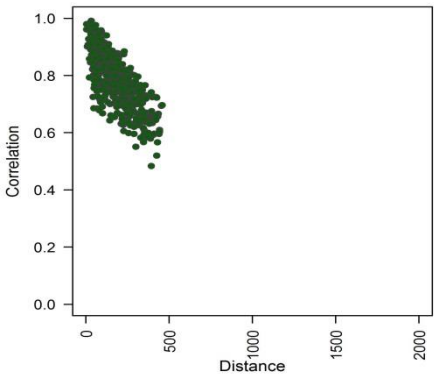
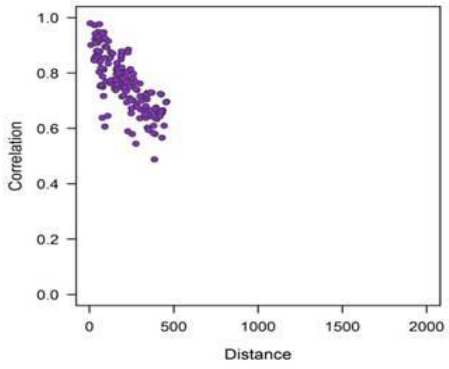


Figure 6.13: Communities identified using the MDEB for two different sizes of networks in the US: (a) 639 stations and (b) 300 stations. Each colour represents a community with at least 20 stations (a) and 10 stations (b), while the open circles represent all communities with less than these.

Figure 6.14 shows the distance-correlation relationships of the selected communities for the base classification with 639 stations (Figure 6.14 of column (a)) and the reduced network of 300 randomly selected stations (Figure 6.14 of column (b)) using the MDEB method. For the communities shown in Figure 6.14 at row 1, comparison is difficult because only very limited number of stations is randomly selected to compare with the base classification. The situation is also similar for Figure

6.14 at row 6, with the distribution is apparently similar. Overall, there are some other cases that are worth mentioning, such as communities that are mostly located in the northeast to mid-west area. For instance, Figure 6.14 at row 3, 4, 5, and 7 shows that the MDEB method manages to identify a similar set of divisions, regardless of the size of the network considered, based on their behaviour in the distance-correlation relationships. In the western region, a similar behaviour is observed, as that illustrated in Figure 6.14 at row 2. This shows that the MDEB method performs well in community detection, regardless of the difference in the network size.

	(a) 639 stations	(b) 300 stations
1.		
2.		

3.	 Correlation Distance 0.0 0.2 0.4 0.6 0.8 1.0 0 500 1000 1500 2000	 Correlation Distance 0.0 0.2 0.4 0.6 0.8 1.0 0 500 1000 1500 2000
4.	 Correlation Distance 0.0 0.2 0.4 0.6 0.8 1.0 0 500 1000 1500 2000	 Correlation Distance 0.0 0.2 0.4 0.6 0.8 1.0 0 500 1000 1500 2000
5.	 Correlation Distance 0.0 0.2 0.4 0.6 0.8 1.0 0 500 1000 1500 2000	 Correlation Distance 0.0 0.2 0.4 0.6 0.8 1.0 0 500 1000 1500 2000

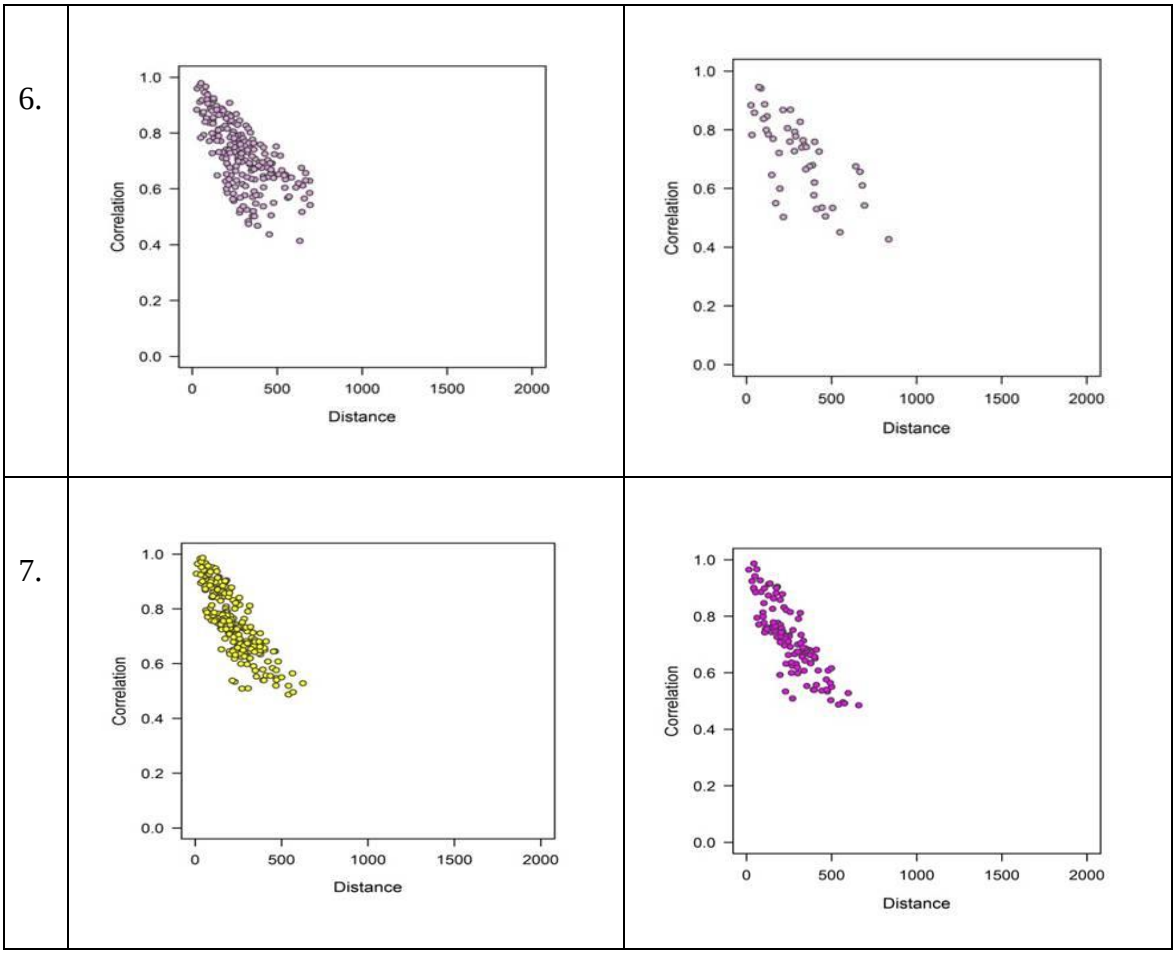


Figure 6.14: Distance-correlation relationship for communities from seven selected regions in the US by the MDEB method (see Table 6.4, ((a)(1-7)) base classification and ((b)(1-7)) 300 randomly selected stations.

6.3.3 Networks of 18 Hydrologic Unit Code (HUC) Regions

In addition to the above analysis of the influence of network size based on reduced number of stations with random selection, an attempt is also made to study the influence of network size by considering the regional impact. To this end, stations within each Hydrologic Unit Code (HUC) are considered independently. The 639 stations considered in the present study in the contiguous US come under HUC 1–18. These 1–

18 HUC regions are shown in Figure 6.15 (the figure also shows the 19–21 HUC regions – Hawaii, Alaska, and Puerto Rico). Pink boundaries and a two-digit number represent each HUC region. The Missouri River Basin (HUC region 10) is very large and is separated into the Upper Missouri region (HUC region 10A) and the Lower Missouri region (HUC region 10B). However, in the present analysis, all catchments belonging to HUC 10 are examined as one whole region. More information on the spatial coverage (i.e., the percentage of each HUC region that is gauged) and the related analysis based on HUC regions can be found in Kiang et al. (2013).

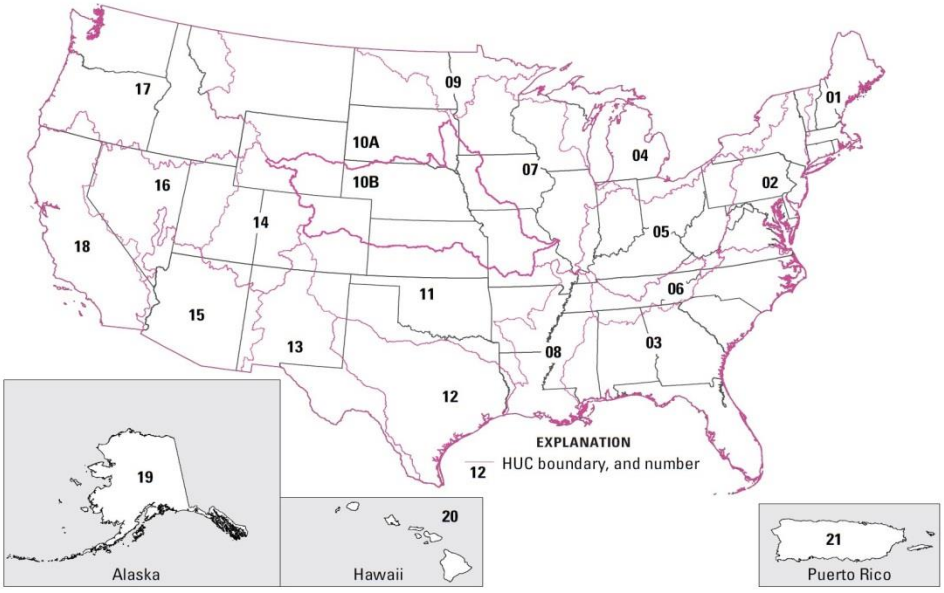


Figure 6.15: Regions according to hydrologic unit code (HUC) in the United States (**source of the map**: (Kiang et al., 2013)).

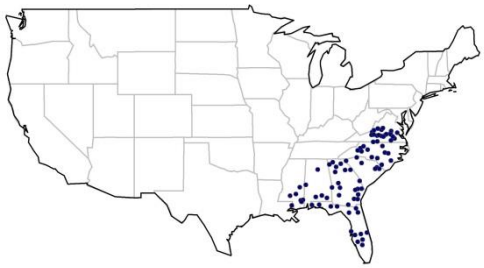
For a clear visualization of the streamflow stations within each HUC region, Figure 6.16 shows the stations within each of the 18 HUC regions independently on the

map of the contiguous US. For an easier understanding, for each HUC region, the stations are also indicated in different colours. These are important and useful to indicate the accurate locations of catchments in each HUC region. For instance, there are some regions that are compact in the distribution of catchments with a large number of catchments at close distances as a network, such as those in the Mid-Atlantic, South Atlantic-Gulf, Ohio, and the Upper Mississippi regions (Figure 6.16(b), (c), (e), and (g)). Meanwhile, stations in some other regions span large distances between each other, such as Souris-Red-Rainy, Missouri, Arkansas-White-Red, Texas-Gulf, Rio Grande, Upper Colorado, Great Basin, Pacific Northwest, and California, as shown in Figure 6.16(i), (j), (k), (l), (m), (n), (p), (q), and (r). The locations of stations are also important in offering explanations related to the variability of the distance-correlation relationships, as shown in Figure 6.17 (more details in below).

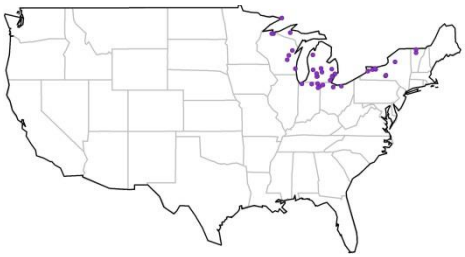
a) New England (HUC 01)

b) Mid-Atlantic (HUC 02)

c) South Atlantic-Gulf (HUC 03)

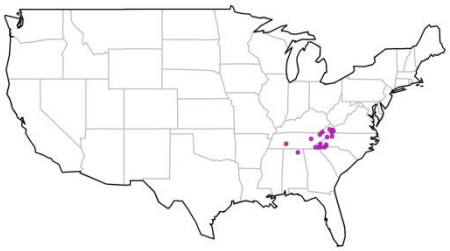


d) Great Lakes (HUC 04)

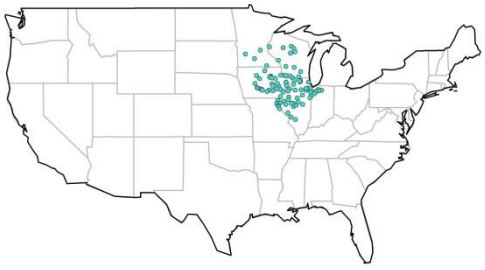


e) Ohio (HUC 05)

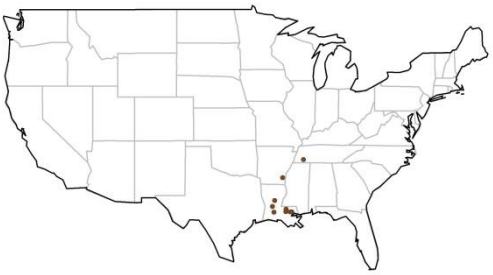
f) Tennessee (HUC 06)



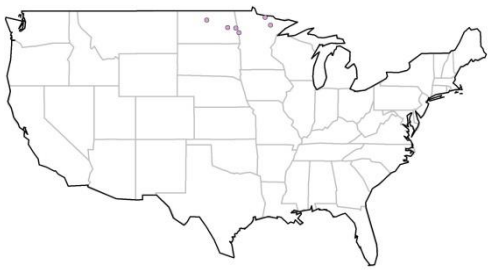
g) Upper Mississippi (HUC 07)



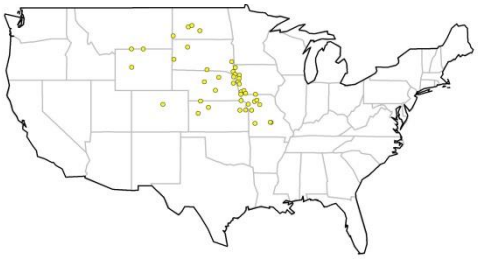
h) Lower Mississippi (HUC 08)



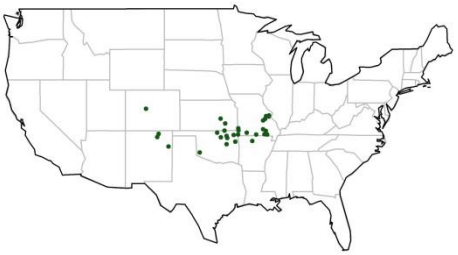
i) Souris-Red-Rainy (HUC 09)



j) Missouri (HUC 10)

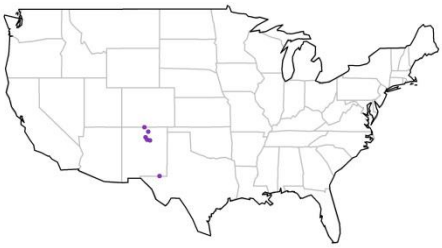


k) Arkansas-White-Red (HUC 11)

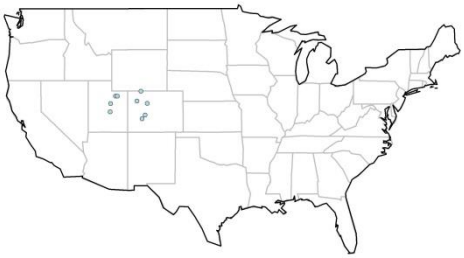


l) Texas-Gulf (HUC 12)

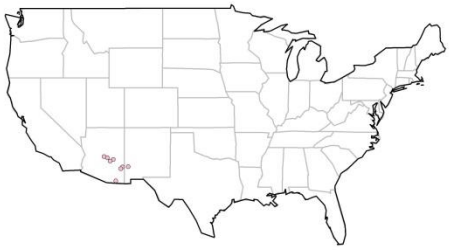
m) Rio Grande (HUC 13)



n) Upper Colorado (HUC 14)

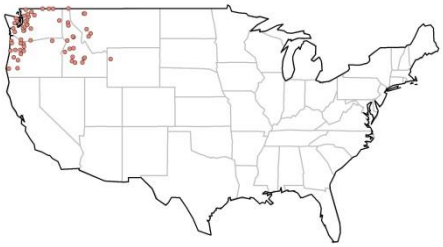


o) Lower Colorado (HUC 15)



p) Great Basin (HUC 16)

q) Pacific Northwest (HUC 17)



r) California (HUC 18)

Figure 6.16: Streamflow station locations in 18 HUC regions in the United States considered in this study.

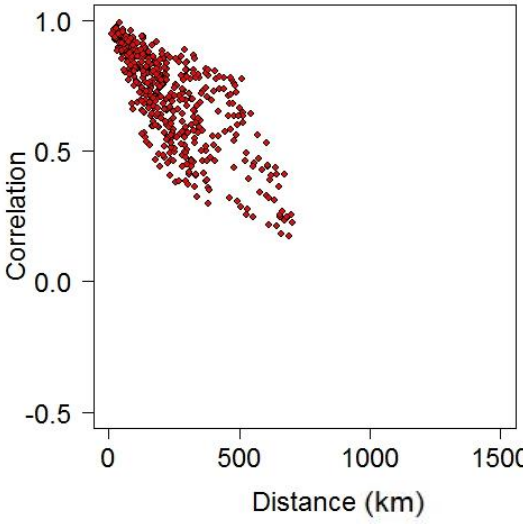
Figure 6.17 (a-r) presents the distance-correlation relationships for stations in the 18 HUC regions. It appears that most of the regions have different patterns of relationship. For instance:

(1) Catchments in regions HUC 01, 03, 04, 10, 11, 12, 17, and 18 exhibit a decrease in correlation with an increase in distance and are also more sparsely distributed than that in the rest of the regions (Figure 6.17(a), (c), (d), (j), (k), (l), (q) and (r)). However, such a characteristic may not be valid with regions that have a smaller number of catchments, such as HUC 08, 09, 13, 14, 15, and 16 (Figure 6.17(h), (i), (m), (n), (o), and (p)). The region HUC 10 is slightly different (Figure 6.17(j)) and is worth mentioning, where the correlation decreases when the distance increases until at about 700 km, and the correlation decreases again when the distance exceeds 700 km. This variability of the connection between these catchments over large distances might be because they are from the same stream/river network that is stretched over long distances, since the region represents the Missouri basin (Figure 6.16(j)); and

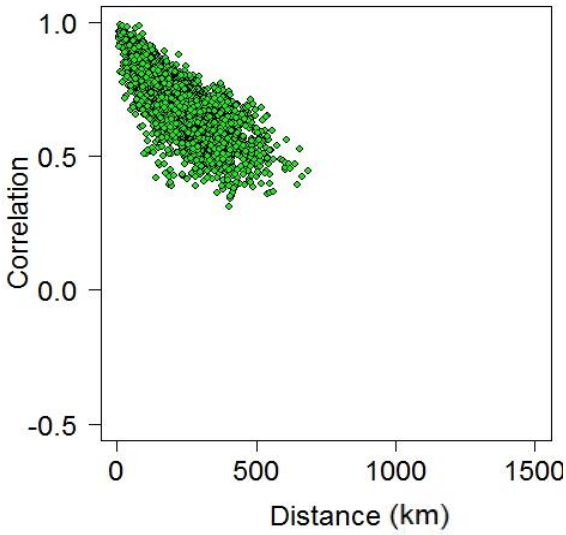
(2) Catchments in regions HUC 02, 05, 06, and 07 are much more compact in terms of correlations, meaning that the catchments are strongly connected even when the catchments span large distances (Figure 6.17(b), (e), (f), and (g)).

The variability of the distance-correlation relationships summarised by the HUC regions is helpful to obtain a better understanding of how the community detection method works and in identifying communities, regardless of geographic proximity or distance factors.

