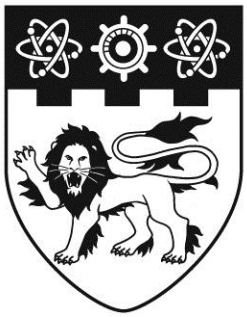


NANYANG TECHNOLOGICAL UNIVERSITY

UNCERTAINTY QUANTIFICATION FRAMEWORK FOR COMBINED STATISTICAL SPATIAL DOWNSCALING AND TEMPORAL DISAGGREGATION FOR CLIMATE CHANGE IMPACT STUDIES ON HYDROLOGY

RAJENDRAN QUEEN SURAAJINI

SCHOOL OF CIVIL AND ENVIRONMENTAL ENGINEERING

2017

UNCERTAINTY QUANTIFICATION FRAMEWORK FOR COMBINED STATISTICAL SPATIAL DOWNSCALING AND TEMPORAL DISAGGREGATION FOR CLIMATE CHANGE IMPACT STUDIES ON HYDROLOGY

RAJENDRAN QUEEN SURAAJINI

School of Civil and Environmental Engineering

A thesis submitted to the Nanyang Technological University

in partial fulfillment of the requirement for the degree of

Doctor of Philosophy

2017

ACKNOWLEDGEMENTS

I would like to acknowledge the School of Civil and Environmental Engineering (CEE) and Earth Observatory of Singapore (EOS) for the great opportunity offered to me to do research at NTU with full scholarship. I express my heartfelt gratitude to my supervisor Dr. Cheung Sai Hung, Joseph for the opportunity given to take up research under his supervision at NTU. I always felt his valuable guidance throughout my Ph.D. work as exemplary. He had facilitated me to experience constant support, confidence, freedom, motivating working environment and inspiration for independent research through various discussions. This work would not have been possible without my supervisor who pushed me beyond limits to realize my full potential in carrying out research.

I sincerely thank Assoc. Professor Dr. Qin Xiaosheng for sharing his academic support and expertise for my research. It's my immense pleasure to thank my research group members Dr. Sahil Bansal, Dr. Shao Zhe, Ms. Yidan Gao and Mr. Fanming Gong for the professional comfort and logistics support extended to me. My thanks are due to Dr. Lu Yan and Dr. Pradeep Mandapaka of my school for their moral support.

I am blessed to have my beloved parents Advocate M. Rajendran and Dr. S. Suganthi who have let me go behind my dreams and have always been my inspiration. I deeply thank them for their unconditional love, blessing and patience for raising me to the level that I have always dreamt of. Special thanks to my sibling R. Prince Krishna for being best brother ever who is very generous of his love and technical support at any time.

I am indebted to my Grandmother Muthulakshmi Sellakkutti who deserves my appreciation for her love towards me and stimulation on my higher studies. I am thankful to Professor Dr. S. Raghavan of National Institute of Technology, Trichy for his encouragement, motivation and his unconditional love towards professional growth.

I record special appreciations to my best friends Sharanya, Petchiyappan, Jayakrishnan, Padmanabhan who had been my pillars of support through all the hard times and cheered me up even for all the little accomplishments despite staying miles apart. I also thank my other friends at NTU for their motivation and happy memories to cherish.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS iii

TABLE OF CONTENTS v

ABSTRACT xi

LIST OF PUBLICATIONS xiv

LIST OF TABLES xvi

LIST OF FIGURES xix

LIST OF SYMBOLS xxv

LIST OF ABBREVIATIONS xxviii

CHAPTER 1 INTRODUCTION 1

1.1 Bayesian inference framework 8

1.2 Objectives of the thesis 10

1.2.1 Single site statistical downscaling model with coupled uncertainty quantification 10

1.2.2 Integrated multi-site statistical downscaling and disaggregation models with uncertainty quantification tool 11

1.2.3 Hydrologic impact studies using integrated downscaling model, disaggregation model and data driven hydrologic models 12

1.3 Outline of the thesis 13

CHAPTER 2 KNN-BUQSDM - A Bayesian Updating Uncertainty Quantification Framework for Statistical Downscaling of Precipitation 15

2.1 Abstract 15

2.2 Introduction 16

2.2.1 Review of regression based SDM 18

2.2.2 Sources of uncertainty in SDM 24

2.2.3 Uncertainty Quantification using GPR 27

2.3 Data and Study area 32

2.4 K-Nearest Neighbour – Bayesian Uncertainty Quantification Statistical Downscaling Model (KNN-BUQSDM) 37

2.4.1 K-Nearest Neighbour (KNN) for occurrence modelling 38

2.4.2 K-means clustering 40

2.4.3 Cluster validation 42

2.5 Precipitation amount estimation using GPR 44

2.5.1 Parameter optimization 47

2.5.2 Predictive distribution 49

2.5.3 Algorithm 1 – SGP-SDM model selection 49

2.5.4 Algorithm 2 – SGP-SDM prediction 50

2.6 Results and Discussion 50

2.6.1 Evaluation criteria for SDM 50

2.6.2 Precipitation occurrence determination and K-means clustering 52

2.6.3 Precipitation amount estimation using GPR 55

2.6.4 Discussions on KNN-BUQSDM structure 67

CHAPTER 3 SGP-SDM – An advanced statistical downscaling model with uncertainty assessment of coupled classification and precipitation amount estimation using Gaussian Process Error Coupling 71

3.1 Abstract 71

3.2 Introduction 72

3.2.1 Uncertainty in statistical downscaling models 73

3.2.2 Uncertainty quantification in precipitation occurrence determination 76

3.2.3 Uncertainty quantification in precipitation amount estimation 80

3.3 Data and study area 83

3.4 Methodology 88

3.4.1 Bayesian updating framework 90

3.4.2 Precipitation occurrence determination model using GPC 92

3.4.3 Precipitation amount determination using GPR 101

3.5 Results and discussion 105

3.5.1 Validation period result (2005-2010) using CFSR reanalysis data 107

3.5.2 Comparison of results from CanESM2 RCP 4.4 and 8.5 113

3.5.3 Discussions on SGP-SDM structure 120

CHAPTER 4 Integrated MGP-SDM (A Bayesian uncertainty quantification framework for multisite statistical downscaling with residual coupling) and disaggregation to generate hourly precipitation at a station scale 123

4.1 Abstract 123

4.2 Introduction 124

4.3 Data and study area 132

4.4 Multi-site precipitation occurrence determination using multi-output GPC 134

4.4.1 Expectation Propagation (EP) approximation for inference 135

4.4.2 Predictive distribution 137

4.5 Multisite Precipitation amount estimation using multi-output GPR 137

4.6 KNN Disaggregation 142

4.7 Results and Discussions 143

4.7.1 MGP-SDM precipitation occurrence determination 144

4.7.2 MGP-SDM precipitation amount estimation 146

4.7.3 KNN disaggregation 147

4.7.4 HadCM3 A2 scenarios future precipitation projection (2011-2099) 148

CHAPTER 5 Integrated MGP-SDM and BUQ-SDDHM for uncertainty quantification to study the impact of climate change on future flood events 167

5.1 Introduction 167

5.1.1 Integrated multi-site SDM and DDHM for runoff simulation 168

5.1.2 Hydrological models 170

5.1.3 Uncertainty analysis in hydrological models 171

5.2 Data and Study area 174

5.3 Methodology 176

5.3.1 MGP-SDM 176

5.3.2 Bayesian Uncertainty Quantification framework for Stochastic Data-Driven Hydrological Model (BUQ-SDDHM) 180

5.3.3 KNN runoff disaggregation model 182

5.4 Results and Discussion 183

5.4.1 Multi-site downscaling using MGP-SDM 183

5.4.2 Future Climate Projections for the period 2011-2099 186

5.4.3 BUQ-SDDHM flow simulation 193

5.4.4 Flood frequency analysis 197

CHAPTER 6 Conclusions and future directions 201

6.1 Chapter 2 conclusions 201

6.2 Chapter 3 conclusions 203

6.3 Chapter 4 conclusions 205

6.4 Chapter 5 conclusions 206

6.5 Suggested future works 207

REFERENCES 209

ABSTRACT

The International panel on climate change (IPCC) reported that the impact of climate change is considered as one of the major reasons for the increase in the extreme flood events especially the intense precipitation. The intense flood in a short period of time can cause damage to the properties and affect daily life of human. Singapore has witnessed the increased flood events which occurred in a short period of time in the recent past and the documental increasing trend in the rainfall amount is in agreement with the IPCC report. Thus, the strategies to assist future adaption planning and risk mitigation during extreme flood events need to be developed based on the predicted future climate conditions. Fine spatial and temporal resolutions in climate data are needed to simulate the change in future flood events and to study the climate change impact on hydrology. The downscaling model combined with the temporal disaggregation can generate high spatial and temporal resolution of future climate data. The Statistical Downscaling Model (SDM) is the bridging model which is used to downscale the output from the General Circulation Model (GCM) for increasing the spatial resolution of future climate scenarios. The temporal disaggregation model increases the temporal scale, for example, from daily to hourly or minute scale. The information on the expected future change in the precipitation is needed to make efficient decisions. However, the predictions obtained from numerical climate model have uncertainties due to the generalized representation of the complex climate system. The sources of uncertainty include natural variability, uncertainties in the climate model(s), the downscaling model, the disaggregation model and the model inadequacy.

This research focuses on quantifying the uncertainty in the downscaled and disaggregated future climate variables by adopting a full Bayesian updating model framework for the statistical downscaling and the data-driven hydrological models. Bayesian updating framework provides a principled probabilistic way to quantify uncertainty in the model calibration and prediction. This research investigation has been carried out in two levels. The first goal was to develop a combined stochastic

statistical downscaling and disaggregation model coupled with uncertainty quantification tool that captures both aleatory and epistemic uncertainties in downscaled climate variables from large scale climate model data. The second goal was to develop a stochastic process based data-driven hydrological model, integrated with the uncertainty quantification tool to simulate the river flow using the downscaled and disaggregated climate variables as inputs.

Initially, a single site statistical downscaling model has been considered where a stochastic process-based SDM is proposed to couple the uncertainty quantification tool with model calibration and model prediction. The classical SDM has three steps such as 1) precipitation occurrence determination 2) precipitation amount estimation and 3) residual fitting. The contemporary regression based SDM assumes different distribution for precipitation amount estimation and residual fitting. Two new SDM approaches was developed for single site downscaling in this research. The first approach was named K-nearest neighbor-Bayesian Uncertainty Quantification for Statistical Downscaling Model (KNN-BUQSDM). In KNN-BUQSDM, KNN was used to determine the precipitation occurrence and to classify the wet days into different rainfall types based on the rainfall magnitude. For each rainfall type, the rainfall amount was estimated using a Gaussian Processes (GP) model. The GP model is based on stochastic error coupling method wherein the dependency between the residuals were used for prediction. The stochastic SDM couples the amount estimation model and the residual fitting under same distribution assumption using a Bayesian framework (in this thesis Gaussian distribution). The GP model enables to simulate the posterior predictive distribution for precipitation amount. The study results demonstrated that the classifying rainfall into several types and coupling the precipitation amount estimation and residual fitting was helpful to capture the characteristics of precipitation in downscaling.

The second approach proposed for SDM was named Single site Gaussian Process-Statistical Downscaling Model (SGP-SDM), a methodology to quantify the uncertainty in the precipitation occurrence model as well as the precipitation amount estimation

model. In SGP-SDM, GP was used for both precipitation occurrence determination and amount estimation. SGP-SDM gives the posterior predictive distribution for both the precipitation occurrence and the precipitation amount. The rainfall was not classified into several types; however, the results were comparable to KNN-BUQSDM without classifying rainfall into different types. The local characteristics of the rainfall was also captured well by SGP-SDM.

The extension of the single site SDM to multi-site SDM has been considered as the next step to downscale the climate variable observations at multiple sites simultaneously. The proposed multisite downscaling model was named MGP-SDM. The spatial correlation between the sites and the uncertainty quantification tool was coupled with the model calibration and prediction using Bayesian framework in MGP-SDM. The posterior predictive distribution of the climate variables at multiple sites can be estimated using MGP-SDM simultaneously. KNN disaggregation model was then integrated with MGP-SDM to simulate hourly precipitation at multiple sites in Singapore. The proposed combined multi-site downscaling and disaggregation model was used to project hourly precipitation under future climate conditions. From the literature study, it is learnt that the data-driven hydrological models are widely used to simulate the river flow based on the data rather than using the physical relationship between the variables for river flow prediction. A GP data-driven hydrological model named BUQ-SDDHM (Bayesian Uncertainty Quantification for Stochastic Data Driven Hydrological Model) is the one proposed in this research for the simulation of the river flow using the downscaled and disaggregated climate data. This method couples the uncertainty quantification tool with model calibration and prediction of streamflow. The posterior predictive distribution of the streamflow can be obtained from BUQ-SDDHM. In the last step, MGP-SDM and BUQ-SDDHM was integrated with the KNN disaggregation model to simulate high resolution streamflow under future climate conditions. The proposed method for climate change impact studies on hydrology makes use of the full Bayesian framework to propagate the uncertainty in projecting flood frequencies in future using GCM data.

LIST OF PUBLICATIONS

1) RAJENDRAN Queen Suraajini and CHEUNG Sai Hung (2015). "BUASCSDSEC –Uncertainty Assessment of Coupled Classification and Statistical Downscaling Using Gaussian Process Error Coupling." International Journal of Environment Science and Development 6(3): 211.

2) RAJENDRAN Queen Suraajini and CHEUNG Sai Hung, "KNN-BUQSDM- A Bayesian Updating Uncertainty Quantification Framework for Statistical Downscaling of Precipitation." International Journal of Climatology. (Under review).

3) RAJENDRAN Queen Suraajini and CHEUNG Sai Hung, "SGP-SDM – An Advanced statistical downscaling model with uncertainty assessment of coupled classification and precipitation amount estimation using Gaussian Process Error Coupling." (under review)

4) RAJENDRAN Queen Suraajini and CHEUNG Sai Hung, "Integrated SGP-SDM with temporal disaggregation model for high resolution precipitation simulation for climate change impact studies." (to be submitted)

5) RAJENDRAN Queen Suraajini and CHEUNG Sai Hung, "MGP-SDM – A Bayesian uncertainty quantification framework for multisite statistical downscaling with error coupling." (to be submitted)

6) RAJENDRAN Queen Suraajini and CHEUNG Sai Hung, "Integrated MGP-SDM and multisite temporal disaggregation model for high resolution precipitation simulation for studying climate change impact on hydrology." (to be submitted)

7) RAJENDRAN Queen Suraajini and CHEUNG Sai Hung, "BUQ-SDDHM – A Bayesian data-driven hydrological model runoff simulation using stochastic error coupling for uncertainty assessment." (to be submitted)

8) RAJENDRAN Queen Suraajini and CHEUNG Sai Hung, "Integrated MGP-SDM and BUQ-SDDHM for uncertainty quantification to study the impact of climate change on future flood events." (to be submitted)

9) Queen Suraajini Rajendran, Sai Hung Cheung (2014), "BUASCSDSEC - Uncertainty Assessment of Coupled Classification and Statistical Downscaling Using Gaussian Process Error Coupling." 2014 International Conference on Environmental Engineering and Development (ICEED 2014), Sydney, Australia, May 27-28.

10) Queen Suraajini Rajendran, Sai Hung Cheung (2014), "Statistical Classification, Downscaling and Uncertainty Assessment for Global Climate Model Outputs." 18th International Conference on Environmental and Ecological Systems (ICEES 2014), Sydney, Australia, December 15-16.

LIST OF TABLES

Table 2-1 Comparison of ASD, GLM and KNN-BUQSDM statistical downscaling model 19

Table 2-2 List of CFSR predictors for Singapore 34

Table 2-3 Correct wet day and dry day classification by KNN 53

Table 2-4 Cluster validation index for the month of February 54

Table 2-5 Cluster validation index for the month of December 54

Table 2-6 Comparison of dry-day proportion estimated by KNN and GPR with the observed dry day proportion 57

Table 2-7 Accuracy of downscaled precipitation for the month of December 58

Table 2-8 Comparison of evaluation statistics of ASD, GLM, KN-BNN and KNN-BUQSDM for the month of December 58

Table 2-9 Monthly mean squared error 59

Table 2-10 MAPBE values of the downscaled results using ASD, GLM, KNN-BNN and KNN-BUQSDM at S44 61

Table 2-11 p-values of ks-test value of the distribution of the precipitation 62

Table 2-12 MAPBE values of the simulated results using three methods at S24 62

Table 3-1 CanESM2 predictors 89

Table 3-2 Log marginal likelihood for the models with different covariance functions calculated using CFSR reanalysis data for the validation period 108

Table 3-3 Accuracy, correct percentage of wet and dry days calculated by GPC using the CFSR reanalysis data for the validation period 109

Table 3-4 Comparison of dry-day proportion estimated by GPC and GPR with the observed proportion 110

Table 3-5 Mean Square Error (MSE) of the average of the evaluation statistics 113

Table 3-6 Comparison of observed and simulated NEE for validation period (2005-2010) using CFSR data 114

Table 3-7 Accuracy of downscaled precipitation for the month of December 114

Table 3-8 MAPBE value of downscaled precipitation envelop obtained from 50 ensembles 118

Table 3-9 Minimum, average and maximum evaluation statistical indicators obtained using 50 ensemble simulated by SGP-SDM for the month of December 119

Table 3-10 Comparison of observed and simulated NEE for validation period (2006-2010) using CanESM2 data 120

Table 4-1 Accuracy of the MGP-SDM classification ensembles at three stations 150

Table 4-2 Cross-correlation of the downscaled rainfall between the sites 150

Table 4-3 Cross-correlation of the disaggregated rainfall between the sites 150

Table 5-1 Vermillon climate station information 176

Table 5-2 Vermillon hydrometric station information..... 176

Table 5-3 Comparison of simulated minimum, average and maximum monthly rainfall with the observed rainfall at three stations during the validation period (1976-1980) 185

Table 5-4 Comparison of simulated minimum, average and maximum monthly minimum temperature with the observed minimum temperature at three stations during the validation period (1976-1980) 186

LIST OF FIGURES

Figure 1-1 Cascade of uncertainty in assessment of climate change impacts on hydrology 3

Figure 2-1 Rain gauge location and CFSR data for Singapore 33

Figure 2-2 KNN-BUQSDM Framework 38

Figure 2-3 Quantile-Quantile plot for the month of December to compare the observed and downscaled precipitation with one and two classes 55

Figure 2-4 Monthly mean evaluation statistics for the rain gauge station at S44. The shaded area represents the 5th and 9th percentile of the ensembles. The dashed line represents the minimum and maximum values of the ensemble statistics 63

Figure 2-5 Observed and downscaled precipitation at S44 for the month of February and December 64

Figure 2-6 Comparison of CDF of the observed and downscaled precipitation for all the months at station S44 66

Figure 2-7 Monthly mean evaluation statistics of the downscaled precipitation at the rain gauge station S24. The shaded area represents the 5th and 9th percentile of the ensembles. The dashed line represents the minimum and maximum values of the ensemble statistics. 70

Figure 3-1 Study area and rain gauge location in Singapore 85

Figure 3-2 SGP-SDM statistical downscaling framework 91

Figure 3-3 Monthly mean evaluation statistics for the rain gauge stations S44. The shaded area represents the 5th and 9th percentile of the ensembles. The dashed line represents the minimum and maximum values of the ensemble statistic 115

Figure 3-4 Monthly mean evaluation statistics for the rain gauge stations S44 using CanESM2 RCP 4.5 scenarios. The shaded area represents the 5th and 9th percentile of the ensembles. The dashed line represents the minimum and maximum values of the ensemble statistics 116

Figure 3-5 Monthly mean evaluation statistics for the rain gauge stations S44 using CanESM2 RCP 8.5 scenarios. The shaded area represents the 5th and 9th percentile of the ensembles. The dashed line represents the minimum and maximum values of the ensemble statistics 117

Figure 4-1 Location of rain gauge stations at Singapore used in the study 133

Figure 4-2 Comparison of observed and simulated dry day proportion by MGP-SDM at three stations a) S46, b) S55 and c) S69 for the validation period (1980-2010) 149

Figure 4-3 Comparison of observed and simulated dry day transition probability at three stations a) S46, b) S55 and c) S69 for the validation period (1980-2010) 151

Figure 4-4 Comparison of observed and simulated wet-day transition probability at three stations a) S46, b) S55 and c) S69 for the validation period (1980-2010) 152

Figure 4-5 Comparison of cdf of the downscaled precipitation ensembles with observed cdf for all the months at Station S46 154

Figure 4-6 Comparison of cdf of the downscaled precipitation ensembles with observed cdf for all the months at station S55 156

Figure 4-7 Comparison of cdf of the downscaled precipitation ensembles with observed cdf for all the months at station S69 158

Figure 4-8 Evaluation Statistics for the station S46. The line with square represents the observed data. The shaded region shows the 5^{th} and 95^{th} percentile of the prediction ensembles. The dotted lines represent the maximum and minimum range. 159

Figure 4-9 Evaluation Statistics for the station S55. The line with square represents the observed data. The shaded region shows the 5^{th} and 95^{th} percentile of the prediction ensembles. The dotted lines represent the maximum and minimum range. 160

Figure 4-10 Evaluation Statistics for the station S69. The line with square represents the observed data. The shaded region shows the 5^{th} and 95^{th} percentile of the prediction ensembles. The dotted lines represent the maximum and minimum range. 161

Figure 4-11 Disaggregated precipitation projection at station S46 for the validation period 1980 -2010. The average hourly disaggregated data for the baseline period calculated from 20 ensembles 162

Figure 4-12 Disaggregated precipitation projection at station S55 for the validation period 1980 -2010. The average hourly disaggregated data for the baseline period calculated from 20 ensembles. 163

Figure 4-13 Disaggregated precipitation projection at station S69 (4-13a-4-13d) for the validation period 1980 -2010. The average hourly disaggregated data for the baseline period calculated from 20 ensembles. 164

Figure 4-14 Mean hourly precipitation projection for future periods (2011-2099)..... 165

Figure 4-15 Maximum hourly precipitation projection for future periods (2011-2099) 166

Figure 5-1 Climate station and hydrometric data location for Vermillon watershed, Quebec, Canada 175

Figure 5-2 Integrated MGP-SDM and BUQ-SDDHM workflow framework for hydrological impact studies 177

Figure 5-3 Comparison of (a) observed and downscaled precipitation ensembles and (b) observed and downscaled minimum temperature ensembles at La Turque 187

Figure 5-4 Comparison of (a) observed and downscaled precipitation ensembles and (b) observed and downscaled minimum temperature ensembles at Barrage Mattawin 187

Figure 5-5 Comparison of (a) observed and downscaled precipitation ensembles and (b) observed and downscaled minimum temperature ensembles at St-Michel-des-Saints 188

Figure 5-6 Downscaled precipitation ensembles of downscaled (a) monthly rainfall and (b) minimum temperature ensembles for future HadCM3 scenarios (2011-2040) at La Turque 188

Figure 5-7 Downscaled precipitation ensembles of downscaled (a) monthly rainfall and (b) minimum temperature ensembles for future HadCM3 scenarios (2011-2040) at Barrage Mattawin 189

Figure 5-8 Downscaled precipitation ensembles of downscaled (a) monthly rainfall and (b) minimum temperature ensembles for future HadCM3 scenarios (2011-2040) at St-Michel-des-Saints 189

Figure 5-9 Downscaled precipitation ensembles of downscaled (a) monthly rainfall and (b) minimum temperature ensembles for future HadCM3 scenarios (2041-2070) at La Turque 190

Figure 5-10 Downscaled precipitation ensembles of downscaled (a) monthly rainfall and (b) minimum temperature ensembles for future HadCM3 scenarios (2041-2070) at Barrage Mattawin 190

Figure 5-11 Downscaled precipitation ensembles of downscaled (a) monthly rainfall and (b) minimum temperature ensembles for future HadCM3 scenarios (2041-2070) at St-Michel-des-Saints..... 191

Figure 5-12 Downscaled precipitation ensembles of downscaled (a) monthly rainfall and (b) minimum temperature ensembles for future HadCM3 scenarios (2071-2099) at La Turque..... 191

Figure 5-13 Downscaled precipitation ensembles of downscaled (a) monthly rainfall and (b) minimum temperature ensembles for future HadCM3 scenarios (2071-2099) at Barrage Mattawin 192

Figure 5-14 Downscaled precipitation ensembles of downscaled (a) monthly rainfall and (b) minimum temperature ensembles for future HadCM3 scenarios (2071-2099) at St-Michel-des-Saints 192

Figure 5-15 Comparison of observed and simulated monthly river flows for the validation period 1976-1982 using BUQ-SDDHM 195

Figure 5-16 Simulated monthly river flows for future HadCM3 A2 scenarios 2011-2040 using BUQ-SDDHM 196

Figure 5-17 Simulated monthly river flows for future HadCM3 A2 scenarios 2041-2070 using BUQ-SDDHM 196

Figure 5-18 Simulated monthly river flows for future HadCM3 A2 scenarios 2071-2099 using BUQ-SDDHM 197

Figure 5-19 Comparison of observed runoff with one of the simulated runoffs from SGP-SDM 197

Figure 5-20 Comparison of CDF of observed runoff with simulated runoff from BUQ-SDDHM 198

Figure 5-21 Flood frequency analysis of the flows predicted using BUQ-SDDHM for the validation period (1976-1982). The median of the results is represented as middle line of the box, the 25^{th} and 75^{th} percentile is presented at the top and bottom lines and the whiskers are represented as the bars at the top and the bottom 199

Figure 5-22 Flood frequency analysis of the flows predicted using BUQ-SDDHM for 2011-2040. The median of the results is represented as middle line of the box, the 25^{th}

and 75th percentile is presented at the top and bottom lines and the whiskers are represented as the bars at the top and the bottom 200

LIST OF SYMBOLS

$\simeq$	approximately equal to
$\triangleq$	an equality which acts as a definition
$\otimes$	Convolution operator
$\lvert K_a \rvert or \lvert K_c \rvert$	determinant of K_a or K_c matrix
$\sim$	distributed according to
$\lVert . \rVert$	Euclidean distance
∇	partial derivatives
$\nabla\nabla$	the (Hessian) matrix of second derivatives
$\mathbf{1}$	vector of all 1's (of length n)
$\boldsymbol{\beta}_a$	Gaussian process mean function parameters
$\boldsymbol{\beta}_a^*$	optimal Gaussian process mean function parameters
C	k-means clusters
C_{dry}	total number of correctly classified dry days
C_{wet}	total number of correctly classified wet days
D_c or D_{mc}	single site (or multi-site) data for precipitation occurrence determination
D_{EUC}	Euclidian distance
D_a or D_{ma}	single site (or multi-site) data for precipitation amount estimation
δ_{ij}	Kronecker delta function, $\delta_{ij} = 1$ if $i = j$ and 0 otherwise
$\boldsymbol{\varepsilon}_a$	SDM additive errors/residuals
$f_a(X_a)$ or $\mathbf{f}_a$	precipitation amount estimation model function (or vectors)
f_{a*}	Gaussian process posterior of precipitation amount model
$f_c(X_c)$ or $\mathbf{f}_c$	precipitation occurrence determination model function (or vectors)

$\bar{f}_{c*}$	posterior mean of the precipitation occurrence model
F_D	Dunn's index
F_{DB}	Davies-Bouldin index
F_S	sillouette index
I	the identity matrix of size n
$K(X_a, X_a')$ or K_a or K_c	$n \times n$ Gaussian process covariance matrix without noise
K_m	number of K-means clusters in clustering problem
K_y	$n \times n$ Gaussian process covariance matrix with noise
$K_{mc}^{y_{mc}}$ or $K_{mc}^{x_{mc}}$	multi-site between-site (or within-site) covariance matrix for precipitation occurrence determination
L_a or L_c	single site covariance function lengthscale/correlation length matrix (diagonal positive symmetric matrix) precipitation amount (or precipitation occurrence) model
l_{ad}^* or l_{cd}^*	single site optimal correlation length value of covariance function precipitation amount (or precipitation occurrence) model
l_{m_a} or l_{m_c}	multi-site covariance function lengthscale precipitation amount (or precipitation occurrence) model
$m(X_a)$	Gaussian process mean function
m_a	dimension of predictors for precipitation amount estimation
m_c	dimension of predictors for precipitation occurrence determination
m_{cgj}	cluster centroids
$\mu(\mathbf{x}_a)$ or $\boldsymbol{\mu}_a$	Gaussian process mean function (or vectors) for the calibration period
$\mu(\mathbf{x}_{a*})$ or $\boldsymbol{\mu}_{a*}$	Gaussian process posterior mean function (or vectors) for the prediction period

n	number of historical data for calibration
$\pi(\mathbf{x}_c)$ *or* π	the sigmoid of the latent value, $\pi(\mathbf{x}_c) = \sigma(f_c(\mathbf{x}_c))$, $\pi(\mathbf{x}_c)$ is stochastic as $f_c(\mathbf{x}_c)$ is stochastic
$\bar{\pi}_{c*}$	Gaussian process precipitation occurrence prediction
σ_{fa}^2 *or* σ_{fc}^2	signal variance of the covariance function
σ_{fa}^{2*} *or* σ_{fc}^{2*}	optimal signal variance of the covariance function
σ_j	average distance of data in each clusters to its center m_{cgj}
σ_{na}^{2*}	optimal signal noise values of covariance function
s	number of downscaling sites
Σ_{ma}	multi-site block (between-site and within-site) covariance matrix for precipitation amount estimation
TP_{dry}	total number of dry days
TP_{wet}	total number of wet days
$\boldsymbol{\theta}_a$ *or* $\boldsymbol{\theta}_c$	vector of hyperparameters of precipitation amount model (or precipitation occurrence model)
$\boldsymbol{\theta}_a^*$ *or* $\boldsymbol{\theta}_c^*$	vector of optimal hyperparameters of precipitation amount model (or precipitation occurrence model)
X_a *or* X_{ma}	single site (or multi-site) predictors for precipitation amount model calibration
X_c *or* X_{mc}	single site (or multi-site) predictors for precipitation occurrence model calibration
$\mathbf{x}_{a*}$ *or* $\mathbf{x}_{c*}$	future predictors for precipitation amount model (or precipitation occurrence model)
$\mathbf{y}_a$ *or* $\mathbf{y}_{ma}$	single site (or multi-site) precipitation amounts
$\mathbf{y}_c$ *or* $\mathbf{y}_{mc}$	single site (or multi-site) precipitation occurrences

LIST OF ABBREVIATIONS

ARD	Automatic Relevance Determination
ANN	Artificial Neural Network
ASD	Automated Statistical Downscaling tool
CC	Correlation Coefficient
CanESM2	The Second Generation Earth System Model
CFSR	Climate Forecast System Reanalysis
EP	Expectation Propagation
GCM	General Circulation Model
GLM	Generalized Linear Model
GP	Gaussian Processes
GPC	Gaussian Process Classification
GPR	Gaussian Process Regression
HadCM3	Hadley Centre for Climate Model version 3
IPCC	Intergovernmental Panel on Climate Change
KNN	K-nearest neighbour
MAPE	Mean Absolute Percentage Error
MSE	Mean Square Error
MLR	Multiple Linear Regression
NCEP	National Centers for Environmental Prediction
OBS	Observed data
SDSM	Statistical Downscaling Model
SDM	Statistical Downscaling Model
SIM	Simulated data by the model

CHAPTER 1 **INTRODUCTION**

Climate change is considered as one of the causes for the increase in duration and intensity of rainfall which in turn causes intense flood events. The fourth assessment report on Inter-governmental Panel on Climate Change states that one of the reasons for increasing trend of intense rainfall is due to climate change (Pachauri, 2004). However, the response of each region to climate change signal varies (Mujumdar and Nagesh Kumar, 2012); the increase in the rainfall amount affects the drainage system in an urban area like Singapore the most. The expert panel on drainage design and flood protection measures of Singapore showed that the trends for rainfall intensities and frequency of intensive rainfall have been increasing for the last 30 years (MEWR, 2012). The complex urban drainage systems are built in the metropolitan cities like Singapore to convey urban runoff to protect the city from flood inundation. The city becomes vulnerable to flood events when there is excess amount of rainfall than the designed capacity of the drainage system (Willems, 2012). The increased amount of intense rainfall in the short period of time causes the failure of drainage system to hold huge amount of rain water leading to urban inundation which Singapore has been experiencing for the past few years. There is an increasing need to assess the impact of climate change on hydrology using reliable and accurate fine resolution prediction of the change in the future rainfall events for effective mitigation and decision making for incorporating new drainage system design in an urban area.

The climate change impact on hydrology is commonly assessed by integrating the emission scenarios from the GCM with the hydrological models to simulate the corresponding hydrologic projections. The issue of scale mismatch occurs since the spatial and temporal resolution of GCM is designed to operate at a lower resolution than the higher resolution required by the hydrological models (Wilby and Wigley, 1997). The spatial and temporal resolution of the simulated future climate scenario from GCM is too coarse (large scale climate variables) that it does not match the characteristics of local variables like precipitation for a smaller area. This is because

the change in large scale variables happens smoothly; hence large scale variable does not account for changes at the local scale (Benestad *et al.*, 2008). This issue is generally addressed by combining spatial downscaling model and temporal disaggregation model to increase the spatial resolution and the temporal resolution required by the hydrological models that simulate hydrological projections (Mezghani and Hingray, 2009). The assessment of climate change impact on hydrology comprises of huge amount of uncertainty due to the choice of GCMs and emission scenarios, downscaling models, disaggregation models, hydrological models and the historical data used for model calibration (Foley, 2010). The climate system is a complex physical process due to the interaction circulation of the atmospheric, ocean, ice and land surface components. GCMs represent the dynamics of the climate system using mathematical equations and are represented numerically through the computer program (Graham *et al.*, 2011). GCMs have been used for simulating future climate scenarios to project future climate change and its impact on hydrologic cycle, health and economy (Randall, 2000). The GCMs which are used for future climate scenarios always consist of uncertainties due to incomplete knowledge about the climate system and imperfect generalization of the climate by using parameterizations in the model (Woldemeskel *et al.*, 2014). The uncertainties are also introduced due to downscaling, disaggregation and hydrological model structure and parameters (Dobler *et al.*, 2012; Refsgaard *et al.*, 2013; Clark *et al.*, 2016). In climate change studies, the uncertainty is cascaded from GCM to hydrologic projections as shown in Figure 1-1 (Graham *et al.*, 2007; Oyebode *et al.*, 2014).

As the transfer of climate projections to stream flow simulation is a nonlinear process, it is difficult to obtain similar results when different combinations of models are used (Chen *et al.*, 2011). Prudhomme and Davies (2009) investigated three main sources of uncertainty in river flow projections. Three GCMs (HadCM3, CCGCM2 and CSIRO-mk2) and two downscaling models (Statistical Downscaling Model (SDSM) and dynamic downscaling model, HadRM3) were considered for the analyses. Their study results concluded that the GCM is the largest contributor of uncertainty in impact

studies compared to statistical downscaling and is greater than the uncertainty from the hydrological models. However, the results could not conclude the best performing GCM or downscaling model for impact studies. In another study by Prudhomme and Davies (2009), the uncertainty due to Green House Gases (GHG) emission scenarios, GCM, downscaling model and hydrological model were studied and it is concluded that the GCM is the greatest source of uncertainty, the uncertainty from downscaling techniques and emission scenarios are of similar magnitude and is smaller than GCM; the uncertainty due to hydrological modelling is small. Similar results were also seen in the studies by Wilby and Harris (2006) and Kay *et al.* (2009). The uncertainty from GCMs and emission scenarios were investigated in several studies and these studies also concluded that the GCM is the major contributor of uncertainty, for example (Jenkins and Lowe, 2003; Rowell, 2006). It is recommended in all the studies to use multiple GCMs and emission scenarios to quantify the uncertainty in the prediction. From the literature studies on assessing the main sources of uncertainty, it is found that GCM is the main source of uncertainty (Hawkins and Sutton, 2009), statistical downscaling model is the second largest contributor of uncertainty while the uncertainty due to hydrological model and its parameters is less significant (Chen *et al.*, 2011).