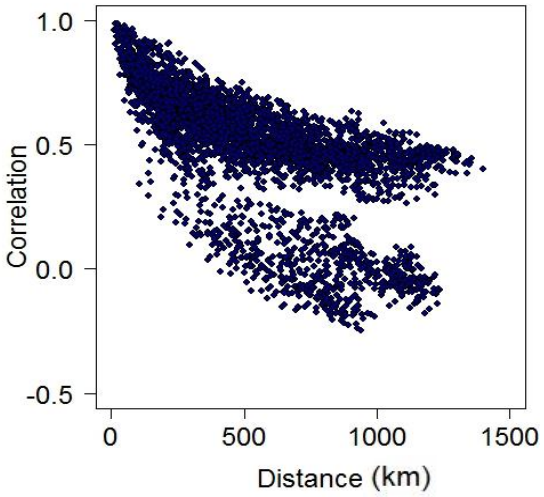
(a) HUC 01



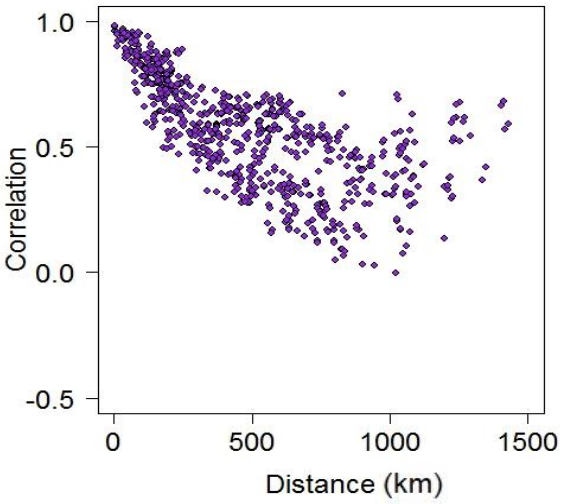
(b) HUC 02



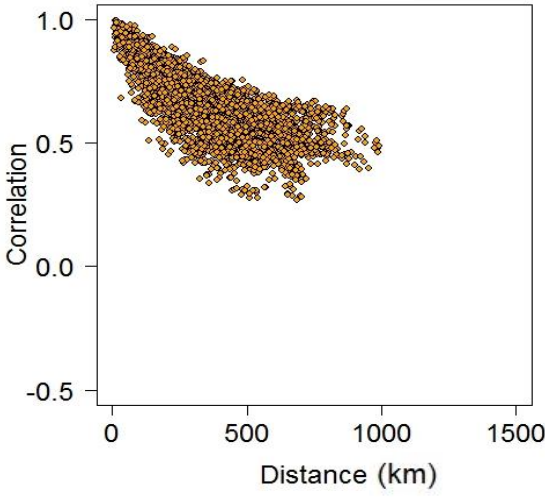
(c) HUC 03



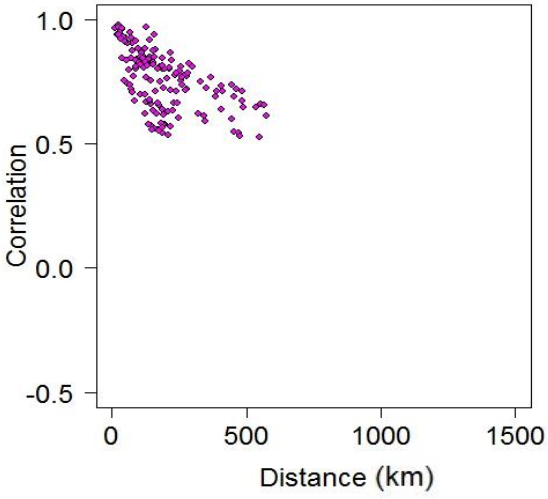
(d) HUC 04



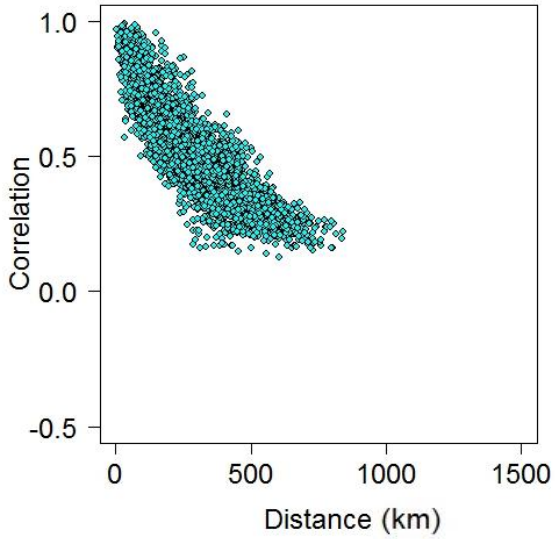
(e) HUC 05



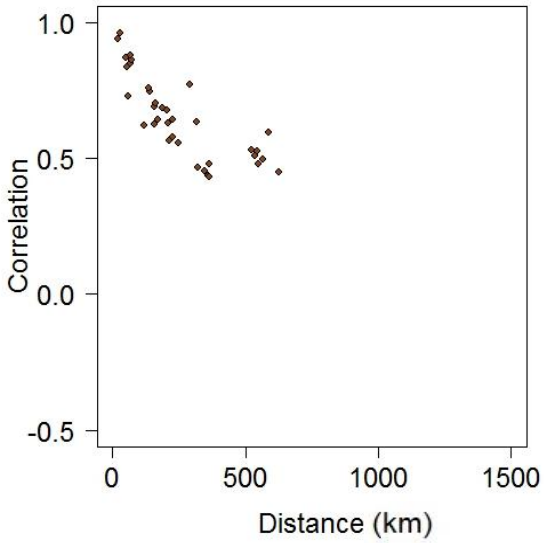
(f) HUC 06



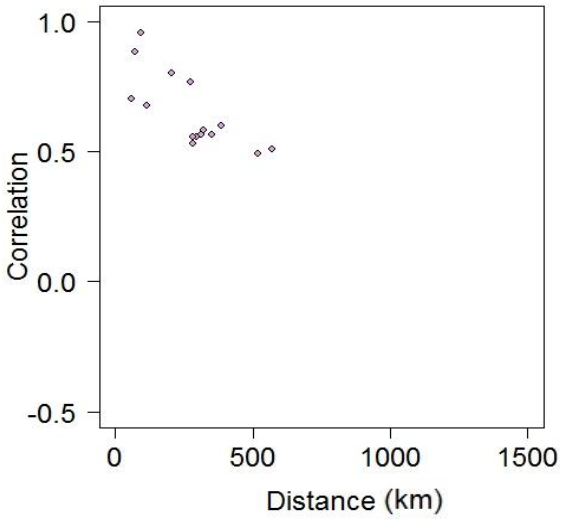
g) HUC 07



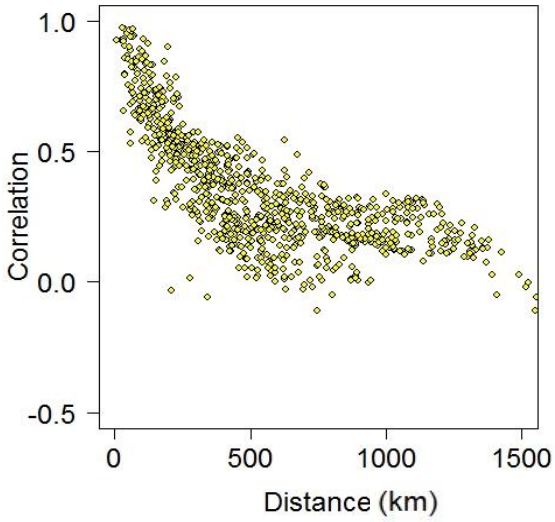
h) HUC 08



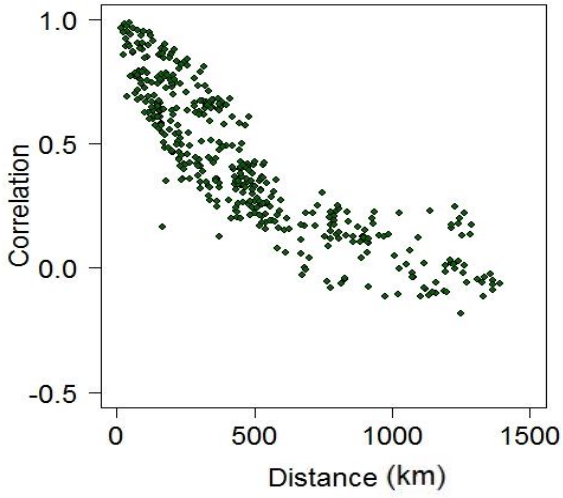
i) HUC 09



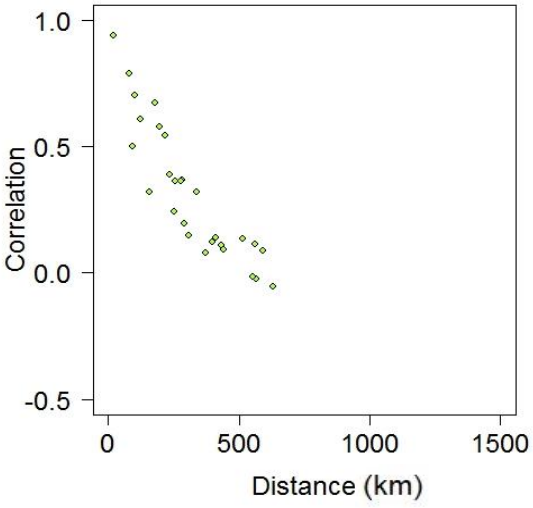
j) HUC 10



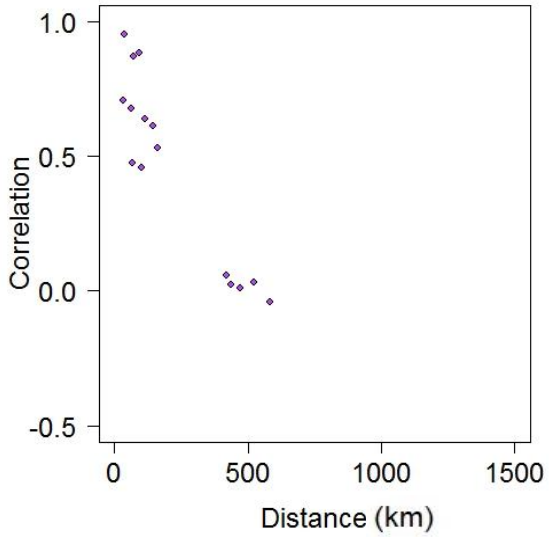
k) HUC 11



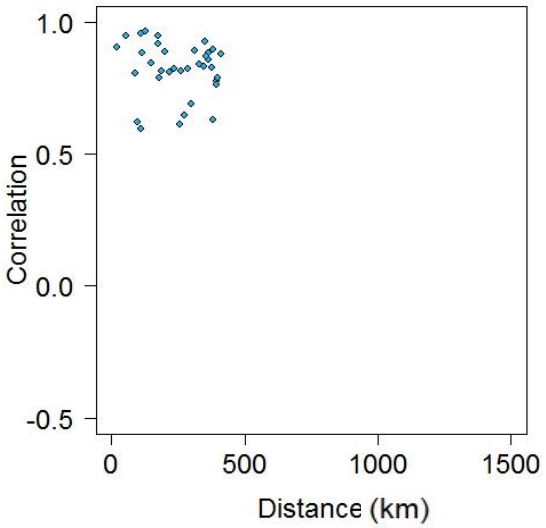
l) HUC 12



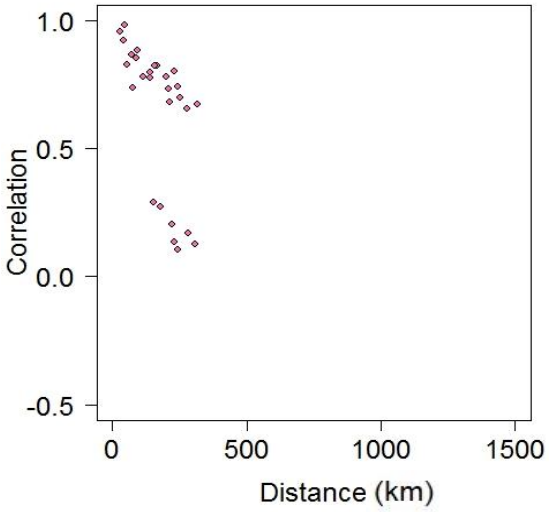
m) HUC 13



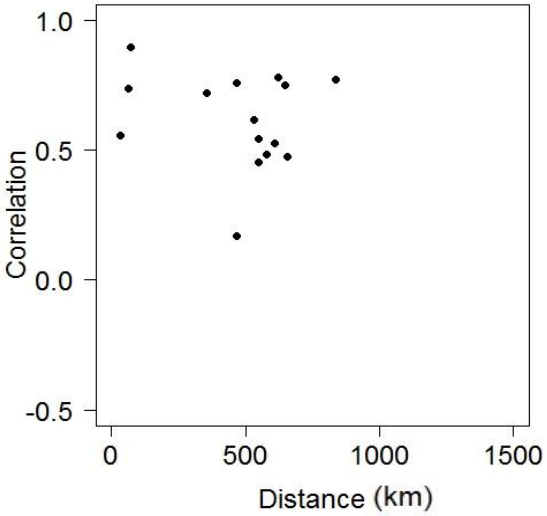
n) HUC 14



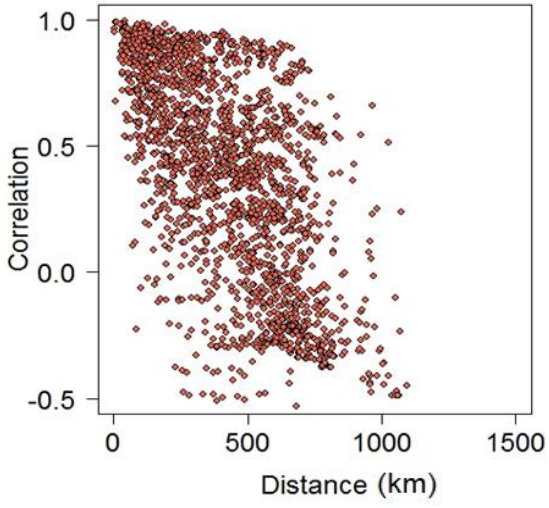
o) HUC 15



p) HUC 16



q) HUC 17



r) HUC 18

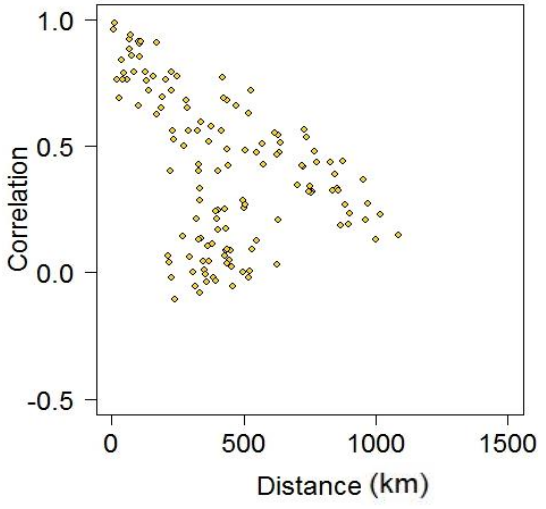


Figure 6.17: Distance-correlation relationship for stations within the 18 HUC regions in the US. The colors correspond to those in Figure 6.16.

6.3.4 Comparison between EB and MDEB methods

Considering network size through random realization, Figure 6.18 presents the number of communities identified for all iterations using (a) the EB method and (b) the MDEB method. The figure also presents the horizontal red solid lines that represent the number of communities identified with the entire network of 639 stations (i.e., base classification). As discussed earlier, the number of communities identified with the 639 stations is 61 using the EB method and 76 using the MDEB method. Intuitively, it could be expected that the range of variability in the number of communities identified with a smaller size of network (i.e., 300 stations) would be correspondingly smaller than that of the base network (639 stations). As seen in Figure 6.18(a), the range of differences in the number of communities identified using the EB method is larger (i.e., most of the iterations have nearly similar number of communities and one of the iterations has a greater number) than using the MDEB method. The MDEB method shows more differences, as illustrated by the gap between the black circles and the red line, as seen in Figure 6.18(b).

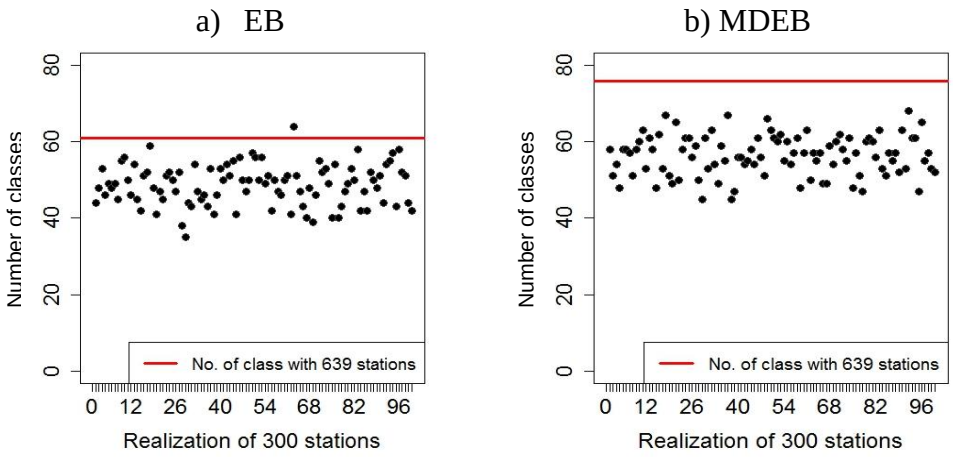


Figure 6.18: Number of communities identified for all 100 random realizations using (a) the EB method and (b) the MDEB method for the US. The red horizontal lines represent the number of communities with the base classification.

Figure 6.19 shows the difference in the communities identified between the random realizations and the base classification for (a) the EB and (b) the MDEB methods. It can be clearly seen, in Figure 6.19(a), that the number of communities from the EB method is similar in variability with the base classification, i.e., mostly range from 0 to 20 communities and the MDEB method results mostly range from 10 to 30 communities (Figure 6.19(b)). However, as shown previously (Figure 6.18), the EB method forms large number of communities in which almost similar to the identified number of communities with the base classification (639 stations) and in fact, there is seen one realization has larger than that. The performance of the MDEB method is investigated further in more detail in terms of the number of stations that change (merge or form as other communities) and the percentage of the number of changed stations, in order to compare the tendency of changes by each method for catchment classification.

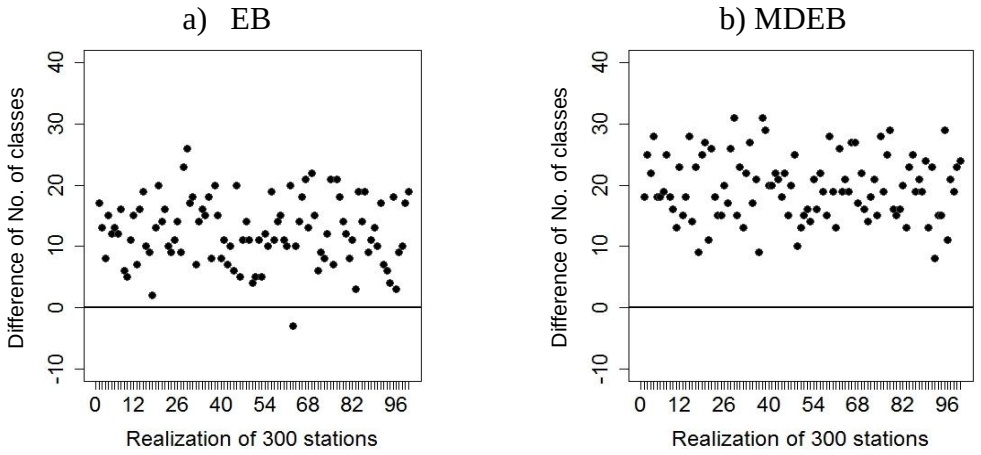


Figure 6.19: Difference in the number of communities identified for all 100 random realizations for the catchments in the US: (a) EB method; and (b) MDEB method.

Figure 6.20 presents a bar plot to compare the number of stations that are found to change at each iteration of random realizations using the EB method (represented in blue boxes) and the MDEB method (represented in red boxes), with coloured horizontal lines to represent the average of the number of stations changed – corresponding to the legend. In addition, the percentage of stations that are found to change is also considered, as shown in Figure 6.21, to identify the proportion of the count of the number of stations changed at each random realization by both methods. Despite the apparent randomness in the number and percentage of stations changed (Figures 6.20 and 6.21), the purpose is also to count the number of iterations that have the fewer number of stations changed between the methods. It can be seen that the MDEB method produces fewer stations changed, since the red boxes are mostly lower than the blue ones (which represent the EB method). An accurate count of the number of stations changed and the average of the number of stations changed based on the 100 random realizations is also obtained, as shown in Table 6.5.

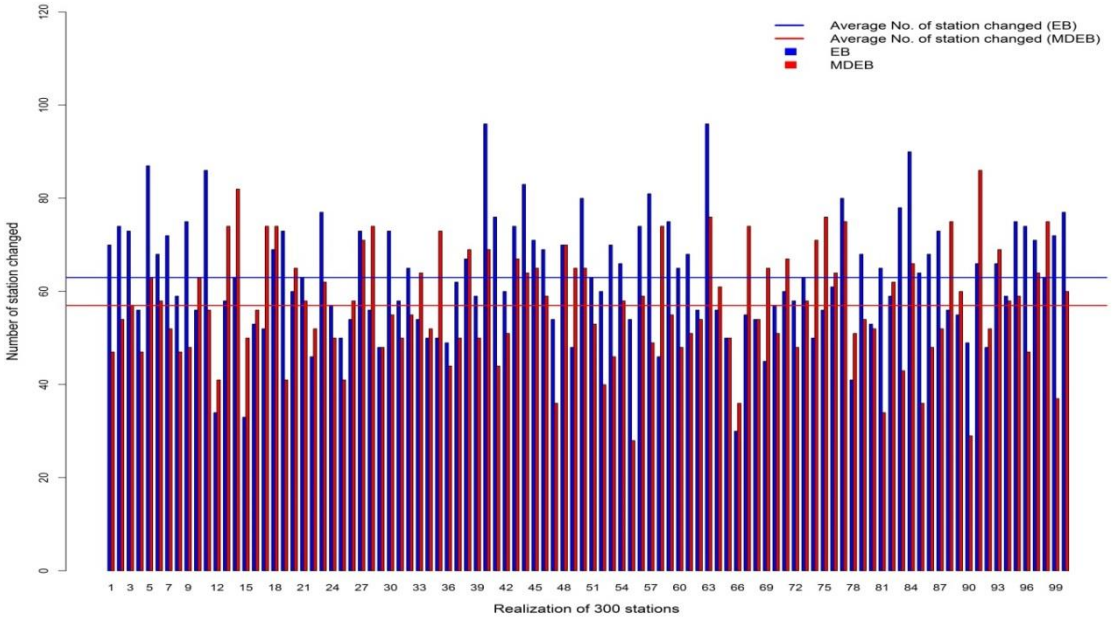


Figure 6.20: Bar plot to compare the number of stations changed for 100 random realizations between the EB and MDEB methods for the US. Horizontal lines represent the average number of stations changed using 100 random realizations.

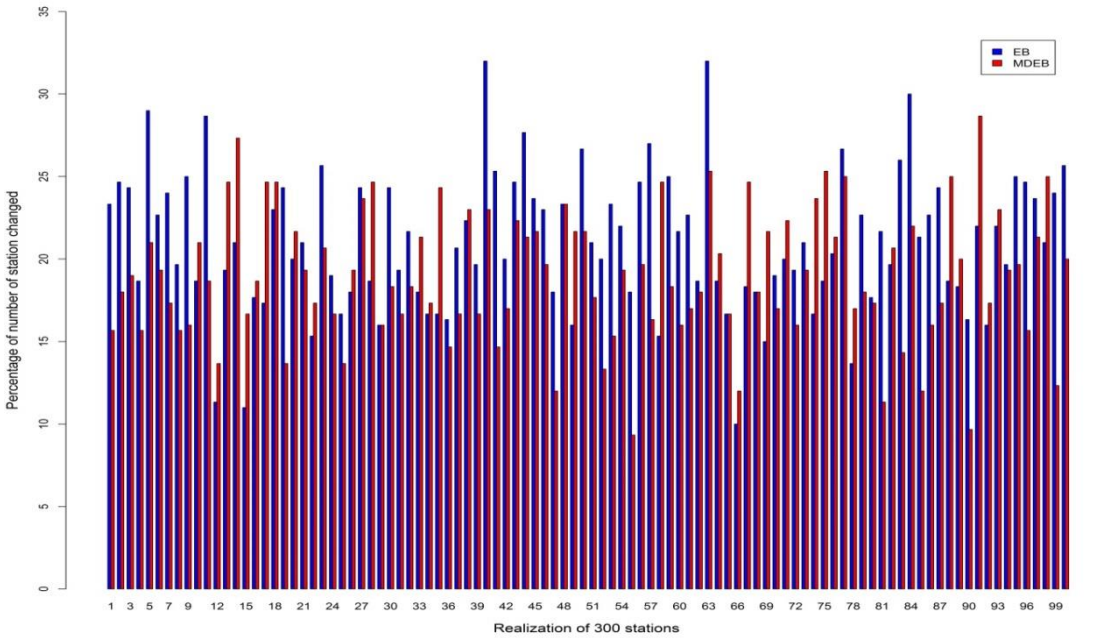


Figure 6.21: Bar plot to compare the percentage of stations changed for 100 random realizations between the EB and MDEB methods for the US.

In regards to the above, it will also be helpful to have a specific count of the number of realizations with the fewer number of changed stations by each method as indicated in Table 6.5. For instance, out of 100 realizations, the MDEB method is found to have fewer changed stations at 66 times when compared to the EB method, which only manages to have fewer changed stations based on the iterations with count of 34. Furthermore, the average number of stations changed at each iteration is 63 by the EB method and 57 by the MDEB method, as represented by the blue horizontal lines (for the EB method) and the red horizontal lines (for the MDEB method) (Figure 6.20).

Table 6.5: Number of random realizations and the average number of stations for EB and MDEB methods based on the stations changed in classification for the US.

	Method	
	EB	MDEB
Number of realizations with a fewer number of stations changed (refer Figure 6.20)	**34**	**66**
Average number of stations changed with 100 realizations (refer Figure 6.20)	**63**	**57**

Finally, by forming small-size networks based on HUC boundaries, the regional classification is carried out by applying the EB and the MDEB methods and the communities identified are compared with the base classification. Table 6.6 presents an accurate count of the number of stations changed (i.e., how many of the stations merged to or formed as another community) for both methods.

Table 6.6: Number of stations changed using EB and MDEB methods according to HUC regions in the US.

	Number of stations changed	
Hydrologic Unit Code (HUC)	EB method	MDEB method
01 - New England (31 stations)	13	2
02 - Mid Atlantic (84 stations)	24	37
03 - South Atlantic-Gulf (85 stations)	21	14
04 - Great Lakes (39 stations)	8	4
05 – Ohio (85 stations)	17	17
06 – Tennessee (18 stations)	8	1
07 - Upper Mississippi (92 stations)	26	14
08 - Lower Mississippi (9 stations)	1	1
09 - Souris-Red-Rainy (6 stations)	0	0
10 – Missouri (43 stations)	11	9
11 - Arkansas-White-Red (30 stations)	5	4
12 - Texas-Gulf (8 stations)	0	2
13 - Rio Grande (6 stations)	0	1

14 - Upper Colorado (9 stations)	0	1
15 - Lower Colorado (8 stations)	2	0
16 - Great Basin (6 stations)	0	1
17 - Pacific Northwest (63 stations)	13	3
18 – California (17 stations)	7	3
Total number of stations changed	156	114

According to Table 6.6, the MDEB method results in fewer number of stations changed with a count of 114 stations from all 18 different HUC regions across the US, while the EB method has a count of 156 stations. The MDEB method appears to perform well with either large or small networks when compared to the EB method. Both methods also give similar results with regions from Ohio (with 85 stations), Lower Mississippi (9 stations), and Souris-Red-Rainy (6 stations).

6.4 Summary

This chapter has presented the application of an improved edge betweenness (EB) method, called the Modularity-Density based EB (MDEB) method, for classification of catchments in Australia and in the United States. Three network scenarios have been studied: (1) base classification (i.e., 218 stations in Australia and 639 stations in the US); (2) through 100 random realizations with a network of 100 randomly selected stations out of 218 stations for Australia and 300 randomly selected stations out of 639 stations from the US, to purely to address the problem of network size; and (3) through regional classification based on nine drainage divisions and river regions for Australia

and 18 Hydrologic Unit Code (HUC) regions for the US, to address the network size and regional similarity and influence. Due to the issue of resolution (or scale) limit problem inherent in the modularity function in the EB method (as explained in Section 3.4.2), the classification outcomes from the MDEB method based on the three scenarios considered in this study (i.e., entire network size, smaller network size based on random realizations, smaller network size based on drainage divisions or hydrologic regions) were compared with the classification based on the EB method. This comparison is crucial to evaluate the efficacy of the two methods, and possible identification of the superior one, towards reliable and accurate classification. Hydrologic monitoring networks, to represent catchments, often change in size for various purposes. For instance, due to difficulty in maintenance, one or more streamflow stations may be removed from an existing network; similarly, to measure data at ungauged locations, one or more new streamflow monitoring stations may be installed. The three scenarios were essentially considered to illustrate these changes and their effects on catchment classification. From the results, the MDEB method was found to generally perform better than the EB method in catchment classification, as evaluated in terms of the number of communities identified and the number of stations that changed communities (based on two scenarios of random realizations and drainage divisions or hydrologic regions). The superior performance of the MDEB method was essentially due to its ability to take into account the resolution or scale issue in classification. With these results suggesting the superiority of the MDEB method, an attempt is also made to use the MDEB method in the context of a multi-variable approach, involving rainfall and PET, in addition to streamflow, for the 218 catchments in Australia. Details of such an analysis and results are presented in Chapter 7.

Chapter 7

Catchment Classification based on Multiple Variables

7.1 Introduction

The fundamental idea of catchment (and other hydrologic) studies, including catchment classification, is to establish connections between the different elements or items (known or assumed) that generally exist within the underlying system. Depending upon the conditions (e.g., catchment, purpose, problem), these elements include catchment characteristics, hydroclimatic variables, model parameters, and others (and their combinations) with respect to space, time, and space–time. With all these influencing the functions of catchments, studying only one variable (or component) that represents or influences catchments, as was the case using only streamflow in Chapters 5 and 6, is often not adequate for catchment classification. It is also important to consider any other variable(s) that influence the catchments, especially those that affect streamflow. The use of multiple variables would also help put more stringent conditions on classification and, thus, the outcomes would be more reliable.

In view of this, an attempt is made here to present a multi-variable approach for classification of catchments. In addition to streamflow, rainfall and potential evapotranspiration (PET) are considered for the multi-variable analysis. With the classification analysis presented with streamflow alone in Chapters 5 and 6, and with the observation that the MDEB performs slightly better when compared to the EB method, the multi-variable approach is presented only using the MDEB method. The classification is performed for the 218 catchments from Australia. In the implementation of the MDEB method in a multi-variable sense, and with three variables (streamflow, rainfall, and PET), four different combinations of multiple variables are considered: (1) streamflow and rainfall; (2) streamflow and PET; (3) rainfall and PET; and (4) streamflow, rainfall, and PET. For each of these combinations, the identification of the connections between the stations and assignment of links is done as follows. First, the correlations between any two stations are obtained from the average of correlations of the respective (two or three) variables between the stations. Next, these correlations are compared against the (assumed) threshold values to check the existence/non-existence of connections. The classification results from the multi-variable analysis are also compared with those obtained from the single-variable analysis, including streamflow, rainfall, and PET independently. However, the case of streamflow is given particular importance, especially considering the extensive amount of results obtained using streamflow (Chapters 5 and 6). The comparisons are done using distance-correlation relationship, and the classification outcomes are also interpreted in terms of accurate station count and connection link count for each case. For a more systematic presentation of the analysis, details of the single variables

(streamflow, rainfall, and PET) are presented first, with particular focus on correlations between the stations considered.

7.2 Single-variable correlation analysis

7.2.1 Streamflow

Figure 7.1 presents the correlations in monthly streamflow among the 218 stations (Figure 7.1(a)) as well as the distance-correlation relationship for the 218 stations (Figure 7.1(b)). As may be seen, the correlations are generally low and range from -0.5 to 0.5 for most stations. However, there also appear to have some higher correlations with values of greater than 0.5 for stations 150 to 210. For the distance-correlation relationship (Figure 7.1(b)), it can be seen that the correlations decrease when the distance increases. Some stations tend to have high correlations at long distances, as shown by the small cluster at distance 2500 to 3000 km with strong correlations between 0.5 to 1. This is not surprising, since strong correlations may also be observed over large distances due to many factors (Fang et al., 2017).

a)

b)

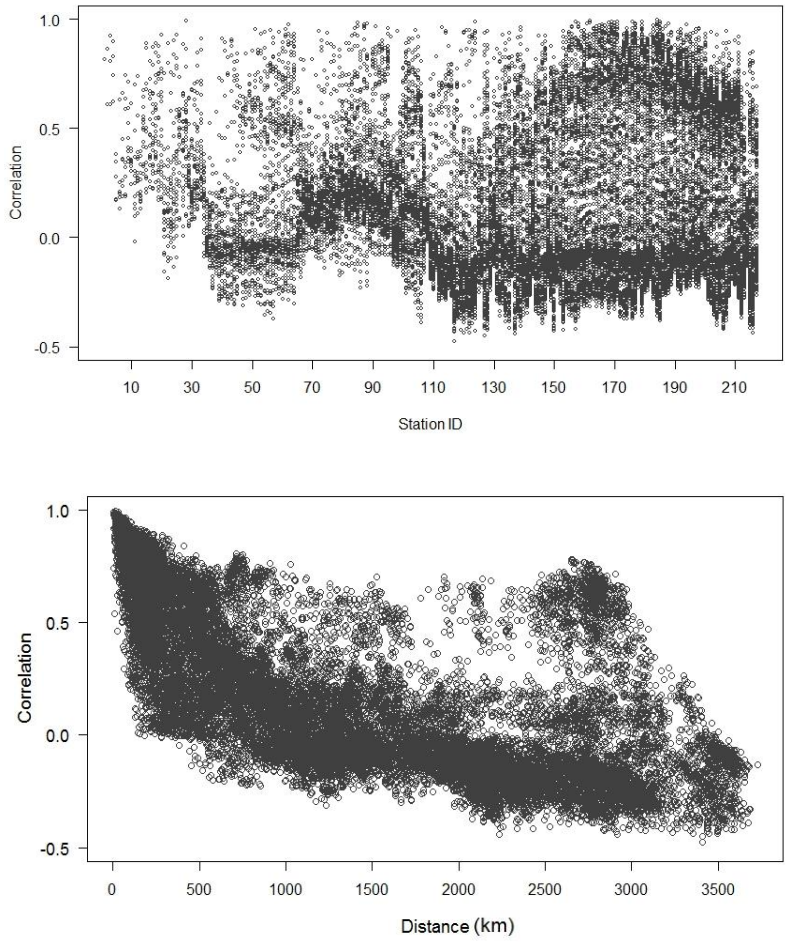


Figure 7.1: Correlation analysis for streamflow from 218 stations in Australia: (a) Correlation of each station to other stations; and (b) distance-correlation relationship.

7.2.2 Rainfall

Figure 7.2 presents the correlations in monthly rainfall among the 218 stations (Figure 7.2(a)) as well as their distance-correlation relationship (Figure 7.2(b)). As may be seen, the distribution of correlations in rainfall between stations is more scattered when compared to that in streamflow (Figure 7.1(a)). The rainfall correlations (Figure 7.2(a)) appear to divide the stations quite significantly, which is essentially because of the

greater variability in rainfall distribution over Australia. The distance-correlation relationship, shown in Figure 7.2(b), indicates that the correlations decrease when the distance increases (similar to that for streamflow data), although with even lower minimum correlation values that are less than -0.5. There also appear to be stations that retain relatively higher correlations even as the distance increases. However, this is sparser when compared to the distribution in the case of streamflow.

a)

b)

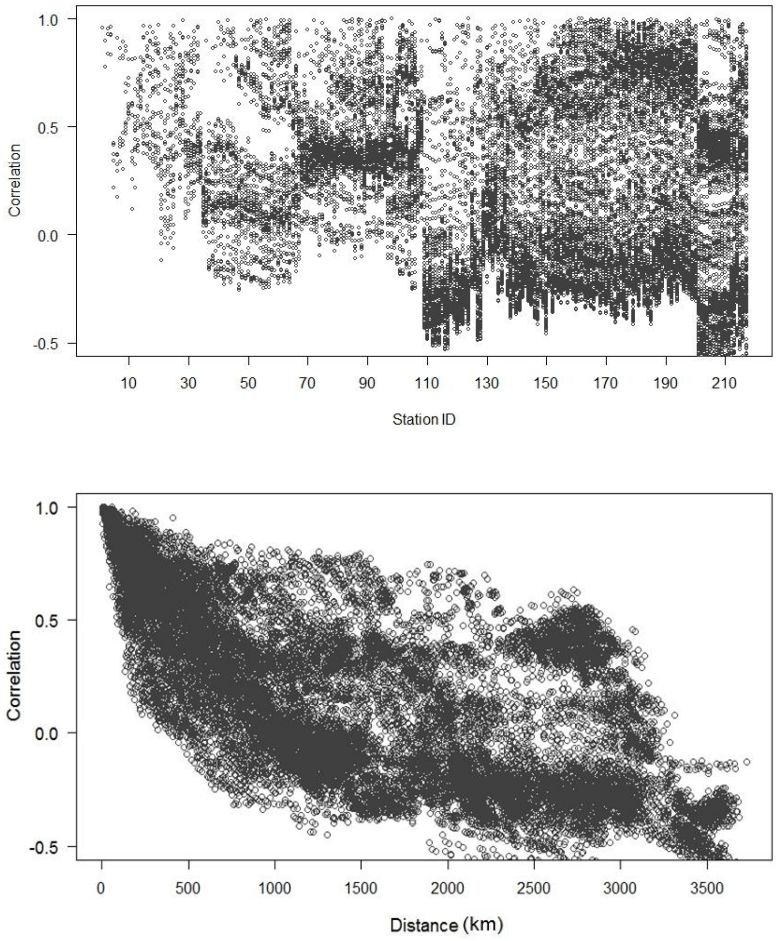


Figure 7.2: Correlation analysis for rainfall from 218 stations in Australia: (a) Correlation of each station to other stations; and (b) distance-correlation relationship.

7.2.3 Potential Evapotranspiration (PET)

Figure 7.3 presents the correlations in monthly PET among the 218 stations (Figure 7.3(a)) as well as their distance-correlation relationship (Figure 7.3(b)). The case of PET is quite interesting, as the correlations mostly have high values, with only a very small number of stations have low correlations. Thus, practically, the use of PET may not really help in classification, since most of the stations will be connected even when very high thresholds are used to identify/assign links and, hence, there will be only a very few classes. Nevertheless, this remains to be seen. The distance-correlation relationship, shown in Figure 7.3(b), indicates that the correlations decrease as the distance increases, similar to that observed for streamflow and rainfall. Nevertheless, the distributions of correlations are more scattered when compared to those obtained for streamflow and rainfall.

a)

b)

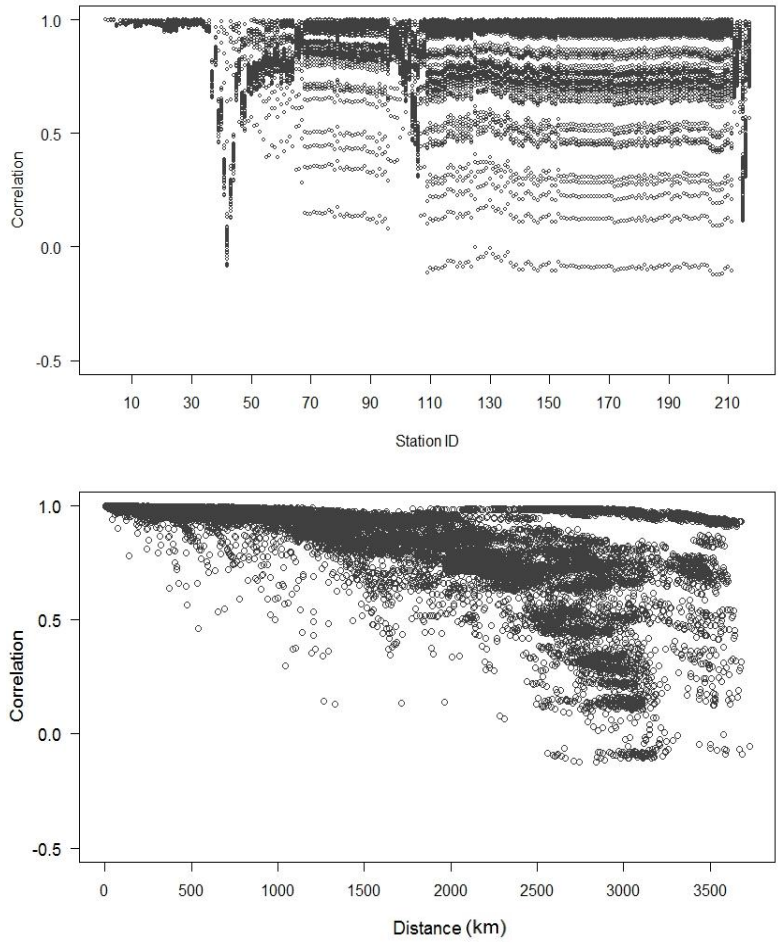


Figure 7.3: Correlation analysis for potential evapotranspiration from 218 stations in Australia: (a) Correlation of each station to other stations; and (b) distance-correlation relationship.

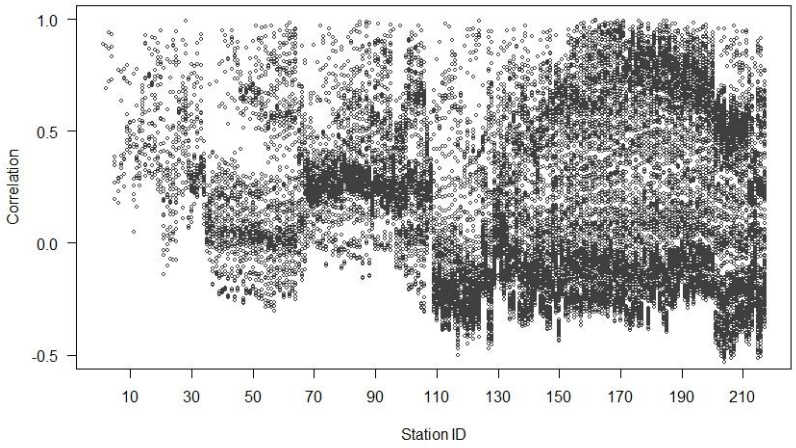
7.3 Multi-variable correlation analysis

7.3.1 Streamflow and Rainfall

Figure 7.4 presents the correlations between the 218 stations (Figure 7.4(a)) and their distance-correlation relationship (Figure 7.4(b)) for the combination of streamflow and rainfall data, i.e., average of correlations for streamflow and rainfall. As can be seen, the

correlations between stations appear to be more like that observed for rainfall (Figure 7.2(a)), especially by looking at the partition of stations based on the combined correlations. However, from the distance-correlation relationship, it can be seen that the distribution is more linear and compact in formation, which is not entirely similar to rainfall.

a)



b)

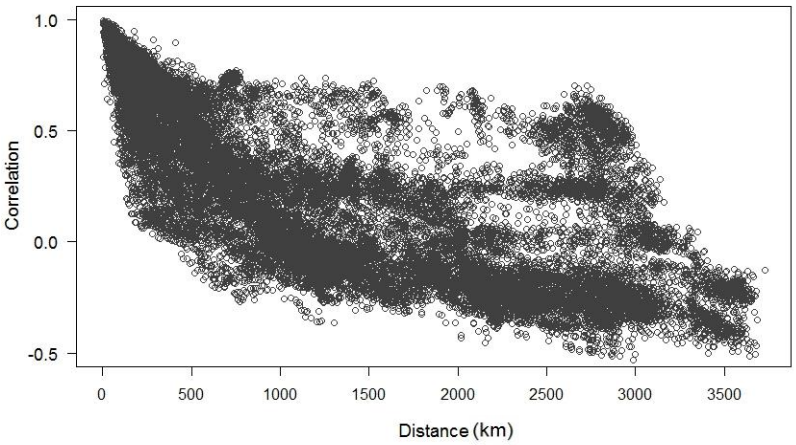
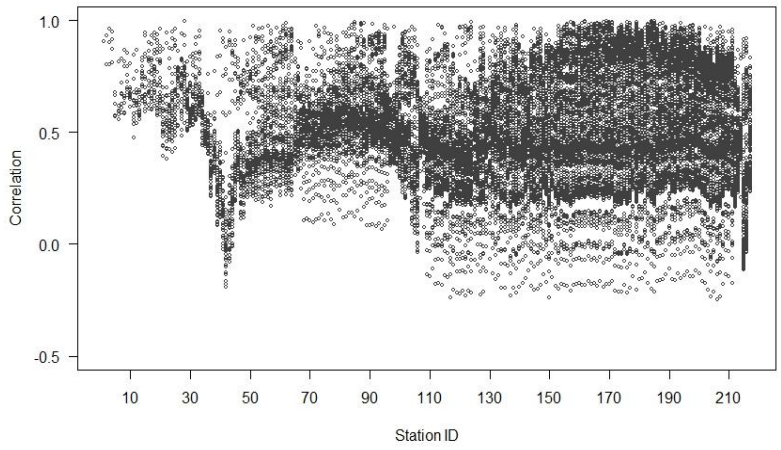


Figure 7.4: Correlation analysis for streamflow-rainfall combination for 218 stations in Australia: (a) Correlation of each station to other stations; and (b) distance-correlation relationship.