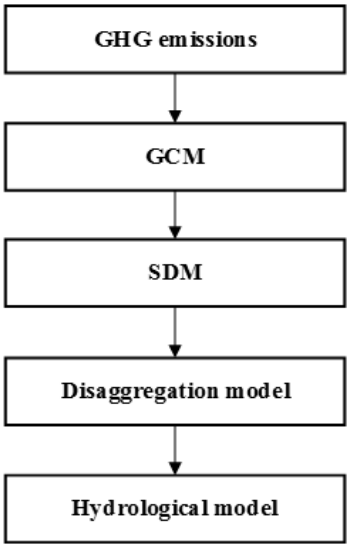


Figure 1-1 Cascade of uncertainty in assessment of climate change impacts on hydrology

Several downscaling techniques have been proposed in the literature to increase the spatial resolution of the GCM scenarios for climate change impact study on hydrology. The two types of downscaling techniques are Dynamical Downscaling Model (DDM) and Statistical Downscaling Model (SDM). The detailed explanation about the downscaling methods can be found in the literature review by Fowler *et al.* (2007) and Maraun *et al.* (2010). Dynamic downscaling involves nesting regional meteorological models into GCM to simulate local variables. As the time slice for DDMs is only 30 years, it is difficult to do future climate change impact studies as it becomes computationally intensive to get output for a longer period. Thus, the projection of climate using RCM is computationally intensive which makes it less preferable compared to statistical downscaling methods. SDMs use the statistical relationship between the large scale predictors and local weather variables to simulate climate variables at a station (point) scale (Wilby and Wigley, 1997). The choice of GCM for statistical downscaling contributes to the uncertainty in the downscaled variable prediction. It is important to acquire adequate knowledge of advantages and disadvantages about the applicability of GCM to the corresponding study area to facilitate the reduction of uncertainty from GCM (Oyebode *et al.*, 2014). However, the assumptions in statistical downscaling are considered as a major cause for uncertainty in the downscaled variables; the assumption of stationarity relationship between the current and future climatic conditions may not hold true in the future. There are numerous studies on assessing inherent uncertainty in the statistical downscaling approaches which have been conducted by comparing the downscaling approaches. In STARDEX project, the uncertainty from SDM has been analyzed. Twenty two statistical downscaling techniques to downscale temperature and precipitation have been assessed using ten indices. It was suggested that a range of downscaling models should be used for accurate projections of future scenarios as there is no best downscaling model exists (Goodess *et al.*, 2007). Wilby *et al.* (1998) compared Weather Generators (WGEN) and Artificial Neural Network (ANN) in downscaling precipitation. The results indicated that the ANN showed poor simulation of wet-day occurrences compared to WGEN. Khan *et al.* (2006) compared SDSM, Long Ashton

Research Station Weather Generator (LARS-WG) model and ANN. The results showed that the SDSM reproduced the observed statistics with 95% confidence intervals compared to ANN and LARS-WG. Two SDMs which were used to downscale daily precipitation and temperature and monthly precipitation were compared by Schoof and Pryor (2001). It was found out that the regression and ANN models yielded similar results. It was also shown that the accuracy of temperature downscaling was better than the precipitation downscaling. The detailed review of all the uncertainty analyses on the downscaling model can be found in Fowler *et al.* (2007), Maraun *et al.* (2010) and Quintana Seguí *et al.* (2010). The choice of model structure for the statistical downscaling will impact the amount of uncertainty in the streamflow predictions using hydrological models (Oyebode *et al.*, 2014). The assessment of uncertainty due to choice of downscaling model in streamflow predictions was studied in a number of research works; those works comprise of comparison of several downscaling models combined with the hydrological model in simulation of the streamflow. Tisseuil *et al.* (2010) conducted a study to select a suitable downscaling technique for predicting river flow by linking the downscaling model to the river flows. They compared Generalized Linear Model (GLM), Generalized Additive Model (GAM), aggregated boosted trees (ABT) and ANN. The results indicated that the nonlinear models such as GAM, ABT, and ANN perform better compared to the linear models. Wilby and Harris (2006) suggested probabilistic framework to combine ensembles from different models such as four GCMs, two emission scenarios, two SDMs, two hydrological models and hydrological parameters for assessing uncertainty in climate change impact studies. Each GCMs and hydrological models were assigned weights based on the reliability in the predicted results while the emission scenarios and the downscaling models were not weighted. It was observed that the GCM and the downscaling model contribute more uncertainty compared to hydrological models and emission scenarios. Similar results have also been found in other research studies (Khan *et al.*, 2006; Maraun *et al.*, 2010; Ghosh and Katkar, 2012). Even with the availability of several approaches, the comparison results are inconclusive in choosing the best approaches. One of the reasons is that all the classical downscaling approaches

have similar assumptions in model structure formulation. The basic flaw in the assumptions are 1) there is no consistency between the distribution assumption since Gaussian distribution is assumed for model parameter estimation and an extreme value distribution for residual fitting 2) the model calibration and the residual parameter estimation are implemented separately and 3) the residuals are assumed to be independent in model formulation. Modification of these assumptions is necessary to improve the confidence in the predictions. The residual can be coupled with the model calibration and prediction and residual calibration simultaneously using probabilistic approach (Beck and Cheung, 2009). Probabilistic approach has been explored in the literature to reduce uncertainty due to statistical downscaling model assumptions (Ghosh and Mujumdar, 2008). Further research is needed to find out the applicability of probabilistic approach in quantifying uncertainty in SDM.

The complex and non-linear processes related to climate conditions and water resources are explained using hydrological models. The downscaled meteorological variables are given as inputs in hydrological models to simulate the required hydrological outputs. The hydrological models can be classified based on the model function, its objective and their structure (Fung *et al.*, 2011). They are physics based models (PBM), conceptual models, process based models (PRBM) and data-driven models (DDM). As it is impossible to prevent floods owing to several natural complex factors, it is critical to develop a model that is able to predict the hazardous flood events with high accuracy. In order to manage flood, the flow prediction is being used to assess the duration of flood and spatial extent. The primary area of research in the field of hydroinformatics is flood management and forecasting (Abbott, 1991; Price, 2000; Solomatine and Ostfeld, 2008). In hydrological modelling, the availability of large amount of data through information and communication technologies, the number of measurement for climate variables opened up new modelling techniques. Data-driven hydrological modelling is an active area of research. With respect to the data driven hydrological models, there are many speculations on their ability to represent the flood flows accurately compared to physics based hydrological models (Abrahart

and See, 2000; Solomatine and Ostfeld, 2008). The recent advancement in DDM has proved that the prediction can be obtained accurately by considering the data driven approaches as the alternate models of the physics based hydrological models (Kamp and Savenije, 2007). The uncertainty in the streamflow prediction arises from the model and the sample size used for training the model (Shrestha and Solomatine, 2008). There are two factors that cause model uncertainty. They are 1) the choice of model used for prediction to capture the sufficient details (Booij, 2005) and 2) model order. The uncertainty due to model order in case of ANN can be reduced by using the appropriate architecture for the networks (hidden layers and hidden nodes) and careful model calibration and validation. The suitability of data-driven hydrological model is generally analyzed by utilizing the study results from the application of different models in the research works (Bergström *et al.*, 2001). There needs further exploration of suitability of advanced data-driven stochastic process models for the river flow simulation and climate change impact assessment.

While there are several studies on uncertainty analysis for statistical downscaling models, there are limited studies on uncertainty assessment for integrated statistical downscaling, disaggregation and hydrological models for climate change impact studies on hydrology. It is critical to provide accurate degree of confidence of the predictions obtained from the climate model for efficient mitigation and policy making with respect to climate change impact. It is an essential task to develop a combined downscaling and disaggregation model coupled with uncertainty quantification framework for studying the impact of climate change on hydrology for a smaller urban region like Singapore as the climate change impact studies on urban drainage systems are limited. The term 'coupled' herein refers to the process of calibrating the model, residuals and prediction simultaneously in contrast to the classical methods where these steps are implemented separately. In this research study, a full Bayesian updating framework developed by Cheung *et al.* (2011) is proposed to tackle the above mentioned problems in cascade of uncertainty in climate change impact studies to generate reliable predictions with a more realistic representation of uncertainty.

1.1 Bayesian inference framework

Probabilistic representation of the real world system using Bayesian framework is considered as an effective and rational methodology to represent the uncertainty and variability in the model calibration and prediction (Beck and Katafygiotis, 1991; Beck and Katafygiotis, 1998; Cheung and Beck, 2009). Bayesian framework provides a technique to update the model in consideration whenever new knowledge about the system is obtained. The objective of Bayesian model updating is to provide accurate predictions along with the quantitative estimation of the accuracy of the prediction by reducing the discrepancies between the observed and model output. These discrepancies are caused by different sources of uncertainties in the model referred as model uncertainties. In the advanced research works on uncertainty quantification in the field of structural dynamics (Cheung and Beck, 2009; Cheung *et al.*, 2011) and Geostatistics (Isaaks and Srivastava, 1989), Bayesian updating methods have been implemented successfully to get accurate and robust predictions along with error bars (Beck and Katafygiotis, 1998; Cheung and Beck, 2009). Bayesian approach allows to quantify all types of uncertainty such as epistemic and aleatory in the mathematical models within a single framework (Cheung, 2007; Cheung and Beck, 2007; Cheung and Beck, 2007; Cheung and Beck, 2008; Cheung and Beck, 2008; Cheung and Beck, 2009; Cheung and Beck, 2009; Cheung and Beck, 2010). The stochastic system analysis is considered to be an important step to take into account all the essential types of uncertainties in the model formulation for accurate and reliable prediction by coupling the residual fitting and model parameter calibration.

Bayesian framework provides a robust mechanism for model calibration which also accounts for uncertainty quantification of the parameter estimation and modelling errors (Beck and Cheung, 2009). The prior assumption about the parameters and model function is updated by the likelihood function using Bayes' theorem to obtain the posterior probability distribution for the parameters for the given data. Posterior probability of the parameters is used to compute the posterior predictive distribution of the predictand instead of point estimate for parameters. This posterior probability

distribution can be propagated to obtain the updated uncertainty in the model predictions (Cheung and Beck, 2008). The predictive probability distribution can be used to simulate the ensembles of the predictions to compute the confidence interval (Beck and Katafygiotis, 1991).

$$p(\boldsymbol{\theta} \mid D, M) = c^{-1} p(D \mid \boldsymbol{\theta}, M)\, p(\boldsymbol{\theta} \mid M) \tag{1.1}$$

where $\boldsymbol{\theta}$ is the model parameter, D is the data, M is the model class, $p(D \mid \boldsymbol{\theta}, M)$ is the likelihood to represent the conditional probability of the data given the parameters, the prior pdf $p(\boldsymbol{\theta} \mid M)$, $c = p(D \mid M) = \int p(D \mid \boldsymbol{\theta}, M)\, p(\boldsymbol{\theta} \mid M) d\boldsymbol{\theta}$ is the normalizing constant and $p(D \mid M)$ is the evidence of the model class M. The posterior pdf $p(\boldsymbol{\theta} \mid D, M)$ is proportional to the product of prior pdf and the likelihood function. The initial plausibility of the model defined by the parameters is represented by the prior pdf, the residual of the model class in consideration is incorporated in the likelihood function and the posterior pdf is the updated plausibility of the model class using the information from the data D.

In a lot of cases, the posterior distribution cannot be obtained analytically. Laplace asymptotic approximation method (Beck and Katafygiotis, 1998) is used to solve multi-dimensional integrals numerically. When the availability of the data is less or model class is unidentifiable, stochastic simulation of samples from the posterior PDF can be used to solve the multi-dimensional integrals. The state of the art Markov chain Monte Carlo (MCMC) sampling methods (Cheung and Beck, 2010) can be used for stochastic sample simulation to significantly reduce the computation time in evaluating likelihood function. Statistical averaging of Markov chain samples is used to estimate the integral. The Bayesian analysis also helps in the selection of the plausible model that best represents the system. The evidence of the model is used for selecting the best performing model thus enabling to consider model uncertainty in choosing the model. Bayesian model class averaging can then be used to combine all the models together to make combined prediction (Cheung and Beck, 2010).

The posterior predictive probability distribution for the test/future data X_f is given by the *total theorem of probability* (Cheung and Beck, 2010).

$$p(X_f \mid D, M) = \int p(X_f \mid \boldsymbol{\theta}, D, M)\, p(\boldsymbol{\theta} \mid D, M) d\boldsymbol{\theta} \tag{1.2}$$

The posterior probability $p(\boldsymbol{\theta} \mid D, M)$ from each model is used as weight for prediction $p(\mathbf{X}_f \mid \boldsymbol{\theta}, D, M)$ for each model class.

1.2 Objectives of the thesis

1.2.1 Single site statistical downscaling model with coupled uncertainty quantification

In the recent years, Gaussian Processes (GPs) have become important for machine learning. This methodology is also referred to Kriging in Geostatistics. GP is a non-parametric Bayesian technique which has been applied for spatial stochastic process modelling, robotics and control problems (Rasmussen and Williams, 2006). In those fields, the improved accuracy in the prediction has been shown by considering the dependencies between the residuals. GP offers flexibility to propagate uncertainty in the input variables to the output variables along with the error bars for the prediction of the future data. GP has the ability to calibrate model function parameters and error parameters simultaneously to propagate uncertainty.

The first objective of this thesis to develop a stochastic error coupling based rainfall occurrence and rainfall amount model to estimate the uncertainty in the binary classification model and the downscaling model. In this model, the dependence between the residuals will be taken into account by the model with an assumption that the residuals are stochastic processes following Gaussian distribution. The idea is to consider the uncertainties in the model function, parameters, and residuals and to treat them simultaneously within a single framework. This will be achieved by coupling model parameter calibration and residual parameter calibration in a Bayesian

framework and fitting the data with the same distribution assumption that is, Gaussian distribution. The model will be developed by using the state-of-the-art Bayesian updating approach to quantify uncertainty (Cheung *et al.*, 2011; Cheung and Bansal, 2013). This approach will be useful in the characterization of uncertainty in the structure of the models and in the projections of future climate.

1.2.2 Integrated multi-site statistical downscaling and disaggregation models with uncertainty quantification tool

The simulation of rainfall at a single site is not sufficient as the fine resolution spatial and temporal meteorological variables at multiple sites are needed for hydrological simulations. Single site GP model for downscaling can be extended to multi-site GP model for downscaling using multi-output GP. Multi-output GP is also referred as dependent GP or Co-kriging in other fields. In robotics and control problems, it is proved that by considering the dependencies between multiple outputs the prediction accuracy have been improved (Rasmussen and Williams, 2006). In addition, the timescale of the downscaled model output such as daily precipitation is not sufficient to be used as inputs in hydrological model. The downscaled precipitation also needs to be disaggregated to generate hourly or minute by minute rainfall.

The second objective is to downscale the meteorological variables at multiple sites by using multi-output GP. The multi-output GP will automatically capture the spatial dependencies between the sites; this eliminates the need to implement the downscaling model and the spatial dependencies separately. GP for multi-site downscaling also encapsulates the advantages of GP in single site downscaling models mentioned above.

The third objective is to integrate the multi-site statistical downscaling model with the K-nearest neighbor (KNN) disaggregation model to increase the temporal resolution of the downscaled results to hourly scale at multiple sites simultaneously.

1.2.3 Hydrologic impact studies using integrated downscaling model, disaggregation model and data driven hydrologic models

Empirical models or DDM predict output based on the input data characterizing the system such as meteorological parameters rather than the physical process of the watershed (Solomatine *et al.*, 2008). The examples are unit hydrograph method, regression methods and autoregressive integrated moving average (ARIMA). The ability of neural network to capture the nonlinear hydrological applications was explored and was successful (Bates and Campbell, 2001; Abrahart and See, 2007). There cannot be one model that can best suit the problem due to inadequacies and incomplete knowledge. DDM is not widely accepted among the hydrologic and water resources researchers since physics that is the basis for hydrological model is not involved for implementing the model. The complex nature of the hydrological model errors makes it difficult to parameterize the real world system. This poses a challenge in estimating the uncertainty in hydrological models.

The fourth objective is to develop a nonparametric technique for DDM along with uncertainty quantification tool. Over the past years, there has been advancement in stochastic model based on GP for uncertainty quantification coupled with prediction. The ability of GP to emulate the real world system has been explored in the field of climate models and has shown that the results are comparable to the physical models (Robin, 2012). The GP model will give point estimates along with the error bars; thus there is no need for separate uncertainty analysis methods after calibration and prediction of the models. Within this approach, the uncertainty is the model function and the model errors are coupled and calibrated simultaneously to predict the hydrological outputs such as stream flow.

The fifth objective is to integrate the multisite statistical downscaling and hydrological model developed in this thesis with the disaggregation model to simulate high resolution meteorological variables for analyzing future flood events.

1.3 Outline of the thesis

In this thesis, the focus is on developing uncertainty quantification tool for integrated statistical downscaling model, disaggregation model and empirical hydrological model which have huge uncertainties due to incomplete knowledge about the climate system and hydrological processes. New methods are developed to couple uncertainty quantification tool with statistical downscaling models and data-driven hydrological models. A novel methodology for rainfall occurrence determination and precipitation amount estimation (single site and multiple sites) in downscaling is also presented.

Chapter 2 presents the stochastic error coupling to quantify model and parameter uncertainty simultaneously in single site downscaling approaches. This coupled stochastic error enables simulation of ensembles as model output instead of point estimates. The proposed methodology consists of a two-step classification approach where the KNN classification model is used to determine the wet days and the wet days are further classified into two classes based on the rainfall amount. Gaussian Process Regression (GPR) is then calibrated for each class of rainfall separately to predict rainfall amount for each class. The proposed approach is implemented for Singapore using Climate Forecast System Reanalysis (CFSR) dataset. The comparison results of the proposed approach with the existing models such as Automated Statistical Downscaling (ASD) and GLM are also presented.

In Chapter 3, an approach combining Gaussian Process Classification (GPC) and GPR is proposed for rainfall occurrence determination and rainfall amount estimation for single site precipitation downscaling. This approach takes uncertainty from the classification of wet and dry days into account in downscaling precipitation. This approach is used to downscale precipitation from two sets of large scale predictors such as CFSR and the Second Generation Earth System Model (CanESM2) data. The prediction results from these two data are compared to analyze the uncertainty from large scale climate data and emission scenarios in simulating precipitation.

Chapter 4 proposes a multisite statistical downscaling model framework based on multi-output GPC and multi-output GPR to determine the wet days and the precipitation amount along with uncertainty quantification at all sites simultaneously. The proposed approach is used to downscale the precipitation at multiple sites in Singapore using HadCM3 A2 emission scenarios. The downscaling model is integrated with disaggregation model to simulate hourly precipitation at multiple sites. The future climate scenarios simulated from the downscaling and the disaggregation models are also presented.

In Chapter 5, GP based DDM for uncertainty quantification is proposed. The meteorological variables such as minimum temperature and precipitation are simulated using the integrated multisite statistical downscaling methods proposed in Chapter 4 and disaggregation (KNN) model; the proposed DDM is used to generate future river flow for a watershed in Quebec, Canada using disaggregated climate variables. The potential impact of climate change on hydrology especially extreme events such as flood using the simulated runoff magnitude from the data driven hydrological model is assessed. The DDM technique is similar to the GP used for single site downscaling as both are regression based downscaling models.

Chapter 6 presents the conclusions to each chapter. This chapter also presents the future directions of this thesis work.

CHAPTER 2 KNN-BUQSDM - A Bayesian Updating Uncertainty Quantification Framework for Statistical Downscaling of Precipitation

2.1 Abstract

The content of this chapter is extracted from the submitted journal paper. An approach integrating KNN and GPR is developed to quantify uncertainty in regression based statistical downscaling model structure and its parameters and residuals within a single framework. The proposed method is named as K-nearest neighbor-Bayesian Uncertainty Quantification for Statistical Downscaling Model (KNN-BUQSDM). KNN is implemented to determine the occurrence of the precipitation. This step is followed by classification of rainfall into clusters based on the rainfall magnitude. Subsequently, the precipitation amount is determined for each cluster using GPR. GPR is a non-parametric regression model (developed from Bayesian statistics) which captures model inadequacy and quantifies uncertainty in the model structure and the residuals simultaneously. In GPR, the modelling function is assumed to be a GP with a mean and a covariance function; the model errors are dependent (correlated) and they are treated as a stochastic process following Gaussian distribution.

The uncertainty in the model and the residuals are determined by the prior distribution assumption over the functions rather than the model's parameters. The prior over the functions is updated using the likelihood function to obtain the full posterior distribution of the function. The marginal likelihood of the model is the objective function. The gradient based optimization technique is used to optimize the objective function to find the optimal model function and residual parameters. The posterior predictive distribution of the function is obtained by integrating over the posterior distribution of the function. Since the model follows multivariate Gaussian distribution, the posterior predictive distribution is also the joint distribution of the training and the prediction outputs. The significance of GPR over other methods for estimating

precipitation amount is the uncertainty quantification of the predictions in the form of confidence interval which is essential for climate change impact studies. This enables the model to simulate downscaled future precipitation ensembles directly to assess the reliability of the predictions for decision making. The precipitation is downscaled from the coarse resolution CFSR predictors for Singapore to assess the prediction ability of KNN-BUQSDM. The proposed statistical downscaling model is implemented for each month in order to capture the monthly variations. It is shown that KNN-BUQSDM provides better accuracy in the downscaled precipitation compared to ASD, GLM and KNN-BNN. The results show that the statistics such as mean, standard deviation, proportion of wet days, 90^{th} percentile and maximum of the predicted ensembles from KNN-BUQSDM are closer to the observed monthly statistics for most of the months. There is also a significant reduction noticed in the monthly Mean Squared Error (MSE) when compared to existing downscaling techniques such as ASD, GLM and K-Nearest Neighbour-Bayesian Neural Network (KNN-BNN).

2.2 Introduction

One of the major impacts of climate change on hydrological cycle is that there is an increase in extreme rainfall in a short period of time (Madsen *et al.*, 2009). The Intergovernmental Panel on Climate Change (IPCC) Assessment Report assessed the future climate scenarios using different GCM scenarios and concluded that there would be a high chance of continuation of increase in extreme rainfall in the future (Solomon *et al.*, 2007). Such a huge amount of rainfall water causes flood events that exceed the designed drainage capacity which results in large social and economic damage and losses (Willems, 2012). The large scale climate predictors (e.g. GCM, Regional Climate Model (RCM) and reanalysis data such as National Centers for Environmental Prediction (NCEP) or CFSR are utilized in climate change analysis. However, the resolution of large scale predictors is so coarse that it cannot be used for planning and designing drainage system in an urban area. In an urban area like Singapore, large scale predictors which are reliable and have high resolution are required to analyze the

change in future precipitation amount. This in turn is critical for studying climate change impact on hydrology and decision making accordingly.

The SDMs are a bridging models between the large scale climate predictors and local climate variables such as precipitation to increase the resolution of future precipitation predictions. This is needed for decision making and planning in case of extreme events (Wilby and Wigley, 1997). The knowledge about the physics that drive and link the large scale climate system and the local climate system is not known. This causes uncertainty in the downscaled predictions. The challenge in downscaling precipitation is that the reliability of future precipitation predictions cannot be verified. A comprehensive uncertainty quantification framework to handle such problems is needed. In particular, the uncertainty in the calibration of the model, residuals, validation of the model and predictions need to be treated simultaneously. In this chapter, the development of statistical downscaling technique based on Bayesian technique proposed by Cheung *et al.* (2011) to quantify uncertainty in SDM is considered. Due to the limitations in uncertainty quantification methodology and lack of availability of a specific uncertainty quantification framework for SDM formulation (as presented in Table 2-1), a novel method is proposed in this study to improve the confidence and accuracy of the predictions from SDM. The calibration of SDM is formulated as a Bayesian inference problem in which a posterior probability distribution for model parameters is computed by conditioning the future predictors on the historic predictors and predictand using the prior knowledge and the information from the observed data. This formulation helps to represent uncertainties arising from model inadequacy (or structural uncertainty which arises from the lack of knowledge about the underlying physics) by expressing the SDM as a joint probability distribution of the predictors and predictand. The ensembles of the downscaled predictions are computed by using the posterior predictive distribution of the parameters and the joint distribution of the model. The following section explains the issues in the existing regression based statistical downscaling techniques followed by the description of proposed Bayesian framework for statistical downscaling.

2.2.1 Review of regression based SDM

The detailed review of downscaling methods and uncertainty quantification for hydrological applications can be found in Fowler *et al.* (2007) and Maraun *et al.* (2010). The readers can refer to Hessami *et al.* (2008) and Chandler and Wheater (2002) for detailed discussion about ASD and GLM respectively. The following section discusses about the important issues that need to be addressed in the model structure in the existing statistical downscaling models.

Let the historic data for the precipitation occurrence determination model be, D_c of n observations, $D_c = \{(\mathbf{x}_{ci}, y_{ci}) \mid i = 1,...,n\}$ where $\mathbf{x}_c$ represents the GCM predictors with dimension m_c and $\mathbf{y}_c$ is the binary classification output (wet/dry day). In vector form, the historic data can be represented as a matrix X_c with dimension $m_c \times n$ and the wet/dry day classification output vector is denoted as $\mathbf{y}_c$. The historic data D_a for the precipitation amount estimation model of n observations is represented as $D_a = \{(\mathbf{x}_{ai}, y_{ai}) \mid i = 1,...,n\}$ where $\mathbf{x}_a$ are the GCM predictors with dimension m_a and $\mathbf{y}_a$ represents the rainfall amount. In vector form, the predictor data can be represented as a matrix X_a with dimension $m_a \times n$ and the wet/dry day classification output vector is denoted as $\mathbf{y}_a$. The regression model can be written as (2.1):

$$\mathbf{y}_a = f_a(X_a) \tag{2.1}$$

where f_a is modelling function which can be linear/non-linear which links the predictors and predictand. In traditional regression models, the function describes the mean of the predictand, which can be written as (2.2).

$$f(X_a) = X_a^{\mathrm{T}} \mathbf{b}_a \tag{2.2}$$

where $\mathbf{b}_a$ represents the coefficient of the regression model.

Table 2-1 Comparison of ASD, GLM and KNN-BUQSDM statistical downscaling model

	ASD	GLM	KNN-BUQSDM
Assumptions	Linear relationship between the predictor and predictand	Linear relationship between the predictor and predictand	Linear or non-linear relationship between the predictors and predictand can be captured.
	Predictand is assumed to follow Gaussian distribution	Predictand can be of any distribution (e.g., Gamma (continuous) or Bernoulli (discrete))	Predictand is assumed to follow Gaussian distribution
	Residuals are assumed to follow Gaussian distribution. Errors are independent.	Pearson residuals are obtained. Errors are independent.	Residuals are assumed to follow GP. Errors are dependent.
	Least square fitting to estimate the model parameters	Iterative reweighted least squares (IWLS) to estimate the model parameters	Bayesian inference to estimate model function parameters and residual parameters simultaneously
	Parametric model	Parametric model	Non-parametric model
	Precipitation occurrence and precipitation amount are computed using Linear regression.	Precipitation occurrence is determined by assuming Bernoulli distribution to $\mathbf{y}_c$ and precipitation amount is estimated by assuming Gamma distribution to $\mathbf{y}_c$. Precipitation amount is estimated by assuming Gamma distribution.	Hybrid model of weather types and regression to reduce uncertainty. KNN (non-parametric) is used to determine the precipitation occurrence and classifying rainfall types. GPR is used to estimate the precipitation amount which is a predictive posterior distribution.
	Predictions are not probabilistic	Predictions are not probabilistic	Prior is assumed over the model function. Thus, the posterior distribution of the model function is obtained. The posterior predictive mean and variance are used to simulate the downscaled precipitation ensembles
Ensemble simulations	Errors are simulated by fitting the residual with an extreme value distribution (random part) and added to systematic part to get realisations	Errors are simulated by fitting the residual with an extreme value distribution (random part) and added to systematic part to get realisations	No need to implement separate error fitting as the output of the model is the predictive mean and predictive variance which can be used to simulate ensembles

Limitations	Assumes different distribution for model parameter calibration and error simulation. Model parameter and Model errors are calibrated separately Model parameter and Model errors are calibrated with different distribution assumptions (that is, Gaussian for the former and Extreme Value distribution for the latter) Predictands must be independent Only linear model function can be used. Predictand need to be transformed to make the distribution to follow Gaussian distribution.	Assumes different distribution for model parameter calibration and error simulation. Model parameter and Model errors are calibrated separately Model parameter and Model errors are calibrated with different distribution assumptions (that is, Gaussian for the former and Extreme Value distribution for the latter) Predictands and errors are independent Only linear function can be used	The precipitation needs to be transformed before using for downscaling. This can be solved by assuming non-Gaussian likelihood to compute posterior distribution for the systematic and random part. (see discussion section) In large scale computation, sparse approximation is needed to reduce the computation cost (Banerjee *et al.*, 2012).

The function in (2.1) is assumed to be *deterministic* (that is, it yields 'point estimates' to the output $\mathbf{y}_a$ when the model function parameters are known). As the downscaled precipitation only using the function (2.1) cannot reproduce the variance and the extreme values, Karl *et al.* (1990) suggested 'inflated regression method' (to scale the variation). The inflated regression method could not represent the local variation completely leading to the development of additive randomized error approach by Zorita and Storch (1999). In this approach, the error is added to the modelling function as shown in (2.3):

$$\mathbf{y}_a = \underbrace{f_a(X_a)}_{epistemic} + \underset{aleatory}{\boldsymbol{\varepsilon}_a} \tag{2.3}$$

where $\boldsymbol{\varepsilon}_a$ is the random error (that is, *stochastic*) or residual noise in the model and the error is assumed to follow zero mean and Gaussian noise $\boldsymbol{\varepsilon}_a \sim N(0, \sigma_n^2)$. In linear regression model for statistical downscaling, the functional form in (2.3) describes the linear relationship between the predictor and predictand. Multiple Linear Regression

(MLR) is an example for linear function which is being used in SDSM (Wilby *et al.*, 2002) and ASD model (Hessami *et al.*, 2008). The example for non-linear regression function is ANN (McCulloch and Pitts, 1943) which was first proposed by Cavazos and Hewitson (2005) for downscaling and since then has been widely used for downscaling (Olsson *et al.*, 2001).

The function and the residual of SDM are calibrated separately to estimate the parameter values. At first, the precipitation is modelled in two steps 1) Precipitation occurrence determination (determining wet or dry days) and 2) precipitation amount estimation. After determining the future wet-days using a classification method, the downscaling model function is calibrated using the wet days of historical data to obtain the optimal parameter values of the function. The point estimates of the predictand for the wet days in future are then obtained using optimized parameters in the model function. In the downscaling model (2.3), the daily precipitation is assumed to follow Gaussian distribution. The problem with this assumption is that the precipitation at a smaller time scale such as monthly or daily does not follow Gaussian distribution. Anscombe transformation is suggested as one of the solutions to this issue (Yang *et al.*, 2005) to make the distribution of precipitation closer to the normal distribution. This transformation was applied in Multivariate Multiple Linear Regression (MMLR) for downscaling precipitation at multiple sites (Jeong *et al.*, 2012). SDSM facilitates other transformations such as Fourth root, Natural log and Inverse normal (Wilby *et al.*, 2002). In ASD, fourth root transformation is used to make the precipitation closer to Gaussian (Hessami *et al.*, 2008).

GLM was proposed to model precipitation using Gamma distribution instead of assuming Gaussian distribution (Katz, 1977). GLM is generalized representation of the linear models (Dobson, 2001) and it has been implemented as statistical downscaling model in many research studies (Yang *et al.*, 2005). Capturing the non-normal error distribution in the model is possible with the availability of GLM. The GLM can be modelled using the link function that relates the predictand and the predictors through appropriate distribution. The precipitation is modelled in two steps, determining the

precipitation occurrence using Bernoulli distribution with logit link and estimating the amount of precipitation using Gamma distribution with logarithmic link function (Chandler and Wheater, 2002).

The 'Residue Analysis' is performed as a second step in SDM to simulate ensembles of random noise which is then added to the point estimates of the function $f_a(X_a)$ obtained from regression to compensate for the discrepancies in the observed precipitation versus predicted precipitation; this step helps to simulate reasonable extreme values (Chandler and Wheater, 2002; Wilby *et al.*, 2002; Hessami *et al.*, 2008; Lu and Qin, 2014). The residual obtained in the calibration stage is fitted by using an extreme value distribution for example, Generalized Extreme Value (GEV) distribution (Lu and Qin, 2014) to estimate the parameters for the residuals fitting. This step can also be referred to Uncertainty Quantification (UQ) since the simulation of errors helps to obtain the confidence interval of the model predictions. The confidence interval also helps to evaluate the reliability of SDM predictions.

There are mainly two drawbacks in the aforementioned regression based statistical downscaling models formulation. The covariance of the errors is expressed as (2.4):

$$\operatorname{cov}\begin{bmatrix} \varepsilon_{a1} \\ \varepsilon_{a2} \\ \vdots \\ \varepsilon_{an} \end{bmatrix} = \sigma_n^2 I_{n\times n} = \begin{bmatrix} \sigma_n^2 & 0 & 0 & 0 \\ 0 & \sigma_n^2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \sigma_n^2 \end{bmatrix} \tag{2.4}$$

In the equation (2.4), the off-diagonal terms of the error matrix are zero since the errors are assumed to be independent. However in real situations, the errors are not independent and this is referred as '*serial correlation*' or '*autocorrelation*'. Chandler and Wheater (2002) proposed GLM in which the autocorrelation is modelled by including the covariates of the previous days' within the model that determines the precipitation occurrence and estimates the precipitation amount. This makes the assumption of independently distributed errors (e.g., Pearson residuals) in GLM valid.

In the advanced statistical models developed in the machine learning (Williams, 1999; Rasmussen and Williams, 2006) and Geostatistics (Krige, 1951; Isaaks and Srivastava, 1989) it has been proved that the regression model can be implemented by assuming correlated errors and also the accuracy of the results can be improved. When the errors are assumed to be dependent, the cross-covariance terms are represented in the off-diagonal elements of the error matrix in equation (2.4). The dependent error assumption gives the flexibility to include the serial correlation implicitly in the model calibration as mentioned before. The cross-covariance terms are estimated by using a covariance function (e.g. squared exponential covariance function) in equation (2.21). The assumption and representation of errors in SDM need to be modified to quantify uncertainty by coupling the uncertainty quantification tool with the downscaling model function.

The second issue is that the different distribution is assumed for function parameter estimation, $f_a(X_a)$ (Gaussian distribution or Gamma distribution) and residual parameter estimation, $\boldsymbol{\varepsilon}_a$ (Generalized Extreme Value distribution) to simulate ensembles. The challenge with this assumption is that there is a need to calibrate twice to estimate the function parameters and the residual fitting parameters. The underestimation of extreme precipitation values predicted by the regression based SDM are compensated by adding the random errors simulated using GEV. There is also lack of consistency in the distribution assumption between the model function and the error function. The statistical downscaling model needs to be implemented with same distribution assumption for the model function as well as for the residual analysis (that is, ensemble simulation) (Beck and Katafygiotis, 1991; Beck and Katafygiotis, 1998; Rasmussen and Williams, 2006; Cheung *et al.*, 2011) to improve the model performance. An efficient framework to treat the above mentioned issues in SDM requires methodology and ideas from other disciplines such as structural dynamics (Beck and Katafygiotis, 1991; Beck and Katafygiotis, 1998; Beck and Cheung, 2009) and machine learning (Rasmussen and Williams, 2006). The following section describes the sources of uncertainties in SDM and the Bayesian inference based SDM

framework that resolves the above mentioned issues along with uncertainty quantification.

2.2.2 Sources of uncertainty in SDM

The purpose of UQ is to develop a methodology to characterize and reduce the uncertainty and variability on the model output from the model input. UQ helps to determine the probability of the outcomes even with the less available information about the real world system that is being modelled. In the case of SDM, the uncertainty in the model structure and random noise in the data causes variations in the predictions. There are two types of uncertainties namely aleatory and epistemic uncertainties (Matthies, 2007; Kiureghian and Ditlevsen, 2009).