7.3.2 Streamflow and Potential Evapotranspiration

Figure 7.5 presents correlations between the 218 stations (Figure 7.5(a)) and their distance-correlation relationship (Figure 7.5(b)) for the combination of streamflow and PET. It appears that the correlations (Figure 7.5(a)) range similarly to that for the PET (Figure 7.3(a)) but the distribution pattern seems to be nearly similar to that for streamflow (Figure 7.1(a)). From the distance-correlation relationship, shown in Figure 7.5(b), it can be seen that the distribution is more compact in formation with correlations decreasing as the distance increases.

a)



b)

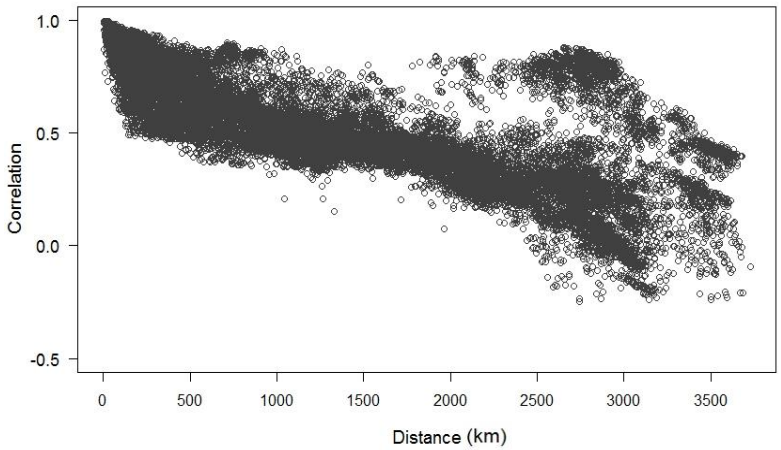


Figure 7.5: Correlation analysis for streamflow-PET combination for 218 stations in Australia: (a) Correlation of each station to other stations; and (b) distance-correlation relationship.

7.3.3 Rainfall and Potential Evapotranspiration

Figure 7.6 presents the correlations between the 218 stations (Figure 7.6(a)) and their distance-correlation relationship (Figure 7.6(b)) for the combination of rainfall and PET. As seen from Figure 7.6(a), the rainfall-PET combination results in slightly different correlations to either that of rainfall or that of PET. The results for this combination are slightly different from those for the combination of streamflow and PET (Figure 7.5(a)) in terms of the partition shown by the rainfall (Figure 7.2(a)), including the comparison in distance-correlation (Figure 7.6(b)) that is more sparse led by lower correlations from the rainfall connections.

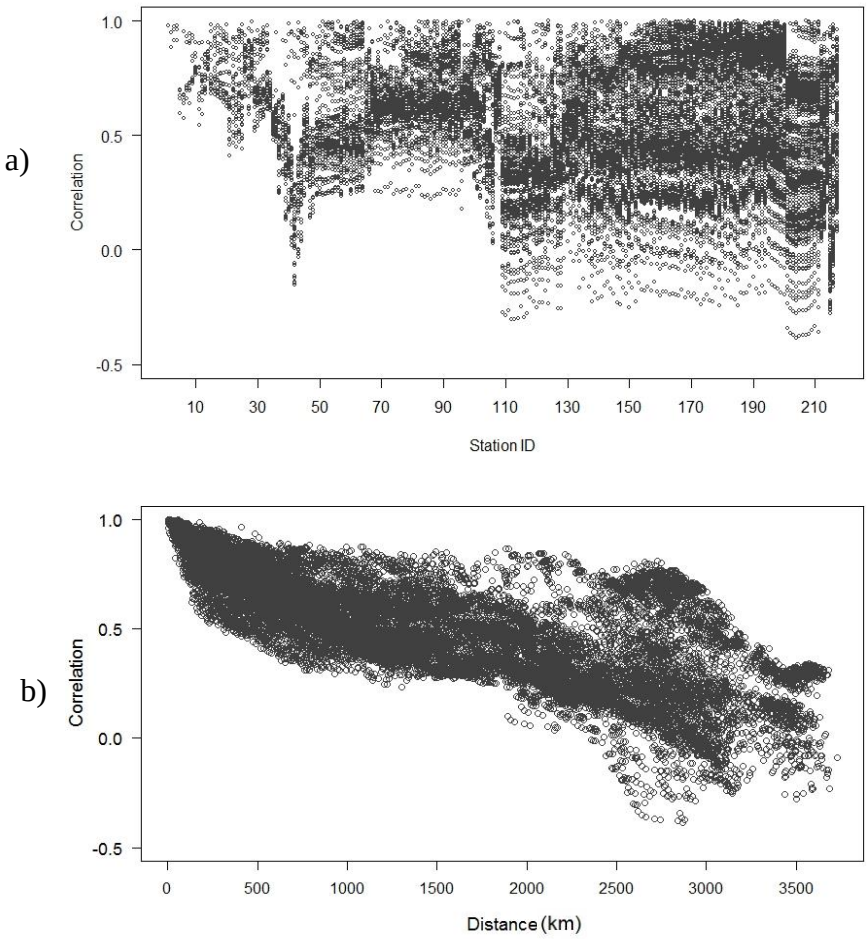


Figure 7.6: Correlation analysis for rainfall-PET combination for 218 stations in Australia: (a) Correlation of each station to other stations; and (b) distance-correlation relationship.

7.3.4 Streamflow, Rainfall, and Potential Evapotranspiration

Figure 7.7 presents the correlations between the 218 stations (Figure 7.6(a)) and their distance-correlation relationship (Figure 7.7(b)) for the combination of streamflow, rainfall, and PET. The correlations shown in Figure 7.7(a) clearly seem to represent the influence of all three variables, with the compactness observed from streamflow, the

divisions observed for rainfall, and also the lower range of correlations observed for PET. In terms of the distance-correlation relationship, shown in Figure 7.7(b), the correlations tend to decrease when the distance increases with a wider distribution when compared to that observed for the other combinations (i.e., with two variables).

a)

b)

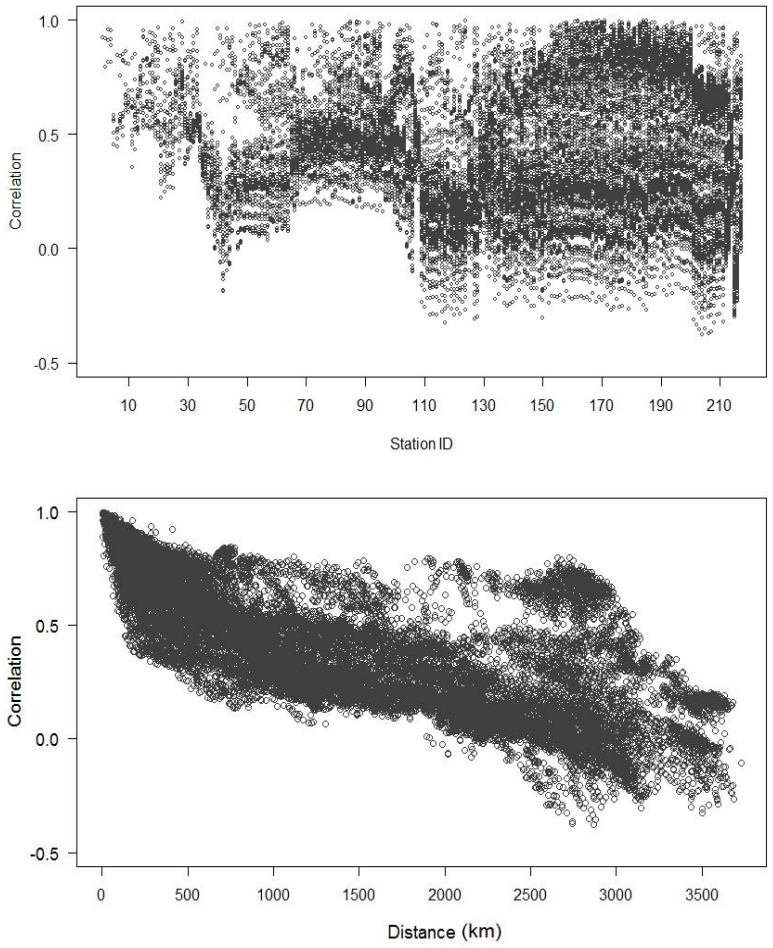


Figure 7.7: Correlation analysis for streamflow-rainfall-PET combination for 218 stations in Australia: (a) Correlation of each station to other stations; and (b) distance-correlation relationship.

7.4 Single-variable vs. Multi-variable Classification

7.4.1 Station Count with the Influence of Variables and Threshold

Having discussed the correlations of the variables both in a single-variable sense and in a multi-variable sense as well as distance-correlation relationships for such, an attempt is now made to examine further the usefulness of multiple variables in the implementation of the MDEB method for classification of the above 218 catchments in Australia. In this regard, it would be helpful, first of all, to have an accurate count of the number of communities identified with a specific number of stations, since such a count can help identify whether a given catchment has a certain level of similarity with other catchments and to the other cases (to different threshold values for that matter), and how. Tables 7.1 to 7.6 present the classification results, including the number of communities identified and the number of stations, for all the above seven cases of single variables and multiple variables for six selected threshold values, respectively: i.e., T = 0.65, 0.7, 0.75, 0.8, 0.85, and 0.9. In these tables, the number of communities (NC) are arranged according to the number of stations in each community (NSC).

There are some worth-mentioning outcomes in a broader sense, with the formation of communities definitely varying depending on the variable and the combinations under certain threshold values. For instance, the results indicate that:

(1) Streamflow data results in the highest number of communities and PET data results in the lowest number of communities, as compared to the other five cases, regardless of the threshold values. These results are shown in Figure 7.8 (details are explained below) with line graphs, where the ones for streamflow data and PET data are coloured

with green and dark blue, respectively. This seems to suggest that the catchments are divided solely based on the level of connectivity within them (this is supported by the results in Figures 7.5(a) and 7.7(a));

(2) Since the streamflow data offers the greatest number of communities that have only few catchments within them (the least at T = 0.65 with more than 50% out of total number of communities) regardless of the threshold, cases (i.e., combinations) associated with streamflow tend to also form more number of communities with only few catchments, especially at threshold value T = 0.7 and above. For instance, the communities identified with only a few catchments in cases of streamflow and rainfall, streamflow and PET, and streamflow, rainfall and PET result in at least one-third of the total number of communities identified (2 out of 6 communities) (in the case by the combination of streamflow and PET at T = 0.7) and keep increasing as the threshold value increases;

(3) Since the PET data results in the greatest number of stations in a community (217 catchments out of 218 total number of catchments at T = 0.65), any case associated with PET data tends to form larger-sized community as well, regardless of the threshold. For instance, the cases of streamflow and PET, rainfall and PET, and streamflow, rainfall, and PET result in at least one community with more than 100 catchments (120, 117 and 112 catchments, respectively, out of 218 catchments at T = 0.65), and the number of catchments in any given community then gradually decreases when the threshold value is increases. This seems to suggest that the correlations from the PET are able to affect the community formation, even when they are combined with the other variables. This means that each of the catchments within a given large community has strong

connections with the rest of the catchments, regardless of the distance between them or whether they are part of different basins or regions.

Table 7.1: Sizes of communities identified using the MDEB method for all single- and multi-variable cases for Australian catchments at threshold value T = 0.65. (NSC is the number of stations in community, NC is the number of communities and NS is the number of stations)

Streamflow (S)			Rainfall (R)			PET (P)			S & R			S & P			R & P			S, R & P		
NSC	NC	NS	NSC	NC	NS	NSC	NC	NS	NSC	NC	NS	NSC	NC	NS	NSC	NC	NS	NSC	NC	NS
1	11	11	1	1	1	1	1	1	1	4	4	2	1	2	1	2	2	2	1	2
2	3	6	2	1	2	217	1	217	2	2	4	48	2	96	46	1	46	3	1	3
3	2	6	3	1	3	**Total**	**2**	**218**	3	1	3	120	1	120	53	1	53	13	1	13
5	1	5	8	1	8				7	1	7	**Total**	**4**	**218**	117	1	117	40	1	40
6	1	6	11	1	11				11	2	22				**Total**	**5**	**218**	48	1	48
7	1	7	15	1	15				15	1	15							112	1	112
9	1	9	40	1	40				30	1	30							**Total**	**6**	**218**
13	2	26	48	1	48				44	1	44									
19	1	19	90	1	90				89	1	89									
29	1	29	**Total**	**9**	**218**				**Total**	**14**	**218**									
94	1	94																		
Total	**25**	**218**																		

Table 7.2: Sizes of communities identified using the MDEB method for all single- and multi-variable cases for Australian catchments at threshold value T = 0.7. (NSC is the number of stations in community, NC is the number of communities and NS is the number of stations).

Streamflow (S)			Rainfall (R)			PET (P)			S & R			S & P			R & P			S, R & P		
NSC	NC	NS	NSC	NC	NS	NSC	NC	NS	NSC	NC	NS	NSC	NC	NS	NSC	NC	NS	NSC	NC	NS
1	14	14	1	2	1	1	1	1	1	5	5	1	2	2	46	1	46	2	3	6
2	6	12	2	2	4	217	1	217	2	3	6	13	1	13	50	1	50	17	1	17
3	1	3	3	1	3	**Total**	**2**	**218**	3	3	9	41	1	41	122	1	122	40	1	40
5	1	5	10	1	10				7	1	7	47	1	47	**Total**	**3**	**218**	47	1	47
7	2	14	11	1	11				9	1	9	113	1	113				108	1	108
9	1	9	15	1	15				11	2	22	**Total**	**6**	**218**				**Total**	**7**	**218**
10	1	10	39	1	39				15	1	15									
14	1	14	48	1	48				16	1	16									
18	1	18	86	1	86				19	1	19									
26	1	26	**Total**	**11**	**218**				26	1	26									
93	1	93							84	1	84									
Total	**30**	**218**							**Total**	**20**	**218**									

Table 7.3: Sizes of communities identified using the MDEB method for all single- and multi-variable cases for Australian catchments at threshold value T = 0.75. (NSC is the number of stations in community, NC is the number of communities and NS is the number of stations).

Streamflow (S)			Rainfall (R)			PET (P)			S & R			S & P			R & P			S, R & P		
NSC	NC	NS	NSC	NC	NS	NSC	NC	NS	NSC	NC	NS	NSC	NC	NS	NSC	NC	NS	NSC	NC	NS
1	24	24	1	1	1	1	1	1	1	8	8	1	2	2	2	1	2	1	4	4
2	3	6	2	3	6	217	1	217	2	2	4	2	3	6	10	1	10	2	3	6
3	1	3	3	1	3	**Total**	**2**	**218**	3	2	6	3	1	3	41	1	41	3	1	3
4	2	8	4	1	4				5	1	5	14	1	14	51	1	51	6	1	6
5	1	5	5	1	5				10	2	20	38	1	38	114	1	114	10	1	10
9	1	9	6	1	6				11	2	22	44	1	44	**Total**	**5**	**218**	11	1	11
10	2	20	9	1	9				14	1	14	111	1	111				15	2	30
17	1	17	11	2	22				15	1	15	**Total**	**10**	**218**				29	1	29
19	1	19	16	1	16				17	1	17							30	1	30
26	1	26	30	1	30				26	1	26							89	1	89
81	1	81	42	1	42				81	1	81							**Total**	**16**	**218**
Total	**38**	**218**	74	1	74				**Total**	**22**	**218**									
			Total	**15**	**218**															

Table 7.4: Sizes of communities identified using the MDEB method for all single- and multi-variable cases for Australian catchments at threshold value T = 0.8. (NSC is the number of stations in community, NC is the number of communities and NS is the number of stations).

Streamflow (S)			Rainfall (R)			PET (P)			S & R			S & P			R & P			S, R & P		
NSC	NC	NS	NSC	NC	NS	NSC	NC	NS	NSC	NC	NS	NSC	NC	NS	NSC	NC	NS	NSC	NC	NS
1	35	35	1	3	3	1	1	1	1	21	21	1	7	7	2	1	2	1	5	5
2	3	6	2	3	6	217	1	217	2	2	4	2	2	4	4	1	4	2	3	6
3	3	9	3	1	3	**Total**	**2**	**218**	3	4	12	3	1	3	16	1	16	3	2	6
4	1	4	5	1	5				4	2	8	6	1	6	40	1	40	5	1	5
7	1	7	8	1	8				5	1	5	11	1	11	47	1	47	8	1	8
9	1	9	10	1	10				6	1	6	14	1	14	109	1	109	9	1	9
10	3	30	11	1	11				10	4	40	15	1	15	**Total**	**6**	**218**	11	2	22
11	1	11	12	2	24				11	1	11	29	1	29				15	1	15
12	1	12	16	1	16				16	1	16	30	1	30				16	1	16
13	1	13	17	1	17				26	1	26	99	1	99				19	1	19
16	1	16	18	1	18				69	1	69	**Total**	**17**	**218**				26	1	26
66	1	66	26	1	26				**Total**	**39**	**218**							81	1	81
Total	**52**	**218**	71	1	71													**Total**	**20**	**218**
			Total	**18**	**218**															

Table 7.5: Sizes of communities identified using the MDEB method for all single- and multi-variable cases for Australian catchments at threshold value T = 0.85. (NSC is the number of stations in community, NC is the number of communities and NS is the number of stations).

Streamflow (S)			Rainfall (R)			PET (P)			S & R			S & P			R & P			S, R & P		
NSC	NC	NS	NSC	NC	NS	NSC	NC	NS	NSC	NC	NS	NSC	NC	NS	NSC	NC	NS	NSC	NC	NS
1	51	51	1	11	11	3	1	3	1	37	37	1	15	15	2	3	6	1	15	15
2	6	12	2	7	14	215	1	215	2	4	8	2	7	14	4	1	4	2	3	6
3	5	15	3	3	9	**Total**	**2**	**218**	3	4	12	3	1	3	11	1	11	3	4	12
4	3	12	4	2	8				4	1	4	5	2	10	12	1	12	4	1	4
5	3	15	8	2	16				5	1	5	7	1	7	16	1	16	5	2	10
6	1	6	9	1	9				8	1	8	8	2	16	18	1	18	7	1	7
7	2	14	11	3	33				9	2	18	9	2	18	27	1	27	10	4	40
10	1	10	12	2	24				10	1	10	14	1	14	39	1	39	11	1	11
12	2	22	16	1	16				11	1	11	18	1	18	85	1	85	17	1	17
14	1	14	20	1	20				12	1	12	21	1	21	**Total**	**11**	**218**	26	1	26
45	1	45	24	1	24				13	1	13	82	1	82				70	1	70
Total	**76**	**218**	34	1	34				16	1	16	**Total**	**34**	**218**				**Total**	**34**	**218**
			Total	**35**	**218**				20	1	20									
									44	1	44									
									Total	**57**	**218**									

Table 7.6: Sizes of communities identified using the MDEB method for all single- and multi-variable cases for Australian catchments at threshold value T = 0.9. (NSC is the number of stations in community, NC is the number of communities and NS is the number of stations).

Streamflow (S)			Rainfall (R)			PET (P)			S & R			S & P			R & P			S, R & P		
NSC	NC	NS	NSC	NC	NS	NSC	NC	NS	NSC	NC	NS	NSC	NC	NS	NSC	NC	NS	NSC	NC	NS
1	82	82	1	26	26	1	1	1	1	55	55	1	35	35	1	3	3	1	37	37
2	22	44	2	6	12	7	1	7	2	7	14	2	3	6	2	3	6	2	4	8
3	3	9	3	4	12	32	1	32	3	6	24	3	2	6	3	2	6	3	4	12
4	3	12	4	7	28	178	1	178	4	5	20	4	3	12	4	1	4	4	1	4
6	1	6	5	1	5	**Total**	**4**	**218**	5	1	5	6	1	6	8	1	8	5	2	10
5	1	5	6	1	6				6	2	12	8	1	8	10	1	10	7	1	7
7	1	7	7	3	21				8	1	8	10	2	20	11	1	11	8	1	8
9	1	9	9	2	18				9	3	27	11	2	22	12	2	24	9	2	18
10	1	10	11	3	33				10	1	10	16	1	16	15	1	15	10	1	10
34	1	34	14	1	14				11	2	22	20	1	20	17	2	34	11	1	11
Total	**116**	**218**	19	1	19				27	1	27	67	1	67	26	1	26	13	1	13
			24	1	24				**Total**	**84**	**218**	**Total**	**52**	**218**	71	1	71	16	1	16
			Total	**56**	**218**										**Total**	**19**	**218**	20	1	20
																		44	1	44
																		Total	**58**	**218**

Figure 7.8 presents the line graph of the communities identified with all seven single-variable and multi-variable cases under the selected threshold values. Different colours are used to represent each case for comparison. As seen, there are clear differences in terms of community formation when different variables are used, such as from streamflow (coloured in green), streamflow and rainfall (coloured in purple), and PET (coloured in dark blue). Despite this, there are also some cases, such as rainfall (coloured in red), streamflow and PET (coloured in brown), and streamflow, rainfall, and PET (coloured in light blue), that tend to have similar number of communities among them at certain thresholds, especially at T = 0.85. These observations show some kind of similarity between single-variable and multi-variable cases. Therefore, an accurate count of connection links is also needed in order to examine the likeliness in community formation, i.e., for catchment classification for each case from single-variable sense and multi-variable sense.