Aleatory uncertainty is also called irreducible uncertainty or inherent uncertainty or in some instances as stochastic uncertainty and is due to natural variability and randomness. This type of uncertainty cannot be reduced. The noise in the data causes aleatory uncertainty which is described by $\boldsymbol{\varepsilon}_a$ in (2.3) (Rasmussen and Williams, 2006; Hermkes *et al.*, 2014). A residue analysis is used to simulate random noise as explained in the previous section. There are limited studies on reducing aleatory uncertainty in downscaling due to noise in the data as the similar error structure is assumed in all the regression based SDM.

In order to reduce aleatory uncertainty and to improve the downscaling model performance for studying the impact of climate change on hydrology, Ghosh and Katkar (2012) proposed multiple downscaling methods and implemented for the climate data from North-east India. They classified monthly rainfall into three groups based on the magnitude (that is, low, medium and high) using Classification and Regression Tree method (CART) and compared the estimated the precipitation amount using three regression methods such as linear regression, ANN and Support Vector Machines (SVM). Their results showed that SVM predicted high magnitude rainfall better and ANN was better in predicting the low and medium monthly rainfall. Since

the applicability of multiple downscaling methods by Ghosh and Katkar (2012) was not explored for daily rainfall, Lu and Qin (2014) proposed KNN-BNN for multi-class downscaling to improve predictions and reduce uncertainty for daily precipitation. In their work, they found out that the drawback of using residue fitting to compensate for model errors in regression-based statistical downscaling was that the extreme values were over- or under-estimated as the whole range of rainfall had been used for downscaling. This issue was resolved by classifying the rainfall into eight groups using KNN and the precipitation amount for each group of rainfall was estimated using BNN along with residue fitting. They also compared downscaling results using daily precipitation data for all the years and daily precipitation data for each month; the results showed that the downscaling model implemented for each month separately showed good performance. Their results also proved that multiclass downscaling using KNN-BNN approach reduced uncertainty and improved the extreme value prediction (Lu and Qin, 2014). However, their approach suffered from reduced sample size for some of the rainfall classes as the number of classes was high especially when implemented for daily precipitation for each month. Even though the classification of the rainfall types may improve the prediction results, the large number of classes in climate change impact study application is not desirable. An approach with reduced number of classes while being efficient in improving the predictions at the same time for statistical downscaling is needed.

Epistemic uncertainty is also called systematic uncertainty and can be caused by inaccurate or incorrect measurements, model structure and parameters, generalization of the real world and lack of knowledge. This type of uncertainty can be reduced with the new knowledge about the system being modelled. The epistemic uncertainty arises from the functional form $f_a(X_a)$ in (2.3) which is used to represent the relationship between the predictor and predictand. Since there are several types of the model function available, it is difficult to choose one function as the best for downscaling. There are many research studies on inter-comparison of the statistical downscaling models to analyze the uncertainty. Khan *et al.* (2006) compared SDSM, Long-Ashton

Research Station Weather Generator (LARS-WG) (Semenov *et al.*, 2002) and ANN to analyze the uncertainty contributed by each model. SDSM reproduced the observed statistics well compared to LASRS-WG and ANN. ANN showed poor performance when compared to other models (Khan *et al.*, 2006). Samadi *et al.* (2013) compared SDSM and ANN to downscale precipitation and temperature. Their study results showed that the uncertainty in the downscaled precipitation was high compared to temperature. It was also concluded that SDSM was efficient in reproducing the observed data of both temperature and precipitation compared to ANN. This study results are also in agreement with the results from ASD that there is a large uncertainty in the downscaled precipitation results. The Statistical and Regional dynamical Downscaling of Extremes for European regions (STARDEX) project compared the statistical downscaling models, dynamical models and statistical-dynamical downscaling methods to assess the downscaling model's ability in downscaling extremes (Goodess, 2003). The results of comparison of the STARDEX project suggested that there was no single method that can be chosen as the best model for downscaling. Thus, it is necessary to use a range of downscaling to get a wide range of uncertainties in the prediction results. All of the comparison results concluded that the performance of the statistical downscaling model in prediction varies with different GCM predictors, geographic location and modeling function used. Despite all these attempts to quantify uncertainty in SDM, the uncertainty contributed by the downscaling model function structure remains unaccounted. Fitting of residuals to generate ensembles has a direct consequence in the prediction results. The quantification of aleatory uncertainty needs to be given equal importance as epistemic uncertainty. Thus it is essential to develop a framework which estimates the amount of uncertainty contributed to downscaled precipitation by the choice of downscaling function and its parameters along with noise simultaneously. In summary, the three important issues in statistical downscaling that need to be improved are 1) Assumption of independency between the residuals 2) Implementation of the downscaling and error analysis separately and 3) The distribution for downscaling and residuals are assumed differently.

2.2.3 Uncertainty Quantification using GPR

GPR has several advantages to solve the issues with the precipitation amount estimation in regression based statistical downscaling models which have been mentioned above. The advantages of GPR (Rasmussen and Williams, 2006) are

1. It is possible to quantify uncertainty in the predictions both from the parameters and in the error terms. Hermkes *et al.* (2014) implemented GPR in Ground Motion Prediction Equation (GMPE) problem and showed that the GPR was efficient in quantifying aleatory and epistemic uncertainty simultaneously. Another issue with the large scale predictors is that the climate model grids of GCMs do not even cover the study area or the rain gauge station under consideration. GPR is proved to predict good results even with the availability of less calibration data (Rasmussen and Williams, 2006).
2. The errors in GPR are assumed to dependent and are assumed to be a stochastic process following Gaussian distribution. When the dependency between the errors is captured, the prediction accuracy can be improved.
3. The predictions from GPR are probabilistic which help to simulate ensembles to represent empirical confidence interval.
4. GPR is a non-parametric model. The examples for parametric models are linear regression models, GLM and non-linear regression models. When the functional form that defines the relationship between the predictors and predictand are known, the parametric models are efficient. In the case of SDM, the relationship that links the large scale predictors and local climate variables are not known. A model that is flexible and captures the non-linear relationship and the unknown function can be obtained by using non-parametric regression model. Non-parametric regression model does not predetermine the fundamental relationship between the predictors and predictand but adjusts the model to capture the non-linear features in the data.

The two types of computational problems involved are *forward problem* and *inverse problem*. In forward problem, the uncertainties in the input are propagated through the model to represent the uncertainty in the output given the parameters. The forward problem more is straightforward to implement as the parameters and the model function are known. In inverse problem, the unknown function and its parameters that best fit the output are estimated given the noisy observed data/inputs. Inverse problems are ill-posed as the small variation in the input may cause error in the estimates. The estimation of parameters is referred as *Model calibration* or *Model updating* (Cheung *et al.*, 2011).

SDM is considered as an inverse problem as the parameters that link the large scale GCM predictors and the local precipitation cannot be measured directly. A number of approaches have been developed to quantify uncertainty in the inverse problem and proved to be efficient in quantifying uncertainty in model calibration and prediction (Biegler *et al.*, 2011). Bayesian inference is adopted in this research study to solve the statistical inverse problem (that is, SDM) to obtain the posterior distribution of the model parameters and the output (MacKay, 1992). By Bayes' theorem (MacKay, 1992), the posterior pdf $p(\boldsymbol{\theta}_a \mid \mathbf{y}_a)$ of the model given the data D_a is expressed in (2.5):

$$p(\boldsymbol{\theta}_a \mid \mathbf{y}_a) = \frac{p(\mathbf{y}_a \mid \boldsymbol{\theta}_a)\, p(\boldsymbol{\theta}_a)}{p(\mathbf{y}_a)} \tag{2.5}$$

where $p(\mathbf{y}_a \mid \boldsymbol{\theta}_a)$ is the likelihood of the model, $p(\boldsymbol{\theta}_a)$ is the prior and $p(\mathbf{y}_a)$ is the marginal likelihood (evidence). The prior probability distribution accounts for the uncertainty in the model parameters (Beck and Katafygiotis, 1998; Cheung *et al.*, 2011). The expression (2.5) describes the posterior distribution of the parameters by utilizing the prior information along with the data providing a mechanism to deal with uncertainties in the model even though it can be computationally costly if not handled properly. The posterior predictive distribution represents the uncertainty in the predictions. The Bayesian framework provides an intuitive approach for uncertainty quantification where the prior belief about the system is updated using the new

knowledge from the data resulting in posterior distribution for the prediction coupled with uncertainty information. Hence, the application of Bayesian uncertainty quantification tool is critical in downscaling to project future climate scenarios and impact studies.

The probabilistic uncertainty quantification approach has gained interest in statistical downscaling recently. Ghosh and Mujumdar (2008) proposed a Principle Component Analysis (PCA) and fuzzy clustering integrated with Relevance Vector Machine (RVM) to downscale precipitation and to capture non-linear relationship between the climate variables and streamflow. Their approach can be considered as a hybrid of weather typing and transfer function. PCA was used to reduce the dimensions and the principle components were classified into clusters using fuzzy clustering. They used ten principle components and two clusters for downscaling using SVM and RVM respectively. Their research study compared RVM prediction results with SVM results. The results showed that RVM with fewer relevant vectors had lesser chance for over-fitting compared to SVM. An important property of RVM is the predictive distribution from which the uncertainty range of the predictions can be obtained. The precipitation distribution accounts for aleatory and epistemic uncertainty simultaneously. However, the RVM model was not efficient in capturing the extreme events. Rasmussen and Williams (2006) showed that the uncertainty in the future data prediction increases when the future data and the training data are far away. These problems can be solved by adopting advanced Bayesian technique to solve the above mentioned issues in statistical downscaling.

A comprehensive state of the art Bayesian uncertainty representation approach was implemented in the other fields (Vanik *et al.*, 2000). Cheung *et al.* (2011) proposed a Bayesian statistical framework in which model calibration, prediction and uncertainty quantification were coupled in a single framework and treated simultaneously. They also showed that the different errors for example, additive or multiplicative could be incorporated in the modelling.

The concept of GPR will be difficult to understand without the understanding of the fundamentals of Bayesian Linear Regression Model (BLRM). The readers can refer to Rasmussen and Williams (2006) for more details on the relationship of BRLM to GPR and the detailed explanation of GPR. In GPR, a kernel function is used to capture the non-linear relationship between the predictand and predictors (predictors refer to large scale predictors throughout this thesis) implicitly, instead of using high dimensional basis function in the model explicitly. This makes it easy to implement the non-linear regression using large dimensional kernel efficiently without computational difficulty. This is known as *kernel trick* and this idea has been implemented in machine learning widely. A natural method to get the full posterior predictive distribution along with kernel trick is also possible by placing the prior over the functions. This technique is called GPR.

GPs are defined as a collection of random variables based on the property that the joint distribution of any of its subset is joint Gaussian distribution (Gibbs, 1998; Rasmussen and Williams, 2006). The distinction between Gaussian distribution and GP is that the Gaussian distribution is a continuous probability distribution characterized by a mean vector and a covariance matrix. A GP is a generalization of the Gaussian probability distribution to infinitely many random variables. GP are stochastic processes characterized by their mean function and covariance function. In GPR, the prior is placed directly over the statistical model functions (O'Hagan and Kingman, 1978; MacKay, 1992; Neal, 1996; Bernardo *et al.*, 1998; Rasmussen and Williams, 2006). GP can also be considered as multilayer perceptron (neural network) with infinitely many hidden units and Gaussian priors on the weights (Neal, 1996). RVM can be viewed as a special case of GP; the correlations between the errors are not present in RVM (Tipping, 2003). The relationship between GP and RVM can be found in (Rasmussen and Williams, 2006):

$$f_a(X_a) \sim GP(m(X_a), K(X_a, X_a')) \tag{2.6}$$

The representation in (2.6) means that $f_a(X_a)$ is a GP with mean function, $m(X_a)$ and covariance function, $K(X_a, X_a')$. One needs to specify the mean and the covariance function to define GP. Generally, in a lot of applications the mean function is assumed to be zero for simplicity and also because there is no prior knowledge about the mean function (Bishop, 2006). The mean function can be specified explicitly if there is any information about the process being modelled. O'Hagan and Kingman (1978) extended developed GPR with mean function and this research follows their framework for implementing GPR. A linear mean function is adopted for downscaling in this study.

The covariance function is chosen based on the underlying assumption about the model such as smoothness, periodicity and stationarity. The covariance function $K(X_a, X_a')$ generates a positive definite covariance matrix. The squared exponential covariance function (2.21) is the most commonly used covariance function. The parameters of the covariance function are called *hyperparameters* and can be learned from the data. It can be seen that from the equation (2.21) that the covariance between any two inputs is closer to one, if the inputs are closer to each other and the covariance decreases exponentially when the distance between the inputs decreases. The parameters of the covariance function have significant effect on the performance of GP. The use of squared exponential covariance function is equivalent to the use of infinitely many Gaussian basis function located everywhere not just at the training data in regression.

A brief introduction to implement Gaussian Process for statistical downscaling was presented in (Rajendran and Cheung, 2015). In this study, KNN and GPR are integrated to develop statistical downscaling model named K-Nearest Neighbour-Bayesian Uncertainty Quantification Statistical Downscaling Model (KNN-BUQSDM). A case study to downscale daily rainfall using KNN-BUQSDM to rain gauge station scale in Singapore using CFSR will be presented for demonstration. A comparison of the proposed method with three existing statistical downscaling approaches ASD, GLM and KNN-BNN will be provided. The results for ASD, GLM, KNN-BNN are obtained

from (Lu and Qin, 2014) for comparison since the study area and the data are the same in this research.

2.3 Data and Study area

Singapore is located near equator and is classified as a tropical climate between 1° N and 2° N latitudes and 103.8° E and $104^{\circ}E$ longitudes. There is abundant rainfall with an average of 2331.2 mm, high and uniform temperature, and high humidity throughout the year. Singapore's climate is represented by two monsoon seasons - Northeast monsoon (December to early March) and southwest monsoon (June to September). Based on the historical rainfall record, December is the wettest month (average rainfall is 230 mm) and February is the driest month (average rainfall of 160 mm) (MSS, 2016; NEA, 2016). The rainfall is higher in the western parts compared to the eastern part of Singapore. Thus, two rain gauge stations located at the western part (S44) and eastern part (S24) are chosen for assessing the performance of KNN-BUQSDM. The selected rain gauge stations for downscaling had good quality of data for the study period (1980-2010). At S44 station, the percentage of missing dataset is 0.09% and at S24 station, the dataset is complete. Since the western region of rainfall has complex pattern of rainfall, the downscaling for precipitation for S44 is completely presented. Partial results of S24 are presented for comparison and to assess the consistency in the performance of the model. The observed precipitation data for the years from 1980 to 2010 are obtained from National Environmental Agency, Singapore. The calibration data period is from 1980 to 2004 and the validation data is from 2005 to 2010. The location of rain gauge stations and the CFSR grid is shown in Figure 2-1. The cubic root is used to transform the daily precipitation to make it closer to normal distribution (Yang *et al.*, 2005). The wet-day threshold is set to 0.1 mm. The prediction was tested with various threshold values. The results best agreed with the observed data when the threshold is set to 0.1 mm. The same threshold 0.1 mm was used in other studies such as (Liu *et al.*, 2011; Lindau and Simmer, 2013; Taye and Willems, 2013; Pervez and Henebry, 2014).

The large scale predictors which are obtained from the CFSR (Saha *et al.*, 2010) are used as climate predictors for downscaling precipitation in Singapore to capture the complexity in the climate. CFSR gives an estimate of coupled atmosphere-ocean-land surface-sea ice system with the spatial resolution of CFSR is $0.5^{\circ} \times 0.5^{\circ}$ and a global high-resolution 31-year period data. In terms of spatial and temporal resolution, the CFSR predictors (1-h instantaneous and $0.5^{\circ} \times 0.5^{\circ}$) are superior to National Center for Atmospheric Research (NCEP/NCAR) reanalysis data set (6-h instantaneous and $2.5^{\circ} \times 2.5^{\circ}$) (Liléo and Petrik, 2011; Wang *et al.*, 2011). Singapore is covered by two CFSR grids (that is, CFSR-East and CFSR-west). The central point coordinates of the grid are $1.5^{\circ}N, 103.5^{\circ}E$ (east) and $1.5^{\circ}N, 104^{\circ}E$(west). The available CFSR predictors for Singapore are listed in Table 2-2. CFSR grid predictors in the east are used for downscaling the precipitation for S24 rain gauge station and the CFSR grid predictors in the west are used for downscaling the precipitation for S44 rain gauge station.

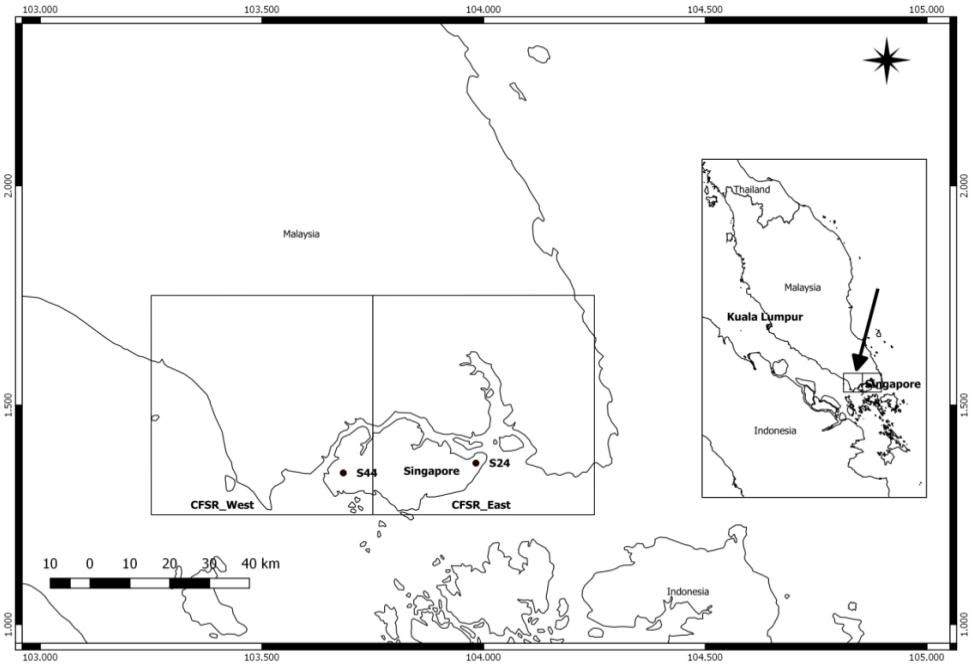


Figure 2-1 Rain gauge location and CFSR data for Singapore

The CFSR predictors are standardized before using it for classification and regression using (2.7):

$$z_i = \frac{r_i - r_{\min}}{r_{\max} - r_{\min}} \tag{2.7}$$

where r_i is the i^{th} input variable; $r_{\min}$ and $r_{\max}$ are the minimum and maximum values of the input variables, respectively.

Table 2-2 List of CFSR predictors for Singapore

CFSR predictors	Symbol
Mean sea level pressure	**prmsl**
Specific humidity at 500 hPa	**q500**
Specific humidity at 850 hPa	**q850**
Specific humidity at 925 hPa	**q925**
Geopotential at 500 hPa	**z500**
Geopotential at 850 hPa	z850
Geopotential at hPa	z1000
Zonal wind at 500 hPa	**u-w 500**
Meridional wind at 500 hPa	v-w 500
Zonal wind at 850 hPa	u-w 850
Meridional wind at 850 hPa	v-w 850

The selection of predictors is important to develop a reliable downscaling model and to get accurate prediction results (Wilby, 1998; Fowler *et al.*, 2007; Maraun *et al.*, 2010). The methods employed for selecting the predictors for determining the rainfall occurrence and estimating the rainfall amount are Two-sample Kolmogorov-Smirnov test and backward stepwise regression, respectively. The predictor that shows the significant difference between a dry day and a wet day is chosen in the Two-sample

Kolmogorov-Smirnov test by analyzing the sensitivity of the predictor to a wet day and a dry day in the model (Chen *et al.*, 2010). The selection of predictors using backward stepwise regression (Hocking, 1976) is similar to the predictor selection method in ASD. In this study, the predictors for regression need to be selected for each group of rainfall for implementing KNN-BUQSDM. The types of stepwise regression include forward selection, backward elimination and bidirectional elimination. Backward stepwise regression is used in this study to select predictors for rainfall amount estimation. As a first step, the selection methodology includes all the predictors in the large scale predictors in the model. This is followed by eliminating the predictors which are not significant step by step until the predictors that are significant in the model remain (Hessami *et al.*, 2008). The elimination of predictors is based on F-test which can be accomplished using (2.8):

$$F = \frac{\left(R_p^2 - R_{p-1}^2\right)\left(n-p-1\right)}{1-R_p^2} \tag{2.8}$$

where n is the number of observed data; p is the number of predictors; R_q is the correlation coefficient between the criterion variables and the predictions with p predictors. The predictors should be removed, if the F values are smaller than a threshold. The threshold for the test is calculated by using equation (2.9) (Hessami *et al.*, 2008).

$$a = 1 - \left(1 - \frac{a}{2}\right)^{1/p} \tag{2.9}$$

where a is the significance level which is 95% in this study.

In GLM, the likelihood ratio test (or comparison of deviances) is used to select the input predictors (or covariates) (Chandler and Wheater, 2002). The comparative results for predictor selection by Lu and Qin (2014), show that the backward stepwise regression, ASD and GLM method of predictor selection lead to the same set of

predictors for regression. The similar performance in choosing the predictors by all the three methods is due to complex rainfall pattern and the availability of limited number of predictors in CFSR database for Singapore.

The predictor that holds relevant information and is efficient in predicting precipitation is necessary to be included in implementing the downscaling model. Since the downscaled precipitation is used to project climate change, the predictors that capture the variation in global warming should be selected (Wilby, 1998). The humidity predictor contains the information about the capacity of the atmosphere to hold water under global warming and the temperature predictor contains the information about the long-term changes in the precipitation (Wilby and Wigley, 1997). Xoplaki *et al.* (2000) and Cavazos and Hewitson (2005) found that the precipitation and the large scale 500 hPa geopotential height were correlated in separated studies. The relationship between sea-level pressure and precipitation consistently link in different time scales such as seasonal, monthly and daily scales in the study conducted by Thompson and Green (2004). In order to downscale precipitation in Taiwan, Chen *et al.* (2010) used the predictors such as humidity predictors (R500, R850, Rhum), vorticity predictors (P8_z and P_z) and 850 hPa Geopotential height (P850), velocity (P5_v) and divergence (P5zh). The study by Lu and Qin (2014) used Mean sea level pressure, Specific humidity at 500 hPa, Specific humidity at 850 hPa, Specific humidity at 925 hPa, Geopotential at 500 hPa and Zonal wind at 500 hPa for downscaling precipitation in Singapore. Based on the studies that analyze the physics based relationship between large scale predictors and precipitation, the predictors selected for our investigation are highlighted in the Table 2-2. As the results of this investigation are compared with Lu and Qin (2014) care is taken to use almost similar predictors as theirs for fair comparison of the model performance.

2.4 K-Nearest Neighbour – Bayesian Uncertainty Quantification Statistical Downscaling Model (KNN-BUQSDM)

The objective of this study is to develop an uncertainty quantification approach for SDM. The proposed approach integrates the KNN, K-means clustering technique and GPR to improve the prediction ability of the model. Figure 2-2 depicts the flowchart for KNN-BUQSDM. The implementation algorithm of KNN-BUQSDM is explained in Algorithm 1 and Algorithm 2. The various steps involved in the predicting and quantifying future rainfall are as follows:

Step 1: Use KNN classification technique to determine the occurrence of rainfall (dry and wet days). This chapter uses KNN for occurrence determination and classification of rainfall types. The choice of model for occurrence determination is based on the previous research literature which used the same study area and the data as in the paper by Lu and Qin (2014). In their study, they compared ANN, Linear Classification (LC) and KNN to determine the wet and dry days. The prediction is assessed based on the accuracy and the dry day proportion. Their results show that KNN gives better accuracy and dry day proportion compared to ANN and LC.

Step 2: Perform K-means clustering to identify the threshold for classifying the wet days (determined in the previous step) into rainfall types. The number of rainfall types and a threshold for each group of rainfall are computed using K-means algorithm. KNN classification is then used to classify the wet days into groups using the threshold from the K-means algorithm.

Step 3: Calibrate the GPR model for each rainfall group corresponding to the wet days to estimate the hyperparameters of the model. The calibrated hyperparameters are then used to compute the prediction ensembles for the future climate predictors for each rainfall type. The dry days and the ensembles of predicted rainfall types are then combined to get the downscaled precipitation realisations. The detailed description of

ASD, GLM and KNN-BNN can be found in Hessami *et al.* (2008), Chandler and Wheater (2002) and Lu and Qin (2014), respectively.

2.4.1 K-Nearest Neighbour (KNN) for occurrence modelling

KNN was first implemented for precipitation occurrence determination and rainfall type classification in statistical downscaling by Lu and Qin (2014). KNN based simulator was also used by Rajagopalan and Lall (1999) to simulate daily precipitation and other weather variables. KNN classification is a supervised and non-parametric classification method (Altman, 1992). KNN approach uses the training data directly to classify the future data. It is based on the assumption that the data are related to each other. In order to classify the unknown data, the algorithm searches for its K-nearest neighbors based on the distance in the training data to assess the similarity metric. The

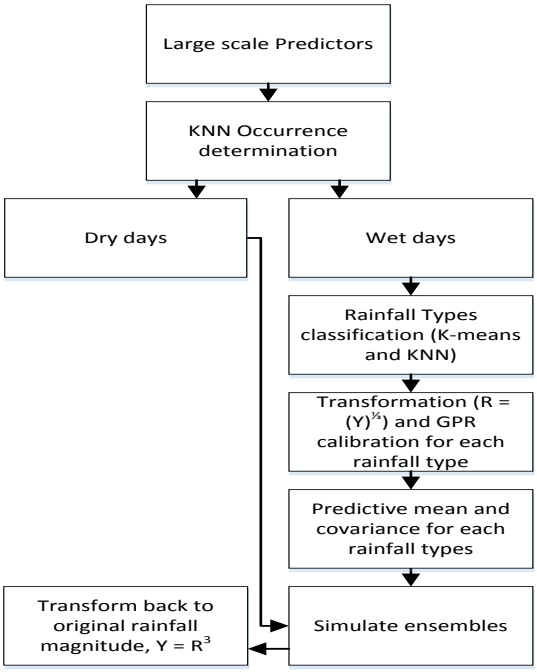


Figure 2-2 KNN-BUQSDM Framework

class to which the majority of the K-nearest neighbors belong in the training data is assigned to the future data. There are three types of distance functions used such as

Euclidean, Manhattan and Minkowski (Celisse and Mary-Huard, 2012). Let $\mathbf{x}_{ci}$ be an input training data with the dimension m_c, where $(i = 1,...,n)$. The Euclidean distance between $\mathbf{x}_{ci}$ and $\mathbf{x}_{cj}$ $(j = 1,...,n)$ is used in this study and is expressed in (2.10):

$$D_{EUC} = \sqrt{(x_{c1} - x_{c1})^2 + (x_{c2} - x_{c2})^2 + ... + (x_{cm_c} - x_{cm_c})^2} \tag{2.10}$$

KNN is used to determine the precipitation occurrence (wet/dry) as well as the classification of rainfall types. The choice of k affects the performance of the classification algorithm. If k is too small, the noise in the data causes over fit of the KNN classifier. The classifier may not classify the data accurately when k is too large; as the nearest neighbour contains the data that are far away from their neighbour (Celisse and Mary-Huard, 2012). The values of k that are suitable to this study lies between 1 to 20 (Nieminen *et al.*, 2012). The steps to identify the proper k values as summarised in Lu and Qin (2014) are as follows.

The k value can be estimated using two approaches such as 1) by estimating expected proportion and 2) by trial and error method. As the first step, 1) compute the Expected Proportion Interval (EPI) using 'simple moving average (SMA)' (Bali *et al.*, 2007). SMA given in (2.11) is used in this study to calculate EPI; 2) using the observed precipitation in the calibration period, compute an Expected average proportion of rainfall number (EPRN) for each rainfall type; 3) calculate the ratio of the calculated rainfall number in a class to the total number of rainfall events over the required time period which is referred as Calculated Proportion of Rainfall Number (CPRN). In single K- scheme, the k value that produces the closest distance between CPRN and EPRN is selected and also single K- scheme gives lesser uncertainty in the results (Lu and Qin, 2014). Thus, single K- scheme is used in this study to classify wet/dry days and rainfall types. The expressions for EPI, EPRN and CPRN are given in (2.12), (2.13) and (2.14), respectively.

$$\mathrm{SMA}(e_{yr}) = \frac{\sum_{i=1}^{n_v} o_{yr-i} + 1}{n_v} \tag{2.11}$$

$$\mathrm{EPI} = \{\min(e_{yr+1}, e_{yr+2}, ..., e_{yr+n_v}), \max(e_{yr+1}, e_{yr+2}, ..., e_{yr+n_v})\} \tag{2.12}$$

$$\mathrm{EPRN} = \frac{\sum_{i=1}^{n_v} e_{yr+i}}{n_v} \tag{2.13}$$

$$\mathrm{CPRN} = \frac{\sum_{i=1}^{n_v} c_{yr+i}}{n_v} \tag{2.14}$$

where n_v is the number of years in the validation period, e_{yr} is the proportion of wet days in the observed data, c_{yr} is the wet days proportion in the KNN classification results, o_{yr} is the proportion of wet days in yr^{th} year and o_{yr+1} is the expected proportion at the $(yr+1)^{th}$ year, $(i=1,...,n_v)$.

2.4.2 K-means clustering

K-means clustering is a non-hierarchical clustering algorithm and was first used by MacQueen (1967) to classify similar data points together. The advantage of k-means is its ease of implementation to find clusters. This algorithm classifies the data into K_c clusters in a way that each data point in the clusters looks similar to each other and dissimilar to the data in other clusters. Each cluster is defined by its mean (or center or centroid). K-means is an unsupervised classification method; it requires the number of clusters (or classes) to be specified in advance.

In statistical downscaling, K-means was first introduced and implemented by Kannan and Ghosh (2011) to identify rainfall states. The k-means clustering is also used to extract the spatial pattern of rainfall (Pelczer and Cisneros-Iturbe, 2008). Let the observations be $\mathbf{x}_{c1}, \mathbf{x}_{c2}, ..., \mathbf{x}_{cn}$ and each observation can be a m_c dimensional real

vector. The aim of K-means clustering (or classification) is to divide n observations into K_m clusters. The clusters $C = \{C_1, C_2, ..., C_{K_m}\}$ are obtained by minimizing the sum of distances of each observed data point in the cluster to the K_m cluster center. Mathematically it is represented as (2.15):

$$J = \sum_{l=1}^{K_c} \sum_{i=1}^{n} || \mathbf{x}_{ci}^{(l)} - m_{cgl} ||^2 \tag{2.15}$$

where m_{cgj} is the cluster centroid, K_m is the number of clusters, n is the number of observations, $\|.\|$ is the Euclidean distance function and J is the function to be minimized.

K-means is an iterative algorithm where the data points (or feature vectors) move from one cluster to another by minimizing the objective function (2.15). The final results of the algorithm are sensitive to the initial number of clusters. The function (2.15) is minimized by alternating between two steps to obtain clusters (MacKay, 1992). The steps involved in implementing k-means clustering and validation are summarized in Kannan and Ghosh (2011) as follows:

1) Initialize cluster centroids $m_{cg1}, ..., m_{cgK_m}$. Assign each observed data point to the cluster with nearest centroid. This is achieved by calculating the Euclidean distance between the observed data and the cluster mean. The data are assigned to the cluster to which the strength of the similarity is the highest compared to the other cluster.

2) Recalculate the new clusters as the centroids of the new clusters using the observed data and compute the function value J using (2.15). For each iteration, compute the difference in the function value J. If there is a change in the function value, iterate steps 1 and 2, else terminate the iteration.

2.4.3 Cluster validation

The problem with clustering is that the number of clusters in the data is not known. This issue is solved by checking the validity of the number of clusters using indices to assess the quality of clustering. These indices give the measure of compactness, connectedness and separation of cluster partitions. In order to identify the correct number of clusters, the k-means algorithm is run with different number of clusters (for example, 2 to 10). The validation indices are computed for each cluster. The cluster number that yields the best indices is considered as the final number. The three commonly used cluster validation indices (Kannan and Ghosh, 2011) are 1) Dunn's index (Dunn, 1973) 2) the Davies- Bouldin index (Davies and Bouldin, 1979) and 3) Silhouette index (Rousseeuw, 1987).

a) Dunn's index

Dunn's index F_D is defined by the ratio between the minimum distance between the clusters to the maximum distance within the clusters. The Euclidean distance is used in this cluster. The Dunn's index is given by (2.16):

$$F_D = \frac{s_{\min}}{s_{\max}} \tag{2.16}$$

where $s_{\min}$ the smallest distance between two data points from different clusters and $s_{\max}$ is the largest distance between two data points from the same cluster. Dunn's index lies between $[0\ \infty]$. The cluster number for which the Dun's index is high is considered to be the best cluster.

b) Davies-Bouldin index

The Davies-Bouldin index, F_{DB} is defined as the function of the ratio of the sum of scatters within the cluster to the separation between the cluster and is given by (2.17). The Euclidean distance is used to calculate the distance in this study.

$$F_{DB} = \frac{1}{K_m} \sum_{i=1, i \neq j}^{K_c} \max\left(\frac{\sigma_i + \sigma_j}{s(m_{cgi}, m_{cgj})} \right) \tag{2.17}$$

where K_c is the number of clusters, σ_i is the average distance of all scatters in cluster i to the cluster center m_{cgi}, σ_j is the average distance of all scatters in cluster j to the cluster center m_{cgj} and $s(m_{cgi}, m_{cgj})$ is the separation between clusters m_{cgi} and m_{cgj}. The value of F_{DB} represents the compactness. When the F_{DB} values are small, the corresponding clusters are compact and their centers are far away from each other. The clusters that give smaller F_{DB} are considered the best.

c) Sillhoutte index

Sillhoutte index, F_S in equation (2.18) measures the similarity of the data in its own cluster (that is, cohesion) when compared to other cluster separation. The value ranges from -1 to +1. +1 for F_S represents that the data in the cluster match with their own cluster and are less matched to neighboring clusters. When 0 is assigned to F_S, it indicates that the data cannot be classified in one cluster. When the value of F_S is -1, then the corresponding number of clusters is not appropriate for the data.

$$F_S = \frac{b_i - a_i}{\max(a_i, b_i)} \tag{2.18}$$

For the i^{th} data point, the average distance to all other data points is given by a_i, the minimum of the average distance to all data points in all the clusters except the cluster containing the data points i is given by b_i. The distance is calculated using squared Euclidean index in this study. The Silhouette coefficient for the i^{th} data point is given by F_S. These three indices give an idea about how well the clusters are divided for the given number of classes.

2.5 Precipitation amount estimation using GPR

The proposed framework for precipitation amount estimation follows the methodology developed by Rasmussen and Williams (2006). As expressed in (2.6), the model function $f_a(X_a)$ is assumed to be a GP with mean function $\mu(X_a)$ and covariance function $K(X_a, X_a')$. $f_a(X_a)$ is also referred as latent function at the input points $\mathbf{x}_a$ in GP. The matrix form of $m(X_a)$ and $K(X_a, X_a')$ are $\boldsymbol{\mu}_a$ and K_a respectively.

$$\mathbf{y}_a = f_a(\mathbf{x}_a) + \boldsymbol{\varepsilon}_a, \boldsymbol{\varepsilon}_a \sim N(0, \sigma_{na}^2) \tag{2.19}$$

where $\boldsymbol{\varepsilon}_a$ is normally distributed with zero mean and variance, σ_{na}^2. In Gaussian Process view, the model function $f_a(\mathbf{x}_a)$ is assumed to follow GP with a mean function (2.20).

$$\boldsymbol{\mu}_a = \varphi(\mathbf{x}_a)^{\mathrm{T}} \boldsymbol{\beta}_a \tag{2.20}$$

where $\boldsymbol{\mu}_a$ is vector form of mean function, $\varphi(\mathbf{x}_a)$ is the linear or non-linear basis function and $\boldsymbol{\beta}_a$ are the coefficients for each of the vectors in the basis function. The relationship between the predictors and predictand is not always linear. Thus, the complicated non-linear mean functions can capture the non-linear relationship between the predictors and predictand to improve the prediction results. The basis function can be linear or for example, any order of polynomial function (O'Hagan and Kingman, 1978). The central tendency of the mean function is represented by the mean function.