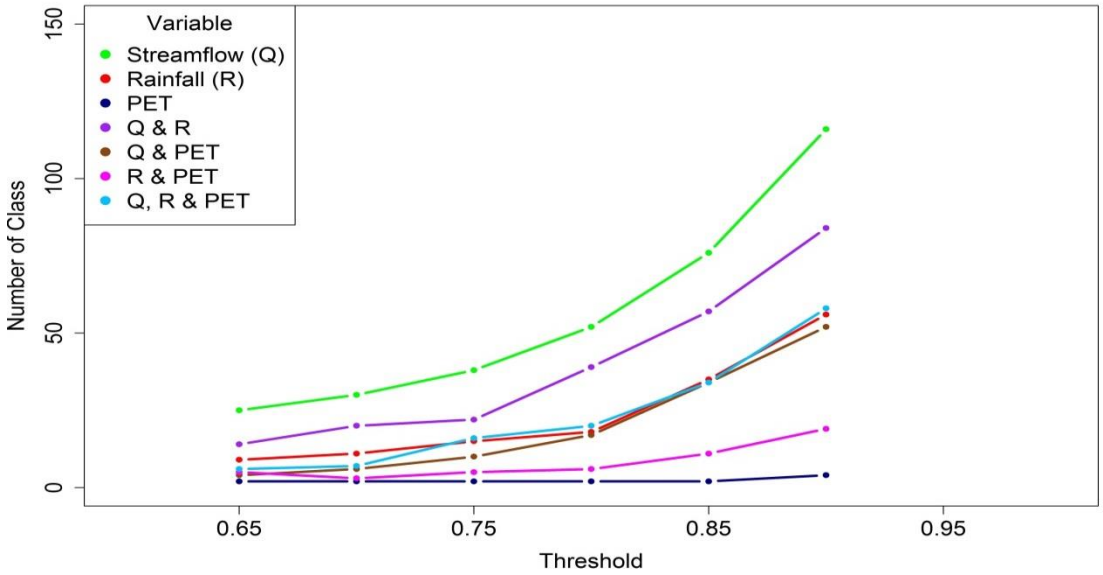


Figure 7.8: Number of communities identified for six selected threshold values (T = 0.65, 0.70, 0.75, 0.80, 0.85, and 0.90) for all seven single-variable and multi-variable cases. Each case is indicated with a different colour.

7.4.2 Connection Link Count

Table 7.7 presents the number of connection links for each of the above seven cases at the six selected threshold values. For better visualization, Figure 7.9 presents the line graph to represent the number of connection links based on the selected threshold values for all the seven cases, with different colours corresponding to the different cases, similar to that shown in Figure 7.8. As seen from Table 7.7, the PET data has the highest count of links (21319 links) at T = 0.65, and the count then gradually decreases as the threshold value increases. The lowest count of the connection links is observed for the streamflow data, with 3095 links, which seems to suggest that this could be the cause of the high number of communities identified (Table 7.1) from the streamflow data. In the case of streamflow-rainfall combination, the number of connection links is almost similar to that of the streamflow data as represented with colours in purple and green, respectively (Figure 7.9). However, it also results in a significantly different number of communities identified within them (Figure 7.8). In addition, in spite of the similarity in the number of communities identified by certain cases, such are rainfall (red), streamflow and PET (brown), and streamflow, rainfall, and PET (light blue) from Figure 7.8, there still exist clear differences in terms of the number of connection links obtained by the respective cases, which suggests that similarity in the number of communities identified is not significantly related to the number of connections links,

and vice versa. Therefore, comparison in terms of catchment classification results based on these cases is needed in order to examine if there is any similarity (or dissimilarity) could be achieved by either similarity in the number of communities or if the number of connection links really can effectively divide the catchment network.

Table 7.7: Number of connection links for seven different single-variable and multi-variable cases at six selected threshold values.

Threshold Values (T)	Stream-flow (S)	Rainfall (R)	PET (P)	S & R	S & P	R & P	S, R & P
0.65	3095	4078	21319	3304	7086	8523	5735
0.7	2347	3344	20582	2616	5941	7050	4600
0.75	1686	2537	19137	1970	4834	5632	3479
0.8	1178	1816	17852	1357	3621	4295	2529
0.85	747	1133	16672	832	2211	2990	1622
0.9	348	651	15500	431	1138	1680	809

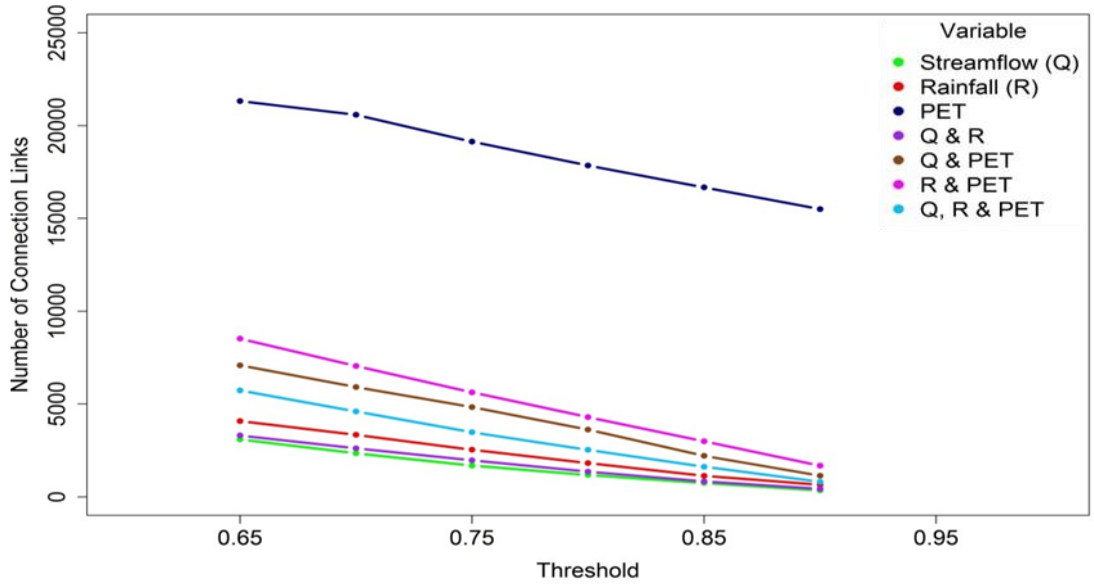
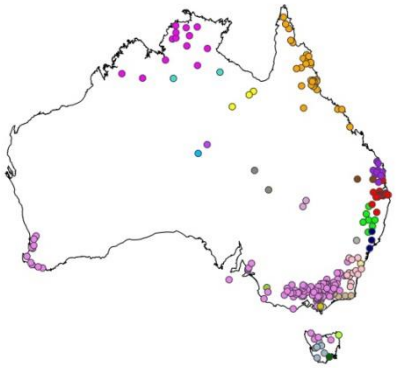


Figure 7.9: Line graph of the number of connection links based on six selected threshold values for all seven single-variable and multi-variable cases. Each case is indicated by a colour corresponding to the legend in Figure 7.8.

Finally, one must also note that it is possible to have similar classification based on different variables, but at different thresholds. An example of this is presented in Figure 7.10. As seen, communities identified using streamflow data (Figure 7.10(a), (c), and (e)) at T = 0.65, 0.75 and 0.8, respectively, are almost similar to the classification results obtained using the combination of streamflow and PET (Figure 7.10(b), (d), and (f)) at T = 0.8, 0.85 and 0.9, respectively. The similarities among them are considered mostly in larger communities and also depend on the number of communities identified and the number of connection links. From Figures 7.10(e) and (f), the similarities can be found based on the same number of communities identified (52) and almost similar in the number of connections links (1178 and 1138), as indicated at T = 0.8 and 0.9,

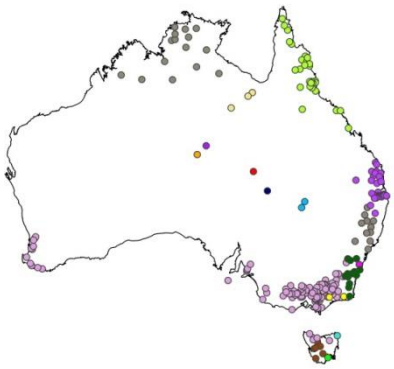
respectively. This seems to suggest that the single-variable streamflow data tends to form the communities at lower threshold values than those formed by the multi-variables of streamflow and PET, which need higher threshold value in order to form such communities.

a)



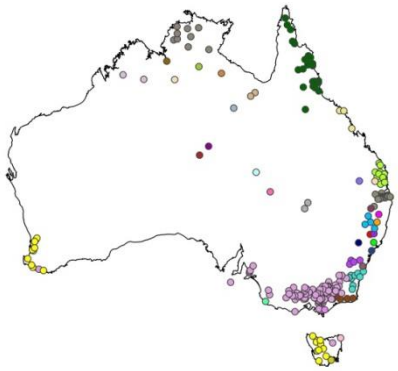
Streamflow variable
at T = 0.65.
No. of communities=25
Connection links= 3095

b)



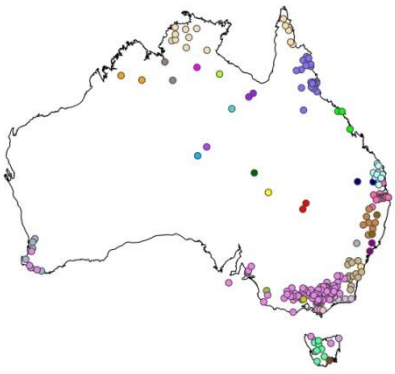
Streamflow & PET
variables at T = 0.8.
No. of communities=17
Connection links= 3621

c)



Streamflow variable
at T = 0.75.
No. of communities=38
Connection links= 1686

d)



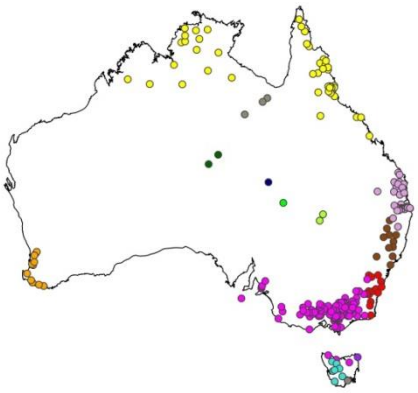
Streamflow & PET
variables at T = 0.85.
No. of communities=34
Connection links= 2211

Figure 7.10: Communities identified by: (a, c, and e) streamflow; and (b, d, and f) streamflow and PET with respect to different correlation threshold values using the MDEB method. Each colour represents a community, and different colours are used only to distinguish the communities and hold no meaning when comparing across plots.

Apart from the above, the results presented in Figure 7.11 indicate that communities identified by the combination of streamflow and rainfall (Figure 7.11(a), (c) and (e)) at T = 0.65, 0.75 and 0.85, respectively, are almost similar to the classification results obtained using the combination of streamflow, rainfall, and PET (Figure 7.11(b), (d) and (f)) at T = 0.75, 0.8 and 0.9, respectively. As seen, there is quite a bit of similarity, especially the ones indicated by Figure 7.11(e) and (f), by considering the number of communities identified (57 and 58) and the number of connection links (832 and 809), as they could form similar community structure. This comparison also suggests that the two-variable combination only requires a low value of

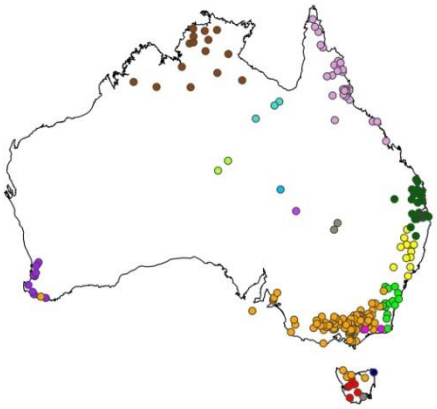
threshold when compared to three-variable combination, which requires a higher threshold value to form such division.

a)



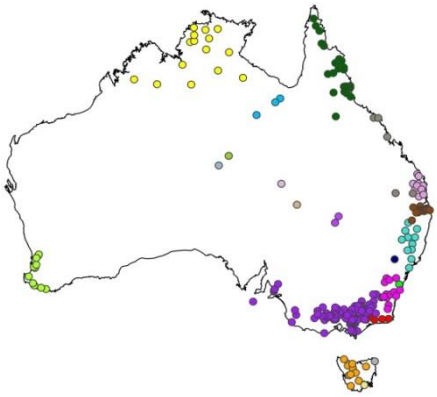
Streamflow & Rainfall variables at T = 0.65.
No. of communities=14
Connection links= 3304

b)



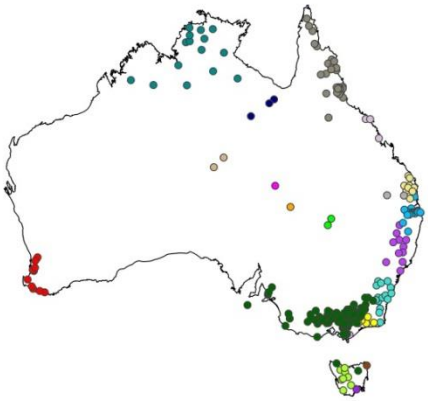
Streamflow, Rainfall & PET variables at T = 0.75.
No. of communities=16
Connection links= 3479

c)



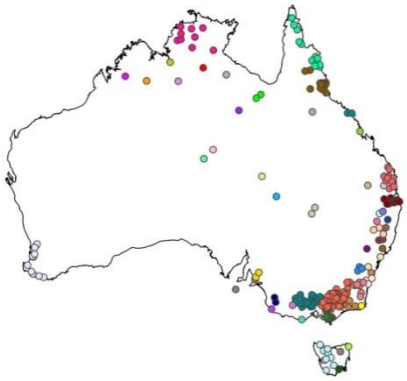
Streamflow & Rainfall variables at T = 0.75.
No. of communities=22
Connection links= 1970

d)



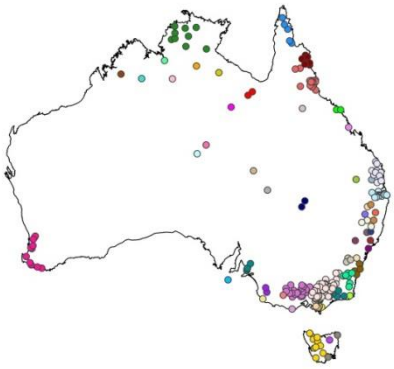
Streamflow, Rainfall & PET variables at T = 0.8.
No. of communities=20
Connection links= 2529

e)



Streamflow & Rainfall
variables at T = 0.85.
No. of communities=57
Connection links= 832

f)



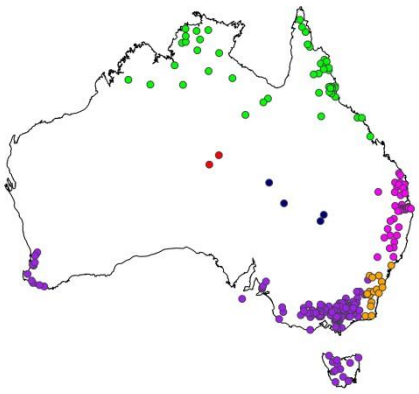
Streamflow, Rainfall &
PET variables at T = 0.9.
No. of communities=58
Connection links= 809

Figure 7.11: Communities identified by: (a, c, and e) streamflow and rainfall; and (b, d, and f) streamflow, rainfall and PET with respect to different correlation threshold values using the MDEB method. Each colour represents a community, and different colours are used only to distinguish the communities and hold no meaning when comparing across plot.

Figure 7.12, presenting the communities identified using two variables (i.e., rainfall and PET) and three variables (i.e., streamflow, rainfall and PET), also deserves discussion. A comparison of the results indicates that higher threshold values are required by the rainfall and PET case, with T = 0.8, 0.85 and 0.9 (Figure 7.12 (a), (c) and (e)) than by the streamflow, rainfall, and PET case, with T = 0.7, 0.75 and 0.85, (Figure 7.12 (b), (d), and (f)), respectively. This seems to suggest that classification based on hydroclimatic variables alone in a multi-variable sense is also found useful to

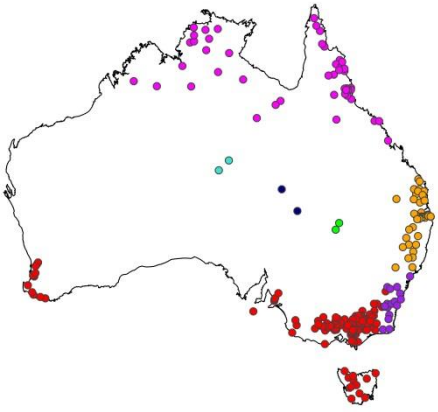
assess the catchment classification. In spite of this, the rainfall associated correlations are found to need higher threshold values to divide the network when compared to the case of streamflow, rainfall, and PET classification result.

a)



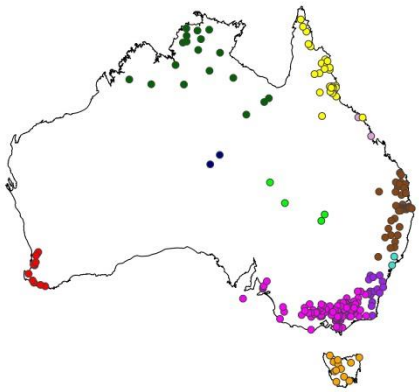
Rainfall & PET
variables at T = 0.8.
No. of communities=6
Connection links= 4295

b)



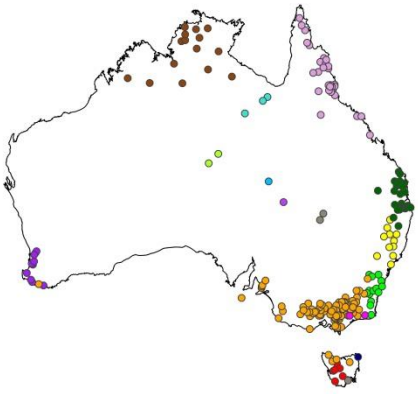
Streamflow, Rainfall &
PET variables at T = 0.7.
No. of communities=7
Connection links= 4600

c)



Rainfall & PET
variables at T = 0.85.
No. of communities=11
Connection links= 2990

d)



Streamflow, Rainfall &
PET variables at T = 0.75.
No. of communities=16
Connection links= 3479

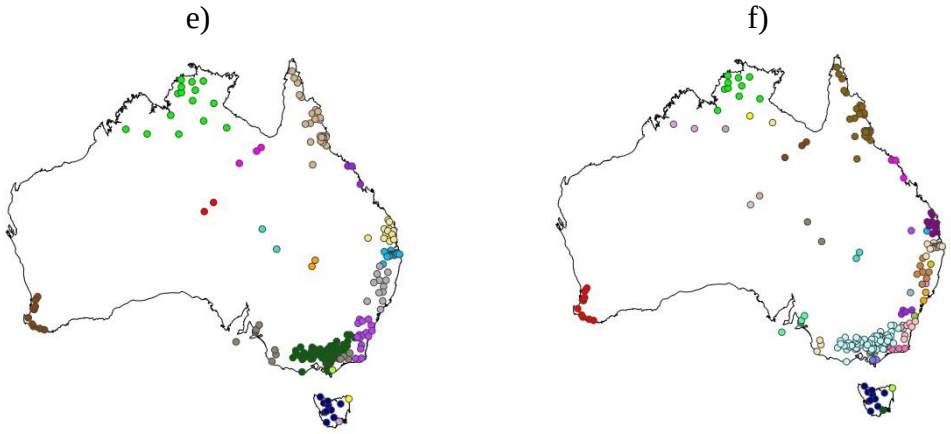


Figure 7.12: Communities identified by: (a, c, and e) rainfall and PET; and (b, d, and f) streamflow, rainfall, and PET with respect to different correlation threshold values using the MDEB method. Each colour represents a community, and different colours are used only to distinguish the communities and hold no meaning when comparing across plot.

7.5 Summary

This chapter has presented the application of the Modularity Density-based Edge Betweenness (MDEB) method for catchment classification in a multi-variable sense. The approach has been implemented for classification of 218 catchments across Australia. In addition to streamflow (as presented in Chapters 5 and 6), rainfall and PET have also been used. Four different combinations of multiple variables (i.e., streamflow and rainfall, streamflow and PET, rainfall and PET, as well as streamflow, rainfall, and

PET) have been used for classification, with six different thresholds considered. The results are also compared with those obtained from the three single-variable cases (i.e., streamflow, rainfall, and PET independently). Each of these cases was examined in terms of correlations between stations and distance-correlation relationship. Differences in the number of communities and in the number of connection links were explored to assess the similarity of classification outcomes based on the different combinations of variables. The results generally indicated that the multi-variable approach for catchment classification is useful and effective, and that variables or combinations of variables can provide similar classification outcomes, but at different correlation threshold values. The present results are certainly encouraging as to the usefulness of the community structure methods, especially when appropriate modifications and improvements are made to the existing community structure methods and appropriate approach is adopted by using the most important variables influencing the catchments, either individually or in combination.

Chapter 8
Conclusions

The main focus of this thesis was the development of an improved community structure method for catchment classification. The Edge Betweenness (EB) method, due to its widespread use, was considered as a representative community structure method for further improvement. The improved method, called the Modularity Density-based Edge Betweenness (MDEB) method, was applied to hydrologic data from a large number of catchments in Australia and in the United States, for classification within the respective countries. Both single-variable and multi-variable cases were considered in this study. In a single-variable sense, streamflow was the main and common variable considered for both Australia and the United States. The multi-variable approach was attempted only for data from Australia, with rainfall and potential evapotranspiration (PET) also considered, in addition to streamflow. In the multi-variable approach, combinations of any two and all three variables were considered.

8.1 Edge Betweenness method for catchment classification

To study the general suitability of the Edge Betweenness (EB) method, the method was applied to classify a large number of catchments in two regions: 218 catchments in Australia and 639 catchments in the United States. These two regions and, thus, the associated catchments cover a wide range of possibilities, in terms of hydroclimatic, topographic, geomorphic, land use, and other relevant properties. In the single-variable streamflow case (see Chapter 5), a threshold value (i.e., correlation threshold) of T = 0.8 was considered for Australia and T = 0.75 was considered for the US. In each case, a total of 61 communities were identified by the EB method. The results generally indicated that: (1) a very small number of catchment communities had a large number of catchments within them — for instance, 11 largest catchment communities from Australia and ten largest communities from the US combined to represent as much as 70% of the total number of catchments in the respective cases; and (2) in both cases, a significantly large number of catchment communities had only a very few catchments within them — for instance, almost 70% of the total number of communities identified had only one or two stations within them and, thus, represented only about 20% and 10% of the total number of catchments from Australia and the US, respectively. Additionally, the catchment classification results offered some interesting interpretations when the catchment communities were compared with the catchment properties (i.e., drainage area, stream length, elevation) and the flow properties (i.e., mean, coefficient of variation, correlation-distance).

8.2 Modularity Density-based EB (MDEB) method for catchment classification

As the EB method is susceptible to scale problem due to the modularity function that is used to measure the strength of the community structure (Fortunato and Barthelemy, 2007), an improvement to the EB method was proposed. The new method used a modularity density function (or D value) by maximization, to obtain the best split of the network, and the method was termed as the Modularity Density-based Edge Betweenness (MDEB) method.

The MDEB method was applied to the same 218 catchments from Australia and 639 catchments in the United States for catchment classification. Considering the single-variable streamflow case (Chapter 6), three different scenarios in network sizes were studied: (1) the entire network – i.e., 218 catchments in Australia and 639 stations in the US; (2) smaller network sizes, based on 100 and 300 randomly selected stations (with 100 different realizations) for Australia and the US, respectively – purely to address the network size; and (3) smaller network sizes, based on 9 different drainage division regions in Australia and 18 different hydrologic units in the US – to address the network size and regional similarity. The results indicated that the MDEB method generally performed better than the EB method, for both Australia and the US. Furthermore, the superiority of the MDEB method over the EB method was assessed in terms of the number and percentage of stations that changed (based on classification from random realizations and drainage divisions as well as hydrologic units) from the base (original) classification results.

8.3 Multi-variable approach for catchment classification

In addition to the mainly streamflow-based single-variable approach for catchment classification, a multi-variable approach using the MDEB method was also proposed and applied for the 218 Australian catchments (Chapter 7). The multi-variable approach included rainfall and PET, in addition to streamflow. Different combinations of these three variables (any two of the three as well as all three) were considered for implementation. In each combination, the correlations between the different stations were estimated by the average of the summation of the correlations of the variables under consideration (i.e., two or three, as appropriate). The classification results from the multi-variable-based approach were also compared with those from the single-variable approach (with each of the three considered separately), through assessing the distance-correlation relationship of the stations, a count of the number of stations within the communities identified, and count of the connection links that occurred within the network at different (six) threshold values (i.e., T = 0.65, 0.7, 0.75, 0.8, 0.85, and 0.9).

The results indicated that classification based on the multi-variable approach was almost similar to the classification based on the single-variable approach, especially streamflow, but at different correlation thresholds. Considering the similarities in the number of stations in the communities, number of communities identified, and number of connection links for the purpose of comparison, the following observations were made: (1) classification of catchments based on streamflow alone was found to be somewhat similar to that found for streamflow and PET together; (2) classification based on streamflow and rainfall together was found to be somewhat similar to that obtained for streamflow, rainfall and PET together; and (3) classification based on

rainfall and PET together tended to be similar to that obtained for streamflow, rainfall, and PET together as well.

8.4 Limitations and future work

The limitations of the present thesis and scope for future work relate largely to further improve the community structure-based methods for catchment classification and also covering a wide range of other catchments and associated properties towards a more generally acceptable classification framework.

Development and application of the MDEB method (Chapters 3 and 6) as an improvement to the traditional EB method certainly provided a more reliable catchment classification. However, the method is still not completely satisfactory, since a great number of stations were found to change communities (either merged with other communities or formed new ones). This was particularly the case for Australia, perhaps because of the distribution of the catchments in the study area — most of the stations are located along the coastal area and only a very few in the middle region. The MDEB method performed better for the catchments in the US. It would be interesting to see if this (i.e., locations and density of stations) is indeed the case for any and every region, by studying many different regions around the world. Further modification to the modularity density function may also be possible, by considering the external links that are associated with external nodes, instead of only the internal links that are within the internal nodes of a subgraph, in order to address the imbalance of the fraction of the density of connection links for network partition.

The multi-variable approach for catchment classification, using the MDEB method in particular, presented in Chapter 7, has offered some useful insights, especially in terms of the combination of variables considered and the 'equivalent' correlation thresholds with the single-variable approach. Nevertheless, since the PET data has very strong correlations among the stations, it is not really useful in identifying and separating the communities when correlations are used as a basis to identify the connections. Other climatic/catchment variables, which exhibit more variability across the catchments, may turn out to be more appropriate for consideration. Regions and catchments for which data for a large number of climatic/catchment variables are available are particularly suited for such an analysis. Even then, it would be important and interesting to see which combination(s) of variables would offer the most reliable classification, and also whether such a combination(s) would provide a classification that would be substantially different and better than the one that can be obtained from a single-variable approach. With the importance of catchment classification in hydrology (and beyond) and the emerging ideas of complex networks and community structure concepts and their applications, it is hoped that research in the above directions will assume particular significance in the coming years.

REFERENCES

Abarbanel, H. D. I. (1996). Analysis of observed chaotic data, Springer, New York.

Albert, R., Jeong, H., and Barabasi, A-L., 1999. Internet: diameter of the world wibe web. Nature 401:130-131.

Ali, G., Tetzlaff, D., Soulsby, C., McDonnell, J. J., and Capell, R. (2012). "A comparison of similarity indices for catchment classification using a cross-regional dataset." Adv. Water Resour., 40, 11–22.

Archfield, S. A. and Vogel, R. M. 2010. Map correlation method: Selection of a reference streamgage to estimate daily streamflow at ungaged catchments, Water Resour. Res., 46, W10513, doi:10.1029/2009WR008481.

Bak, P., 1996. How nature works: the science of self-organized criticality. Springer-Verlag, New York, 212 pp.

Barabási, A.-L., 2002. Linked: the new science of networks. Perseus, Cambridge, MA.

Barabási, A.-L., and Albert, R. (1999). "Emergence of scaling in random networks." Science, 286(5439), 509–512.

Barrat, A., Barthélemy, M., and Vespinani, A. (2008). Dynamical processes on complex networks, Cambridge University Press, Cambridge, U.K.

Beckinsale, R. P. (1969). "River regimes." Water, earth, and man, R. J. Chorley, ed., Methuen, London, 455–471.

Berge, C.: The Theory of Graphs and Its Applications, Matheun, Ann Arbor, MI, USA, 1962.

Beven, K. J. (2002). "Uncertainty and the detection of structural change in models of environmental systems." Environmental foresight and models: A manifesto, M. B. Beck, ed., Elsevier, Oxford, U.K., 227–250.

Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. J. Stat. Mech. 2008 (10), P10008.

Bocchiola, D., et al. (2011). "Prediction of future hydrological regimes in poorly gauged high altitude basins: The case study of the upper Indus, Pakistan." Hydrol. Earth Syst. Sci., 15(7), 2059–2075.

Boers, N., Bookhagen, B., Marwan, N., Kurths, J., and Marengo, J.: Complex networks identify spatial patterns of extreme rainfall events of the South American Monsoon System, Geophys. Res. Lett., 40, 4386–4392, doi:10.1002/grl.50681, 2013.

Bollobás, B.: Modern Graph Theory, Springer, New York, USA, 1998.

Bondy, J. A. and Murty, U. S. R.: Graph Theory with Applications, Elsevier Science Ltd, New York, USA, 1976.

Botta, F. and Del Genio, C. I., 2016. “Finding network communities using modularity density.” J. Stat. Mech - Theory E., doi:10.1088/1742-5468/2016/12/123402

Bouchaud, J.-P. and Mézard, M.: Wealth condensation in a simple model of economy, Physica A, 282, 536–540, 2000.

Bouma, J., et al. (2011). “Hydropedological insights when considering catchment classification.” Hydrol. Earth Syst. Sci., 15(6), 1909–1919.

Bower, D., Hannah, D. M., and McGregor, G. R. (2004). “Techniques for assessing the climatic sensitivity of river flow regimes.” Hydrol. Processes, 18(13), 2515–2543.

Braga, A. C., Alves, L. G.A., Costa, L. S., Ribeiro, A. A., De Jesus, M. M.A., Tateishi, A. A., and Ribeiro, H. V. (2016). “Characterization of river flow fluctuations via horizontal visibility graphs.” Physica A: Statistical Mechanics and its Applications., 444, 1003-1011. https://doi.org/10.1016/j.physa.2015.10.102

Brutsaert, W. (2008). Hydrology: an introduction. 3rd ed. Hydrology: An Introduction. http://doi.org/10.2277/0521824796

Budyko, M. I. (1974). “Climate and life.” Academic Press, New York.

Burn, D. H., and Boorman, D. B. (1992). “Catchment classification applied to the estimation of hydrological parameters at ungauged catchments.” Institute of Hydrology Rep. No. 118, Wallingford, U.K.

Carillo, G., Troch, P. A., Sivapalan, M., Wagener, T., Harman, C., and Sawicz, K. (2011). “Catchment classification: Hydrological analysis of catchment behavior through process-based modeling along a climate gradient.” Hydrol. Earth Syst. Sci., 15(11), 3411–3430.

Carr, J., D’Odorico, P., Laio, F., and Ridolfi, L. (2012) On the temporal variability of the virtual water network. Geophys., Res., Lett., 39:L06404, doi: 10,1029/2012GL051247.

Casper,M. C., et al. (2012). “Analysis of projected hydrologic behavior of catchments based on signature indices.” Hydrol. Earth Syst. Sci., 16(2), 409–421.

Castellarin, A., Claps, P., Troch, P. A.,Wagener, T., andWoods, R. (2011). “Catchment classification and PUB.” Hydrol. Earth Syst. Sci.

Castiglioni, S., Castellarin, A., Montanari, A., Skøien, J. O., Laaha, G., and Blöschl, G. (2011). “Smooth regional estimation of low-flow indices: Physiographical space based interpolation and top-kriging.” Hydrol. Earth Syst. Sci., 15(3), 715–727.

Cayley, A., 1857. On the theory of the analytical forms called tress. Philos Mag, Ser IV 13(85):172-176

Chapman, T. (1989). “Classification of regions.” Comparative hydrology: An ecological approach to land and water resources, M. Falkenmark and T. Chapman, eds., UNESCO, Paris, 67–74.

Chen, M., Kuzmin, K., and Szymanski, B. K., 2014. Extension of Modularity Density for overlapping community structure. ASONAM 2014 - Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. 2, 856-863.

Chiew, F., Wang, Q. J., McConarchy, F., James, R., Wright, W., and deHoedt, G. (2002) Evapotranspiration Maps for Australia.

Clark, M.P., Kavetski, D. & Fenicia, F., 2011. Pursuing the method of multiple working hypotheses for hydrological modeling. Water Resources Research, 47(9), pp.1–16.

Clauset, A., Newman, M.E.J., Moore, C., 2004. Finding community structure in very large networks. Phys. Rev. E 70, 066111.

Coscia, M., Giannotti, F., and Pedreschi, D. (2012) A Classification for Community Discovery Methods in Complex Networks. Statistical Analysis and Data Mining journal, 4: (5): 512-546.Doi: 10.1002/sam.10133

D'Odorico, P., Carr, J., Laio, F., and Ridolfi, L. (2012) Spatial organization and drivers of the virtual water trade: A community structure analysis. Environ., Res., Lett., 7:034007, doi: 10.1088/1748-9326/7/3/034007.

Dalin, C., Konar, M., Hanasaki, N., Rinaldo, A., and Rodriguez- Iturbe, I.: Evolution of the global virtual water trade network, P. Natl. Acad. Sci. USA, 109, 5989–5994, 2012.

Dawdy, D.R., 2007. Prediction versus Understanding. Journal of Hydrologic Engineering, 12(1), pp.1–3.

Detenbeck, N. E., et al. (2000). "A test for watershed classification systems for ecological risk assessment." Environ. Toxic. Chem., 19(4), 1174–1181.

Di Prinzio, M., Castellarin, A., and Toth, E. (2011). "Data-driven catchment classification: Application to the pub problem." Hydrol. Earth Syst. Sci., 15(6), 1921–1935.

Erdös, P., and Rényi,, A., 1960. On the evolution of random graphs. Publ Math Inst Hung Acad Sci 5:17-61

Estrada, E., 2012. The Structure of Complex Networks: Theory and Applications. Oxford University Press, Oxford, UK.

Euler, L. (1741) Solutio problematis ad geometriam situs pertinentis. Comment Acad Sci Petropolitanae 8:128-140

Fang, K., Sivakumar, B. & Woldemeskel, F.M., 2017. Complex networks , community structure , and catchment classification in a large-scale river basin. Journal of Hydrology, 545, pp.478–493. Available at: http://dx.doi.org/10.1016/j.jhydrol.2016.11.056.

Fortunato, S. & Barthelemy, M., 2007. Resolution limit in community detection. *Proc Natl Acad Sci USA*, 104(1).

Ghassemi, F., White, I., 2007. Inter-Basin Water Transfer: Case Studies From Australia, United States, Canada. Cambridge University Press, China and India.

Girvan, M. & Newman, M.E.J., 2002. Community structure in social and biological networks. Proceedings of the National Academy of Sciences of the United States of America, 99(12), pp.7821–6. Available at: http://arxiv.org/abs/cond-mat/0112110.

Goerner, S. J. (1994). Chaos and the evolving ecological universe, Gordon and Breach, New York.

Gottschalk, L., Jensen, J. L., Lundquist, D., Solantie, R., and Tollan, A. (1979). "Hydrologic regions in the Nordic countries." Hydrol. Res., 10(5), 273–286.

Grayson, R. B., and Blöschl, G. (2000). Spatial patterns in catchment hydrology: Observations and modeling, Cambridge University Press, Cambridge, U.K.

Grigg, D. (1965). The logic of regional systems. Annals of the Association of American

Gupta, V. K., Rodriguez-Iturbe, I., and Wood, E. F., 1986. : Scale Problems in Hydrology: Runoff Generation and Basin Response,Water Science and Technology Library Series, Springer, Dordrecht, the Netherlands.

Gupta, V.K. (2004). "Emergence of statistical scaling in floods on channel networks from complex runoff dynamics." Chaos, Solitons and Fractals, 19(2), pp.357–365.

Haines, A. T., Finlayson, B. L., and McMahon, T. A. (1988). "A global classification of river regimes." Appl. Geogr., 8(4), 255–272.

Hall, M. J., and Minns, A.W. (1999). "The classification of hydrologically homogeneous regions." Hydrol. Sci. J., 44(5), 693–704.

Halverson, M.J. & Fleming, S.W., 2015. Complex network theory, streamflow, and hydrometric monitoring system design. Hydrology and Earth System Sciences, 19(7), pp.3301–3318.

Han, X., Sivakumar, B., Woldemeskel, F. M., and Guerra de Aguilar, M. (2018). Temporal dynamics of streamflow: application of complex networks. Geoscience Letter. 5:10. doi: 10.1186/s40562-018-0109-8

Harris, N. M., Gurnell, A. M., Hannah, D. M., and Petts, G. E. (2000). "Classification of river regimes: A context for hydroecology." Hydrol. Processes, 14(16–17), 2831–2848.

Hauhs, M. & Lange, H., 2008. Classification of Runoff in Headwater Catchments: A Physical Problem? Geography Compass, 2(1), pp.235–254. Available at: http://doi.wiley.com/10.1111/j.1749-8198.2007.00075.x.

Hrachowitz, M., Soulsby, C., Tetzlaff, D., Dawson, J. J. C., and Malcolm, I. A. (2009). "Regionalization of transit time estimates in montane catchments by integrating landscape controls." Water Resour. Res., 45(5), W05421.

Isik, S., & Singh, V. P. (2008). Hydrologic Regionalization of Watersheds in Turkey. Journal of Hydrologic Engineering. http://doi.org/10.1061/(ASCE)1084-0699(2008)13:9(824)

Jakeman, A. J., and Hornberger, G. M. (1993). "How much complexity is warranted in a rainfall-runoff model?" Water Resour. Res., 29(8), 2637–2649.

Jha, S.K. et al., 2015. Network theory and spatial rainfall connections: An interpretation. Journal of Hydrology, 527, pp.13–19. Available at: http://dx.doi.org/10.1016/j.jhydrol.2015.04.035.

Kahya, E. and Dracup, J. A.: U.S. streamflow patterns in relation to the El Niño/Southern Oscillation, Water Resour. Res., 29, 2491–2503, 1993.

Kampichler, C., Wieland, R., Calmé, S., Weissenberger, H., and Arriaga-Weiss, A. (2010). "Classification in conservation biology: A comparison of five machine-learning methods." Ecol. Inform., 5(6), 441–450.

Kantz, H., and Schreiber, T. (1997). Nonlinear time series analysis, Cambridge University Press, Cambridge, U.K.

Kennard, M. J., Mackay, S. J., Pusey, B. J., Olden, J. D., and Marsh, N. (2010a). "Quantifying uncertainty in estimation of hydrologic metrics for ecohydrological studies." River Res. Appl., 26(2), 137–156.

Kennard, M. J., Pusey, B. J., Olden, J. D., Mackay, S. J., Stein, J. L., and Marsh, N. (2010b). "Classification of natural flow regimes in Australia to support environmental flow management." Freshwater Biol., 55(1), 171–193.

Kiang, J. E., Stewart, D. W., Archfield, S. A., Osborne, E. B., and Eng., K.: A national streamflow network gap analysis, US Ge- ological Survey Scientific Investigations Report 2013-5013, Re- ston, Virginia, USA, 2013.

Kim, U., and Kaluarachchi, J. J. (2008). "Application of parameter estimation and regionalization methodologies to ungauged basins of the Upper Blue Nile River Basin, Ethiopia." J. Hydrol., 362(1–2), 39–56.

Konar, M., Dalin, C., Suweis, S., Hanasaki, N., Rinaldo, A., and Rodriguez-Iturbe, I. (2011) Water for food: the global virtual water trade network. Water Resources Res. 47:W05520. Doi: 10.1029/2010WR010307

Krasovskaia, I. (1995). "Quantification of the stability of river flow regimes." Hydrol. Sci. J., 40(5), 587–598.

Krasovskaia, I. (1997). "Entropy-based grouping of river flow regimes." J. Hydrol., 202(1–4), 173–191.

Krasovskaia, I., and Gottschalk, L. (2002). "River flow regimes in a changing climate." Hydrol. Sci. J., 47(4), 597–609.

Krasovskaia, I., Gottschalk, L., and Kundzewicz, Z. W. (1999). "Dimensionality of Scandinavian river flow regimes." Hydrol. Sci. J., 44(5), 705–723.

L'vovich, M. I. (1979). World water resources and their future, LithoCrafters, Chelsea, U.K.

Lavery, B.M., Joung, G., Nicholls, N., 1997. An extended high-quality historical rainfall dataset for Australia. Aust. Meteorol. Mag. 46, 27e38.

Lavery, B.M., Kariko, A.P., Nicholls, N., 1992. A historical rainfall data set for Australia. Aust. Meteorol. Mag. 40, 33e39.

Ley, R., Casper, M. C., Hellebrand, H., and Merz, R. (2011). "Catchment classification by runoff behaviour with self-organizing maps (SOM)." Hydrol. Earth Syst. Sci., 15(9), 2947–2962.

Li, C., Singh, V. P., and Mishra, A. K. 2012. Entropy theory-based criterion for hydrometric network evaluation and design: maximum information minimum redundancy, Water Resour. Res., 48, W05521, doi:10.1029/2011WR011251.

Li, T., Wang, G., Chen, J., 2010. A modified binary tree codification of drainage networks to support complex hydrological models. Comput. Geosci. 36 (11), 1427–1435.