The covariance matrix corresponding to the covariance function should be positive semi-definite. The shape and structure of the covariance between the predictors are described by the covariance function. The commonly used one is the Squared Exponential (SE) covariance function expressed in (2.21). The SE kernel depends on the distance for the dimension m_a, $\|x_a - x'_a\|_2^2$ between the inputs, where $\|.\|^2$ is the Euclidean Norm.

$$\mathbf{k}(\mathbf{x}_a, \mathbf{x}'_a) = \sigma_{fa}^2 \exp\{-\frac{1}{2}(\mathbf{x}_a - \mathbf{x}'_a)^T L_a (\mathbf{x}_a - \mathbf{x}'_a)\} + \sigma_{na}^2 \delta_{ij},\ \sigma_{fa}^2 \geq 0; \sigma_{na}^2 \geq 0; \quad diag(L_a) > 0 \tag{2.21}$$

where σ_{fa}^2 is the signal variance, σ_{na}^2 is the noise variance, δ_{ij} is kronecker delta function and L_a is the diagonal positive symmetric matrix consisting of the characteristic correlation length, l_a. The lengthscale characterizes the distance between inputs that change the function value significantly. The predictive variance moves away from the data points when the lengthscale is shorter and the predictions are correlated with each other. If the same characteristic lengthscale is assumed for all the dimensions, then L_a is expressed as (2.22). In this case, the contribution from each predictor is considered equal.

$$L_a = \begin{bmatrix} l_a^{-2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & l_a^{-2} \end{bmatrix} = l_a^{-2} I \tag{2.22}$$

where I is the identity matrix. When different characteristic length scale is assumed for each dimension, L_a is expressed as (2.23).

$$L_a = \begin{bmatrix} l_{a1}^{-2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & l_{am_a}^{-2} \end{bmatrix} \tag{2.23}$$

This type of lengthscale definition is also called automatic relevance determination (ARD), where the covariance function automatically computes input predictors for the model. This also gives the idea of suitable predictors for precipitation amount estimation. The detailed description of various covariance functions, its formulation, properties and derivatives can be found in Rasmussen and Williams (2006). The noise term σ_{na}^2 is added to the covariance function to model the noise in the output. The noise variance is present only in the diagonals of the covariance matrix as the noise process that affects the model predictions are random. The squared exponential kernel (SE) captures the uncertainty (epistemic) contributed by the predictors to the output. The noise variance captures the aleatory uncertainty in the model.

$$f_a(\mathbf{x}_a) \sim GP(\boldsymbol{\mu}_a, \mathbf{k}(\mathbf{x}_a, \mathbf{x}'_a)) \tag{2.24}$$

The GP prior has parameters called *hyperparameters* associated with the mean and covariance function. These hyperparameters are not known *apriori* and they need to be optimized. The hyperparameters that are needed to be learnt are $\boldsymbol{\theta}_a = \{\boldsymbol{\beta}_a, \sigma_{fa}^2, \sigma_{na}^2, l_{ad}\}, d = 1,...,l_{m_a}$.

In Bayesian inference, the important step is to choose the priors. The choice of priors over the parameters $\boldsymbol{\beta}_a$ affects the posterior distribution. Generally, the information about the prior is not known and is vague; in this case a non-informative and/or flat prior is assumed (Beck and Katafygiotis, 1998). In this chapter, a uniform prior is assumed to reduce the effect of the prior distribution on $\boldsymbol{\beta}_a$ over the posterior distribution. By Bayes' theorem, the integral likelihood multiplied by the prior yields the marginal likelihood of the model (2.25). For notational convenience, the vector form of $f_a(X_a)$ is represented as $\mathbf{f}_a$.

$$p(\mathbf{y}_a \mid X_a) = \int p(\mathbf{y}_a \mid \mathbf{f}_a, X_a)\, p(\mathbf{f}_a \mid X_a)\, \mathrm{d}\mathbf{f}_a \tag{2.25}$$

As the Gaussian prior is placed on the modelling function the log of the prior is expressed as (2.26):

$$\log p(\mathbf{f}_a \mid X_a) = -\frac{1}{2}\mathbf{f}_a^{\mathrm{T}}(K_a + \sigma_{na}^2)^{-1}\mathbf{f}_a - \frac{1}{2}\log \mid K_a + \sigma_{na}^2 \mid -\frac{n}{2}\log 2\pi \tag{2.26}$$

The likelihood $\mathbf{y}_a \mid \mathbf{f}_a$ follows $N(\mathbf{f}_a, \sigma_{na}^2 I)$. The log marginal likelihood is given in (2.27):

$$\begin{aligned} \mathrm{Q} = \log p(\mathbf{y}_a \mid X_a) = -\frac{1}{2}(\mathbf{y}_a - \boldsymbol{\mu}_a)^T (K_a + \sigma_{na}^2 I)^{-1}(\mathbf{y}_a - \boldsymbol{\mu}_a) - \frac{1}{2}\log \mid K_a + \sigma_{na}^2 I \mid \\ -\frac{N}{2}\log 2\pi \end{aligned} \tag{2.27}$$

The posterior distribution in equation (2.27) is the objective function. The values of the hyperparameters which maximize (2.27) are the optimal parameters of the model which means that they are the most probable parameters.

2.5.1 Parameter optimization

It is straightforward to simulate predictions for the future data once the covariance function and its parameters are known. The ability of GP to learn hyperparameters from the training data directly is considered as one of the major advantages of GP. This is possible because GP is a full probabilistic model. In order to learn the optimal values of hyperparameters, a principled method for inferring them from the data is needed.

The maximization of log marginal likelihood called type-II maximum likelihood method (Berger, 1985) is generally used to find the optimal parameter values. In several cases, the prior belief about the hyperparameters are known; this information is incorporated in modelling by placing the prior belief over the hyperparameters. The prior belief placed on the hyperparameters changes the posterior inference output (Snelson, 2008; Cheung *et al.*, 2011). In case of precipitation amount estimation, the prior information is not known. In this case, the prior information on the parameters

should have minimal influence on the Bayesian inference. Thus a non-informative uniform prior is placed over hyperparameters of model mean function. The uniform prior means that all possible values of hyperparameters have equal probability. The negative log marginal likelihood is minimized to give optimal estimates of the parameters. The expression for $\boldsymbol{\beta}_a$ can be obtained analytically by differentiating negative of (2.27) with respect to $\boldsymbol{\beta}_a$ and equating to zero.

$$\boldsymbol{\beta}_a = (X_a^{\mathrm{T}} K_a^{-1} X_a)^{-1} (X_a^{\mathrm{T}} K_a^{-1} \mathbf{y}_a) \tag{2.28}$$

The optimal parameters corresponding to the covariance function does not have analytical expression similar to the mean function and it is obtained by minimizing negative log marginal likelihood (2.27) using gradient based optimization methods until convergence. The optimization needs the partial derivative of the marginal likelihood with respect to the covariance hyperparameters. The gradient of the log marginal likelihood with respect to $\boldsymbol{\theta}_a = \{\sigma_{fa}^2, \sigma_{na}^2, l_{ad}\}, d = 1,...,l_{m_a}$ is expressed in (2.29).

$$\frac{\partial Q}{\partial \boldsymbol{\theta}_a} = -\frac{1}{2}(\mathbf{y}_a - \boldsymbol{\mu}_a)^T (K_y^{-1} \frac{\partial K_y}{\partial \boldsymbol{\theta}_a} K_y^{-1})(\mathbf{y}_a - \boldsymbol{\mu}_a) - \frac{1}{2} tr(K_y^{-1} \frac{\partial K_y}{\partial \boldsymbol{\theta}_a}) \tag{2.29}$$

where $K_y = K_a + \sigma_{na}^2 I$

In order to alleviate the problem of obtaining local maxima as optimal values, in this paper, the optimization needs to be repeated with different initial values. Gradient Descent is commonly used to find the optimal hyperparameters that maximizes the log marginal likelihood (Rasmussen and Williams, 2006). The cross-validation or generalized cross-validation algorithm is an alternative to gradient descent for optimization. However, the cross-validation is difficult to use when there is a large number of parameters (Williams, 1999).

2.5.2 Predictive distribution

Bayesian model updating can be applied to make robust predictions for the future events conditioned on the past data. By the Theorem of Total Probability, the information for future prediction of the model is given by weighting the predictive pdf, $p(f_{a^*} \mid X_a, \mathbf{y}_a, \mathbf{x}_{a^*})$ using the posterior probability, $p(\mathbf{f}_a \mid X_a, \mathbf{y}_a)$ and the future large predictor data $\mathbf{x}_{a^*}$. The future predictive pdf is given by (2.30).

$$p(f_{a^*} \mid X_a, \mathbf{y}_a, \mathbf{x}_{a^*}) = \int p(f_{a^*} \mid X_a, \mathbf{x}_a, \mathbf{f}_a)\, p(\mathbf{f}_a \mid X_a, \mathbf{y}_a) d\mathbf{f}_a \tag{2.30}$$

The predictions for the future data obtained are conditioned on the past data using the property of the conditional Gaussian distribution is given in (2.31).

$$\begin{bmatrix} \mathbf{f}_a \\ \mathbf{f}_{a^*} \end{bmatrix} \sim N\left(\begin{pmatrix} \mu(\mathbf{x}_a) \\ \mu(\mathbf{x}_{a^*}) \end{pmatrix}, \begin{pmatrix} K_a & K_{a^*} \\ K_{a^*}^{\mathrm{T}} & K_{a^{**}} \end{pmatrix} \right) \tag{2.31}$$

$$M(\mathbf{x}_{a^*}) = f_a(\mathbf{x}_{a^*}, \boldsymbol{\theta}_a) + K_{a^{**}} K_a^{-1} (\mathbf{y}_a - f_a(\mathbf{x}_a, \boldsymbol{\theta}_a)) \tag{2.32}$$

$$Cov(\mathbf{x}_{a^*}) = K_{a^{**}} - K_{a^*}^T K_a^{-1} K_{a^*} \tag{2.33}$$

where $K_{a^{**}}$ is the covariance between the future predictors and K_{a^*} between the covariance between the calibration predictors data and the future predictors data. The ensembles can be simulated using the mean vector (2.32) and the covariance matrix (2.33) following the multivariate normal distribution. The simulated ensembles in the previous step are transformed to the real precipitation values by taking cubic power.

Algorithm 1 and Algorithm 2 explain the implementation of GPR in KNN-BUQSDM model calibration and prediction.

2.5.3 Algorithm 1 – SGP-SDM model selection

1. INTPUTS: X_a (Predictors), $\mathbf{y}_a$ (Predictand).

2. OUTPUTS: Optimal hyperparameters (estimates of the $\boldsymbol{\theta}_a = \{\boldsymbol{\beta}_a, \sigma_{fa}^2, \sigma_{na}^2, l_{ad}\}, d = 1,...,l_{m_a}$).
3. Initialize the hyperparameters value.
4. Compute covariance matrix using (2.21) and mean function coefficients using (2.28).
5. Estimate the negative marginal likelihood using (2.27).
6. Obtain the derivatives of the negative log marginal likelihood with respect to the hyperparameters using (2.29).
7. The conjugate gradient optimization routine can be used to find the optimal hyperparameters using the negative log marginal likelihood and the gradient.

2.5.4 Algorithm 2 – SGP-SDM prediction

1. INPUT: X_a (Training Predictors), X_a^* (Future Predictors), $\mathbf{y}_a$ (Predictand), optimal hyperparameters ($\boldsymbol{\theta}_a = \{\boldsymbol{\beta}_a^*, \sigma_{fa}^{2*}, \sigma_{na}^{2*}, l_{ad}^*\}, d = 1,...,l_{m_a}$).
2. OUTPUT: predictive mean $M(\mathbf{x}_{a^*})$, prediction covariance, $Cov(\mathbf{x}_{a^*})$ and $\mathbf{y}_a^*$ (future predictand).
3. Compute $K_a^* = \begin{pmatrix} K_a & K_{a^*} \\ K_{a^*}^{\mathrm{T}} & K_{a^{**}} \end{pmatrix}$ using both X_a and X_a^*.
4. Obtain the predictive mean and predictive covariance using the (2.32) and (2.33)
5. Simulate ensembles using the predictive mean and predictive covariance obtained in the previous step.

2.6 Results and Discussion

2.6.1 Evaluation criteria for SDM

The downscaled precipitation using the large scale predictors needs to be evaluated in comparison with the observed precipitation for the validation period. In this study, the results are evaluated based on the precipitation intensity metrics and the temporal characteristics of the downscaled precipitation. The KNN, K-means algorithm and GPR are implemented in MATLAB (The MathWorks; Martinez *et al.*, 2011; Ramos, 2012).

The evaluations statistics such as mean (Mean) and standard deviation (STD) are used to assess the precipitation (Hessami *et al.*, 2008; Fowler and Ekström, 2009; Maraun *et*

al., 2010). The extreme values in the precipitation are assessed by the 90th percentile (PERC90) of the precipitation on wet days (Haylock *et al.*, 2006; Goodess *et al.*, 2007). The maximum amount (Max) (Hessami *et al.*, 2008) and the proportion of wet days (Pwet) can be used to assess the temporal metrics (Semenov *et al.*, 1998). These five evaluation metrics were used to compare the downscaled precipitation with the observed data. The distribution of the observed and the downscaled precipitation is evaluated using Kolmogorov-Smirnov nonparametric goodness of fit test to assess whether the two distribution has the same distribution (Khan *et al.*, 2006). The accuracy of the downscaled precipitation can be evaluated using the MSE (Armstrong and Collopy, 1992) given in (2.34):

$$MSE = \frac{1}{N}[\sum_{m=1}^{12}(y_{m,obs} - y_{m,sim})^2] \tag{2.34}$$

where N is the data length, $y_{m,obs}$ represents the daily precipitation observed for the month, m and $y_{m,sim}$ represents the daily precipitation simulated for the month, m. The accuracy (*acc*) of the wet and the dry day classification (Chen *et al.*, 2010) is given in (2.35):

$$acc = \frac{C_{dry} + C_{wet}}{TP_{dry} + TP_{wet}} \tag{2.35}$$

where C_{dry} is the total number of is correctly classified dry days, C_{wet} is the total number of correctly classified wet days, TP_{dry} is the total number of dry days and TP_{wet} is the total number of wet days.

APE is the Absolute Percentage Error (Ghosh and Katkar, 2012) which represents the accuracy of the evaluation statistics indicators. APE is given by (2.36):

$$\text{APE} = \frac{|P_{i,obs} - P_{i,sim}|}{P_{i,obs}} \tag{2.36}$$

where $P_{i,obs}$ is the observed precipitation for the i^{th} month and $P_{i,sim}$ is the simulated precipitation for the i^{th} month. The uncertainty range of the downscaled results is assessed by a mean absolute percentage boundary error (MAPBE) (Lu and Qin, 2014) which is given by (2.37):

$$\text{MAPBE} = \frac{1}{n_m}\sum(\text{APE}_{L,i} + \text{APE}_{U,i}) \tag{2.37}$$

where n_m is the number of months, the lower boundary APE for the i^{th} month is given by $\text{APE}_{L,i}$ and the upper boundary APE for the i^{th} month is given by $\text{APE}_{U,i}$.

2.6.2 Precipitation occurrence determination and K-means clustering

Using KNN, the occurrence of precipitation (wet days) is determined first. The choice for occurrence determination is based on the previous literature study which used the same study area and the data of our research work, by Lu and Qin (2014). They compared ANN, LC and KNN to determine the wet and dry days based on accuracy and dry day proportion. Their results showed that KNN gave better accuracy and dry day proportion compared to ANN and LC. Also since the objective of this study is to develop an uncertainty quantification tool for precipitation amount estimation, the comparison of classification models is not considered. The percentage of correct wet and dry day classification is used to assess the performance of the KNN occurrence determination. Table 2-3 shows the correct wet day and dry percentage predicted by KNN. The results in the table show that the wet days are predicted better than the dry days. The average classification success rate for the dry days is 45.96% and for the wet days is 56.30%. In the next step, the rainfall is classified into two (medium and high) rainfall types. Due to complex nature of precipitation, it is difficult to classify wet and dry days more accurately than the presented results. Olsson *et al.* (2001) used ANN to downscale precipitation and their results showed that the ANN underestimated low

intensity rainfall. Since ANN is a special case of GPR, the occurrence determination and wet day classification steps are implemented in two steps to improve prediction accuracy; otherwise the GPR may not predict dry days and low intensity rainfall efficiently.

The K-means cluster algorithm is used to identify the number of clusters in the wet days. The validation of the algorithm is not possible and not required as it is the unsupervised classification without any output (Kannan and Ghosh, 2011). The objective function in (2.15) is minimized by the algorithm to identify the clusters/classes. The K-means algorithm is executed many times for various numbers of clusters (K_m) in each run. The cluster validation indices are computed for each of the identified rainfall clusters along with the centroid of each class which are obtained as the output for K-means algorithm. The optimum numbers of clusters are found based on the validation indices such as Dunn's index, Davies-Bouldin index and Silhoutte index. These indices are computed for each cluster obtained from the algorithm for the wet days. The aim of this work is to implement the downscaling model with lesser number of rainfall clusters while still achieving better accuracy in the predictions. Thus, class numbers from two to four were tested for each month.

Table 2-3 Correct wet day and dry day classification by KNN

Month	Correct wet day (%)	Correct dry day (%)
January	68.97	57.58
February	42.25	60.20
March	59.22	44.58
April	60.53	31.82
May	47.62	32.10
June	41.86	46.81
July	52.63	47.25
August	46.51	52.04
September	54.12	54.74
October	59.62	43.90
November	71.77	35.77
December	70.59	44.78

The clusters with high Dunn's index are considered as good number, the clusters that give high Sillouette values are considered desirable and the clusters with the smallest Davies-Bouldin index are considered as good number. For illustration, the number of classes and the validation indices for the months of December (wet month) and February (dry month) are shown in Table 2-4 and Table 2-5. From the table results, it can be seen that Dunn's index and Sillouette values are high for the two clusters while Davies-Bouldin index for two clusters is high. Davies-Bouldin index with four clusters is the smallest. The number of rainfall days in each cluster is also considered as a factor in classifying the rainfall. If the cluster number is chosen to be high for the month of December, then there is less number of data in each cluster. Thus the optimal number based on validation indices for the month of December is two. Similarly for the month of February and the other months, the number of clusters is found to be two. Thus there are three rainfall classes such as dry, medium and high similar to number of the rainfall types classified by Kannan and Ghosh (2011) based on the rainfall intensity. KNN is then implemented to classify the wet days into several rainfall types in the future data and the precipitation amount is estimated for the two rainfall types for each month using GPR.

Table 2-4 Cluster validation index for the month of February

Classes	Dunn's index	Sillouette values	Davies-Bouldin index
2	**0.0433**	**0.8909**	**0.4521**
3	0.0088	0.8520	0.5133
4	0.0021	0.8130	0.4887

Table 2-5 Cluster validation index for the month of December

Classes	Dunn's index	Sillouette values	Davies-Bouldin index
2	**0.0055**	**0.8527**	0.5226
3	0.0034	0.8407	0.5231
4	0.0028	0.8024	**0.5067**

Figure 2-3 shows the comparison of the Quantile-Quantile (QQ) plot of the observed and the downscaled precipitation with one class and two classes. It depicts that the downscaled precipitation is underestimated with only one class whereas the downscaled precipitation is predicted well when two classes are used for downscaling. The results indicate that the performance of the downscaling model is improved when two classes are used.

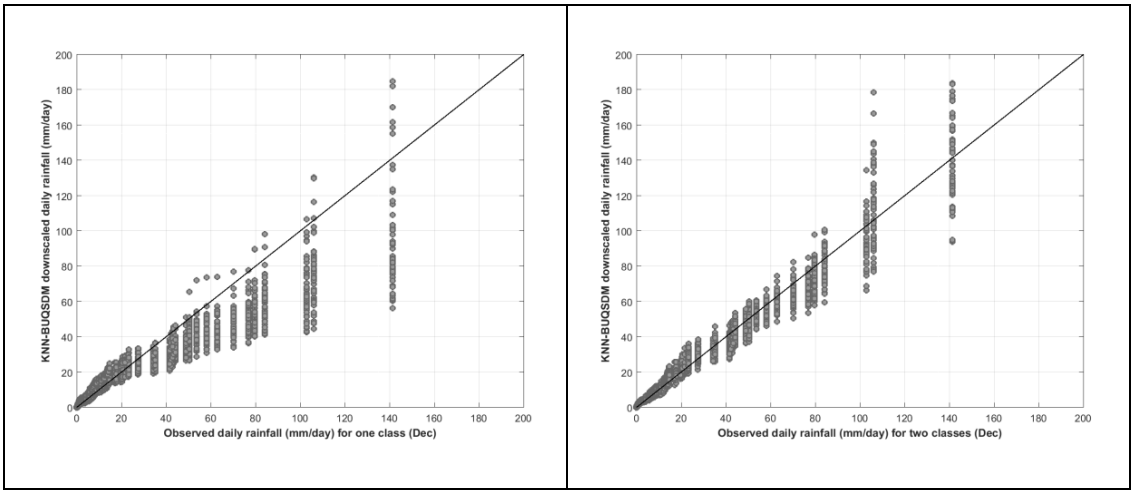


Figure 2-3 Quantile-Quantile plot for the month of December to compare the observed and downscaled precipitation with one and two classes

2.6.3 Precipitation amount estimation using GPR

The downscaling results are obtained by fitting KNN-BUQSDM for individual monthly data to capture the seasonal variation. The GPR is calibrated for each of the clusters identified in the previous step to find the optimal hyperparameters and the prediction is obtained for each rainfall type as shown in Figure 2-3. As the predictive mean and the predictive variance follow Gaussian distribution, there is a high possibility that the predictions will have negative values. The proportion of negative values simulated by GPR is checked by comparing the dry-day proportion simulated by KNN and the dry-day proportion in the precipitation amount estimation as shown in Table 2-6. It shows that the proportion of negative values from GPR is relatively less as the values from GPR is closer to the classification model (KNN). The dry-day proportion of KNN and GPR is also closer to the observed dry-day proportion. This result shows the ability of

GPR to simulate precipitation effectively. The days for which the model gives negative value are considered as dry days (Chen *et al.*, 2010). The accuracy of the rainfall occurrence determination for the month of December is 66.67%. Table 2-7 shows the accuracy of the downscaled precipitation amount for the month of December. The accuracy of the downscaled results is assessed by using equation (2.35). The results show that the accuracy of KNN-BUQSDM is higher than the accuracy of the other downscaling models such as ASD, GLM and KNN-BNN in comparison. Additionally, it proves that the KNN-BUQSDM preserves the accuracy of the classification model by predicting lesser number of negative values. It is also shown that KNN-BUQSDM is able to achieve better accuracy even with lesser number for rainfall types compared to KNN-BNN. It remains important to consider the misclassification rate (probability that the model simulates type 2 (high) rainfall when calibrated for type 1 (medium) and vice versa) by GPR as data are fitted for two rainfall types. The probability that the high rainfall is predicted in the medium rainfall class for the month of April and December is 0.0002% and 0.01% respectively. The results for these two months are shown because the month of December is wet period and the month of April is both dry and wet. The probability is the average misclassification rate of the 50 ensembles. Similarly the probability that the medium rainfall is predicted in the high rainfall type for the month of April and December is 0.1% and 0.0003% respectively. The misclassification rate for the other months is also very small which makes GPR desirable for precipitation prediction. The lower misclassification rate is attributed to several factors such as the use of correlation between the predictors in prediction and the accuracy of the precipitation occurrence determination and the correct rainfall type classification. Several previous studies used different number of downscaled ensembles for example, 40 ensembles were used by Segond *et al.* (2007), 20 ensembles by Samadi *et al.* (2013) and Mezghani and Hingray (2009) used 50. Based on literature study and the comparison of different number of ensembles in this work, the number of downscaled precipitation ensembles is chosen as 50 to represent the confidence interval and evaluate the monthly statistics.

Table 2-8 lists the envelops (minimum and maximum) and the average evaluation statistical indicators obtained using 50 ensemble simulated by KNN-BUQSDM for the month of December. Compared with those results for ASD, GLM, the proposed KNN-BUQSDM demonstrates notable improvement and is on par with KNN-BNN in terms of envelop and average statistics. The 90^{th} percentile is underestimated by ASD, GLM and KNN-BNN while KNN-BUQSDM is able to capture the 90^{th} percentile well. This is attributed to the fact that the GPR is adaptable and can capture aleatory and epistemic uncertainty in the model structure simultaneously to improve the prediction. In general, the performance of KNN-BUQSDM model in reproducing all the required statistics is better in terms of mean value of the statistics. The uncertainty range of all the statistics is better than ASD, GLM and comparable to KNN-BNN. However in KNN-BNN, the number of cluster used was six.

Table 2-6 Comparison of dry-day proportion estimated by KNN and GPR with the observed dry day proportion

Month	Dry-day proportion from KNN (precipitation occurrence determination)	Dry-day proportion from GPR (precipitation amount estimation)	Observed Dry-day proportion
January	0.45	0.46	0.53
February	0.59	0.60	0.57
March	0.42	0.44	0.44
April	0.37	0.37	0.36
May	0.44	0.45	0.43
June	0.52	0.53	0.52
July	0.47	0.49	0.48
August	0.53	0.55	0.52
September	0.52	0.51	0.53
October	0.42	0.43	0.44
November	0.30	0.33	0.31
December	0.34	0.34	0.36

Table 2-7 Accuracy of downscaled precipitation for the month of December

SDM	Min	Max	Average
ASD[1]	0.560	0.605	0.582
GLM[1]	0.550	0.591	0.566
KNN-BNN[1]	0.582	0.597	0.591
KNN-BUQSDM	0.656	0.672	0.666

[1]Results obtained from (Lu and Qin, 2014)

Table 2-8 Comparison of evaluation statistics of ASD, GLM, KN-BNN and KNN-BUQSDM for the month of December

Statistics	ASD 2(min, avg, max)	GLM 2(min, avg, max)	KNN-BNN 2(min, avg, max)	KNN-BUQSDM 2(min, avg, max)	OBS
Mean	6.28, 9.25, 11.65	6.29, 9.96, 12.12	9.85, 10.19, 10.49	9.43, 10.54, 11.95	10.39
SD	10.39, 16.06, 25.34	12.46, 17.34, 24.89	19.80, 20.74, 22.12	17.52, 20.76, 25.62	21.31
Pwet	0.66, 0.71, 0.76	0.55, 0.65, 0.73	0.63, 0.66, 0.68	0.64, 0.65, 0.66	0.64
PERC90	27.54, 33.69, 48.77	26.11, 38.42, 48.61	41.74, 44.84, 48.52	36.78, 46.30, 56.66	50.40
Max	58.84, 109.45, 224.68	68.66, 115.54, 282.27	121.03, 139.39, 158.76	93.77, 141.67, 211.93	141.40

[1]Results obtained from (Lu and Qin, 2014).[2]The three numbers represent the minimum (min), average (avg) and maximum (max) value of the evaluation statistics calculated from 50 ensembles.

When more number of rainfall types used, there was a problem of insufficient data to fit the regression model (Lu and Qin, 2014). This issue is resolved by using two rainfall types in KNN-BUQSDM. Figure 2-4 shows the average evaluation statistics of all the ensembles along with the two uncertainty ranges such as Envelop Range (ER) which represents the lower and the upper range and the 5^{th} and the 95^{th} percentile range (P95R) marked as grey region compared to the observed evaluation statistics. The Mean, STD, Pwet, PERC90 and Max of the simulated precipitation by KNN-BUQSDM corresponding to S44 rain gauge station are presented in Figure 2-4 (2-4a-2.4e) respectively. When the average evaluation statistics are compared with the observed one, the MSE of KNN-BUQSDM is significantly less than that of ASD, GLM and KNN-BNN as shown in Table 2-9. The MSE values for ASD, GLM and KNN-BNN are obtained from the results presented by Lu and Qin (2014). The MSE is calculated using equation (2.34). The MSE of the Mean, STD, Pwet, PERC90 and Max simulated by KNN-BUQSDM is 41.6%, 34.33%, 86%, 55.13% less than ASD, 58.23%, 39.87%, 91.43%, 48.32% and 53.82% less than GLM and 1.49%, 33.91%, 70%, 18.03%, 24.76% less than KNN-BNN. The KNN-BUQSDM method shows that the average of the statistics for the predicted precipitation is closer to the observed one for all the months. This indicates that there is less variability in predicting the rainfall by KNN-BUQSDM.

Table 2-9 Monthly mean squared error

Statistics	MSE-ASD1	MSE-GLM1	MSE-KNN-BNN1	MSE-KNN-BSDM-UQ
Mean	1.13	1.58	0.67	0.65
Std	4.69	4.95	4.66	3.08
Pwet	0.004	0.007	0.002	0.0007
Perc90	33.54	29.12	18.36	16.96
Max	869.33	652.42	400.38	301.26

[1]Results obtained from (Lu and Qin, 2014)

In the results obtained by Lu and Qin (2014), they showed that the prediction uncertainty range by KNN-BNN was less for Southwest monsoon (June –Sep) compared to Northeast monsoon (Nov-Feb). This was attributed to the fact that there was rainfall pattern variability in Singapore. As shown in Figure 2-4, the KNN-

BUQSDM is able to capture the rainfall variability for all the months effectively. This proves the ability of GPR to predict the precipitation even if there is a variation in the rainfall pattern. From Figure 2-4a, the average value of the mean indicator is closer to the observed one and is reproduced well by KNN-BUQSDM. In Figure 2-4c, it can be seen that for the month of January, the Pwet is underestimated. This is due to the availability of less observed data in all the range of rainfall for downscaling. Similar deviation was reported by Lu and Qin (2014). It can be understood that better downscaling results can be obtained even with the Gaussian distribution assumption for model function (epistemic uncertainty) and residual fitting (aleatory uncertainty). Thus by treating the epistemic and aleatory uncertainty simultaneously the prediction results can be improved.

Table 2-10 shows the MAPBE for KNN-BUQSDM to represent the quantitative levels of uncertainty in the downscaled predictions. It is calculated using ER and P95R values. MAPBE calculated using full range of data is represented as ER; when the 5^{th} and 95^{th} percentile of the data are used for calculating MAPBE, it is represented as P95R. The MAPBE is estimated using (2.37). The recorded values in table show reduced uncertainty compared to ASD, GLM while the results are comparable with the KNN-BNN. It should be noted that the KNN-BUQSDM is able to achieve similar results to KNN-BNN with fewer number of classes than KNN-BNN. This study has demonstrated that KNN-BUQSDM has better prediction accuracy and smaller uncertainty range. Figure 2-5 shows the downscaled daily precipitation simulated by KNN-BUQSDM. Only one random sequence from the 50 ensembles in comparison with the observed daily precipitation for the month of December (wet period) and February (dry period) is presented. This figure analyses the ability of KNN-BUQSDM to reproduce the extreme events.

Table 2-10 MAPBE values of the downscaled results using ASD, GLM, KNN-BNN and KNN-BUQSDM at S44

MAPBE	ASD[1]		GLM[1]		KNN-BNN[1]		KNN-BUQSDM	
	ER	P95R	ER	P95R	ER	P95R	ER	P95R
Mean	0.66	0.53	0.67	0.52	0.22	0.21	0.53	0.38
STD	0.94	0.70	0.83	0.60	0.30	0.28	0.69	0.46
Pwet	0.28	0.25	0.67	0.32	0.13	0.13	0.09	0.08
PERC90	0.80	0.60	0.74	0.58	0.29	0.24	0.66	0.48
Max	1.83	1.26	1.58	1. 21	0.61	0.52	1.25	0.89

[1]Results obtained from (Lu and Qin, 2014)

The rainfall magnitude greater than 50 $mmday^{-1}$ is set as the threshold for extreme rainfall. The number of extreme events (NEE) in the observed daily precipitation for the month of December and February is 12 and 3 respectively. In, the NEE for December is 11 and for February it is 3. The average NEE for all the 50 ensembles for December and February are 10.76 and 2.32 respectively. The maximum amount of daily rainfall observed for December and February is 141.4 mm and 76.5 mm respectively. The average maximum amount of daily rainfall simulated by KNN-BUQSDM for December and February is 141.67 and 76.96 respectively. In the results presented by Lu and Qin (2014) it was shown that the NEE simulated by ASD and GLM were 7.3 and 7.9 respectively whereas the NEE simulated by KNN-BNN was 10.2. These results prove that KNN-BUQSDM has the ability to simulate extreme events well compared to ASD, GLM and it is comparable to KNN-BNN. The results presented in our research also show that KNN-BUQSDM can capture rainfall variability as the performance is good for both wet and dry months. The downscaled results for station S24 are also presented in order to verify the robustness of the proposed downscaling approach. The distribution of the observed and downscaled daily precipitation is investigated for the validation period. The cumulative distributions of the observed and downscaled daily precipitation have been estimated for each month and are presented in Figure 2-6. The graphical representation of the ensemble distribution from the proposed downscaling model captures the observed distribution of the precipitation well. It can also be seen that the downscaled precipitation ensembles closer to the observed data even though there are some deviations from the observed

data distribution. Thus the uncertainty range in the prediction is narrower for all the months. The Kolmogorov-Smrinov nonparametric goodness of fit test (ks-test) is used to assess whether the observed and the downscaled precipitation are from the same distribution. The ks-test value is computed between the observed data and each of the downscaled ensembles. The average of the p-values for downscaled precipitation at two stations (S44 and S24) for each month is presented in Table 2-11. The results indicate that the p-value is found greater than 0.05 (95% confidence level) for all the months at two stations. Thus the p-values show that the KNN-BUQSDM model reproduces the distribution of the daily precipitation well at a 95% confidence level consistently for all the months at two stations (S44 and S24).

Table 2-11 p-values of ks-test value of the distribution of the precipitation

	P-Value at S44	P-Value at S24
January	0.0892	0.6248
February	0.7088	0.9670
March	0.8334	0.6982
April	0.7341	0.7615
May	0.8363	0.4018
June	0.8141	0.3925
July	0.8182	0.8184
August	0.8443	0.7658
September	0.8113	0.8164
October	0.7762	0.6078
November	0.4012	0.3505
December	0.5072	0.2021

Table 2-12 MAPBE values of the simulated results using three methods at S24

MAPBE	ASD[1]		GLM[1]		KNN-BNN[1]		KNN-BUQSDM	
	ER	P95R	ER	P95R	ER	P95R	ER	P95R
Mean	0.736	0.613	0.825	0.672	0.313	0.308	0.4951	0.3732
STD	0.786	0.566	0.914	0.732	0.369	0.361	0.6097	0.4404
Pwet	0.430	0.392	0.391	0.307	0.116	0.114	0.1136	0.1050
PERC90	0.855	0.744	0.946	0.789	0.456	0.416	0.6783	0.5024
Max	1.414	1.046	1.777	1.315	0.654	0.598	1.0268	0.7616

[1]Results obtained from (Lu and Qin, 2014)

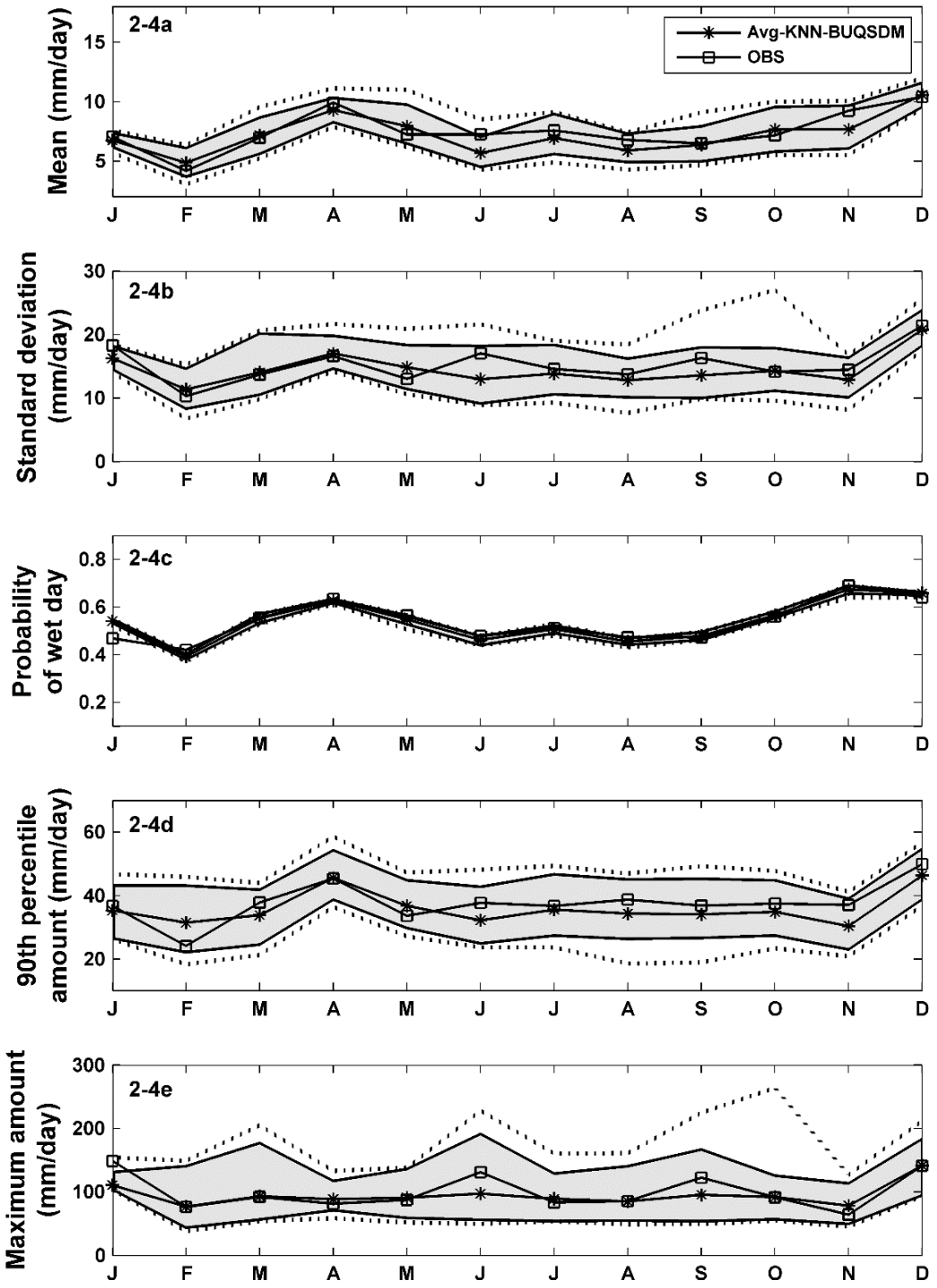


Figure 2-4 Monthly mean evaluation statistics for the rain gauge station at S44. The shaded area represents the 5th and 9th percentile of the ensembles. The dashed line represents the minimum and maximum values of the ensemble statistics

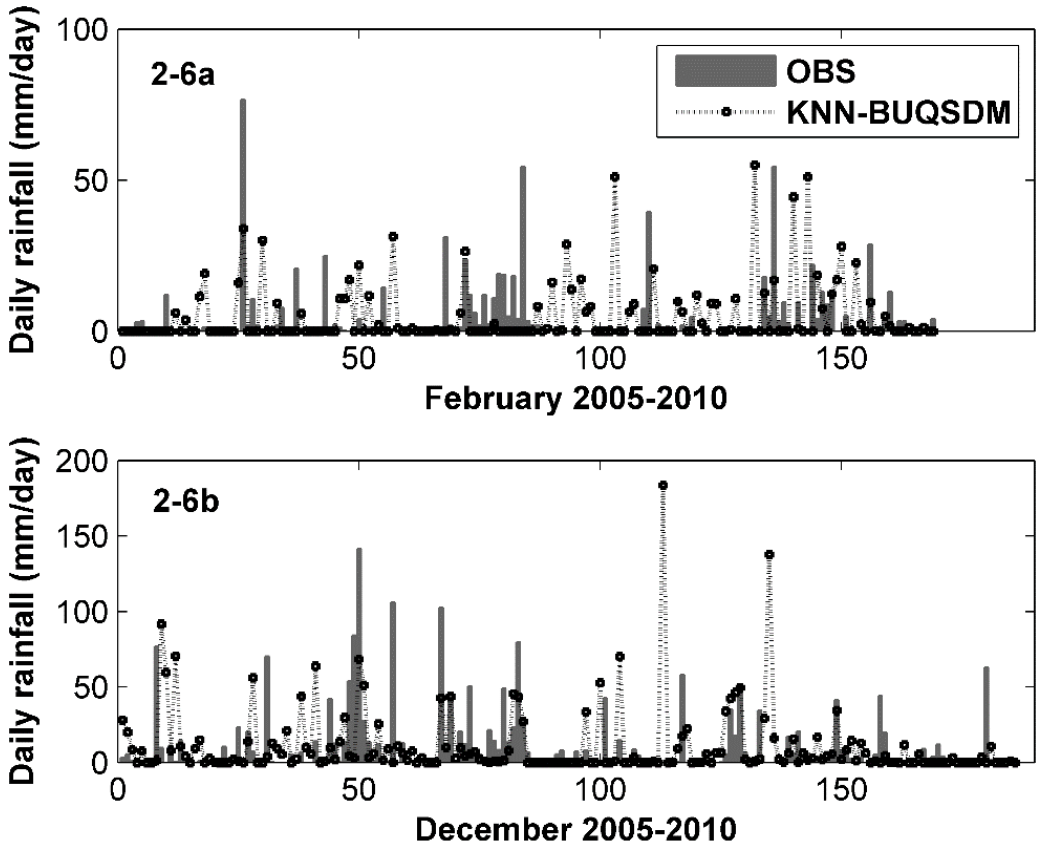
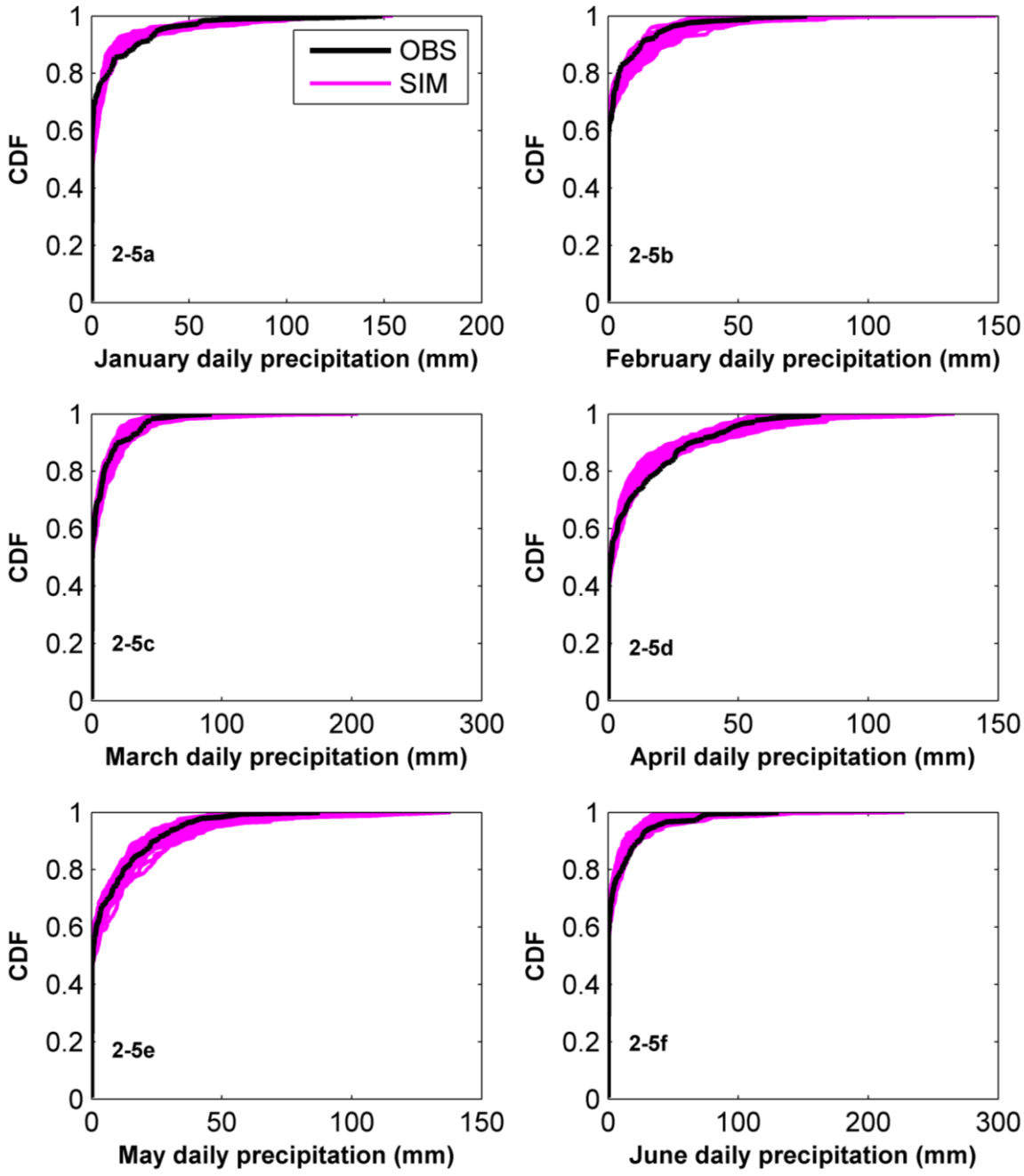


Figure 2-5 Observed and downscaled precipitation at S44 for the month of February and December

Figure 2-7 shows the evaluation statistics indicators for all the months at S24. The weather pattern is drier in the east compared to the west. Thus S24 station which is located at the eastern side is chosen to assess the model performance. The number of rainfall types is identified as two using K-means algorithm for all the months at S24 rain gauge station. From Figure 2-7, it is proved that the average of all the evaluation statistics obtained by KNN-BUQSDM is closer to the observed data as well as within the confidence number of rainfall types for downscaling precipitation using KNN-BNN at S24, the results showed deviations in the average values of the indicators such as mean, STD and maximum rainfall amount interval. This shows the robustness and the consistently good performance of KNN-BUQSDM for statistical downscaling of the precipitation.


OBS
SIM
CDF
2-5a
January daily precipitation (mm)
2-5b
February daily precipitation (mm)
2-5c
March daily precipitation (mm)
2-5d
April daily precipitation (mm)
2-5e
May daily precipitation (mm)
2-5f
June daily precipitation (mm)

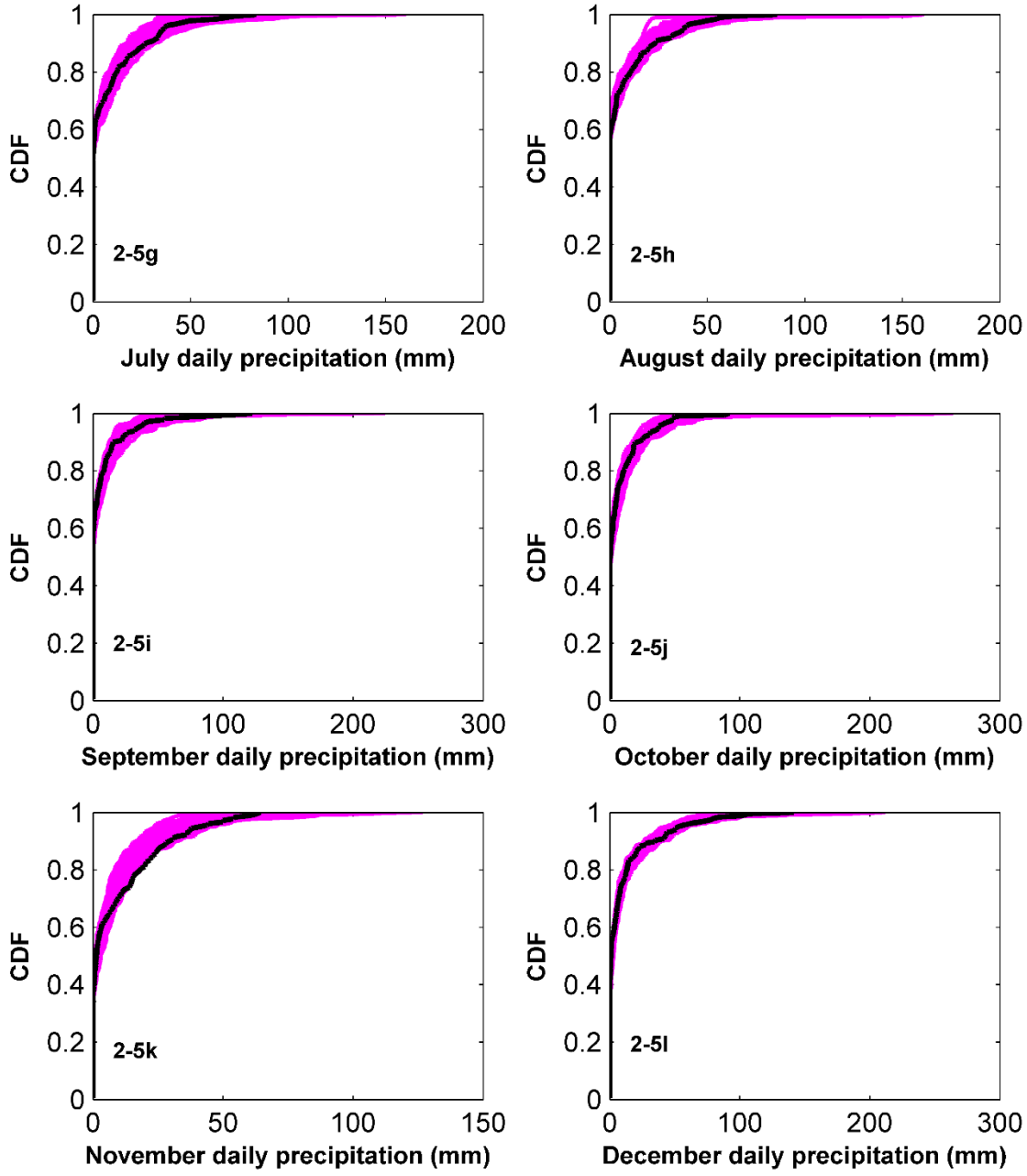


Figure 2-6 Comparison of CDF of the observed and downscaled precipitation for all the months at station S44

The MAPBE values of the downscaled precipitation at S24 based on monthly data are illustrated in Table 2-12. The performance of the model is similar to S44 station. KNN-BUQSDM shows narrower confidence interval for all the indicators compared to ASD, GLM and is comparable to KNN-BUQSDM. When Lu and Qin (2014) used more It is observed that the model also suffered from insufficient data for each rainfall type.

2.6.4 Discussions on KNN-BUQSDM structure

The proposed KNN-BUQSDM can be considered as a hybrid model since it encompasses the features of regression based downscaling method and weather typing. KNN-BUQSDM also couples the uncertainty analysis (residual fitting) within the model framework. The details of KNN-BUQSDM model structure are as follows:

1. The rainfall is classified into different types to reflect the local precipitation distribution in downscaling. This concept is similar to weather typing which is modelled based on classifying local climate occurrence (Fowler *et al.*, 2007; Maraun *et al.*, 2010).
2. The predictions are probabilistic by treating the residuals as stochastic processes (Rasmussen and Williams, 2006); there is no need to calibrate residual fitting separately to simulate ensembles. This feature can also be viewed as coupled weather generator and regression model.
3. The predictive posterior distribution is the joint conditional distribution of the training data and the future data. This model takes the correlation between the past and future data to improve the model performance. The conditional distribution of KNN-BUQSDM model helps to consider the extent to which the past data can affect the performance of the model in the predictions (2.31). It is also pointed out in many research studies that the stationary relationship between the past and the future data is questionable (Dixon *et al.*, 2016).

Non-stationary covariance functions can be used to capture the non-stationary relationship between the historic and the future data. The proposed model framework is

similar to those of GLM and ASD. The major difference is that it predicts the occurrence of rainfall as well as the rainfall type and the precipitation amount is predicted for each class of rainfall. The advantage of KNN-BUQSDM are 1) KNN is a non-parametric supervised classification model which is flexible and can be implemented directly (He and Wang, 2007); 2) The k-means clustering is an unsupervised clustering method which helps to identify the threshold for dividing the rainfall into types (Kannan and Ghosh, 2011) and 3) GPR is the state-of-the-art full probabilistic stochastic model which gives the full posterior distribution of the model parameters for prediction. The over-fitting and under-fitting problem can be solved using Bayesian inference. The relationship of the KNN-BUQSDM with SDSM, ASD and GLM is discussed in the following part. The assumptions, advantages and limitations of the linear regression model, GLM and KNN-BUQSDM are presented in Table 2-1. SDSM/ASD and KNN-BUQSDM follow Gaussian error structure; the contrast between the two is that in KNN-BUQSDM, the errors/residuals are dependent and are assumed to be stochastic processes. The advantages of GLM compared to ASD are that GLM includes features such as autocorrelation, nonlinearities, transformation and spatial structure in the model structure (Chandler and Wheater, 2002). These features are included in GLM through high dimension basis function. The basis function includes constant model, linear model, cyclic trend (to capture periodicity), autocorrelation covariates indicator and seasonal effect (sine and cosine component), Legendre polynomial (Yang *et al.*, 2005) in precipitation occurrence model as well as the precipitation amount model. These features can also be included in GPR through different covariance structure (Rasmussen and Williams, 2006) in the model instead of using complex basis function which is referred as kernel trick. The different types of covariance function can be found in (Rasmussen and Williams, 2006; Roberts *et al.*, 2013). For example, Rasmussen and Williams (2006) used several combinations of covariance function to predict atmospheric CO_2 concentration (Wilson and Adams, 2013). In their study, they utilized the fact that the product/addition of the covariance function is also a covariance function. In their work, 1) Radial Basis kernel was used to model the long term smoothing trend; 2) the periodic exponential sine squared kernel

was used to model the periodic component in the data; 3) the smaller, medium term irregularities were explained by a Rational Quadratic kernel component and 4) noise term to represent the noise components. Their results proved that prediction accuracy had improved with the use of several covariance functions to capture different data properties. The flexible features of GP also enable its use in multi-dimensional weather sensor data (Osborne *et al.*, 2012), the lighting curve modelling of transiting exoplanets (Gibson *et al.*, 2012) and Gaussian Process for predicting climate network discovery (Das and Srivastava, 2011). The non-stationary nature of the climate processes can be captured by using non-stationary covariance function to capture the changing relationship between the past and the future data in the model (Dixon *et al.*, 2016). Linear covariance function, neural network covariance function are few examples of non-stationary covariance function. As GPR uses covariance matrices, there is a possibility that the covariance matrices become ill-conditioned when large dataset are used. Since the climate change field involves large dataset, the usage of sparse approximation techniques can solve the computational issues (Das and Srivastava, 2011). Several research studies have been done to improve the applicability of GPR to handle large data (Vasudevan *et al.*, 2009). The Gaussian Process model can also be used for precipitation amount determination by using logistic likelihood (Rasmussen and Williams, 2006). Thus Gaussian Process Classification (GPC) can also adopt the features of GLM similar to GPR as explained previously in this section. The results from KNN-BUQSDM also indicate that the ensembles simulated by assuming the same distribution for traditional residual analysis, model calibration and the predictions are better than SDMs. The improvement in the results attributes to several factors such as classification accuracy, partitioning into rainfall types before predicting rainfall amount, GPR model structure and Bayesian inference. However, the need to classify the rainfall into different types may not be applicable to all geographical locations and is subjected to vary with data and study areas. Further research is needed to assess whether the classification of rainfall improves prediction for all climate conditions.

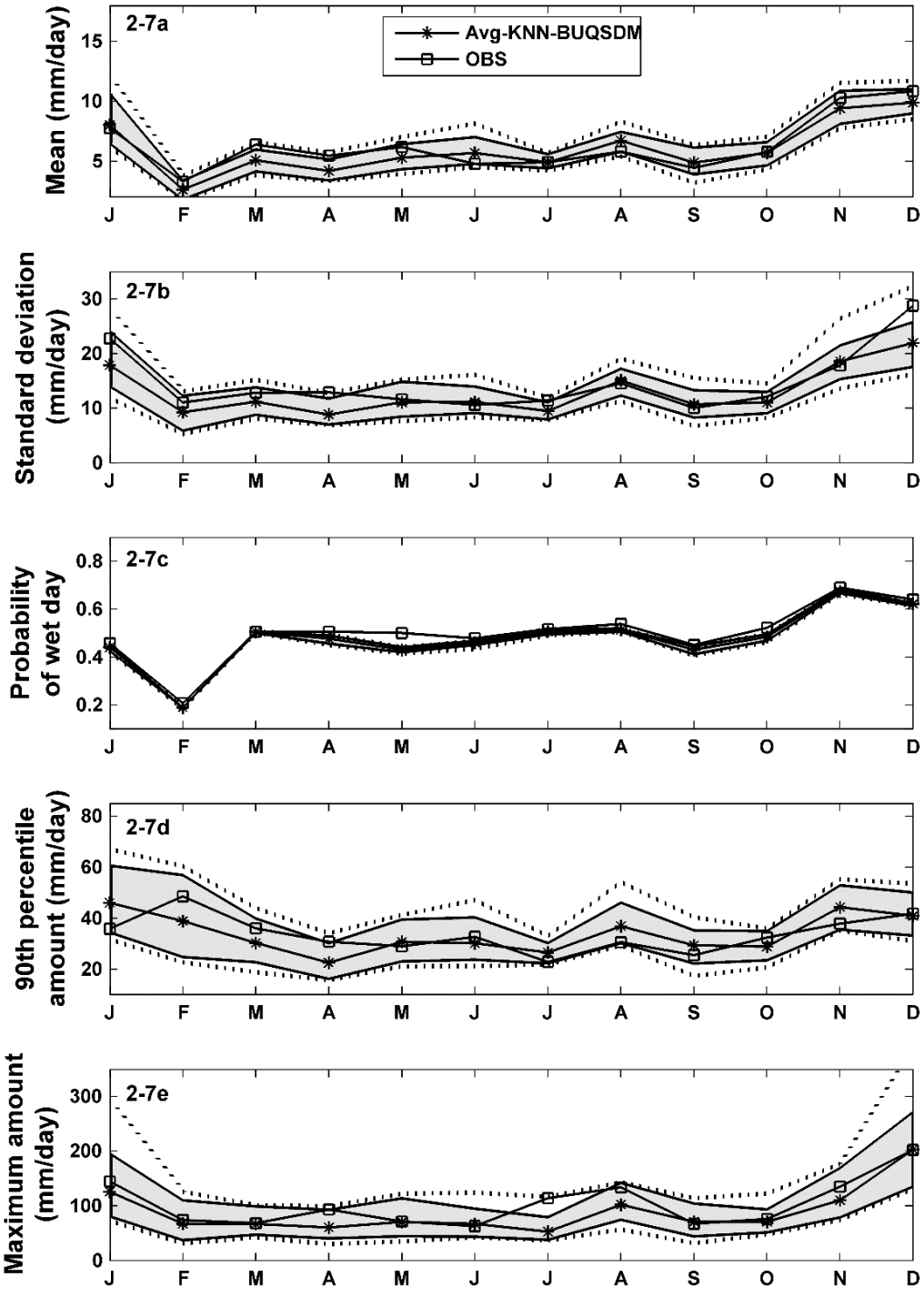


Figure 2-7 Monthly mean evaluation statistics of the downscaled precipitation at the rain gauge station S24. The shaded area represents the 5th and 9th percentile of the ensembles. The dashed line represents the minimum and maximum values of the ensemble statistics.

CHAPTER 3 SGP-SDM – An advanced statistical downscaling model with uncertainty assessment of coupled classification and precipitation amount estimation using Gaussian Process Error Coupling

3.1 Abstract

The content of this chapter is extracted from the submitted journal paper. SGP-SDM (Single site Gaussian Processes-Statistical Downscaling Model) is developed in this chapter to provide a principled way to quantify and propagate uncertainty in statistical downscaling model calibration and prediction within a single framework. This is achieved by coupling the residual error with the model instead of calibrating the model parameters and residual parameters separately. The GP assumes that the residuals are dependent and is modelled as stochastic processes following Gaussian distribution. Gaussian Process Classification (GPC) is adopted to determine the wet day occurrence and GPR is adopted to estimate the wet day rainfall amount. GP facilitates to capture the model adequacy using Bayes' theorem. A GP prior is placed over the modelling function instead of model parameters; thus GP is a non-parametric model. By using Bayes' theorem, the prior over the model function is updated using the likelihood function to obtain the posterior distribution of the model. In case of precipitation occurrence model, the likelihood is logistic distribution; thus the posterior does not follow Gaussian distribution and is analytically intractable. Laplace approximation is adopted to obtain the solution for the posterior distribution of the occurrence model. In the case of precipitation amount estimation, the likelihood is also Gaussian distribution. Thus the posterior distribution follows Gaussian distribution and it can be solved analytically. The marginal likelihood is the objective function and a gradient based optimization technique is used to find the optimal model and residual parameters. This assumption helps to quantify the uncertainty in the model function and in the residuals simultaneously. The posterior predictive distribution of the model function is

computed by integrating over the posterior distribution; the predictive distribution is the joint distribution of the historic and future data. The ensembles of the downscaled precipitation are directly generated by using the predictive distribution instead of point estimates. The efficiency of the proposed model is assessed by downscaling precipitation from two large scale predictors such as CFSR and CanESM2 for Singapore. SGP-SDM is implemented for each month to capture the monthly variations. The result shows that the SGP-SDM captures the observed statistics such as Mean, standard deviation, proportion of wet days, 90^{th} percentile of rainfall and maximum rainfall well compared to ASD, GLM and KNN-BNN for CFSR data. The results for the validation show that downscaled precipitation has less mean square error for CanESM2 data compared to CFSR data. The results cannot conclude the best emission scenarios between RCP 4.5 and 8.5 RCP.

3.2 Introduction

There has been an increase in extreme flood events in the recent years due to climate change. The increased amount of flood in a short duration of time causes huge social and economic losses especially in an urban area (Willems, 2012). Due to various natural complex factors, the future climate is uncertain, leading to the development of several GCMs and GHG emission scenarios (Nakićenović, 2000). In order to understand the hydrological processes that cause flooding, several hydrological models available in (Solomatine *et al.*, 2008; Refsgaard *et al.*, 2013) can be referred. However, the GCMs climate projection cannot be used to represent the local urban climate due to limitations in understanding the complex geophysical process that links the large-scale climate pattern with the local climate as well as the difficulty in representing the fine scale natural processes computationally (Wilby and Wigley, 1997). This makes it difficult to use the GCMs data directly in hydrological models. The simulation of high resolution future climate projection from a coarse resolution GCM is necessary to assess the impact of climate change on hydrology at a smaller urban area for decision making and mitigation planning (Willems *et al.*; Arnbjerg-Nielsen *et al.*, 2013; Taye

and Willems, 2013). Several statistical downscaling models based on regression to act as a link between global climate and local climate variables such as precipitation and temperature are being developed to simulate high resolution climate projections for impact studies. Since the GCM and statistical downscaling model involves numerical computations and simulations, the presence and the cascade of uncertainty is a natural and inevitable part in the downscaled climate predictions due to complex climate interactions, parameterizations in GCM, structure of statistical downscaling model and its parameters as shown in Figure 1-1 (Reilly *et al.*, 2001; Kay *et al.*, 2009; Chen *et al.*, 2011). It is therefore necessary to quantify, represent and propagate such uncertainties in the predictions with an effective mechanism as the predictions with less uncertainty play an important role in decision making. This research study aims to develop a new stochastic process based statistical downscaling model to simulate probabilistic projections of fine scale future climate variables to assess the impact of climate change on hydrology. The proposed method also aims to map uncertainty from GCM outputs to the downscaled high resolution precipitation.

3.2.1 Uncertainty in statistical downscaling models

Quantification of uncertainty in the statistical downscaling models has received a lot of attention recently (Wilby and Harris, 2006; Clark *et al.*, 2016). There are mainly three sources of uncertainty in the climate projections (Foley, 2010; Environmnet and Heritage, 2016). The future emissions of greenhouse gases and aerosols are considered as one of the largest sources of uncertainty. Since the knowledge about them is unknown, it is not possible to quantify this uncertainty mathematically. This uncertainty can be represented by a number of plausible future emission scenarios or projections in GCM simulations to assess the climate change impact. The second source of uncertainty is the response of the climate systems due to changes in the constituents of the atmosphere. This is represented by using different GCM simulations which reflect the model structural and parametric uncertainties. The uncertainty from the emission scenario can be assessed by simulating the high resolution climate projections from several scenarios. The model structural and parametric uncertainties in

the GCM are assessed by using ensembles of possible future climate simulations from many GCMs (Environmnet and Heritage, 2016). The probability of future simulations from each GCM was estimated by combining them in a Bayesian framework (Tebaldi *et al.*, 2005; Buser *et al.*, 2009; Mani and Tsai, 2017) or multi-model ensemble techniques and this is an active area of research. The third source of uncertainty is the response in the local climate due to climate change. This uncertainty is also associated with the techniques used for statistical downscaling of future climate. The confidence in the predictions of future local climate variables is affected by the above mentioned sources of uncertainty (Fowler *et al.*, 2007; Maraun *et al.*, 2010).

Prudhomme and Davies (2009) investigated three main sources of uncertainty in river flow projections. Three GCMs (HadCM3, CCGCM2 and CSIRO-mk2) and two downscaling models (SDSM and HadRM3) were considered for the analyses. Their study results concluded that the GCM was the largest contributor of uncertainty in impact studies compared to statistical downscaling and was greater than the uncertainty from the hydrological models. However, they couldn't conclude the best performing GCM or downscaling model for impact studies. From the literature review on assessing the main sources of uncertainty, it is found that GCM stands to be the main source (Hawkins and Sutton, 2009) whereas the statistical downscaling model is the second largest contributor of uncertainty while the uncertainty due to hydrological model and its parameters is less significant (Chen *et al.*, 2011). Similar results were concluded in the other studies by Wilby and Harris (2006) and Kay *et al.* (2009). This research study focuses on assessing uncertainty due to SDMs.

All the past comparison studies concluded that the statistical downscaling model is the second largest sources of uncertainty in impact studies. From the past studies, it is noted the quantification of uncertainty in the statistical downscaling models is a major concern in impact studies. Each statistical downscaling model has unique advantages and disadvantages which results in variations in the future climate projections. There exists a plethora of comparison studies on the downscaling methods to characterize the uncertainties underlying in the predictions. Weather generators (WGEN) and ANN

were used by Wilby *et al.* (1998) to downscale precipitation. Their results showed that WGEN performed better in simulating wet-day occurrences compared to ANN. Three statistical downscaling models including SDSM, LARS-WG and ANN were compared by Khan *et al.* (2006). The observed statistics were reproduced well by SDSM with 95% confidence interval compared to ANN and LARS-WG. The daily precipitation, daily temperature and monthly precipitation were downscaled using several downscaling methods by Schoof and Pryor (2001). It was shown that the regression models and ANN model yielded similar results. The temperature was downscaled accurately compared to precipitation. Ghosh and Mujumdar (2009) suggested some imprecise probability to represent uncertainty band from several GCMs including GCM with missing output. In their work, it was mentioned that the confidence band from one GCM may not be sufficient to make decisions and the confidence bands from several GCMs were needed. Giorgi and Mearns (2002) proposed Reliability Ensemble Averaging (REA) method that calculated average, uncertainty range and a reliability measure of simulated climate change ensembles. Similar comparison studies can be found in Schoof and Pryor (2001), Haylock *et al.* (2006), Wetterhall *et al.* (2006), Dibike *et al.* (2008) and Chen *et al.* (2011). Fowler *et al.* (2007), Maraun *et al.* (2010) and Quintana Seguí *et al.* (2010) presented a detailed review of all the uncertainty analysis studies on the downscaling model. In order to obtain the robust precipitation downscaling results, it was recommended in all their studies to use different statistical downscaling models rather than one model as the predictions from only one statistical downscaling model since it is not reliable as the performance varies for the different seasons and GCM predictors (Chen *et al.*, 2012; Sunyer *et al.*, 2012; Sunyer Pinya *et al.*, 2015).

All the comparison results are inconclusive as a single model cannot be chosen as the best model for downscaling as the performance varies with the location and seasons. Fowler *et al.* (2007) suggested that further research was needed to characterize uncertainty in statistical downscaling models. While there are many methods for uncertainty quantification, there are very limited number of studies on coupling the

uncertainty quantification tool with the SDM model structure. The first reason for this can be attributed to the fact that the assumptions of statistical downscaling model structure remain the same even though there are many types of regression based models available. The second reason is that a general uncertainty quantification framework for downscaling model to propagate and represent the uncertainty is not available yet.

The two types of downscaling techniques are DDM and SDM. The detailed explanation about the downscaling methods can be found in the literature review by (Wilby and Wigley, 1997; Fowler *et al.*, 2007; Maraun *et al.*, 2010). The projection of climate using RCM is computationally intensive which makes it less preferable compared to statistical downscaling methods (Giorgi and Mearns, 1991; Tang *et al.*, 2016). As regression based statistical downscaling methods are easier to implement and required less computational time, they are considered in this study (Fowler *et al.*, 2007). The statistical downscaling model for precipitation is a two-step process where the precipitation amount is conditioned on the occurrence of the rainfall. Thus the precipitation occurrence is determined as the first step and the amount of precipitation is then estimated as the second step. The commonly used regression based statistical downscaling models are Generalized Linear Model (GLM) in which the wet days are determined using logistic regression (Coe and Stern, 1982; Chandler and Wheater, 2002) and the rainfall amount is estimated using gamma regression, Statistical Downscaling Model (SDSM) and Automated Statistical Downscaling (ASD) where the wet days and the amount of rainfall are estimated using multiple linear regression (Wilby *et al.*, 2002; Hessami *et al.*, 2008). Several other downscaling approaches such as ANN (Hewitson and Crane, 1996), SVM (Tripathi *et al.*, 2006) and RVM (Ghosh and Mujumdar, 2008) have been proposed in the literature.

3.2.2 Uncertainty quantification in precipitation occurrence determination

Precipitation occurrence depends on the regional climate predictors such as mean sea level pressure, specific humidity and geopotential height. The precipitation occurrence is downscaled by relating the daily precipitation occurrence with the large scale climate

predictors (Wilby and Wigley, 1997; Wilby *et al.*, 2002). The determination of correct number of wet and dry days is very important for estimating the rainfall amount with improved accuracy.

Let the historic data for the precipitation occurrence determination model be D_c of n observations, that is, $D_c = \{(\mathbf{x}_{ci}, \mathbf{y}_{ci}) \mid i = 1, ..., n\}$ where $\mathbf{x}_c$ represents the GCM predictors with dimension m_c and $\mathbf{y}_c$ is the binary classification output (wet/dry day). The rainfall occurrence model in SDSM is expressed as follows (3.1) (Wilby *et al.*, 2002; Wetterhall *et al.*, 2006):

$$O_d = a_0 + \sum_{j=1}^{m_c} a_j \hat{x}_c^{(j)} + a_{d-1} O_{d-1} \tag{3.1}$$

where d is the number of days, O_d is the conditional probabilities of the rainfall occurrence on day d, $\hat{x}_c^{(j)}$ is the GCM predictors which are normalized, a_j is the regression parameters which are inferred by using least square method and $d-1$ and O_{d-1} are the regression parameters of lag-1 data and the conditional probabilities of rain occurrence on day $d-1$, respectively. The last two parameters can be optional and it depends on the area of study and predictand. A uniform distribution random number $r_d (0 \le r_d \le 1)$ is used to determine the occurrence of rainfall $\mathbf{y}_c$, if $O_d \le r_d$.