Li, Z., Zhang, S., Wang, R. S., Zhang, X. S., & Chen, L. (2008). Quantitative function for community detection. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, *77*(3), 1–9.

Liljeros, F., Edling, C., Amaral, L. N., Stanley, H. E., and Åberg,Y. (2001). "The web of human sexual contacts." Nature, 411(6840), 907–908.

Lins, H. F. (1997). "Regional streamflow regimes and hydroclimatology of the United States." Water Resour. Res., 33(7), 1655–1667.

Listing, J. B., 1848. Vorstudien zur Topologie. Vandenhoeck und Ruprecht, Göttingen, 811-875

Lorenz, E. N., 1963. Deterministic periodic flow. J. Atmos Sci 20(2):130-141

Loveland, T. R., and Merchant, J. M. (2004). "Ecoregions and ecoregion- alization: Geographical and ecological perspectives." Environ. Manage., 34(Suppl. 1), S1–S13.

Malik, N., Bookhagen, B., Marwan, N., and Kurths, J. (2012), Analysis of spatial and temporal extreme monsoonal rainfall over South Asia using complex networks. Clim Dyn 39: 971-987.

Mandelbrot, B. B., 1983. The fractal geometry of nature. W. H. Freeman and Company, New York.

Mcdonnell, J.J. & Beven, K., 2014. Debates—The future of hydrological sciences: A (common) path forward? A call to action aimed at understanding velocities, celerities and residence time distributions of the headwater hydrograph. Water Resources Research. , pp.5342–5350.

McDonnell, J.J. & Woods, R., 2004. On the need for catchment classification. Journal of Hydrology, 299(1–2), pp.2–3.

McMahon, T. A., and Finlayson, B. L. (1992). Global runoff: Continental comparisons of annual flows and peak discharges, Catena Verlag, Cremlingen-Destedt, Germany.

McMahon, T. A., Peel, M. C., Vogel, R. M., and Pegram, G. G. S. (2007). "Global streamflows—Part 3: Country and climate zone characteris- tics." J. Hydrol., 347(3–4), 272–291.

Merz, B., and Blöschl, G. (2004). "Regionalization of catchment model parameters." J. Hydrol., 287(1–4), 95–123.

Milly, P. C. D., Betancourt, J., Falkenmark, M., Hirsch, R. M., Kundzewicz, Z. W., Lettenmaier, D. P., & Stouffer, R. J. (2008). Climate change: Stationarity is dead: Whither water management? Science. http://doi.org/10.1126/science.1151915

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U.: Network motifs: simple building blocks of com- plex networks, Science, 298, 824–827, 2002.

Mishra, A.K., Özger, M., Singh, V.P., 2009. An entropy-based investigation into the variability of precipitation. J. Hydrol. 370 (1–4), 139–154.

Moliere, D. R., Lowry, J. B. C., and Humphrey, C. L. (2009). "Classifying the flow regime of data-limited streams in the wet-dry tropical region of Australia." J. Hydrol., 367(1–2), 1–13.

Morton FI (1983) Operational estimates of areal evapotranspiration and their significance to the science and practice of hydrology. Journal of Hydrology, 66, 1-76

Nathan, R. J., and McMahon, T. A. (1990). "Identification of homogeneous regions for the purpose of regionalization." J. Hydrol., 121(1–4), 217–238.

Newman, M. E. J., 2010. Networks: An Introduction. Oxford University Press., 18(2), pp.241-242.

Newman, M. E. J., Barabási, A.-L., and Watts, D. J., 2003. The structure and dynamics of networks, Princeton University Press, Princeton, NJ.

Newman, M., Girvan, M., 2004. Finding and evaluating community structure in networks. Phys. Rev. E 69, 026113.

Newman, M.E.J., 2004. Detecting community structure in networks. *The European Physical Journal B - Condensed Matter and Complex Systems*, 38(2), pp.321–330.

Newman, M.E.J., 2006. Finding community structure using the eigenvectors of matrices. Phys. Rev. E 74, 036104.

Nguyen, T. T., Kawamura, A., Tong, T. N., Nakagawa, N., Amaguchi, H., & Gilbuena, R. (2015). Clustering spatio-seasonal hydrogeochemical data using self-organizing

maps for groundwater quality assessment in the Red River Delta, Vietnam. Journal of Hydrology, 522, 661–673. http://doi.org/10.1016/j.jhydrol.2015.01.023

Olden, J. D., Kennard, M. J., and Pusey, B. J. (2012). "A framework for hydrologic classification with a review of methodologies and applications in ecohydrology." Ecohydrology, 5(4), 503–518.

Olden, J.D. & Poff, N.L., 2003. Redundancy and the choice of hydrologic indices for characterizing streamflow regimes. River Research and Applications, 19(2), pp.101–121.

Oudin, L., Andréassian, V., Perrin, C., Michel, C., and Le Moine, N. (2008). "Spatial proximity, physical similarity, regression and ungaged catchments: A comparison of regionalization approaches based on 913 French catchments." Water Resour. Res., 44(3), W03413.

Pardé, M. (1933). Fleuves et rivières, Collection Armond Colin, Paris.

Patil, S. & Stieglitz, M., 2011. Hydrologic similarity among catchments under variable flow conditions. Hydrology and Earth System Sciences, 15(3), pp.989–997.

Patil, S. and Stieglitz, M.: Controls on hydrologic similarity: role of nearby gauged catchments for prediction at an ungauged catchment, Hydrol. Earth Syst. Sci., 16, 551–562, doi:10.5194/hess- 16-551-2012, 2012.

Perrin, C., Michel, C., & Andréassian, V. (2001). "Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments." Journal of Hydrology, 242(3–4), 275–301. http://doi.org/10.1016/S0022-1694(00)00393-0.

Phillips, J. D. (1999). Earth surface systems, complexity, order, and scale, Basil Blackwell, Oxford, U.K.

Poff, N. L., Olden, J. D., Pepin, D. M., and Bledsoe, B. P. (2006). "Placing global stream flow variability in geographic and geomorphic contexts." River Res. Appl., 22(2), 149–166.

Pons, P., Latapy, M., 2005. Computing communities in large networks using random walks. Lect. Notes Comput. Sci. 3733, 284–293.

Raghavan, U.N., Albert, R., Kumara, S., 2007. Near linear time algorithm to detect community structures in large-scale networks. Phys. Rev. E 76, 036106.

Rao, A. R., and Srinivas, V. V. (2006a). "Regionalization of watersheds by fuzzy cluster analysis." J. Hydrol., 318(1–4), 57–79.

Regonda, S., Sivakumar, B., Jain, A., 2004. Temporal scaling in river flow: can it be chaotic? Hydrol. Sci. J. 49 (3), 373–385.

Reichl, J. P. C., Western, A. W., McIntyre, N. R., and Chiew, F. H. S. (2009). "Optimization of a similarity measure for estimating ungauged streamflow." Water Resour. Res., 45(10), W10423.

Rinaldo, A., Banavar, J. R., and Maritan, A.: Trees, net- works, and hydrology, Water Resour. Res., 42, W06D07, doi:10.1029/2005WR004108, 2006.

Rosgen, D. L. (1994). “A classification of natural rivers.” Catena, 22(3), 169–199.

Rosvall, M. and Bergstrom, C. T., 2007. An information-theoretic framework for resolving community structure in complex networks. Proc. Natl. Acad. Sci. U. S. A. 104, 7327-7331.

Salas, J. D., Delleur, J.W., Yevjevich, V., and Lane,W. L. 1995. Applied Modeling of Hydrologic Time Series,Water Resources Publications, Littleton, Colorado, USA.

Salas, J.D., Kim, H.S., Eykholt, R., Burlando, P., Green, T.R., 2005. Aggregation and sampling in deterministic chaos: implications for chaos identification in hydrological processes. Nonlinear Processes Geophys. 12, 557–567.

Sauquet, E., and Catalogne, C. (2011). “Comparison of catchment grouping methods for flow duration curve estimation at ungauged sites in France.” Hydrol. Earth Syst. Sci., 15(8), 2421–2435.

Sawicz, K. et al., 2011. Catchment classification: Empirical analysis of hydrologic similarity based on catchment function in the eastern USA. Hydrology and Earth System Sciences, 15(9), pp.2895–2911.

Scarsoglio, S., Laio, F., and Ridolfi, L.: Climate dynamics: A network-based approach for the analysis of global precipitation, PLoS ONE, 8, e71129, doi:10.1371/journal.pone.0071129, 2013.

Schröder, B. (2006). “Pattern, process, and function in landscape ecology and catchment hydrology—How can quantitative landscape ecology support predictions in ungauged basins?” Hydrol. Earth Syst. Sci., 10(6), 967–979.

Seibert, J., and Beven, K. J. (2009). “Gauging the ungauged basin: How many discharge measurements are needed?” Hydrol. Earth Syst. Sci., 13(6), 883–892.

Serinaldi, F. and Kilsby, C. G. (2016). “Irreversibility and complex network behavior of stream flow fluctuations.” Physica A., 450, 585–600. https://doi.org/10.1016/j.physa.2016.01.043

Shang, R., Zhang, W., Jiao, L., Stolkin, R., and Xue, Y., 2017. A community integration strategy based on an improved modularity density increment for large-scale networks. Physica A. 469, 471-485. http://dx.doi.org/10.1016/j.physa.2016.11.066

Sims, N. C., Chariton, A. A., Jin, H., and Colloff, M. J. (2012). “A clas- sification of floodplains and wetlands of the Murray-Darling basin based on changes in flows following water resource development.” Wetlands, 32(2), 239–248.

Sivakumar, B. & Woldemeskel, F.M., 2014. Complex networks for streamflow dynamics. Hydrology and Earth System Sciences Discussions, 11(7), pp.7255–7289. Available at: http://www.hydrol-earth-syst-sci-discuss.net/11/7255/2014/.

Sivakumar, B. & Woldemeskel, F.M., 2015. A network-based analysis of spatial rainfall connections. Environmental Modelling and Software, 69, pp.55–62. Available at: http://dx.doi.org/10.1016/j.envsoft.2015.02.020.

Sivakumar, B. (2000). "Chaos theory in hydrology: Important issues and interpretations." J. Hydrol., 227(1–4), 1–20.

Sivakumar, B. (2003). "Forecasting monthly streamflow dynamics in the western United States: A nonlinear dynamical approach." Environ. Modell. Software, 18(8–9), 721–728.

Sivakumar, B. (2004). "Dominant processes concept in hydrology: Moving forward." Hydrol. Processes, 18(12), 2349–2353.

Sivakumar, B. (2008a). "Dominant processes concept, model simplification and classification framework in catchment hydrology." Stochastic Environ. Res. Risk Assess., 22(6), 737–748.

Sivakumar, B. (2008b). "The more things change, the more they stay the same: The state of hydrologic modeling." Hydrol. Processes, 22(21), 4333–4337.

Sivakumar, B. (2017). Chaos in hydrology: Bridging determinism and stochasticity.

Sivakumar, B. and Berndtsson, R. (2010). Advances in data-based approaches for hydrologic modelling and forecasting. Doi: https://doi.org/10.1142/7783

Sivakumar, B. and Singh, V. P.: Hydrologic system complexity and nonlinear dynamic concepts for a catchment classification framework, Hydrol. Earth Syst. Sci., 16, 4119–4131, doi:10.5194/hess-16-4119-2012, 2012.

Sivakumar, B., 2015. Networks: a generic theory for hydrology? Stoch. Environ. Res. Risk Assess. 29, 761–771.

Sivakumar, B., and Singh, V. P. (2012). "Hydrologic system complexity and nonlinear dynamic concepts for a catchment classification framework." Hydrol. Earth Syst. Sci., 16(11), 4119–4131.

Sivakumar, B., Jayawardena, A.W., Li, W.K., 2007. Hydrologic complexity and classification: a simple data reconstruction approach. Hydrol. Process. 21 (20), 2713–2728.

Sivakumar, B., Singh, V. P., Berndtsson, R., & Khan, S. K. (2015). "Catchment Classification Framework in Hydrology: Challenges and Directions." Journal of Hydrologic Engineering, (2), 130426211354007. http://doi.org/10.1061/(ASCE)HE.1943-5584.0000837

Sivakumar, B., Woldemeskel, F.M., 2015. A network-based analysis of spatial rainfall connections. Environ. Modell. Softw. 69, 55–62.

Sivakumar, B.: Forecasting monthly streamflow dynamics in the western United States: a nonlinear dynamical approach, Environ. Modell. Softw., 18, 721–728, 2003.

Sivapalan, M. (2005). Pattern, Process and Function: Elements of a Unified Theory of Hydrology at the Catchment Scale. In Encyclopedia of Hydrological Sciences (pp. 193–219). http://doi.org/10.1002/0470848944

Sivapalan, M., et al. (2003). "IAHS decade on predictions in ungauged basins (PUB), 2003–2012: Shaping an exciting future for the hydrological sciences." Hydrol. Sci. J., 48(6), 857–880.

Slack, J. R. and Landwehr, V. M.: Hydro-climatic data network (HCDN): a US Geological Survey streamflow data set for the United States for the study of climate variations, 1847–1988, US Geological Survey Open File Report 92-129, US Geological Survey, Reston, Virginia, USA, 1992.

Snelder, T. H., and Biggs, B. J. F. (2002). "Multi-scale river environment classification for water resources management." J. Am. Water Resour. Assoc., 38(5), 1225–1239.

Snelder, T. H., and Hughey, K. F. D. (2005). "On the use of an ecologic classification to improve water resource planning in New Zealand." Environ. Manage., 36(5), 741–756.

Snelder, T. H., Cattanéo, F., Suren, A. M., and Biggs, B. J. F. (2004). "Is the river environment classification an improved landscape-scale classifica- tion of rivers?" J. N. Am. Benthol. Soc., 23(3), 580–598.

Snelder, T.H., Biggs, B.J.F. & Woods, R.A., 2005. Improved eco-hydrological classification of rivers. River Research and Applications, 21(6), pp.609–628.

Song, X. et al., 2015. Global sensitivity analysis in hydrological modeling: Review of concepts, methods, theoretical framework, and applications. Journal of Hydrology, 523(225), pp.739–757. Available at: http://dx.doi.org/10.1016/j.jhydrol.2015.02.013.

Strogatz, S. H. (1994). Nonlinear dynamics and chaos: With applications to physics, biology, chemistry, and engineering, Addison-Wesley, Reading, MA.

Sugawara, M. (1961). "On the analysis of runoff structure about several Japanese rivers." Jap. J. Geophys., 2(4), 1–76.

Suweis, S., Konar, M., Dalin, C., Hanasaki, N., Rinaldo, A., and Rodriguez-Iturbe, I. (2011). "Structure and controls of the global virtual water trade network." Geophys. Res. Lett., 38(10), L10403.

Tamea, S., Allamano, P., Carr, J, Claps, P., Laio, F., and Ridolfi, L. (2013) Local and global perspectives on the virtual water trade., Hydrol. Earth. Syst. Sci. 17:1205-1215

Tasker, G. D. (1982). "Comparing methods of hydrologic regionalization." Water Resour. Bull., 18(6), 965–970.

Tongal, H. & Sivakumar, B., 2017. Cross-entropy clustering framework for catchment classification. Journal of Hydrology, 552, pp.433–446. Available at: http://dx.doi.org/10.1016/j.jhydrol.2017.07.005.

Tongal, H., Demirel, M.C., Booij, M.J., 2013. Seasonality of low flows and dominant processes in the Rhine River. Stoch. Environ. Res. Risk Assess. 27, 489–503.

Tootle, G. A. and Piechota, T. C.: Relationships between Pacific and Atlantic ocean sea surface temperatures and U.S. streamflow variability, Water Resour. Res., 42, W07411, doi:10.1029/2005WR004184, 2006.

Tsonis, A. A. (1992). Chaos: From theory to applications, Plenum Press, New York.

Tsonis, A. A., and Roebber, P. J. (2004). "The architecture of climate networks." Physica A, 333, 497–504.

Vignesh, R., Jothiprakash, V. & Sivakumar, B., 2015. "Streamflow variability and classification using false nearest neighbor method." Journal of Hydrology., 531.

Vogel, R. M. and Sankarasubramanian, A.: Spatial scaling proper- ties of annual streamflow in the United States, Hydrolog. Sci. J., 45, 465–476, 2000.

Vormoor, K., Skaugen, T., Langsholt, E., Diekkrüger, B., and Skøien, J. O. (2011). "Geostatistical regionalization of daily runoff forecasts in Norway." Int. J. River Basin Manage., 9(1), 3–15.

Wagener, T., and McIntyre, N. (2012). "Hydrological catchment classifi- cation using a data-based mechanistic strategy." System identification, environmental modelling, and control system design, L. Wang and H. Garnier, eds., Springer, London, 483–500.

Wagener, T., Sivapalan, M., and McGlynn, B. (2008). "Catchment classification and services—Toward a new paradigm for catchment hydrology driven by societal needs." Encyclopedia of hydrological sciences, M. G. Anderson, ed., Wiley, Chichester, U.K., 1–12.

Wagener, T., Sivapalan, M., Troch, P., & Woods, R. (2007). Catchment Classification and Hydrologic Similarity. Geography Compass, 1, 1–31. http://doi.org/10.1111/j.1749-8198.2007.00039.x

Wardrop, D. H., et al. (2005). "Use of landscape and land use parameters for classification of watersheds in the mid-Atlantic across five physio-graphic provinces." Environ. Ecol. Stat., 12(2), 209–223.

Watts, D. J. (1999). Small worlds: The dynamics of networks between order and randomness, Princeton University Press, Princeton, NJ.

Watts, D. J., and Strogatz, S. H. (1998). "Collective dynamics of 'small-world' networks." Nature, 393(6684), 440–442.

Winsemius, H.C. et al., 2009. On the calibration of hydrological models in ungauged basins: A framework for integrating hard and soft hydrological information. Water Resources Research, 45(1).

Woods, R. A. (2002). "Seeing catchments with new eyes." Hydrol. Process., 16(5), 1111–1113.

Yang, D., Li, C., Hu, H., Lei, Z., Yang, S., Kusuda, T., Koike, T., and Musiake, K. 2004. Analysis of water resources variability in the Yellow River of China during the

last half century using historical data, Water Resour. Res., 40, W06502, doi:10.1029/2003WR002763.

Yasmin, N., and Sivakumar, B. (2018). Temporal streamflow analysis: Coupling nonlinear dynamics with complex networks. Journal of Hydrology 564: 59–67. Doi: 10.1016/j.jhydrol.2018.06.072

Young, P. C., and Ratto, M. (2009). "A unified approach to environmental systems modeling." Stochastic Environ. Res. Risk Assess., 23(7), 1037–1057.

Zachary, W.W., 1977. An Information Flow Model for Conflict and Fission in Small Groups. *Journal of Anthropological Research*, 33(4), pp.452–473.

Zaliapin, I., Foufoula-Georgiou, F., and Ghil, M. (2010) transport on river networks: A dynamic tree approach. J., Geophys Res., 115: F00A15, doi: 10.1029/2009JF001281.

Zhang, S. H., Ning, X. M., and Ding, C., 2009. Maximizing Modularity Density for Exploring Modular Organization of Protein Interaction Networks. Optimization and Systems Biology Journal. 11, 361-370.

Zhang, X.S. et al., 2016. How streamflow has changed across Australia since the 1950 s: evidence from the network of hydrologic reference stations. Hydrol. Earth Syst. Sci. 20 (9), 3947.

Zhang, Y., Xia, J., Bunn, S. E., Arthington, A. H., Mackay, S., and Kennard, M. (2012). "Classification of flow regimes for environmental flow assessment in regulated rivers: The Huai River Basin, China." River Res. Appl., 28(7), 989–1005.

Zhang,Y., and Chiew, F. H. S. (2009). "Relative merits of different methods for predictions in ungauged catchments." Water Resour. Res., 45(7), W07412.

APPENDICES

Appendix A

A.1 Modularity (Q value) calculation

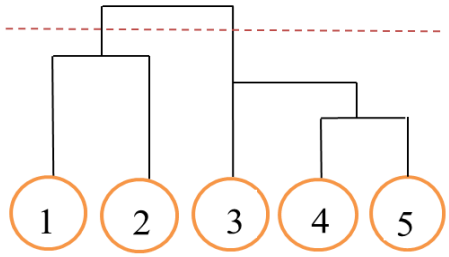
For classification, the connection link between two nodes is identify from the adjacent matrix, where 1 indicated that two nodes is connected, otherwise as 0 as shown in Table 3.A.1. Then, the betweenness of edges are calculated based on the steps (1) to (4) until dendrogram is formed. As mentioned, the communities in the network is then determined by the horizontal level on the dendrogram (as in Figure 3.4(d)) and the membership will keep changing following the position of the horizontal line at each level until each node belongs to each (different) group. By each level of the horizontal line, the Q value is calculated and the level of the horizontal line on dendrogram with the maximum Q value will represent the best split of the network.