In ASD, the occurrence of rainfall is expressed as (3.2):

$$O_d = a_0 + \sum_{i=1}^{m_c} a_j \hat{x}_{ci} \tag{3.2}$$

where d is the number of days, m_c is the number of predictors, O_d is the conditional probabilities of the rainfall occurrence on day d, $\hat{x}_{ci}$ is the GCM predictors which are normalized, a_j is the regression parameters which are inferred by using least square method (Hessami *et al.*, 2008).

The drawbacks with the linear classification methods in SDSM and ASD are 1) it lacks strength to detect outliers 2) cannot be used with all types of data and 3) the decision boundary is based on the assumption that follows Gaussian distribution; however, the binary output does not follow Gaussian. The above mentioned issues were solved by using logistic regression as proposed in GLM (Chandler and Wheater, 2002). The logistic regression used to determine the precipitation occurrence in GLM model is represented as (3.3):

$$\ln\left(\frac{P}{1-P}\right) = B_0 + B_1 x_{c1} + \cdots + B_{p_c} x_{cp_c} \tag{3.3}$$

Equation (3.3) can be represented in terms of probability by rearranging as given in equation (3.4):

$$\frac{P}{1-P} = e^{B_0 + B_1 x_{c_{p_c}} + \cdots + B_{p_c} x_{c_{p_c}}} \tag{3.4}$$

where the probability of the event is represented by P (the values of P lies between 0 and 1), e represents the base of the natural logarithms, $x_{c_{p_c}}$ is the predictors, p_c refers to the predictors used in the model and $B_0, B_1, ..., B_{p_c}$ are the model coefficients. The model parameters are estimated using maximum likelihood method. The logistic regression model is an improvement over the multiple linear regression model since the probability of wet day occurrence could be obtained. The logistic regression model assumes linear relationship between the predictor and the predictand; the predictions from the model are transformed using logit link function.

Lu and Qin (2014) compared the ability of KNN, ANN and LC approaches to determine the precipitation occurrences based on accuracy of determination and dry day proportion. Their results showed that the performance of KNN was better than the other models in comparison. In Chapter 2, KNN-BUQSDM is proposed to downscale precipitation. In KNN-BUQSDM, GPR is integrated with KNN model for downscaling

precipitation in Singapore. In KNN-BUQSDM, KNN is used to determine the wet days; the wet days are in turn classified into two rainfall classes using KNN. Then GPR is used to estimate the precipitation amount for each rainfall class. It is shown that the KNN-BUQSDM shows better performance in terms of reproducing observed statistics and accuracy compared to ASD, GLM and KNN-BNN.

It is noticed that in all the research works on statistical downscaling, the uncertainty due to choice precipitation occurrence model was not given much attention. The standard binary classification methods for rainfall occurrence determination described above do not capture the uncertainty in the model. The predictive probabilities obtained as the softmax output in logistic regression (3.3) are misinterpreted as the model confidence. Even if the softmax output is high, the model predictions can still be uncertain (Gal and Ghahramani, 2015). In their study, it was shown that if a point estimate of a function is passed through softmax function then that gave high confidence for the prediction points which were far away from the training points. In contrast if the distribution of a function is passed through the softmax function, then the output reflects the classification uncertainty for the predictions points. The readers are referred to the illustrated figure in a research study by (Gal and Ghahramani, 2015). The difficulty with KNN classification is the selection of K values and quantification of uncertainty. Thus, developing statistical downscaling model that takes the uncertainty from wet days determination is critical for the determination of correct number of wet days in future projections for decision making.

It was shown in a number of researches in the field of machine learning that GPC could circumvent the drawbacks in the traditional classification methods which rely on kernel trick to circumvent the curse of dimensionality and uncertainty quantification using Bayesian inference. GPC are considered as a potential approach for classification along with uncertainty quantification in machine learning (Rasmussen and Williams, 2006). In a research study by Zhao *et al.* (2008), GPC and KNN were applied to classify the Infrared Imaging Spectroradiometer image. The results proved that GPC performed better than KNN and SVM. They showed that GPC was a competitive tool in

classification for remote sensing applications (Yang *et al.*, 2015). GPC has been implemented in other fields such as medical imaging (Challis *et al.*, 2015) to quantify uncertainty and to improve accuracy. GPC is a full Bayesian statistical kernel based classification model in which the probability of an input belonging to a certain class is related to the value of latent function at the corresponding input. The statistical inference is implemented systematically using the Bayesian framework to estimate the binary class output type along with uncertainty. GPC involves the use of the Gaussian process prior and the logistic likelihood function for the Bayesian inference which involves integration over the latent function values to compute the posterior distribution. An analytical Laplace approximation is needed as the posterior distribution of GPC is non-Gaussian and cannot be solved analytically. Laplace approximation technique is needed to solve the marginal likelihood analytically (Kuss and Rasmussen, 2006). Thus, in our study GPC is adopted to address the above mentioned drawback in the rainfall occurrence determination model.

3.2.3 Uncertainty quantification in precipitation amount estimation

The historic data D_a for the precipitation amount estimation model of n observations are represented as $D_a = \{(\mathbf{x}_{ai}, y_{ai}) \mid i = 1,...,n\}$ where $\mathbf{x}_a$ are the GCM predictors with dimension m_a and $\mathbf{y}_a$ represents the rainfall amount. The simplest form of regression based precipitation amount estimation is given in (3.5):

$$\mathbf{y}_a = f_a(\mathbf{x}_a) + \boldsymbol{\varepsilon}_a \tag{3.5}$$

where $\mathbf{y}_a$ represents local precipitation, $\mathbf{x}_a$ represents the predictors from GCM/reanalysis model output, f_a is the downscaling function which can be linear or non-linear and $\boldsymbol{\varepsilon}_a$ is the additive model error/residual. A range of regression based such as multiple linear regression, gamma regression in Generalized Linear Model (GLM) and non-linear models such as Artificial Neural Network (ANN) have been proposed

for estimating the precipitation amount to assess the climate change impact on urban hydrology.

In a regression model, the function $f_a(\mathbf{x}_a)$ contributes to the epistemic uncertainty and the residual $\boldsymbol{\varepsilon}_a$ contributes to the aleatory uncertainty to the downscaled precipitation as shown in Section 2.5 of Chapter 2 of this thesis. The parameters of the model function $f_a(\mathbf{x}_a)$ are obtained using the historic $\mathbf{x}_a$ and $\mathbf{y}_a$ data. The optimized parameters are used to obtain the point estimates of the model function. In order to simulate ensembles to represent the uncertainty in the point estimates, the error is fitted with the GEV distribution. The simulated error ensembles are then added to the point estimates of the precipitation prediction obtained using $f_a(\mathbf{x}_a)$. These enable to simulate ensembles of the error in the future to represent the variability in the prediction. In all the existing regression based downscaling techniques, the above method is followed regardless of the type of downscaling model function $f_a(\mathbf{x}_a)$ used.

The flaw in this model assumption is explained in detail in Chapter 2.5. The flaws pointed out in their paper are 1) the model function and the error are calibrated separately 2) the distribution assumption for the model function and the residual are different and 3) there is no specified framework for quantifying uncertainty in precipitation amount estimation.

The probabilistic uncertainty quantification approach for statistical downscaling has been explored in the past literature studies to compute predictive distribution. RVM integrated with Principle Component Analysis (PCA) and fuzzy clustering was proposed by Ghosh and Mujumdar (2008) to downscale precipitation. Their method was able to capture the non-linear relationship between the climate variables and streamflow. The downscaled precipitation results were compared with downscaled precipitation from SVM. The results showed that the performance of RVM was better than SVM. Another advantage of RVM is that the aleatory and epistemic uncertainty is quantified simultaneously and the distribution assumption for model function and

residual is Gaussian. However, RVM introduced by (Tipping, 2003) has some disadvantages as pointed out in (Rasmussen and Williams, 2006). In RVM, when the future prediction point is away from the relevance vectors, the model finds it difficult to draw conclusion about the output in extrapolation. The predictive uncertainty at the test points becomes small (Rasmussen and Williams, 2006). RVM is also considered as a special case of GP; the errors in RVM are assumed to be independent. The ability of GPR to estimate precipitation amount by assuming dependency between the errors is shown in Chapter 2 by comparing the downscaled precipitation results with ASD, GLM and KNN-BNN. The results have shown that GPR performs better in reproducing the statistics such as mean, standard deviation, proportion of wet days, 90^{th} percentile and maximum values. The accuracy of the downscaled precipitation is high compared with other models. In a study by Rajendran and Cheung (2015), the above mentioned problem was addressed by using BUASCSDSEC (Bayesian uncertainty analysis for stochastic classification and statistical downscaling with stochastic dependent error coupling), a GPR model in which the residual and model calibration are coupled together with Gaussian distribution assumption. A simpler version of GPC and GPR was used for quantifying uncertainty in the statistical downscaling model. Their results proved that by using the dependency information between the residuals, the model prediction accuracy and confidence interval could be improved significantly. Elaborate use of GP for statistical downscaling is presented in Chapter 2 Section 2.2.3. of this thesis. In this chapter, an advanced version of KNN-BUQSDM developed in Chapter 2 is proposed. GPC is used for precipitation occurrence determination and GPR is used for precipitation amount determination. The proposed idea is also the advancement of BUASCSDSEC technique by Rajendran and Cheung (2015). The proposed approach is named as 'K-Nearest Neighbor-Bayesian Uncertainty Quantification Statistical Downscaling Model (KNN-BUQSDM).'

The objective of this study is to implement the state-of-the-art statistical approaches to develop a framework to quantify uncertainty in statistical downscaling techniques at single site. In this work, a stochastic process based Bayesian inference model proposed

by Cheung *et al.* (2011) is adopted for quantifying uncertainty statistical downscaling framework. The key idea in this model is that the uncertainty in the model parameters and the structures are propagated from precipitation occurrence determination prediction to precipitation amount estimation prediction. A significant aspect of the proposed method is that the model function and the model errors are coupled by assuming that the residuals are stochastic processes following Gaussian distribution. In this research study, the main focus is on downscaling precipitation from large scale climate predictors such as CFSR and CanESM2. Two large scale predictors are used to assess the uncertainty due to different scenarios and different GCMs. One specific aspect of this research work is to develop precipitation occurrence determination model using GPC and to investigate the applicability of different covariance function in wet day determination. The methodology also includes a method to select the model using marginal likelihood.

Chapter 3 is organized as follows. The study area and the data for this study are explained in the data and study area section. The proposed SGP-SDM methodology is explained in detail in the methodology section. The performance the proposed approach is assessed using three datasets such as CFSR and CanESM2. The precipitation is first downscaled using CFSR data and is compared with the classical downscaling approaches such as ASD, GLM and KNN-BNN; the corresponding results are presented in the results and discussion section. The downscaled results for ASD, GLM and KNN-BNN using CFSR are obtained from the research study by Lu and Qin (2014) since the CFSR data and study of their work are the same as the ones in Chapter 2. CanESM2 (RCP4.5 and RCP8.5) are used as GCM predictors for assessing the model's ability in downscaling precipitation from GCM predictors in Singapore.

3.3 Data and study area

The study area for this research work is Singapore which lies between 1° N and 2° N latitudes and 103.8° E and 104° E longitudes. It covers an area of 719.1 square kilometers. The climate of Singapore is classified as a tropical climate and has two

monsoon seasons namely Northeast monsoon (December to early March) and southwest monsoon (June to September). From the historical rainfall record, it is observed that the wettest month is December with the average rainfall of 230 mm and the driest month with the average rainfall of about 160 mm). There is a variation in the spatial distribution of rainfall where the rainfall is higher in the western parts compared to the eastern parts of Singapore. Thus a rain gauge station in the west (S44) is considered in this study to evaluate the performance of the proposed statistical downscaling model. The historically observed precipitation data for Singapore was obtained from (NEA, 2016). The meteorological station (S44) selected in this study have good quality of the data for the period from 1980 to 2010. At S44 station, the missing dataset is 0.09%. The western region of Singapore has complex rainfall pattern; thus the results for precipitation downscaling at S44 is completely presented. Figure 3-1 shows the location of the rain gauge stations along with CFSR and CanESM2 grids.

In order to capture the complexity in the climate of Singapore, the large scale predictors from CFSR are used to downscale precipitation in Singapore (Saha *et al.*, 2010; Dile and Srinivasan, 2014). CFSR data are chosen to compare the proposed downscaling approach with other downscaling model such as ASD, GLM and KNN-BNN. The spatial resolution of CFSR is $0.5^{\circ} \times 0.5^{\circ}$ and the temporal resolution of the predictors is 1-h instantaneous. The resolution of CFSR is superior to National Center for Atmospheric Research (NCEP/NCAR) reanalysis data set (6-h instantaneous and $2.5^{\circ} \times 2.5^{\circ}$). CFSR is a global high-resolution data and is available for 31-year period (Liléo and Petrik, 2011; Wang *et al.*, 2011). Mean sea level pressure (prmsl), Specific humidity at 500 hPa (q500), Specific humidity at 850 hPa (q850), Specific humidity at 925 hPa (q925), Geopotential at 500 hPa (z500), Geopotential at 850 hPa (z850), Geopotential at hPa (z1000), Zonal wind at 500 hPa (u-w 500), Meridional wind at 500 hPa (v-w 500), Zonal wind at 850 hPa (u-w 850) and Meridional wind at 850 hPa (v-w 850). For downscaling precipitation at S44 station, CFSR-West is used.

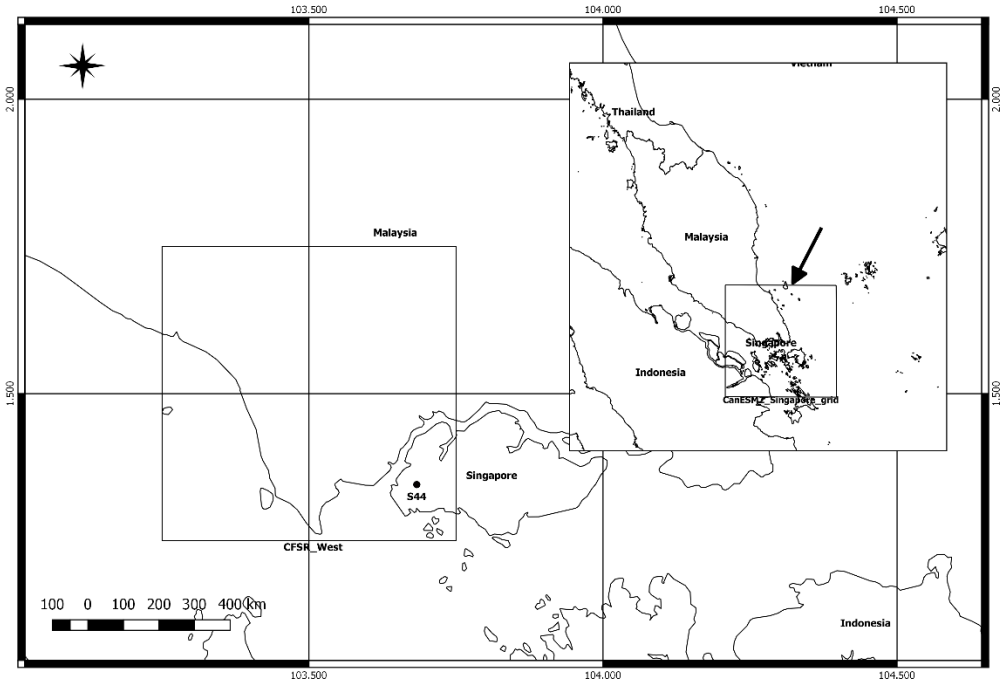


Figure 3-1 Study area and rain gauge location in Singapore

The CFSR predictors need to be standardized using (3.6) before using it for precipitation amount estimation and precipitation occurrence determination.

$$z_i = \frac{r_i - r_{\min}}{r_{\max} - r_{\min}} \tag{3.6}$$

where r_i is the ith input variable; $r_{\min}$ and $r_{\max}$ are the minimum and maximum values of the input variables, respectively.

The uncertainty is high in studying the climate change impact using single GCM as the spatial and temporal resolution of each GCM is different due to model assumptions (Kim and Kaluarachchi, 2009). The availability of several of large scale climate predictors provides an opportunity to evaluate the statistical downscaling model performance and to analyze the uncertainties of the downscaled precipitation corresponding to the different GCMs and emission scenarios. However, it is difficult to

compare all the 13 GCMs performance in downscaling in this study as the objective is to develop a downscaling model coupled with the uncertainty framework. The performance of different GCMs to downscale precipitation in Singapore for climate change analyses were studied by Singapore 2nd National Climate Change Study – Phase 1 based on the presence of biases in the predictions (http://ccrs.weather.gov.sg/publications-second-national-climate-change-study-science-reports/). CanESM2 is a second generation Canadian Earth System Model developed by Environmental Canada of Canadian Center for Climate Modelling and Analysis (CCCma) for IPCC Fifth Assessment Report (AR5) (Chylek *et al.*, 2011). The horizontal resolution of the model is nearly uniform of 2.8125°. CanESM2 output has 26 daily predictors which are listed in the Table 3-1 and the predictors are obtained from (http://ccds-dscc.ec.gc.ca/?page=pred-canesm2). National Centre for Environmental Prediction (NCEP) (Kalnay *et al.*, 1996) reanalysis data set is used as atmospheric large scale predictor variables for representing the present condition. The NCEP-derived predictor data have been interpolated onto the same grid as the GCM.

The selection of predictors is an important step in developing a reliable statistical downscaling model and also to improve the accuracy of the prediction results (Wilby, 1998; Fowler *et al.*, 2007; Maraun *et al.*, 2010). The predictors that are relevant for downscaling precipitation are necessary to be included in developing the downscaling model (Wilby *et al.*, 2004). It is also important to select the predictors that can capture the variation in global warming to project the future climate change (Wilby, 1998). The humidity predictors represent the water holding ability of the atmosphere under global warming and the information about the changes in precipitation in a long-term is represented by the temperature predictor (Wilby and Wigley, 1997). The circulation predictors are also selected as predictors for downscaling precipitation (Cavazos and Hewitson, 2005). It is also shown in other research studies that the circulation variables may not be sufficient to capture the mechanisms that generate precipitation such as thermodynamics and moisture content. Thus, Karl *et al.* (1990) and Wilby and Wigley (1997) used humidity as an important predictor in downscaling precipitation in a

changing climate. The relationship between sea-level pressure and precipitation was studied by Thompson and Green (2004) and shown that there was a link between them at different time scales which include seasonal, monthly and daily scales. The predictors selected in this study are also based on the important predictors for downscaling precipitation in the previous studies.

The wet-day threshold is set to 0.1 mm. The prediction is tested with various threshold values. The results are in the best agreement with the observed data when the threshold is set to 0.1 mm. The same threshold level was used in other studies such as (Liu *et al.*, 2011; Lindau and Simmer, 2013; Taye and Willems, 2013; Pervez and Henebry, 2014). As the daily precipitation does not follow normal distribution, cubic root/fourth root transformation is applied to make the distribution of the precipitation closer to the normal distribution (Jeong *et al.*, 2012).

Two-sample Kolmogorov-Smirnov test is employed to select predictors for determining the rainfall occurrence and backward stepwise regression is employed to select predictors for estimating the rainfall amount. The Two-sample Kolmogorov-Smirnov test chooses the predictors that show significant difference between a dry day and a wet day; this is based on the fact that the predictors are sensitive to a wet and a dry day in the model (Chen *et al.*, 2010).

Three types of stepwise regression are available such as backward elimination, forward selection and bidirectional elimination. For estimating the precipitation amount, backward stepwise regression (Hocking, 1976) similar to ASD is adopted in our study. This regression can choose the predictors for rainfall amount estimation. In the first step, all the large scale predictors are considered in the model. The predictors that are not significant are eliminated in each step until the predictors that are significant in the model remain at the end (Hessami *et al.*, 2008). F-test is used to eliminate the predictors and is given in (3.7):

$$F = \frac{\left(R_p^2 - R_{p-1}^2\right)\left(n - p - 1\right)}{1 - R_p^2} \tag{3.7}$$

where the number of observed data is represented by *n*, the number of predictors is represented by *p*, the correlation coefficient between the criterion variables and the predictions with p predictors is R_q. If the F values are less than the threshold defined by (3.7), the predictors need to be eliminated (Hessami *et al.*, 2008). The threshold for the test can be calculated using equation (2.9).

3.4 Methodology

An uncertainty quantification framework for statistical downscaling model is developed in our research. In the method developed, the downscaling model function is coupled with the modelling error/residual; the error is assumed to be dependent and it is a stochastic process following Gaussian distribution. The dependency between the errors is utilized to capture the uncertainty in the prediction of precipitation occurrence determination and rainfall estimation. For precipitation occurrence determination, a zero-mean function and logistic likelihood with the combination of linear or squared exponential covariance function is used. A linear mean function and squared exponential covariance are used for precipitation amount estimation. The mean function is assumed only for the precipitation amount estimation. As the computational time increases when the mean function is used for precipitation occurrence determination, a zero-mean function is adopted.

The methodology for implementing the model consists of two stages such as rainfall occurrence determination and rainfall amount determination. The framework for implementing SGP-SDM is shown in Figure 3-2. The precipitation occurrence determination implementation algorithm for SGP-SDM consists of implementing three algorithms such as hyperparameter optimization algorithm, computation of model of the model function and prediction.

Table 3-1 CanESM2 predictors

CanESM2 Predictors	Symbol
Mean sea level pressure	Ceshmslpgl
1000 hPa wind speed	ceshp1fpgl
1000 hPa zonal velocity	ceshp1upgl
1000 hPa meridional velocity	ceshp1vpgl
1000 hPa vorticity	ceshp1zpgl
1000 hPa wind direction	ceshp1thpgl
1000 hPa divergence	ceshp1zhpgl
500 hPa wind speed	ceshp5fpgl
500 hPa zonal velocity	ceshp5upgl
500 hPa meridional velocity	ceshp5vpgl
500 hPa vorticity	ceshp5zpgl
500 hPa wind direction	ceshp5thpgl
500 hPa divergence	ceshp5zhpgl
800 hPa wind speed	ceshp8fpgl
800 hPa zonal velocity	ceshp8upgl
800 hPa meridional velocity	ceshp8vpgl
800 hPa vorticity	ceshp8zpgl
800 hPa wind direction	ceshp8thpgl
800 hPa divergence	ceshp8zhpgl
Relative humidity at 500 hPa	ceshp500pgl
Relative humidity at 850 hPa	ceshp850pgl
Total rainfall	Ceshprcppgl
Specific humidity at 500 hPa	ceshs500pgl
Specific humidity at 850 hPa	ceshs850pgl
Surface-specific humidity	ceshshumpgl
Mean temperature at 2 m height	ceshtemppgl

The algorithms from the book by Rasmussen and Williams (2005) are adopted; they can be found in algorithm 5.1, algorithm 3.1 and algorithm 3.2 respectively in the book (Rasmussen and Williams, 2005). The readers are suggested to refer to the book to get the detailed information regarding the algorithms. The detailed implementation algorithm for precipitation amount estimation using GPR can be found in Chapter 2 (Algorithm 1 and Algorithm 2). The methodology also includes a technique to select the model using marginal likelihood.

3.4.1 Bayesian updating framework

This section presents the generalized Bayesian framework for GPC and GPR. In Bayesian optimization, the known knowledge about the functions is presented as prior. A likelihood function $p(\mathbf{y}|\mathbf{f})$ needs to be specified in order to get the posterior distribution. $D, \mathbf{f}, \mathbf{y}, X$ here represent the data, model function, predictand and predictor that can belong to GPC or GPR and θ represents the hyperparameters. By Bayes' theorem, the posterior distribution of the latent function $\mathbf{f}$, $p(\mathbf{f}|X,y)$ given the data D is (3.8):

$$p(\mathbf{f}|D,\theta) = \frac{p(\mathbf{y}|\mathbf{f},X)\,p(\mathbf{f}|X,\theta)}{p(D|\theta)} \tag{3.8}$$

where $p(\mathbf{f}|X,\theta)$ is the GP prior and $p(D|\theta)$ is the marginal likelihood (*model evidence*). The marginal likelihood is also called the *normalizing constant* of the

Bayesian posterior distribution. The integral of the likelihood times the prior yields the marginal likelihood of the model:

$$p(\mathbf{y}|X) = \int p(\mathbf{y}|\mathbf{f},X)\,p(\mathbf{f}|X)d\mathbf{f} \tag{3.9}$$

The marginal likelihood (3.9) gives a measure of how well a model fits the data. The integral does not have analytical form and an approximation is needed to solve the integral (Anzai, 2012; Barber, 2012). The uncertainty in the model parameters is accounted using prior probability (Beck and Katafygiotis, 1998). The posterior pdf changes with the different prior pdf (Cheung *et al.*, 2011). The values of the hyperparameters θ cannot be determined *a priori*; the optimal value of $\hat{\theta}$ based on the training data can be learnt by maximizing the marginal likelihood $p(D|\theta)$ with respect

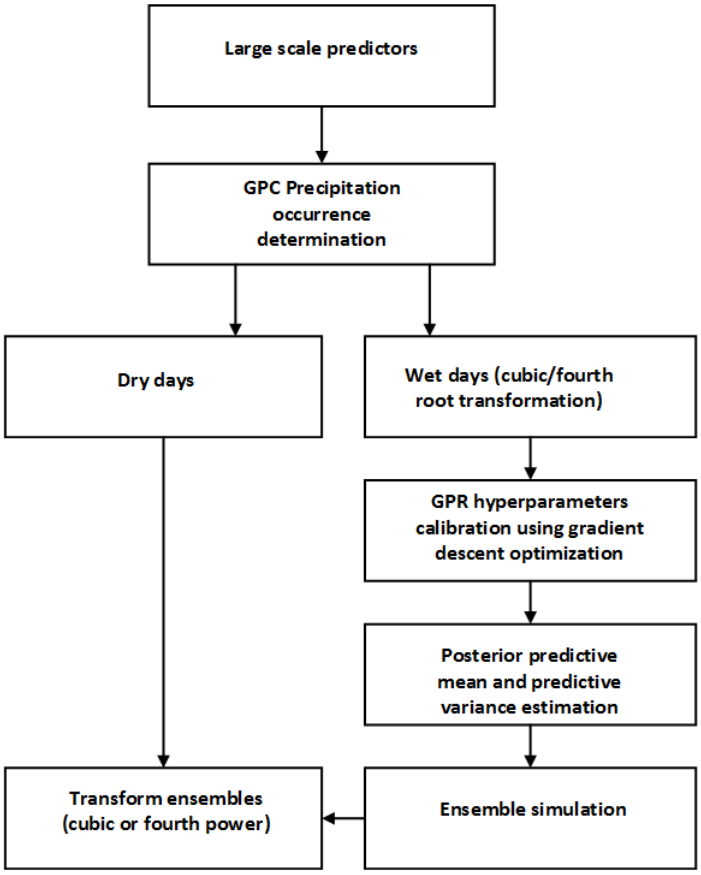


Figure 3-2 SGP-SDM statistical downscaling framework

to θ. As soon as the posterior $p(\mathbf{f} \mid D, \hat{\theta})$ is available, the prediction at a future predictor x_* can be obtained by (3.10):

$$p(f_* \mid x_*, D, \hat{\boldsymbol{\theta}}) = \int p(f_* \mid \mathbf{X}, x_*.\mathbf{f})\, p(\mathbf{f} \mid \mathbf{X}, y) d\mathbf{f} \tag{3.10}$$

The predictive distribution of y_* is given by (3.11):

$$p(y_* \mid x_*, D, \hat{\boldsymbol{\theta}}) = \int p(y_* \mid f_*)\, p(f_* \mid x_*, D, \hat{\boldsymbol{\theta}}) df_* \tag{3.11}$$

Equations (3.8) to (3.11) represent the generalized formulation of GP models which can be used in precipitation occurrence and amount estimation problems by choosing the appropriate likelihood model $p(\mathbf{y} \mid \mathbf{f})$.

3.4.2 Precipitation occurrence determination model using GPC

GPC is a nonparametric classification method based on a Bayesian methodology. In this section, an introduction and overview of the GPC is presented to solve wet day determination binary classification problem. The derivation follows GPC presented in Rasmussen and Williams (2006). GPC assumes Gaussian prior distribution on the model function to guarantee smoothness properties. This helps to determine the final classification that provides a good fit for the observed data while preserving the smoothness in the prediction. The uncertainty in determining wet or dry days is reflected in the posterior probability. In GPC, $y_c \in \{1,-1\}$ denotes the class labels (wet (+1) and dry (-1) day) corresponding to input predictors $\mathbf{x}_c$. A zero mean Gaussian process prior is assumed as prior function over the latent function $f_c(\mathbf{x}_c)$ or $\mathbf{f}_c$.

$$p(\mathbf{f}_c \mid X_c, \boldsymbol{\theta}_c) = \frac{1}{\sqrt{|2\pi K_c|}} \exp(-\frac{1}{2}\mathbf{f}_c^{\mathrm{T}} K_c \mathbf{f}_c) \tag{3.12}$$

where $p(\mathbf{f}_c \mid X_c, \boldsymbol{\theta}_c) \sim N(\mathbf{f}_c \mid 0, K_c)$ is the multivariate Gaussian with zero mean vector $0 \in R^n$ and symmetric positive-semidefinite covariance matrix $K_c \in R^{n \times n}$. The Gaussian process prior with zero mean and covariance function, K_c which depends on the hyperparameters $\theta_c = \{\sigma_{fc}^2, \{l_{cd}\}_{d=1}^{m_c}\}$. There are several choices of covariance functions that can used in Gaussian process models. Moreover, the sum or product of covariance functions is also a covariance function. This study explores three types of covariance function such as linear covariance function, squared exponential covariance function and the addition of the linear and the squared exponential kernels to obtain a new covariance function to improve the wet days determination accuracy for a dataset which has nonlinearities. Since the linear and squared exponential covariance function are commonly used (Rasmussen and Williams, 2006), they are chosen for occurrence determination. Similarly other types of covariance functions and their combinations using sum or products can be used for the model implementation. Readers are referred to Rasmussen and Williams (2006) to get more details about the covariance functions.

The covariance functions used in GP is analogous to the kernel function used in SVM or RVM. The covariance function defines a non-linear and high dimensional relationship between the predictors X_c. The squared exponential covariance function is a stationary covariance function as it depends on the relative locations of the inputs. On the other hand, the linear kernel is a non-stationary function where it allows the smoothness to vary with the inputs. For some data, the two covariance functions can be added together to improve the classification accuracy (Rasmussen and Williams, 2006). The linear Automatic Relevance Determination (ARD) covariance function is expressed in (3.13).

$$\mathbf{k}_{clin}(\mathbf{x}_c, \mathbf{x}'_c) = \sum_{d=1}^{m_c} \mathbf{l}_c^{-2} x_{cd} x'_{cd} \tag{3.13}$$

where $\mathbf{k}_{clin}(\mathbf{x}_c, \mathbf{x}'_c)$ is the linear covariance between the input $\mathbf{x}_c$ and $\mathbf{x}'_c$, m_c is the dimension of the data and $l_c = diag(l_{c1},...,l_{cm_c}) \in R^{m_c}, l_c > 0$ is the correlation length between the inputs to control the contribution of each predictor $\mathbf{x}_c$ to the model output. In matrix form, $\mathbf{k}_{clin}(\mathbf{x}_c, \mathbf{x}'_c)$ is represented as K_{clin}. By definition, the ARD function automatically determines the predictors that are relevant for the model. This process acts as an additional step to select predictors, after using Two-sample Kolmogorov-Smirnov test for choosing predictors. The linear ARD covariance function is non-stationary.

The Squared Exponential (SE) ARD covariance function is expressed in (3.14). The SE kernel depends on the distance for the dimension m_c, $\|x_a - x'_a\|_2^2$ between the inputs, where $\|.\|^2$ is the Euclidean Norm.

$$\mathbf{k}_{cSE}(\mathbf{x}_c, \mathbf{x}'_c) = \sigma_{fc}^2 \exp\{-\frac{1}{2}(\mathbf{x}_c - \mathbf{x}'_c)^T L_{\mathbf{c}}(\mathbf{x}_c - \mathbf{x}'_c)\}, \sigma_{fc}^2 \geq 0; diag(L_{\mathbf{c}}) > 0 \tag{3.14}$$

where σ_{fa}^2 is the signal variance and $L_{\mathbf{c}}$ is the diagonal positive symmetric matrix

consisting of the characteristic correlation length, l_c. The matrix form of $\mathbf{k}_{cSE}(\mathbf{x}_c, \mathbf{x}'_c)$ is written as K_{cSE}. In squared exponential covariance function, the smoothness of the model function prediction is controlled by lengthscale of the covariance function; l_c lengthscale determines how far two points are separated; if the lengthscale is small, the correlation between the predictors is high and vice versa. If the predictors are far away from each other, they are less correlated. This implies that if the predictors are highly correlated, the output of the model will be similar and vice versa. The variance of the function values is determined by the σ_{fc}^2; it therefore controls the relation between the latent function and the classification results. Thus, the hyperparameters for this covariance function is $\theta_c = [\sigma_{fc}^2, l_1, .., l_{m_c}]^{\mathrm{T}} \in R^{m_c+1}$. The two covariance functions can also be combined by addition or multiplication to utilize the advantages of both covariance functions. For precipitation occurrence determination, the linear ARD and SE ARD are added to improve the prediction accuracy.

In the case of GPC, there are many possible forms of likelihood function; thus the posterior does not follow Gaussian distribution anymore. Two commonly used functions are logistic and probit likelihood function. Thus, the posterior of GPC needs to be approximated using Laplace approximation. This study uses the most commonly used logistic function to show the applicability of GPC for wet days determination. The expression for log likelihood of the logistic function and its first and second order derivatives are presented in the equation (3.15):

$$p(y_c = 1 | \mathbf{f}_c) = \frac{1}{1 + e^{-\mathbf{f}_c}} \tag{3.15}$$

and

$$p(y_c = -1 | \mathbf{f}_c) = \frac{1}{1 + e^{\mathbf{f}_c}} \tag{3.16}$$

$$\log p(y_{ci} \mid f_{ci}) = -\log(1 + \exp(y_{ci} f_{ci})) \tag{3.17}$$

$$\frac{\partial}{\partial f_i}\log p(y_{ci} \mid f_{ci}) = t_i - \pi_i \tag{3.18}$$

$$\frac{\partial^2}{\partial f_i^2}\log p(y_{ci} \mid f_{ci}) = \pi_i(1-\pi_i) \tag{3.19}$$

where $\pi_{ci} = p(y_{ci} = 1 \mid f_{ci})$ and $\mathbf{t} = (\mathbf{y}_c + \mathbf{1})/2$.

The posterior over the latent function is expressed using Bayes' rule. The marginal likelihood is the un-normalized posterior and can be written as (3.20).

$$p(\mathbf{y}_c \mid X_c) \propto p(\mathbf{y}_c \mid \mathbf{f}_c)\, p(\mathbf{f}_c \mid X_c) \tag{3.20}$$

The logarithm of the marginal likelihood is (3.21):

$$\begin{aligned} \Psi(\mathbf{f}_c) &\sim \log p(\mathbf{y}_c \mid \mathbf{f}_c) + \log p(\mathbf{f}_c \mid X_c) \\ &= \log p(\mathbf{y}_c \mid \mathbf{f}_c) - \frac{1}{2}\mathbf{f}_c^{\mathrm{T}} K_c \mathbf{f}_c - \frac{1}{2}\log \mid K_c \mid - \frac{n}{2}\log 2\pi \end{aligned} \tag{3.21}$$

The marginal likelihood does not follow Gaussian distribution and cannot be solved analytically. The marginal likelihood needs to be minimized using ML-II hyperparameter optimization to estimate the values of the optimal hyperparameters. The inference can be obtained using approximation methods such as Laplace approximation (Williams and Barber, 1998), Expectation propagation (Minka, 2001; Seeger, 2005), MCMC, variational (Gibbs, 1998) and Kullback-leibler divergence (KL) minimization (Kuss and Rasmussen, 2005) to compute the approximate marginal likelihood. In practice, Laplace approximation is quick to implement and commonly used in several researches (Challis *et al.*, 2015) and is chosen in this study.

3.4.2.1 Newton update iteration for finding mode of the latent function $\hat{\mathbf{f}}_c$

The mode of the latent function $\hat{\mathbf{f}}_c$ and the Hessian matrix at the mode is needed to do Laplace approximation described in the next section. The Newton's method is

generally used to find the mode and the Hessian matrix. The Newton's method needs the first and the second order derivative of (3.21) to estimate the mode and the Hessian (Rasmussen and Williams, 2006).

The first and second derivative of (3.21) with respect to $\hat{f}_c$ is

$$\nabla\Psi(\mathbf{f}_c) = \nabla \log p(\mathbf{y}_c \mid \mathbf{f}_c) - K_c^{-1}\mathbf{f}_c \tag{3.22}$$

$$\nabla\nabla\Psi(\mathbf{f}_c) = \nabla\nabla \log p(\mathbf{y}_c \mid \mathbf{f}_c) - K_c^{-1} = -S - K_c^{-1} \tag{3.23}$$

where $S \triangleq -\nabla\nabla \log p(\mathbf{y}_c \mid \mathbf{f}_c)$ is diagonal, as the factorization of likelihood distribution depends only on f_i and $f_{j\neq i}$.

The maximum of $\Psi(\mathbf{f}_c)$ is computed by equating (3.22) to zero:

$$\nabla\Psi(\mathbf{f}_c) = 0 \Rightarrow \hat{\mathbf{f}}_c = K_c \nabla \log p(\mathbf{y}_c \mid \mathbf{f}_c) \tag{3.24}$$

As $\nabla \log p(\mathbf{y}_c \mid \mathbf{f}_c)$ is a non-linear function of $\hat{\mathbf{f}}_c$, it cannot be solved directly. Newton's method of iteration is used to find the maximum; the iteration is given in (3.25):

$$\begin{aligned} f_c^{new} &= f_c - (\nabla\nabla\Psi)^{-1}\nabla\Psi \\ &= \mathbf{f}_c + (K_c + S)^{-1}(\nabla \log p(\mathbf{y}_c \mid \mathbf{f}_c) - K_c^{-1}\mathbf{f}_c) \\ &= (K_c^{-1} + S)^{-1}(S\mathbf{f}_c + \nabla \log p(\mathbf{y}_c \mid \mathbf{f}_c)) \end{aligned} \tag{3.25}$$

Once the mode is found, the Laplace approximation to the posterior as a Gaussian distribution following mean $\hat{\mathbf{f}}_c$ and covariance matrix as the negative of the inverse Hessian of $\Psi(\mathbf{f}_c)$ can be expressed as in(3.26):

$$q(\mathbf{f}_c \mid X_c, \mathbf{y}_c) = N(\hat{\mathbf{f}}_c, (K_c^{-1} + S)^{-1}) \tag{3.26}$$

The algorithm for implementing the Newton's method to find mode $\hat{\mathbf{f}}_c$ is presented in Chapter 3, Algorithm 3.1 of the book Rasmussen and Williams (2006). The same is followed in this work.

3.4.2.2 Laplace approximation for marginal likelihood

In Laplace approximation method, a Gaussian approximation $q(\mathbf{f}_c \mid X_c, \mathbf{y}_c)$ is computed for the posterior $p(\mathbf{f}_c \mid X_c, \mathbf{y}_c)$ to evaluate the Gaussian integral of the marginal likelihood. The computation of Laplace approximation method requires to determine the maximum a posteriori (MAP) probability which is generally done using a gradient search.

$$p(\mathbf{y}_c \mid X_c) = \int p(\mathbf{y}_c \mid \mathbf{f}_c)\, p(\mathbf{f}_c \mid X_c) d\mathbf{f}_c = \int \exp(\Psi(\mathbf{f}_c)) d\mathbf{f}_c \tag{3.27}$$

A second order Taylor expansion of the un-normalized posterior $\Psi(\mathbf{f}_c)$ around $\hat{\mathbf{f}}_c$ is $\Psi(\mathbf{f}_c) \simeq \Psi(\hat{\mathbf{f}}_c) - \frac{1}{2}(\mathbf{f}_c - \hat{\mathbf{f}}_c)^{\mathrm{T}}(\mathbf{f}_c - \hat{\mathbf{f}}_c)$. Thus, the approximated posterior is a Gaussian with mean $\hat{\mathbf{f}}_c$ is placed at the mode (MAP) and the covariance H_c equals the negative inverse Hessian of the log posterior density at $\hat{\mathbf{f}}_c$.

$$p(\mathbf{y}_c \mid X_c) = q(\mathbf{f}_c \mid X_c, \mathbf{y}_c) = \exp(\Psi(\mathbf{f}_c)) \int \exp(-\frac{1}{2}(\mathbf{f}_c - \hat{\mathbf{f}}_c)^{\mathrm{T}}(\mathbf{f}_c - \hat{\mathbf{f}}_c)) d\mathbf{f}_c \tag{3.28}$$

where $\hat{\mathbf{f}}_c = \arg\max_{\mathbf{f}}\ p(\mathbf{f}_c \mid X_c, \mathbf{y}_c)$ and $H_c = -\nabla\nabla \log p(\mathbf{f}_c \mid X, \mathbf{y}_c)|_{\mathbf{f}_c = \hat{\mathbf{f}}_c}$.

The Gaussian integral in equation (3.28) can be solved analytically once the mode $\hat{\mathbf{f}}$ and the Hessian, H are obtained to determine the approximation for log marginal likelihood.

$$q(\mathbf{f}_c \mid X_c, \mathbf{y}_c) = N(\mathbf{f}_c \mid \hat{\mathbf{f}}_c, H^{-1}) \propto \exp(-\frac{1}{2}(\mathbf{f}_c - \hat{\mathbf{f}}_c)^{\mathrm{T}}(\mathbf{f}_c - \hat{\mathbf{f}}_c))$$
$$\log q(\mathbf{f}_c \mid X_c, \mathbf{y}_c) = -\frac{1}{2}\hat{\mathbf{f}}_c^{\mathrm{T}} K_c^{-1}\hat{\mathbf{f}}_c + \log p(\mathbf{y}_c \mid \hat{\mathbf{f}}_c) - \frac{1}{2}\log |\mathbf{B}_c| \tag{3.29}$$

where $|\mathbf{B}| = |K_c| . |K_c^{-1} + \mathbf{S}| = |I_n + \mathbf{S}^{\frac{1}{2}} K_c \mathbf{S}^{\frac{1}{2}}|$ and $\boldsymbol{\theta}_c$ is a vector of hyperparameters of the covariance function. The GPC model calibration involves deriving the gradients of the approximate marginal likelihood as in equation (3.29) with respect to each of the hyperparameters. The gradient of the marginal likelihood with respect to hyperparameters depends on the hyperparameters explicitly and on the mode $\hat{\mathbf{f}}_c$ and S implicitly because the change in the value of hyperparaemeter $\boldsymbol{\theta}_c$ affects the variation in the optimum value of the posterior mode $\hat{\mathbf{f}}_c$ and S. Thus, the gradient is expressed using chain rule as shown in equation (3.30). The detailed explanation of the derivation of the gradient with respect to hyperparameters and its implementation algorithm can be found in Chapter 5 and Algorithm 5.1 in Rasmussen and Williams (2006).

$$\frac{\partial \log q(\mathbf{y}_c \mid X_c, \boldsymbol{\theta}_c)}{\partial \theta_{cj}} = \frac{\partial \log q(\mathbf{y}_c \mid X_c, \boldsymbol{\theta}_c)}{\partial \theta_{cj}}\bigg|_{\text{explicit}} + \sum_{i=1}^{n} \frac{\partial \log q(\mathbf{y}_c \mid X_c, \boldsymbol{\theta}_c)}{\partial \hat{f}_{ci}} \frac{\partial \hat{f}_{ci}}{\partial \theta_{cj}} \tag{3.30}$$

$$\frac{\partial \log q(\mathbf{y}_c \mid X_c, \boldsymbol{\theta}_c)}{\partial \theta_{cj}}\bigg|_{\text{explicit}} = \frac{1}{2}\hat{\mathbf{f}}_c^{\mathrm{T}} K_c^{-1} \frac{\partial K_c}{\partial \theta_{cj}} K_c^{-1}\hat{\mathbf{f}}_c - \frac{1}{2} tr\left((S^{-1} + K_c)^{-1} \frac{\partial K_c}{\partial \theta_{cj}}\right) \tag{3.31}$$

$$\frac{\partial \log q(\mathbf{y}_c \mid X_c, \boldsymbol{\theta}_c)}{\partial \hat{f}_{ci}} = -\frac{1}{2}[(K_c^{-1} + S)^{-1}]_{ii} \frac{\partial^3}{\partial f_{ci}^3} \log(\mathbf{y}_c \mid \hat{\mathbf{f}}_c) \tag{3.32}$$

$$\frac{\partial \hat{f}_{ci}}{\partial \theta_{cj}} = (I + K_c S)^{-1} \frac{\partial K_c}{\partial \theta_j} \nabla \log p(\mathbf{y}_c \mid \hat{\mathbf{f}}_c) \tag{3.33}$$

3.4.2.3 Prediction

The conditional distribution of the training data $\mathbf{f}_c$ and the future data $\mathbf{f}_{c*}$, given the input X_c and X_{c*}, is expressed as (3.34):

$$p(\mathbf{f}_{c*} \mid \mathbf{f}_c, X_{c*}, X_c, \boldsymbol{\theta}_c^*) = N\left(\begin{bmatrix}\mathbf{f}_c \\ \mathbf{f}_{c*}\end{bmatrix} \mid 0, \begin{bmatrix}K_c & K_{c*} \\ K_{c*}^{\mathrm{T}} & K_{c**}\end{bmatrix}\right) \tag{3.34}$$

By marginalizing over the latent function corresponding to the training data X_c, the prediction can be shown as in equation (3.35):

$$\begin{aligned} p(\mathbf{f}_{c*} \mid X_c, \mathbf{y}_c, X_c, \boldsymbol{\theta}_c^*) &= \int p(\mathbf{f}_{c*}, \mathbf{f}_c \mid X_{c*}, \mathbf{y}_c, X_c, \boldsymbol{\theta}_c^*) d\mathbf{f}_c \\ &= \int p(\mathbf{f}_{c*} \mid \mathbf{f}_c, X_{c*}, X_c, \boldsymbol{\theta}_c^*)\, p(\mathbf{f}_c \mid \mathbf{y}_c, X_c, \boldsymbol{\theta}_c^*) d\mathbf{f}_c \end{aligned} \tag{3.35}$$

where $p(\mathbf{f}_{c*} \mid \mathbf{f}_c, X_{c*}, X_c, \boldsymbol{\theta}_c^*) = N(\mathbf{f}_{c*} \mid K_{c*}^{\mathrm{T}} K_c^{-1} \mathbf{f}_c, K_{c**} - K_{c*}^{\mathrm{T}} K_c^{-1} K_{c*})$

The posterior predictive mean for the future latent function f_{c*} can be expressed as (3.36):

$$E_q[f_{c*} \mid X_c, \mathbf{y}_c, \mathbf{x}_{c*}] = \mathbf{k}(\mathbf{x}_{c*})^{\mathrm{T}} K_c^{-1} \hat{\mathbf{f}}_c \tag{3.36}$$

By using the GP predictive mean using (3.24) in (3.37):

$$E_q[f_{c*} \mid X_c, \mathbf{y}_c, \mathbf{x}_{c*}] = \mathbf{k}(\mathbf{x}_{c*})^{\mathrm{T}} \nabla \log p(\mathbf{y}_c \mid \hat{\mathbf{f}}_c) \tag{3.37}$$

The variance of the future latent function prediction is (3.38):

$$V_q[f_{c*} \mid X_c, \mathbf{y}_c, \mathbf{x}_{c*}] = \mathbf{k}(\mathbf{x}_{c*}, \mathbf{x}_{c*}) - \mathbf{k}_{c*}^{\mathrm{T}} (K_c + S^{-1})^{-1} \mathbf{k}_{c*} \tag{3.38}$$

The predictive probability $p(\mathbf{y}_{c*} = 1 \mid X_{c*}, \mathbf{y}_c, X_c, \boldsymbol{\theta}_c^*)$ of the day being classified as wet is obtained by averaging out the latent function of the data as shown in equation (3.39):

$$p(\mathbf{y}_{c*} \mid X_{c*}, \mathbf{y}_c, X_c, \boldsymbol{\theta}_c) = \int p(\mathbf{y}_{c*} \mid \mathbf{f}_{c*}) p(\mathbf{f}_{c*} \mid X_{c*}, \mathbf{y}_c, X_c, \boldsymbol{\theta}_c) d\mathbf{f}_{c*}$$
$$= \int sig(\mathbf{y}_{c*}, \mathbf{f}_{c*}) p(\mathbf{f}_{c*} \mid X_{c*}, \mathbf{y}_c, X_c, \boldsymbol{\theta}_c) d\mathbf{f}_{c*} \quad (3.39)$$

The predictions of $\mathbf{y}_c$ are given by (3.40):

$$\bar{\pi}_{c*} \sim E_q[f_{c*} \mid X_c, \mathbf{y}_c, \mathbf{x}_{c*}] = \int \sigma(f_{c*}) q(f_{c*} \mid X_c, \mathbf{y}_c, \mathbf{x}_{c*}) \quad (3.40)$$

where $q(f_{c*} \mid X_c, \mathbf{y}_c, \mathbf{x}_{c*})$ is Gaussian with mean in (3.37) and variance in (3.38), $\bar{\pi}_*$ is averaged predictions for the determination of precipitation occurrence. The equation (3.40) cannot be solved directly; logistic sigmoid function can be approximated using the inverse probit function $\Phi(f_{c*})$. The probit function needs to be rescaled to find the best approximation to the logistic function (MacKay, 1992). The rescaled inverse probit function is equation (3.41):

$$\Phi(\lambda f_{c*}); \lambda^2 = \frac{\pi}{8} \quad (3.41)$$

By using the convolution property, (3.41) can be solved analytically as given in (3.42). The derivation to solve this integral can be found in Rasmussen and Williams (2006):

$$\int \sigma(f_{c*}) q(f_{c*} \mid X_c, \mathbf{y}_c, \mathbf{x}_{c*}) = \sigma(\kappa(f_{c*} \mid \mathbf{y}_c) \bar{f}_{c*}) \quad (3.42)$$

where $\kappa^2(f_{c*} \mid \mathbf{y}_c) = \left(1 + \frac{\pi\sigma^2}{8}\right)^{-1}$

The predictions $\bar{\pi}_{c*}$ will have predictions with the probability values ranging from 0 to 1. The decision boundary to divide the classes is 0.5. If the probability is greater than 0.5, it belongs to wet days and the other values belong to dry days.

3.4.3 Precipitation amount determination using GPR

GPR framework for precipitation framework follows the similar derivations presented in Section 2.5 of Chapter 2 where precipitation amount estimation follows the methodology developed by Rasmussen and Williams (2006). As expressed in (3.5), the model function $f_a(\mathbf{x}_a)$ is assumed to be a GP with mean function $\mu(\mathbf{x}_a)$ and covariance function $k(\mathbf{x}_a, \mathbf{x}'_a)$. $f_a(\mathbf{x}_a)$ is also referred as latent function at the input points $\mathbf{x}_a$ in GP. The matrix form of $m(\mathbf{x}_a)$ and $k(\mathbf{x}_a, \mathbf{x}'_a)$ are $\boldsymbol{\mu}_a$ and K_a respectively.

$$\mathbf{y}_a = f_a(\mathbf{x}_a) + \boldsymbol{\varepsilon}_a, \boldsymbol{\varepsilon}_a \sim N(0, \sigma_{na}^2) \tag{3.43}$$

where $\boldsymbol{\varepsilon}_a$ is normally distributed with zero mean and variance, σ_{na}^2. In Gaussian Process view, the model function $f_a(\mathbf{x}_a)$ is assumed to follow GP with a mean function (3.44):

$$\boldsymbol{\mu}_a = \varphi(\mathbf{x}_a)^{\mathrm{T}} \boldsymbol{\beta}_a \tag{3.44}$$

where $\varphi(\mathbf{x}_a)$ is the linear or non-linear basis function and $\boldsymbol{\beta}_a$ are the coefficients for each of the vectors in the basis function. The relationship between the predictors and predictand is not always linear. Thus, the complicated non-linear mean functions can capture the non-linear relationship between the predictors and predictand to improve the prediction results. The basis function can be linear or for example, any order of polynomial function (O'Hagan and Kingman, 1978). The central tendency of the model function is represented by the mean function.

The covariance matrix corresponding to the covariance function should be positive semi-definite. The shape and structure of the covariance between the predictors are described by the covariance function. The commonly used one is the Squared Exponential (SE) covariance function expressed in (3.45). The SE kernel depends on the distance for the dimension m_a, $\|x_a - x'_a\|_2^2$ between the inputs, where $\|.\|^2$ is the Euclidean Norm.

$$\mathbf{k}(\mathbf{x}_a, \mathbf{x}'_a) = \sigma_{fa}^2 \exp\{-\frac{1}{2}(\mathbf{x}_a - \mathbf{x}'_a)^T L_a (\mathbf{x}_a - \mathbf{x}'_a)\} + \sigma_{na}^2 \delta_{ij}, \quad \sigma_{fa}^2 \geq 0; \sigma_{na}^2 \geq 0; diag(L_a) > 0 \tag{3.45}$$

where σ_{fa}^2 is the signal variance, σ_{na}^2 is the noise variance, δ_{ij} is kronecker delta function and L_a is the diagonal positive symmetric matrix consisting of the characteristic correlation length, l_a. The lengthscale characterizes the distance between inputs that change the function value significantly. The predictive variance moves away from the data points when the lengthscale is shorter and the predictions are correlated with each other. If the same characteristic lengthscale is assumed for all the dimensions, then L_a is expressed as (3.46). In this case, the contribution from each predictor is considered equal.

$$L_a = \begin{bmatrix} l_a^{-2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & l_a^{-2} \end{bmatrix} = l_a^{-2} I \tag{3.46}$$

where I is the identity matrix. When different characteristic length scale is assumed for each dimension, L_a is expressed as (3.47):

$$L_a = \begin{bmatrix} l_{a1}^{-2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & l_{ad}^{-2} \end{bmatrix} \tag{3.47}$$

This type of lengthscale definition is also called automatic relevance determination (ARD) (MacKay, 1992; Neal, 1996; Rasmussen and Williams, 2006), where the covariance function automatically computes input predictors for the model. This also gives the idea of suitable predictors for precipitation amount estimation. The detailed description of various covariance functions, its formulation, properties and derivatives can be found in Rasmussen and Williams (2006). The noise term σ_{na}^2 is added to the

covariance function to model the noise in the output. The noise variance is present only in the diagonals of the covariance matrix as the noise process that affects the model predictions is random. The squared exponential kernel (SE) captures the uncertainty (epistemic) contributed by the predictors to the output. The noise variance captures the aleatory uncertainty in the model.

$$f_a(\mathbf{x}_a) \sim GP(\boldsymbol{\mu}_a, k(\mathbf{x}_a, \mathbf{x}'_a)) \tag{3.48}$$

The GP prior has parameters called *hyperparameters* associated with the mean and covariance function. These hyperparameters are not known *a priori* and they need to be optimized. The hyperparameters that are needed to be learnt are $\boldsymbol{\theta}_a = \{\boldsymbol{\beta}_a, \sigma_{fa}^2, \sigma_{na}^2, l_d\}, d = 1,...,l_{m_a}$.

In Bayesian inference, the important step is to choose the priors. The choice of priors over the parameters $\boldsymbol{\beta}_a$ affects the posterior distribution. Generally, the information about the prior is not known and is vague; in this case a non-informative and/or flat prior is assumed (Beck and Katafygiotis, 1998). In this chapter, a uniform prior is assumed to reduce the effect of the prior distribution on $\boldsymbol{\beta}_a$ over the posterior distribution. By Bayes' theorem, the integral likelihood multiplied by the prior yields the marginal likelihood of the model (3.49). For notational convenience, the vector form of $f_a(\mathbf{x}_a)$ is represented as $\mathbf{f}_a$.

$$p(\mathbf{y}_a \mid X_a) = \int p(\mathbf{y}_a \mid \mathbf{f}_a, X_a)\, p(\mathbf{f}_a \mid X_a)\, \mathrm{d}\mathbf{f}_a \tag{3.49}$$

As the Gaussian prior is placed on the modelling function, the log of the prior is expressed as (3.50):

$$\log p(\mathbf{f}_a \mid X_a) = -\frac{1}{2}\mathbf{f}_a^{\mathrm{T}}(K_a + \sigma_{na}^2)^{-1}\mathbf{f}_a - \frac{1}{2}\log|K_a + \sigma_{na}^2| - \frac{n}{2}\log 2\pi \tag{3.50}$$

The likelihood $\mathbf{y}_a \mid \mathbf{f}_a$ follows $N(\mathbf{f}_a, \sigma_{na}^2 I)$. The log marginal likelihood is given in equation (3.51):

$$\begin{aligned} \mathrm{Q} = \log p(\mathbf{y}_a \mid X_a) = & -\frac{1}{2}(\mathbf{y}_a - \boldsymbol{\mu}_a)^T (K_a + \sigma_{na}^2 \mathbf{I})^{-1} (\mathbf{y}_a - \boldsymbol{\mu}_a) \\ & -\frac{1}{2}\log | K_a + \sigma_n^2 \mathbf{I} | - \frac{N}{2}\log 2\pi \end{aligned} \tag{3.51}$$

The marginal likelihood in (3.51) is the objective function. The values of the hyperparameters which maximize (3.51) are the optimal parameters of the model. This means that they are the most probable parameters. The Gaussian process regression algorithm is optimized by following the Algorithm 1and Algorithm 2 in Section 2.5 of Chapter 2. The optimization steps are presented in Section 2.5.1 of Chapter 2.

By the Theorem of Total Probability, the information for future prediction of the model is given by weighting the predictive pdf, $p(f_{a^*} \mid X_a, \mathbf{y}_a, \mathbf{x}_{a^*})$ using the posterior probability, $p(\mathbf{f}_a \mid X_a, \mathbf{y}_a)$. The future predictive pdf is given by (3.52):

$$p(f_{a^*} \mid X_a, \mathbf{y}_a, \mathbf{x}_{a^*}) = \int p(f_{a^*} \mid X_a, \mathbf{x}_a, \mathbf{f}_a) p(\mathbf{f}_a \mid X_a, \mathbf{y}_a) d\mathbf{f}_a \tag{3.52}$$

The predictions for the future data, $\mathbf{x}_{a^*}$ obtained are conditioned on the past data using the property of the conditional Gaussian distribution as in equation (3.53):

$$\begin{bmatrix} \mathbf{f}_a \\ \mathbf{f}_{a^*} \end{bmatrix} \sim N\left(\begin{pmatrix} \mu(\mathbf{x}_a) \\ \mu(\mathbf{x}_{a^*}) \end{pmatrix}, \begin{pmatrix} K_a & K_{a^*} \\ K_{a^*}^{\mathrm{T}} & K_{a^{**}} \end{pmatrix} \right) \tag{3.53}$$

$$M(\mathbf{x}_{a^*}) = f_d(\mathbf{x}_{a^*}, \boldsymbol{\theta}_a) + K_{a^*} K_a^{-1} (\mathbf{y}_a - f_a(\mathbf{x}_a, \boldsymbol{\theta}_a)) \tag{3.54}$$

$$Cov(\mathbf{x}_{a^*}) = K_{a^{**}} - K_{a^*}^{\mathrm{T}} K_a^{-1} K_{a^*} \tag{3.55}$$

The ensembles of $\mathbf{y}_{a^*}$ can be simulated using the mean function (3.54) and the covariance matrix (3.55) following the multivariate normal distribution.

The posterior predictive distribution for the future rainfall prediction $\mathbf{y}_{a^*}$ given the data D_c and D_a is derived by using the theorem of Total probability as shown in equation (3.56):

$$p(\mathbf{y}_{a^*} \mid D_a, D_c, X_{a^*}, X_{c^*}) = \int p(\mathbf{y}_a \mid \mathbf{y}_{a^*}, X_{a^*}, D_a, D_c) p(\mathbf{y}_{c^*} \mid X_{c^*}, D_c) d\mathbf{y}_{a^*} \tag{3.56}$$

where $p(\mathbf{y}_a \mid \mathbf{y}_{a^*}, X_{a^*}, D_a, D_c) = \int p(\mathbf{y}_{a^*} \mid \mathbf{y}_c, X_{a^*}, \boldsymbol{\theta}_a, \boldsymbol{\theta}_c D_a, D_c)\, p(\boldsymbol{\theta}_a \mid D_a) d\boldsymbol{\theta}_a d\boldsymbol{\theta}_c$ in which $p(\mathbf{y}_{a^*} \mid \mathbf{y}_c, X_{a^*}, \boldsymbol{\theta}_a, \boldsymbol{\theta}_c D_a, D_c)$ is obtained from the (3.52) for the future wet days and $p(\mathbf{y}_{c^*} \mid X_{c^*}, D_c)$ is from GPC (3.39). The simulated ensembles $\mathbf{y}_{a^*}$ in the previous step are transformed to the real precipitation values by taking cubic or fourth power.

3.5 Results and discussion

The downscaled precipitation using CFSR reanalysis predictors and CanESM2 predictors are evaluated by comparing it with the observed precipitation for the validation periods. In this chapter, the results are evaluated based on the model's ability to predict precipitation amount and the temporal characteristics of the downscaled precipitation. The GPC is implemented using GPML toolbox (Rasmussen and Nickisch, 2010) in MATLAB for determining the precipitation occurrence. A code for GPR is written in MATLAB for precipitation amount estimation (The MathWorks; Martinez *et al.*, 2011; Ramos, 2012). The downscaled results of SGP-SDM are compared with the results obtained by Lu and Qin (2014) for ASD, GLM and KNN-BNN since the data and study area are the same as those in this study.

The downscaled precipitation is assessed using evaluation statistics such as mean (Mean) and standard deviation (STD) are used to assess the precipitation (Hessami *et al.*, 2008; Fowler and Ekström, 2009; Maraun *et al.*, 2010); maximum amount (Max) (Hessami *et al.*, 2008) and the 90th percentile (PERC90) of the precipitation on wet days are used to evaluate the model's ability to predict extreme precipitation values (Haylock *et al.*, 2006; Goodess *et al.*, 2007). The temporal metrics are assessed using

the proportion of wet days (Pwet) (Semenov *et al.*, 1998). In this chapter, these five evaluation metrics are used to compare the downscaled precipitation with the observed data. The Mean Square Error (MSE) of the downscaled precipitation can be evaluated using (Armstrong and Collopy, 1992). This is expressed as equation in (3.57):

$$MSE = \frac{1}{N}[\sum_{m=1}^{12}(y_{m,obs} - y_{m,sim})^2] \tag{3.57}$$

where N is the data length, $y_{m,obs}$ represents the daily precipitation observed for the month, m and $y_{m,sim}$ represents the daily precipitation simulated for the month, m. The accuracy (*acc*) of the wet and the dry day classification (Chen *et al.*, 2010) is (3.58):

$$acc = \frac{C_{dry} + C_{wet}}{TP_{dry} + TP_{wet}} \tag{3.58}$$

where C_{dry} is the total number of correctly classified dry days, C_{wet} is the total number of correctly classified wet days, TP_{dry} is the total number of dry days and TP_{wet} is the total number of wet days.

APE is the Absolute Percentage Error (Ghosh and Katkar, 2012) which represents the accuracy of the evaluation statistics indicators. APE is given by equation (3.59):

$$\text{APE} = \frac{|P_{i,obs} - P_{i,sim}|}{P_{i,obs}} \tag{3.59}$$

where $P_{i,obs}$ is the observed evaluation statistics of daily precipitation for the i^{th} month and $P_{i,sim}$ is the simulated evaluation statistics of daily precipitation for the i^{th} month. The uncertainty range of the downscaled results is assessed by a mean absolute percentage boundary error (MAPBE) (Lu and Qin, 2014) which is given in equation (3.60):

$$\text{MAPBE} = \frac{1}{n_m}\sum(\text{APE}_{L,i} + \text{APE}_{U,i}) \tag{3.60}$$

where n_m is the number of months, the lower boundary APE for the i^{th} month is given by $\text{APE}_{L,i}$ and the upper boundary APE for the i^{th} month is given by $\text{APE}_{U,i}$.

3.5.1 Validation period result (2005-2010) using CFSR reanalysis data

The proposed SGP-SDM is implemented for each month separately to capture the monthly variations. As the first step, the precipitation occurrence (wet days) is determined using GPC. The performance of GPC in determining wet days is assessed using the percentage of wet and dry days correctly classified by the model. As mentioned in the methodology, GPC is tested with three covariance functions such as linear ARD covariance function including squared exponential ARD covariance function and combined squared exponential and linear ARD covariance function. For each month, the suitability of the covariance function is tested using log marginal likelihood.

Table 3-2 presents the log marginal likelihood of the model combination with three covariance functions for each month computed using CFSR data. The model that gives the highest marginal likelihood is chosen for predicting future wet days. The results in the table show that the linear ARD covariance function is favored for most of the months while few months favor squared exponential ARD covariance function and linear and squared exponential ARD covariance function.

Table 3-3 shows the correct wet day and dry percentage and accuracy predicted by GPC for the CFSR reanalysis dataset validation period (2005 to 2010). The results in the table show that the wet days are predicted better than the dry days. However, for both dry and wet days, the percentage of correct classification is less. The average success rate for the dry days is 30.77% and it is 22.38% for the wet days. The average accuracy of the wet day determination is 53.21%.

GPR is then used to estimate the precipitation amount for the wet days determined using GPC. GPR is calibrated for the wet days to estimate the optimal hyperparameters for prediction. The optimal hyperparameters are then used to compute the posterior predictive mean and variance. The posterior predictive distribution follows Gaussian distribution and thus the predictions may contain negative values. It is important to have fewer number of negative values predicted for more reliability of the model. In Chapter 2, it was shown that GPR predicted with fewer negative values (Chen *et al.*, 2010). The negative values predicted by the model are treated as dry days by making them zero. The proportion of negative values simulated by GPR is checked by comparing the dry-day proportion simulated by GPC and the dry-day proportion by GPR is shown in Table 3-4.

Table 3-2 Log marginal likelihood for the models with different covariance functions calculated using CFSR reanalysis data for the validation period

Month	**Linear ARD covariance function**	**Squared exponential ARD covariance function**	**Linear and squared exponential ARD covariance function**
January	-127.0790	-127.2356	-126.1863
February	-117.1155	-116.5801	-117.1154
March	-124.1504	-124.1503	-124.1504
April	-121.4213	-121.7965	-121.7965
May	-130.2378	-127.9593	-128.2134
June	-124.9134	-125.4688	-125.6507
July	-129.4154	-130.6238	-130.4086
August	-128.7064	-131.7404	-130.2558
September	-124.7665	-125.9205	-125.4779
October	-124.0414	-129.3241	-129.3054
November	-124.7860	-113.9057	-113.8800
December	-112.7226	-113.7003	-113.1148

The dry-day proportion of GPC and GPR is also closer to the observed dry-day proportion. This result shows the ability of GPR to preserve dry day proportion by simulating fewer number of negative values. The number of ensembles to be simulated from GPR is decided based on the previous studies. The investigators (Segond *et al.*, 2007) used 40 ensembles, (Samadi *et al.*, 2013) used 20 ensembles and (Mezghani and

Hingray, 2009) used 50. In this chapter, by comparing various number of ensembles and also based on literature study, 50 ensembles are chosen to represent the confidence interval and to evaluate the monthly statistics. Figure 3-3 shows the average evaluation statistics of all the ensembles (Ghosh and Katkar, 2012) along with the two uncertainty ranges such as Envelop Range (ER) which represents the lower and the upper range and the 5^{th} and the 95^{th} percentile range (P95R) represented as grey region which is compared to the observed evaluation statistics. The figure also presents the evaluation statistics of Mean, STD, Pwet, PERC90 and Max of downscaled precipitation by SGP-SDM corresponding to S44 rain gauge station. It can be seen in the figure that the average of the statistics such as Mean, STD, PERC90 and Max lies within the P95R range and is closer to the observed values. However, the proportion of wet days does not lie within P95R range for all the months and the deviations are seen. This is due to classification error in GPC.

Table 3-3 Accuracy, correct percentage of wet and dry days calculated by GPC using the CFSR reanalysis data for the validation period

Month	Accuracy	Correct dry day	Correct wet day
January	52.69%	25.27%	27.42%
February	45.56%	30.18%	15.38%
March	50.54%	23.66%	26.88%
April	56.11 %	6.67%	49.44%
May	47.31%	16.13%	31.18%
June	51.11%	28.89%	22.22%
July	61.83%	31.72%	30.11%
August	47.85%	22.04%	25.081
September	56.11%	30.56%	25.56%
October	48.39%	26.34%	22.04%
November	53.89%	8.33%	45.56%
December	67.20%	18.82%	48.39%

The MSE comparison of the evaluation statistics of the ensemble average obtained from ASD[1], GLM[1], KNN-BNN[1] ([1]Results obtained from (Lu and Qin, 2014)) and SGP-SDM is shown in Table 3-5. The MSE is calculated using equation (3.57). The MSE for SGP-SDM presented in the Table 3-5 is obtained from three datasets including CFSR reanalysis data, CanESM2 scenarios from two representative pathways

such as 4.5 and 8.5. The table results show that the MSE of Mean from SGP-SDM using CFSR data is significantly less than that of ASD and GLM and is slightly higher than that of KNN-BNN. Similar observation is seen for PERC90 MSE. The MSE of other statistics are significantly less in SGP-SDM (CFSR) compared to ASD, GLM and KNN-BNN. The important aspect is that SGP-SDM is implemented with one class (wet days) instead of predicting rainfall for 8 classes of rainfall. Thus, it is shown that the SGP-SDM predicts rainfall with improved accuracy even with one class compared to KNN-BNN. The Mean, STD, PERC90 and Max simulated by SGP-SDM is 19%, 46.7%, 18.16% and 75.84% less than ASD and Pwet is 25% greater than ASD; compared to GLM, the Mean, STD, Pwet, PERC90 and Max is 42.42%, 49.49%, 28%, 5.73% and 67.8% less; the Mean, Pwet and PERC90 is 35.82%, 60%, 49% greater than KNN-BNN and the STD and Max is 46.35% and 47.54% less than KNN-BNN.

Table 3-4 Comparison of dry-day proportion estimated by GPC and GPR with the observed proportion

Month	Dry-day proportion from GPC	Dry-day proportion from GPR	Observed Dry-day proportion
January	0.4462	0.4510	0.5323
February	0.5680	0.5852	0.5799
March	0.5215	0.5352	0.4462
April	0.2056	0.2356	0.3667
May	0.4140	0.4353	0.4355
June	0.5444	0.5460	0.5222
July	0.5269	0.5448	0.4892
August	0.5108	0.5435	0.5269
September	0.5222	0.5353	0.5278
October	0.6022	0.6120	0.4409
November	0.3167	0.3363	0.3111
December	0.3441	0.3477	0.3602

The MSE is even less than the MSE from KNN-BNN; KNN-BNN uses 8 rainfall classes for downscaling. SGP-SDM is able to downscale the precipitation with one class of rainfall and also with the low resolution GCM predictors. The MSE is even less than SGP-SDM precipitation prediction with CFSR reanalysis data. MSE of the Max statistics from CanESM2 is significantly lower than CFSR reanalysis data. The MSE

from CanESM2 4.5 RCP is 77.11% lower than ASD, 69.5% lower than that from GLM, 50.30% less than KNN-BNN, 5.36% less than SGP-SDM (CFSR). It is also seen that MSE of Max statistics from CanESM2 4.5 RCP is higher than the MSE for CanESM2 8.5 RCP for all statistics except Mean although the values are close. The MSE from CanESM2 8.5 RCP is 80.97% less than ASD, 74.64% less than GLM, 58.68% less than KNN-BNN, 21.24 % less than SGP-SDM (CFSR) and 16.86% less than CanESM2 8.5 predictors.