To calculate Q value, from Equation (1), for instance, at first level of the horizontal line forms two communities of 5 nodes as [1 1 2 2 2]. Each node to every other node in the network need to be considered and the $\delta(c_i, c_j)$ is identify, for example, if group of node 1 (c_1) is not equal to group of node 3 (c_3), then $\delta(c_1, c_3)$ is denoted as 0, so that could save time of calculation i.e. only consider the nodes that are from the same group and ignore the rest.

$$Q = \frac{1}{2m}\sum_{i,j}(A_{ij} - \frac{k_i k_j}{2m})\delta(c_i, c_j)$$

Table 3.A.1: The adjacent matrix of a simple network

A_{ij}	1	2	3	4	5
1	0	1	0	1	0
2	1	0	0	1	0
3	0	0	0	1	1
4	1	1	1	0	1
5	0	0	1	1	0



Calculation for the **first** horizontal cut is explained below. The modularity calculated based on the network of two groups with membership [1,1,2,2,2] is as follows,

$Q = \frac{1}{2\times6}[\left(\left(0-\frac{2\times2}{2\times6}\right)\times1\right)+$ # node 1 to 1 -> absent, same membership

$\left(\left(1-\frac{2\times2}{2\times6}\right)\times1\right)+$ # node 1 to 2 -> present, same membership

$\left(\left(0-\frac{2\times2}{2\times6}\right)\times0\right)+$ # node 1 to 3 -> absent, different membership

$\left(\left(1-\frac{4\times4}{2\times6}\right)\times0\right)+$ # node 1 to 4 -> absent, different membership

$\left(\left(0-\frac{2\times2}{2\times6}\right)\times0\right)+$ # node 1 to 5 -> absent, different membership

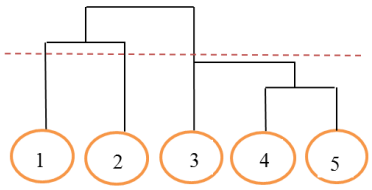
…]

Continuing for the rest of the nodes in the network which then can simplifies to,

Q = 1 / 12 (4* (0 - 2 / 6) + 4 * (1 - 2 / 6) +4 * (1 – 4 / 6) + (0 – 16 / 12)) = 1/9 = **0.11**

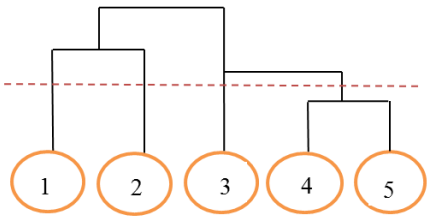
This process is repeated for the next lower level of horizontal cut. Forthe **second** horizontal cut has split the network into 3 groups with membership [1,2,3,3,3] and the modularity calculation,

Q = 1 / 12 (4* (0 – 4 / 12) + 4 * (1 - 8 / 12) +2* (1 – 4 / 12) + (0 – 16 / 12)) = **0.**



Then, the **third** horizontal cut form 4 groups with membership [1,2,3,4,4] is resulted with -0.17 as follows,

Q = 1 / 12 (4* (0 – 4 / 12) + 2 * (1 - 8 / 12) + (0 – 16 / 12)) = **-0.17.**



Clearly exhibits in this case, the value of modularity is started to decreases from the second cut of the dendrogram and therefore, the first cut of the dendrogram with value 0.11 has forms the best split of the network of two groups with membership [1 1 2 2 2] when the modularity measureis applied.

A.2 Modularity Density (D value) calculation

For D value calculation, the number of nodes in subgraphs is focused rather than number of total nodes in network as in the modularity (Q value) measurement. Similar to modularity measure, modularity density measure is started by identify the connection of each pair of nodes based on the adjacent matrix and the membership of each node in the network is also determined by the horizontal line to cut the dendrogram until each node belongs to each group. Then the best split of the network is determined by which level of the horizontal line is located on dendrogram with the maximum D value based on Equation (2).

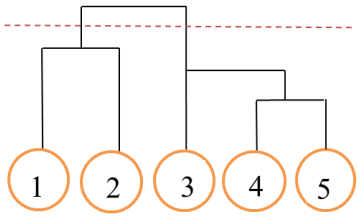
$$D = \sum_{i} \left(\frac{2l_i}{n_i} - \frac{l_i^{ext}}{n_i} \right)$$

for first level of dendrogram, the membership of 5 nodes is [1 1 2 2 2], hence subgraphs *i* are consist of 2 groups and the D value is calculated as shown below,

$D = \left(\frac{2(1)}{2} - \frac{2}{2}\right) +$ # for group 1 that is consists of links associated with nodes 1 and 2

$\left(\frac{2(3)}{3} - \frac{2}{3}\right)$ # for group 2 that is consists of links associated with nodes 3, 4 and 5

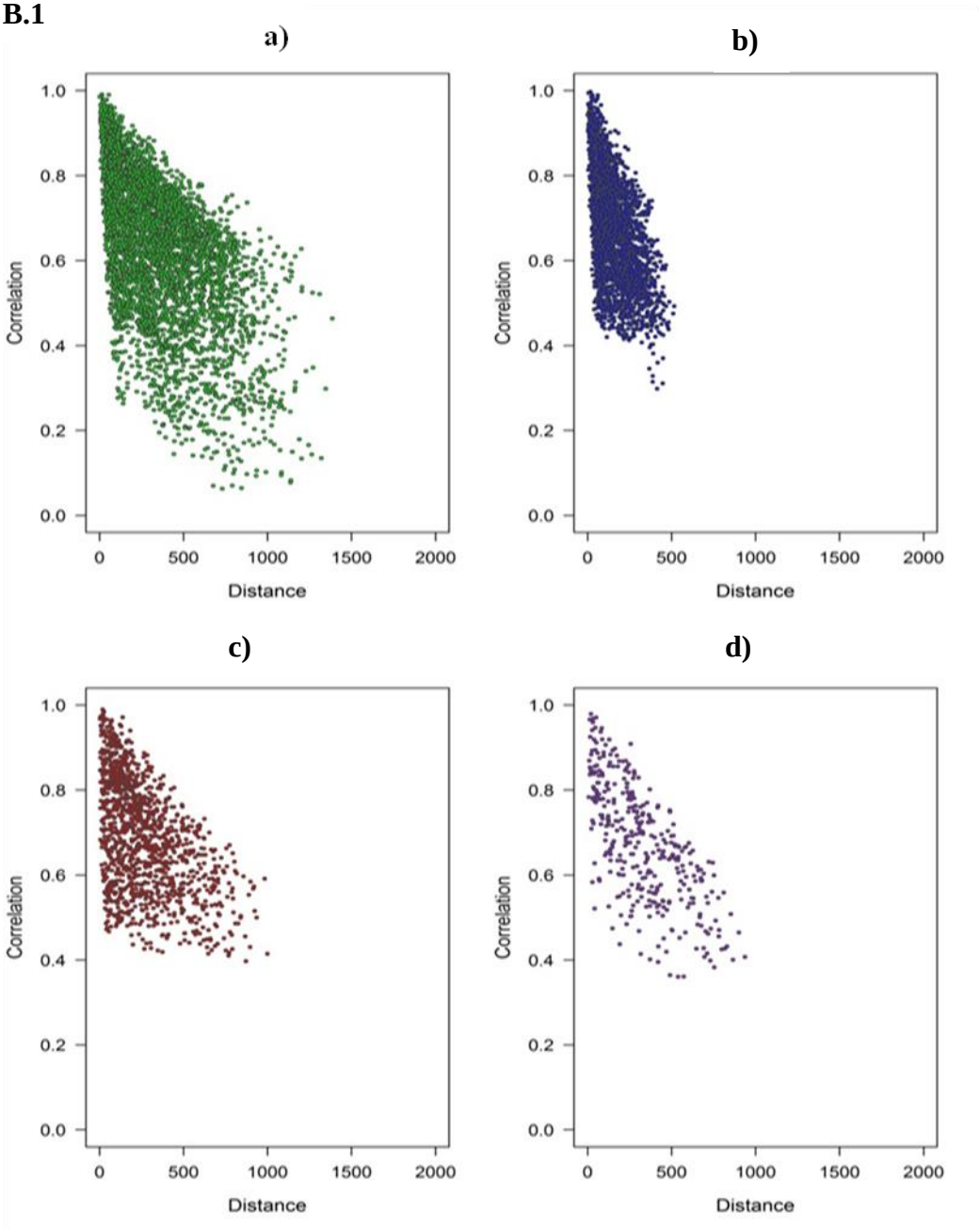
Therefore, D = 1.33 for membership of [1 1 2 2 2].



By repeating the same process for the next lower level of the horizontal line i.e., 2nd, 3rd and 4th cut of dendrogram have resulted the modularity density (D value) as -2.67, -8 and -12, respectively. Thus, in this case, the modularity density measure is also tended to split the network to 2 groups with membership [1 1 2 2 2] as the best split as well.

Appendix B

B.1



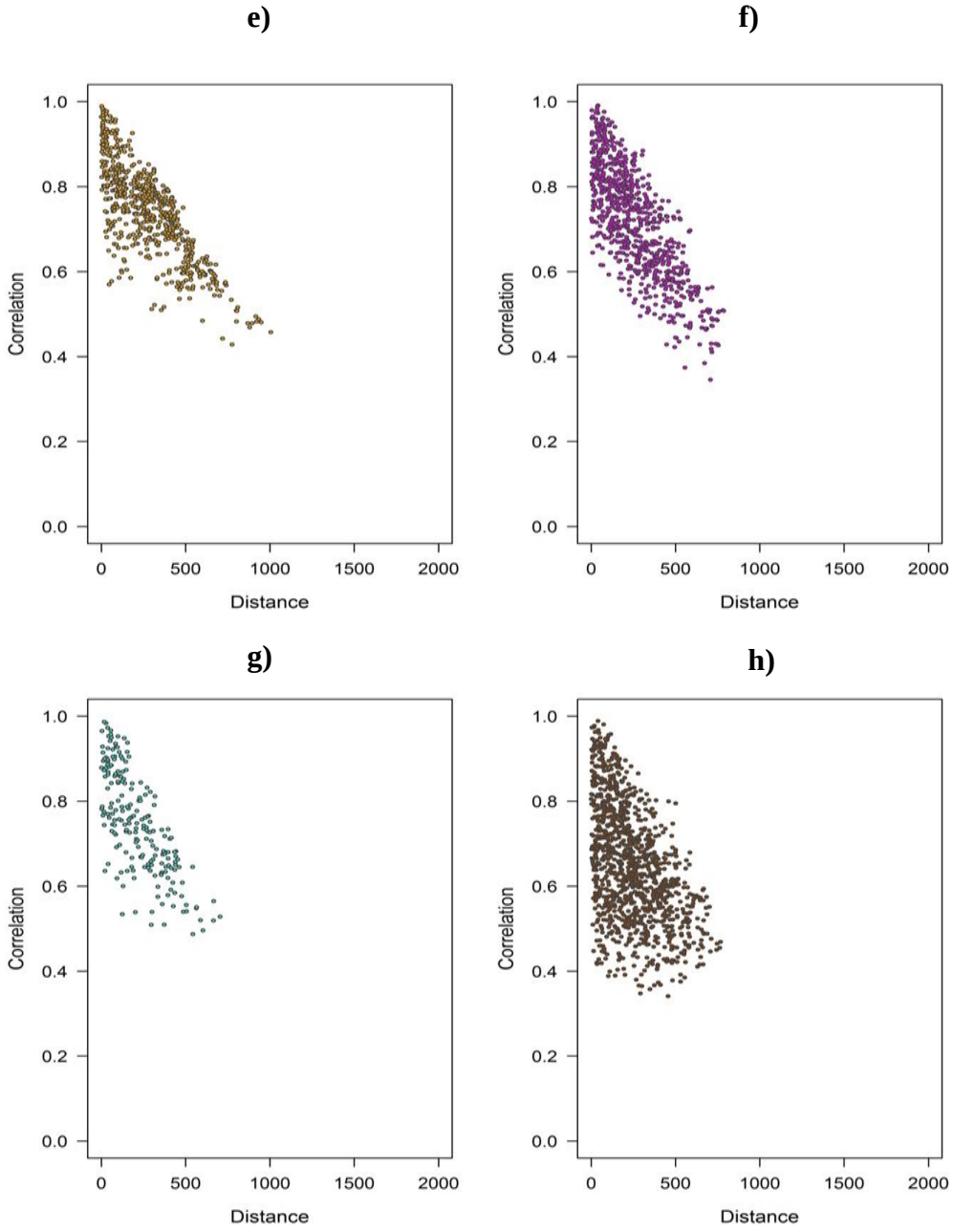
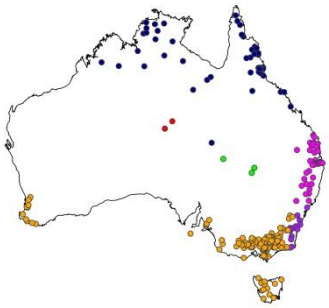


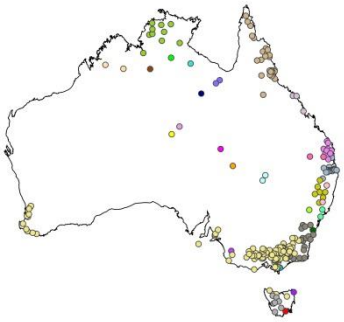
Figure B.1(a-h): Distance-correlation scatterplots for eight communities, corresponding to the coloring scheme in Figure 5.8(d).

B.2

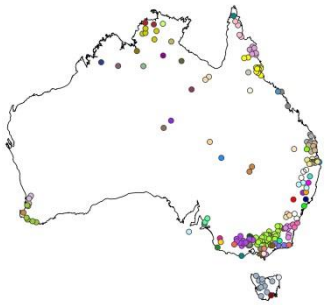
Q, R & PET at T = 0.65



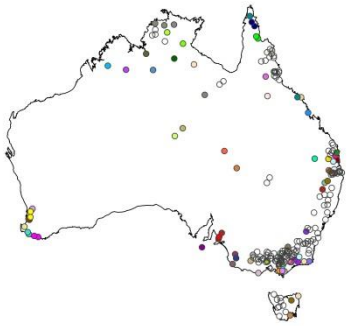
Q at T = 0.7



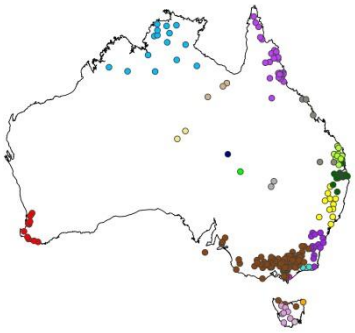
Q at T = 0.85



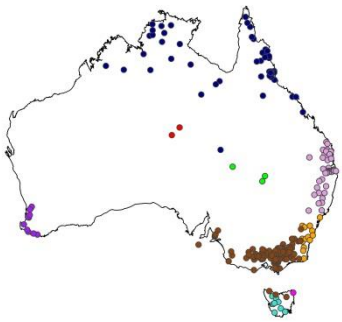
Q at T = 0.9



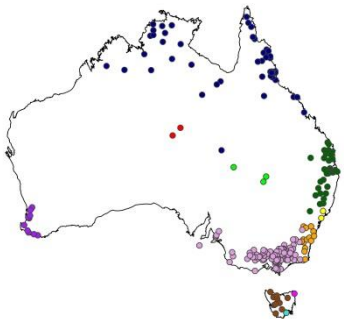
Q & R at T = 0.7



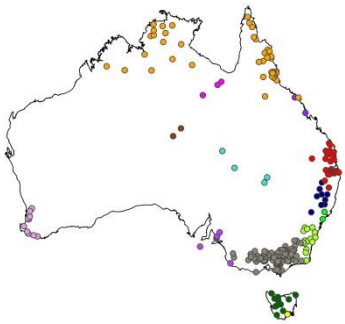
R at T = 0.65



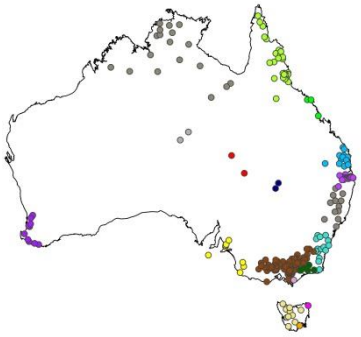
R at T = 0.7



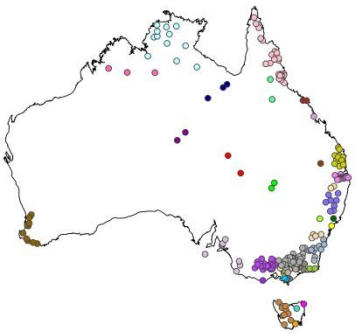
R at T = 0.75



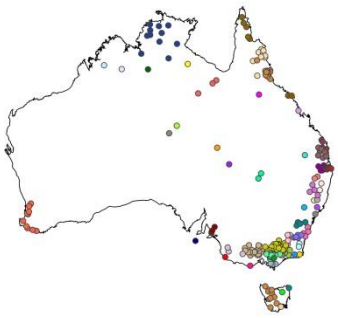
R at T = 0.8



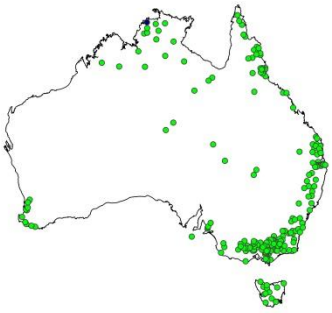
R at T = 0.85



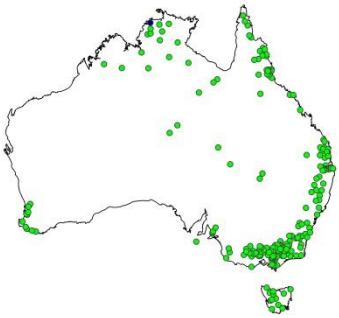
R at T = 0.9



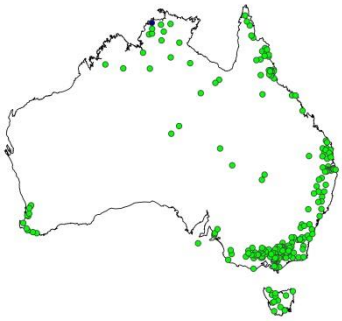
PET at T = 0.65



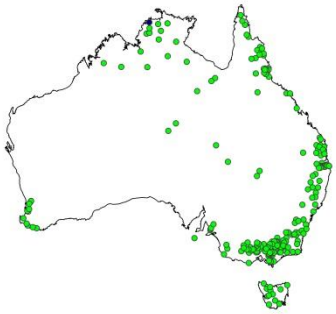
PET at T = 0.7



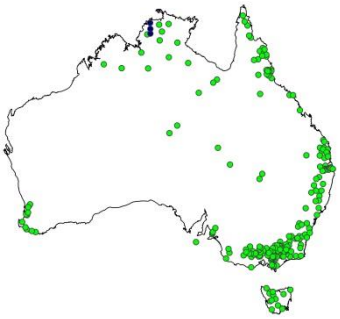
PET at T = 0.75



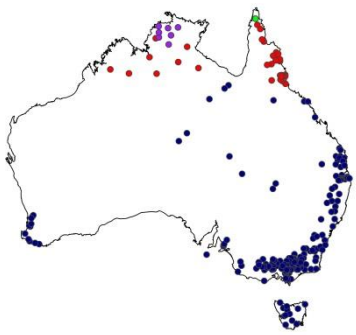
PET at T = 0.8



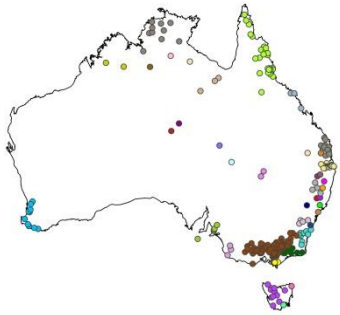
PET at T = 0.85



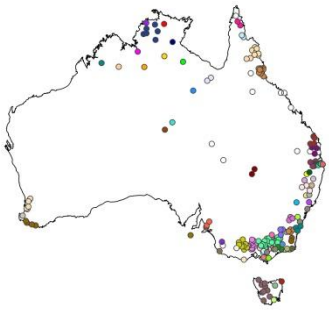
PET at T = 0.9



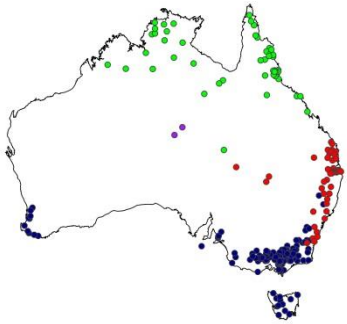
Q & R at T = 0.8



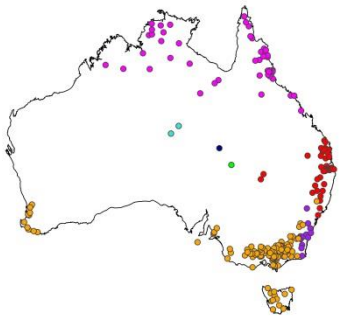
Q & R at T = 0.9



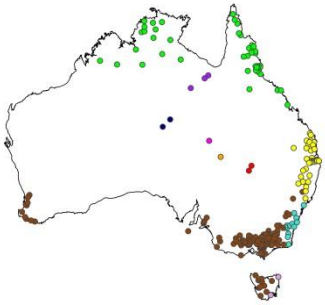
Q & PET at T = 0.65



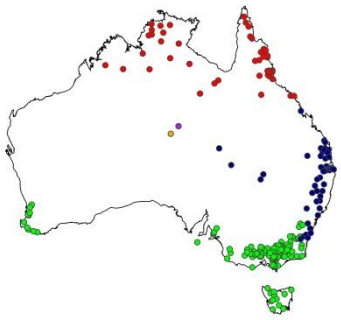
Q & PET at T = 0.7



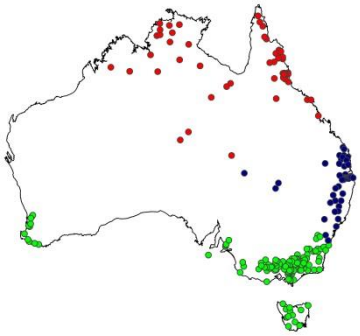
Q & PET at T = 0.75



R & PET at T = 0.65



R & PET at T = 0.7



R & PET at T = 0.75