Table 3-6 presents the comparison of observed and downscaled Number of Extreme Events (NEE) using CFSR data; the average NEE calculated using 50 ensembles for all the months is presented. The magnitude of rainfall greater than 50 $mmday^{-1}$ is set as the threshold for extreme rainfall. The results show that the NEE is predicted well for the months of March, July, August, September and October. For the wet months such as November, December and January, NEE is underestimated. It cannot be concluded about the ability of SGP-SDM in simulating NEE as the variability in the performance is seen for all the months.

The accuracy of the rainfall occurrence determination for the month of December is 67.2%. The comparison of accuracy of the downscaled precipitation amount for the month of December from the models ASD, GLM, KNN-BNN and SGP-SDM is shown in Table 3-7. For SGP-SDM, three datasets such as CFSR, CanESM2 (4.5 and 8.5) are presented. The accuracy results for CanESM2 (4.5 and 8.5) are discussed in the next section of this chapter. Equation (3.58) is used to assess the accuracy of the downscaled results. The accuracy of the precipitation downscaled by SGP-SDM (CFSR) is higher than the accuracy of the precipitation downscaled using ASD, GLM and KNN-BNN. It can also be seen that the accuracy of precipitation downscaling using GPR is closer to the accuracy of the wet days classification from GPC. Thus, the number of negative values simulated by SGP-SDM (CFSR) is small preserving the accuracy of occurrence determination. When compared to the accuracy of KNN-BUQSDM which is presented in Table 2-7, the accuracy of SGP-SDM precipitation downscaling is slightly higher even with the use of single class compared to two classes in the KNN-BUQSDM. Thus,

SGP-SDM (CFSR) shows better prediction accuracy. The accuracy of the precipitation downscaled by SGP-SDM (CanESM2 4.5 and 8.5) is higher than the accuracy of the precipitation downscaled using ASD and GLM. The accuracy of downscaled precipitation from CanESM2 4.5 is comparable to KNN-BNN; downscaled precipitation from CanESM2 8.5 is higher than CanESM2 4.5 and KNN-BNN. The results show that CanESM2 8.5 performs slightly better than CanESM2 4.5 in terms of accuracy.

Table 3-8 presents the MAPBE of the confidence interval of the ensembles to assess the quantitative levels of uncertainty in the downscaled predictions. Two values such as ER and P95R MAPBE are presented in the table. ER values are calculated using the full range of data (that is, using the minimum and maximum of the ensembles) and P95R values are calculated using the 5^{th} and 95^{th} percentile range of the data to calculate MAPE It is calculated using ER and P95R values. The MAPBE is calculates using (3.60). For Mean, STD and Pwet MAPBE, the ER and P95R values from SGP-SDM (CFSR) are less than the ER and P95R MAPBE values from the ASD and GLM and are slightly greater than those from KNN-BNN. The MAPBE values of PERC90 from SGP-SDM (CFSR) are less than ASD and slightly less than those from GLM; when compared with KNN-BNN, the MAPBE values of SGP-SDM (CFSR) are large. Since 8 classes were used for KNN-BNN, the uncertainty range was very less. However, the MAPBE values of Max from GLM and KNN-BNN are less than SGP-SDM (CFSR). The uncertainty range for SGP-SDM (CFSR) is less than the one for ASD. While the uncertainty range for Max is comparable with GLM and KNN-BNN, the MSE is less compared to all other models. The higher uncertainty range is attributed to the dataset used. The MAPBE values of all the statistics for both CanESM2 4.5 and CanESM2 8.5 are comparable to MAPBE values from ASD and GLM. It should be noticed that the results downscaled from low resolution GCM predictors are comparable to the results downscaled using high resolution CFSR data using ASD and GLM. This shows the ability of SGP-SDM in downscaling climate variables from low resolution GCM predictors.

Table 3-9 presents minimum, average and maximum of the evaluation statistics calculated using 50 ensembles from ASD, GLM, KNN-BNN and SGP-SDM using CFSR data for the month of December. The table also presents the results for SGP-SDM results obtained from CanESM2 (4.5 and 8.5 data). The 90^{th} percentile is underestimated by all the models. The Mean and Pwet statistics simulated by SGP-SDM (CFSR) is closer to the observed mean value. Notable prediction results are obtained for Max statistics from SGP-SDM (CFSR); the maximum value predicted by SGP-SDM is 141.37 which is very close to observed value 141.4. The STD is underestimated by SGP-SDM (CFSR) data. As the prediction of extreme value is very important in impact studies, SGP-SDM can be a viable statistical downscaling approach for climate change impact studies. The results show notable improvement in the average evaluation statistics prediction by SGP-SDM. The 90^{th} percentile is underestimated by all other models; however, SGP-SDM (CanESM2 4.5 and 8.5) simulates 90^{th} percentile value that is slightly closer to the observed values. The Mean statistics simulated by SGP-SDM (CanESM2 4.5) is closer to the observed values compared with SGP-SDM (CanESM2 8.5) Mean statistics. Notable prediction results are obtained for Max statistics from SGP-SDM (CanESM2 4.5 and 8.5); the predicted maximum value is very close to the observed value. The STD is slightly underestimated by SGP-SDM (CanESM2 4.5 and 8.5) data while the performance from both models are similar. Pwet is underestimated by CanESM2 4.5 compared to CanEMS2 8.5 predictors.

3.5.2 Comparison of results from CanESM2 RCP 4.4 and 8.5

One rain gauge station (S44) is chosen to illustrate the ability from SGP-SDM in downscaling GCM predictors. The NCEP predictors are re-gridded to CanESM2 grid are used for model calibration. The calibration period is from 1980-2005. For model validation, GCM predictors from two representative pathways such as 8.5 and 4.5 are considered. The validation period is from 2006 to 2010. Figure 3-4 (a-e) presents the Mean, STD, Pwet, PERC90 and Max of the downscaled precipitation using CanESM2

Table 3-5 Mean Square Error (MSE) of the average of the evaluation statistics

Statistics	MSE-ASD1	MSE-GLM1	MSE-KNN-BNN1	MSE-SGP-SDM (CFSR)	MSE-SGP-SDM (CanESM2-4.5)	MSE-SGP-SDM (CanESM2-8.5)
Mean	1.13	1.58	0.67	0.91	0.52	0.81
STD	4.69	4.95	4.66	2.50	2.02	2.01
Pwet	0.004	0.007	0.002	0.005	0.0007	0.002
PERC90	33.54	29.12	18.36	27.45	17.55	17.35
Max	869.33	652.42	400.38	210.05	199.00	165.44

[1]Results obtained from (Lu and Qin, 2014)

Table 3-6 Comparison of observed and simulated NEE for validation period (2005-2010) using CFSR data

Month	**SGP-SDM Avg NEE (CFSR)**	**Observed NEE**
January	4.38	6
February	2.1	3
March	3.94	3
April	6.52	8
May	4.7	3
June	4.76	6
July	4.12	4
August	3.28	4
September	5.14	5
October	3.38	3
November	3.94	5
December	8.12	12

Table 3-7 Accuracy of downscaled precipitation for the month of December

SDM	**Min**	**Max**	**Average**
ASD[1]	0.560	0.605	0.582
GLM[1]	0.550	0.591	0.566
KNN-BNN[1]	0.582	0.597	0.591
SGP-SDM (CFSR)	0.661	0.672	0.683
SGP-SDM (CanESM2 4.5)	0.587	0.597	0.607
SGP-SDM (CanESM2 8.5)	0.60	0.614	0.626

[1]Results obtained from (Lu and Qin, 2014)

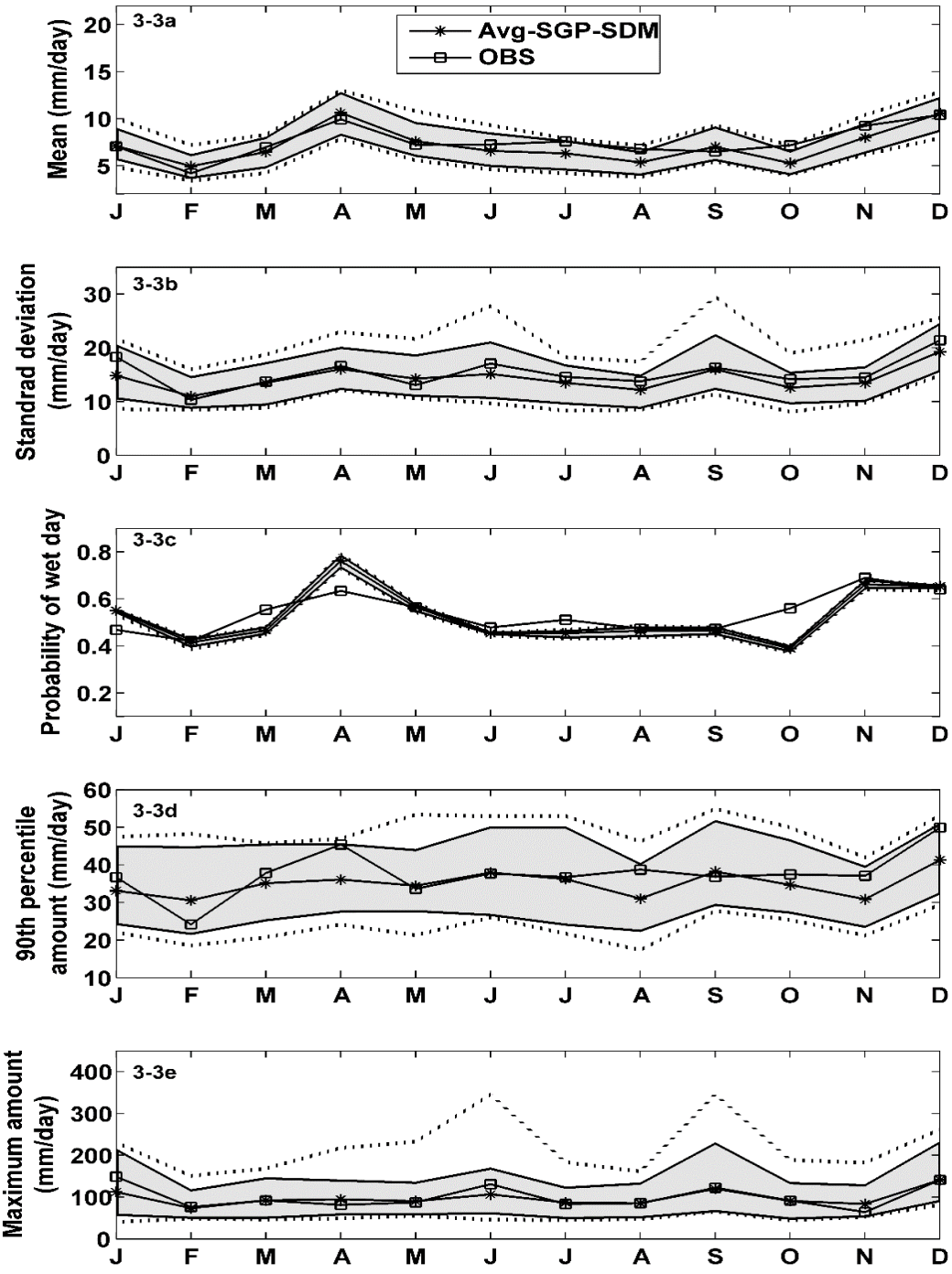


Figure 3-3 Monthly mean evaluation statistics for the rain gauge stations S44. The shaded area represents the 5th and 9th percentile of the ensembles. The dashed line represents the minimum and maximum values of the ensemble statistic

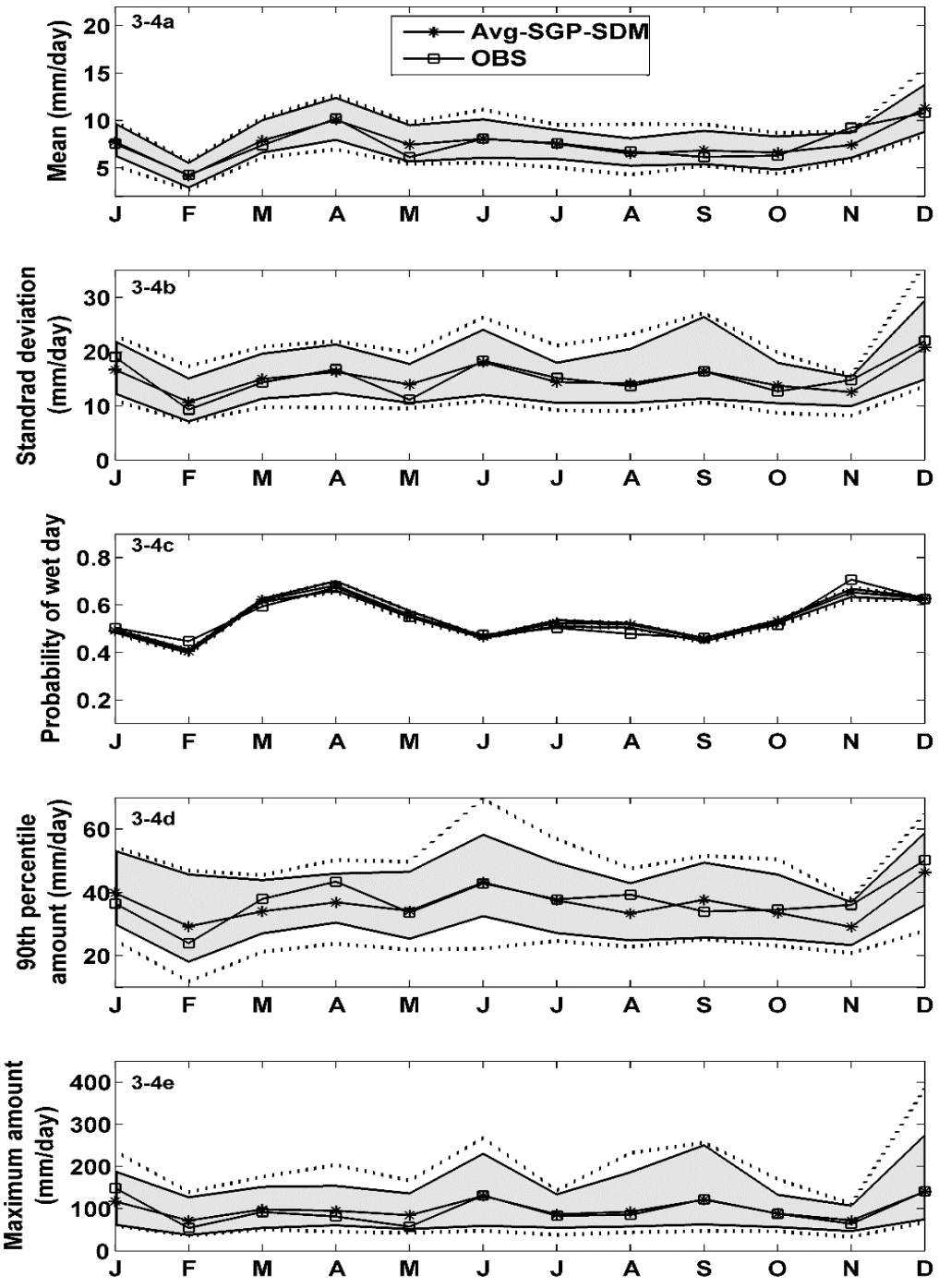


Figure 3-4 Monthly mean evaluation statistics for the rain gauge stations S44 using CanESM2 RCP 4.5 scenarios. The shaded area represents the 5th and 9th percentile of the ensembles. The dashed line represents the minimum and maximum values of the ensemble statistics

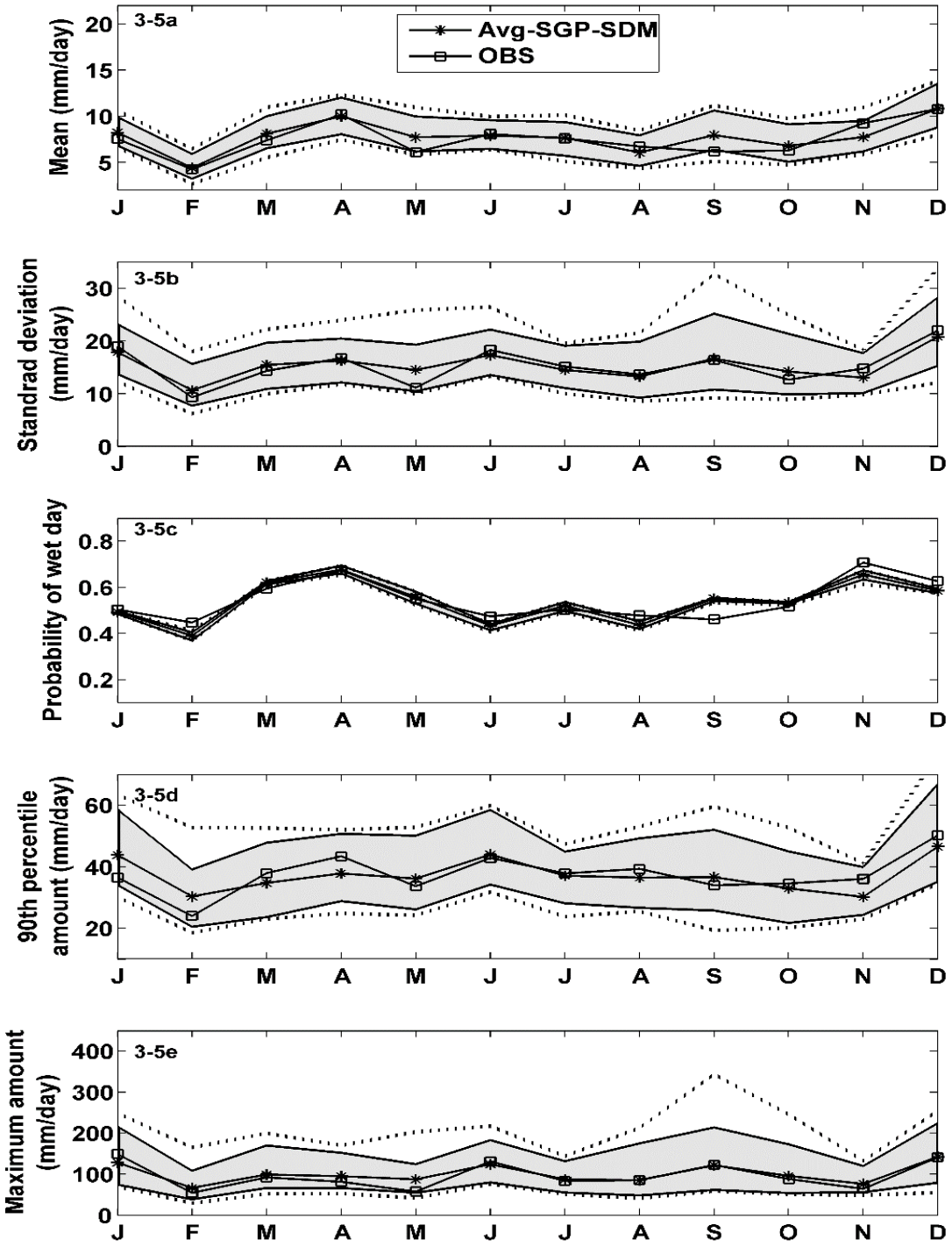


Figure 3-5 Monthly mean evaluation statistics for the rain gauge stations S44 using CanESM2 RCP 8.5 scenarios. The shaded area represents the 5th and 9th percentile of the ensembles. The dashed line represents the minimum and maximum values of the ensemble statistics

Table 3-8 MAPBE value of downscaled precipitation envelop obtained from 50 ensembles

MAPBE	ASD[1]		GLM[1]		KNN-BNN[1]		SGP-SDM (CFSR)		SGP-SDM (CanESM2 4.5 RCP)		SGP-SDM (CanESM2 8.5 RCP)	
	ER	P95R	ER	P95R	ER	P95R	ER	P95R	ER	P95R	ER	P95R
Mean	0.66	0.53	0.67	0.52	0.22	0.21	0.58	0.45	0.64	0.49	0.67	0.51
STD	0.94	0.70	0.83	0.60	0.30	0.28	0.77	0.50	0.85	0.63	0.97	0.66
Pwet	0.28	0.25	0.67	0.32	0.13	0.13	0.21	0.19	0.08	0.08	0.15	0.14
PERC90	0.80	0.60	0.74	0.58	0.29	0.24	0.73	0.55	0.81	0.58	0.84	0.62
Max	1.83	1.26	1.58	1.21	0.61	0.52	1.69	0.98	1.68	1.21	1.74	1.11

[1]Results obtained from (Lu and Qin, 2014)

Table 3-9 Minimum, average and maximum evaluation statistical indicators obtained using 50 ensemble simulated by SGP-SDM for the month of December

	ASD[1] [2](min, avg, max)	GLM[1] [2](min, avg, max)	KNN-BNN[1] [2](min, avg, max)	SGP-SDM (CFSR) [2](min, avg, max)	SGP-SDM (CanESM2 8.5 RCP) [2](min, avg, max)	SGP-SDM (CanESM2 4.5 RCP) [2](min, avg, max)	OBS
Mean	6.28, 9.25, 11.65	6.29, 9.96, 12.12	9.85, 10.19, 10.49	7.91, 10.53, 12.82	7.96, 10.80, 13.83	8.41, 11.28, 15.51	10.39
SD	10.39, 16.06, 25.34	12.46, 17.34, 24.89	19.80, 20.74, 22.12	14.84, 19.24, 25.58	12.06, 20.78, 33.70	13.65, 20.73, 36.12	21.31
Pwet	0.66, 0.71, 0.76	0.55, 0.65, 0.73	0.63, 0.66, 0.68	0.63, 0.65, 0.66	0.57, 0.58, 0.59	0.62, 0.63, 0.63	0.64
PERC90	27.54, 33.69, 48.77	26.11, 38.42, 48.61	41.74, 44.84, 48.52	29.37, 41.26, 53.17	34.46, 46.48, 76.06	27.91, 46.31, 64.89	50.40
Max	58.84,109.45,224.68	68.66,115.54,282.27	121.03,139.39,158.76	81.63,141.37,261.86	55.48,42.57,254.57	68.11,140.76,385.77	141.40

[1]Results obtained from (Lu and Qin, 2014). [2] (Minimum, average and maximum values of 50 ensembles)

RCP 4.5 predictors by SGP-SDM corresponding to S44 rain gauge station. It can be seen in the figure that the average of the statistics such as Mean, STD, PERC90 and Max lies within the P95R range and is closer to the observed values. However, Pwet does not lie within P95R range for 5 months and the deviations are seen. For the remaining months, Pwet is predicted well. This is due to classification errors in GPC. When compared with the downscaling results using CFSR, the evaluation statistics are predicted well by the model. This can be attributed to the predictors' ability in representing the rainfall. Also there is a large number of CanESM2 predictors available for Singapore compared to CFSR predictors. Figure 3-5(a-e) presents the Mean, STD, Pwet, PERC90 and Max of the downscaled precipitation using CanESM2 RCP 8.5 predictors by SGP-SDM corresponding to S44 rain gauge station. It can be seen in the figure that the average of the statistics such as Mean, STD, PERC90 and Max lies within the P95R range and is closer to the observed values. The average evaluation statistics for Pwet Figure 3-5c is simulated well for 9 months except for February, September and November. However, the results are inconclusive to find out the best scenarios to represent Singapore climate from RCP 4.5 and 8.5. Error! Not a valid bookmark self-reference. presents the comparison of observed and downscaled Number of Extreme Events (NEE) using CanESM2 predictors; the average NEE for all the 50 ensembles for all the months is presented. The magnitude of rainfall greater than 50 $mmday^{-1}$ is set as the threshold for extreme rainfall. The results show that the NEE is predicted well for the months of January, February, June and July by CanESM2 4.5 predictors. NEE is slightly underestimated for the months of April, November and December; however, the predicted NEE values are closer to the observed NEE. For the months of May, August, September and October, NEE is overestimated. For CanESM2 8.5 predictors, NEE is underestimated for the months of April, November and December whereas for the months January, February, July and August, NEE is predicted well. From the results, the best CanESM2 predictors cannot be concluded for NEE estimation as the performance is similar.

3.5.3 Discussions on SGP-SDM structure

Probabilistic quantification of uncertainty in downscaling precipitation is one of the most crucial steps in climate change impact studies. SGP-SDM provides a mechanism to quantify uncertainty (both in precipitation occurrence determination and precipitation amount estimation), calibrate and validate the model in a single step by coupling the model function and residual function within a Bayesian framework. The advantages of SGP-SDM are 1) its flexibility in implementation 2) confidence interval in predictions and 3) improved prediction accuracy with less number of calibration data.

Table 3-10 Comparison of observed and simulated NEE for validation period (2006-2010) using CanESM2 data

Months	**SGP-SDM Avg NEE (4.5 RCP)**	**SGP-SDM Avg NEE (8.5 RCP)**	**Observed NEE**
January	4.98	5.7	5
February	1.84	1.84	2
March	4.36	4.62	3
April	5.16	5.52	7
May	3.66	4.26	2
June	5.42	4.96	6
July	4.2	3.82	4
August	3.7	3.34	3
September	3.9	4.66	3
October	3.74	3.7	2
November	2.72	2.98	5
December	7.88	7.72	10

The limitations of SGP-SDM are 1) they are not sparse that is, they use the whole data for model calibration and prediction 2) when the dimension is high, the model may not be efficient and 3) increase in computational time when a large amount of data is used owing to the need of covariance matrix computation (Rasmussen and Williams, 2006). The residuals of SGP-SDM are assumed to be correlated and follow Gaussian distribution. In reality, the precipitation does not follow normal distribution; the assumption of Gaussian distribution is not valid to model precipitation. In order to model precipitation, SGP-SDM transforms the precipitation data using cubic/fourth

root to make it closer to normal distribution similar to SDSM and ASD model. However, transformation may not solve the distribution problem entirely. Thus, the residual should be modelled by assuming a non-Gaussian distribution. There are several ongoing research studies on non-Gaussian Processes model (Gurley, 1997). As SGP-SDM serves as basic step in using GP for downscaling, SGP-SDM can be extended easily to incorporate non-GP models.

CHAPTER 4 **Integrated MGP-SDM (A Bayesian uncertainty quantification framework for multisite statistical downscaling with residual coupling) and disaggregation to generate hourly precipitation at a station scale**

4.1 Abstract

The content of this Chapter is extracted from the article that will be submitted to a journal. An integrated multi-site SDM and temporal disaggregation model is proposed in this Chapter to simulate fine spatial and temporal resolution precipitation for studying the impact of climate change on hydrology. A multi-site regression based statistical downscaling model coupled with both uncertainty quantification tool and spatial dependency function using stochastic processes to downscale precipitation at multiple sites simultaneously is developed and it is named as MGP-SDM (Multi-site Gaussian Processes-Statistical Downscaling Model). MGP-SDM is a full Bayesian inference model in which the multi-output GP classification and regression are used for precipitation occurrence determination and precipitation amount estimation respectively. The between-site spatial dependency is modeled by specifying a cross-covariance function and the dependency between the residuals within each site is captured by specifying an auto-covariance function implicitly in MGP-SDM. The residuals of the model are assumed to be dependent and are treated as the stochastic processes following Gaussian distribution. MGP-SDM is a non-parametric model by assuming GP prior over the model function instead of model parameters for Bayesian inference; MGP-SDM gives joint posterior predictive distribution of precipitation as outputs at all the sites. The precipitation ensembles can be simulated directly from the predictive distribution. MGP-SDM is then integrated with the KNN disaggregation model to increase the temporal resolution of the precipitation required by the hydrological models. The daily precipitation series at the three sites are downscaled jointly in Singapore using NCEP reanalysis data for model calibration and HadCM3

GCM scenarios for generating future daily precipitation ensembles. The performance of the downscaling model is assessed by comparing the downscaled daily precipitation with the observed daily precipitation at three stations. The hourly precipitation obtained from KNN disaggregation model is also compared with the hourly observed rainfall to assess the performance of the disaggregation model. The statistics such as mean, standard deviation, proportion of wet days, 90^{th} percentile of the rainfall and maximum amount are compared for both downscaled and disaggregated precipitation series at three sites. The hourly and daily precipitation is also projected for the future periods 2011-2040, 2041-2070 and 2071-2099.

4.2 Introduction

Precipitation plays an important role in hydrological cycle in the climate system. The change in the climate system affects the occurrence of rainfall pattern and its intensity which in turn affects the hydrological systems at a global and a regional level (Solomon, 2007; Schmocker-Fackel and Naef, 2010). These changes will indirectly impact the management of water resources causing huge social and economic losses in case of extreme weather events especially in an urban area (Grum *et al.*, 2006; De Toffol *et al.*, 2009; Willems, 2012). The assessment of impact of changing climate on water resources has gained a significant interest in the past few years (Chen *et al.*, 2010). As the spatial resolution of GCM is too coarse, SDMs are designed to generate local climate variables such as precipitation, temperature and humidity at station scale from large scale GCM predictors (Wilby and Wigley, 1997; Wilby *et al.*, 1998). There is plethora of SDMs available to downscale the precipitation at a single site. However, in hydrology, the precipitation modelling is considered as a challenge in terms its spatiotemporal intermittence (that is, the precipitation amount at a given site also depends on the precipitation occurrence at the rain gauge stations surrounding it), highly skewed distribution and its complex stochastic dependencies. The stream flow also depends on the spatial distribution of the precipitation within the watershed (Xu, 1999). Many research studies shown that if the spatial dependence is ignored,

significant errors were presented in the impact assessment results (Srikanthan and McMahon, 2001; Qian *et al.*, 2002; Khalili *et al.*, 2013). It is important to consider precipitation at multiple sites to simulate flood events as the downscaled precipitation is used in the hydrological model to simulate stream flow for impact assessment (Xu, 1999). The downscaling of precipitation at multiple sites always consists of uncertainties due to GCM, emission scenarios, downscaling model structure and natural complex process. In addition to the above drawbacks, the daily rainfall downscaled at multiple sites cannot be directly used as input in hydrological models, as the hydrological model requires high temporal resolution such as hourly precipitation for stream flow simulation. The three limitations in the simulating hourly precipitation from GCM predictors are 1) existing multi-site downscaling method do not consider spatial correlation automatically in the model 2) there is no generalized uncertainty quantification framework for multi-site statistical downscaling and 3) there are limited number of studies on integrating downscaling and the disaggregation model for simulating hourly precipitation. In this study, a multi-site SDM based on stochastic processes integrated with disaggregation model is proposed to solve the above mentioned issues.

Statistical downscaling methods can be considered as an alternative to dynamic downscaling methods because of their ease of implementation and less computational requirements (Benestad *et al.*, 2008). Several multisite statistical downscaling approaches have been presented to downscale precipitation at multiple sites. The commonly-used regression based approaches for multisite statistical downscaling are GLM (Generalized Linear Model) (Chandler and Wheater, 2002) and Multivariate Multisite Statistical Downscaling Model (MMSDM) (Jeong *et al.*, 2012).

Let the historic data for the precipitation occurrence determination model, D_{mc} of n observations, $D_{mc} = \{(\mathbf{x}_{mci}, \mathbf{y}_{mci}) \mid i = 1,...,n\}$ where $\mathbf{x}_{mc}$ represents the GCM predictors with dimension m_{mc} and $\mathbf{y}_{mc}$ is the binary classification output (wet/dry day). In vector form, the predictor data can be represented as a matrix X_{mc} with dimension $m_{mc} \times n$ and

the wet/dry day classification output vector is denoted as $\mathbf{y}_{mc}$. The historic data D_{ma} for the precipitation amount estimation model of n observations are represented as $D_{ma} = \{(\mathbf{x}_{mai}, y_{mai}) \mid i = 1,...,n\}$ where $\mathbf{x}_{ma}$ is the GCM predictors with dimension m_{ma} and $\mathbf{y}_{ma}$ represents the rainfall amount. The number of sites is represented as s.

MMSDM is a multi-site hybrid statistical downscaling technique which is developed by integrating the multivariate regression and stochastic weather generator to simulate daily precipitation at several sites. In MMSDM, the multivariate multiple linear regression is used for both precipitation occurrence determination and precipitation amount estimation at multiple sites. Let the precipitation occurrence at s number of multiple sites is represented as $\mathbf{O}$ with dimension $n \times s$. The deterministic series of precipitation amount is modelled by (4.1):

$$\hat{\mathbf{O}} = \hat{a}_0 + X_{mc}\hat{\mathbf{a}} \tag{4.1}$$

where $\hat{\mathbf{O}}$ is the downscaled deterministic series of precipitation occurrence matrix which has elements $\hat{O}_{ij}$, where $i = (1,2,...,n)$ represents the day at the site $j = (1,2,...,s)$.

The model parameters including the constant $\hat{a}_0$ and the parameter matrix $\hat{\mathbf{a}}$ with dimension $m_{mc} \times s$ are estimated using Ordinary Least Square (OLS) estimation. Let the precipitation amount $\mathbf{y}_{ma}$ is transformed to follow normal distribution $\mathbf{R}_{maij} = \mathbf{y}_{maij}^{1/3}$ on a day i at a site j. The determinant series precipitation amount is expressed as (4.2):

$$\hat{\mathbf{R}}_a = \hat{b}_0 + X_{ma}\hat{\mathbf{b}} \tag{4.2}$$

where $\hat{\mathbf{R}}_a$ with dimension $n \times s$ is the downscaled deterministic precipitation amount series; the model parameters such as constant term $\hat{b}_0$ and the parameter coefficients $\hat{\mathbf{b}}$ with dimension $n \times s$ are determined using OLS estimation.

Although regression based statistical downscaling models have been implemented in several studies, they have major drawbacks in simulating observed variability (von Storch, 1999). In MMSDM, all the days including wet and dry are also used for training since all the stations cannot be wet or dry at the same time. The results from MMSDM showed that the downscaled precipitation amount series were underestimated as zero values were used for training the model. This bias was removed by using a statistical adjusting technique called probability distribution mapping. The precipitation occurrence and amount are obtained by (4.1) and (4.2) where precipitation is a deterministic series; in order to simulate the ensembles of the precipitation series, a residual matrix with a multivariate normal distribution matrix is added separately as a stochastic component to the deterministic precipitation series. The multivariate normal distribution matrix helps to reproduce the temporal variability, spatial dependency and to reproduce the variability in the predictions. The detailed explanation of MMSDM can be found in (Jeong *et al.*, 2012). Another drawback is that the spatial dependence between multi-sites cannot be preserved well in regression-based approaches; the interstation correlations of the precipitation amounts are underestimated and the time-domain variability cannot be reproduced well (Wilby *et al.*, 2003). In addition, the precipitation does not follow normal distribution as there can be varying number of dry and heavy rainfall days; this causes difficulty to model precipitation by assuming normal distribution SDM. The daily precipitation is skewed and normality assumptions may not hold for downscaling (Chandler and Wheater, 2002; Segond, 2006). GLM is employed to extend the linear regression to assume logistic regression for precipitation occurrence determination and gamma regression for precipitation amount estimation (Coe and Stern, 1982; Stern and Coe, 1984; Chandler and Wheater, 2002). GLM has been implemented in several research studies to generate precipitation at multiple sites (Yang *et al.*, 2005; Frost, 2007; Frost *et al.*, 2011; Liu *et al.*, 2011). The advantage of GLM is that it can model precipitation even when there are scarce data (Kigobe and Van Griensven, 2010). The Generalized Linear Modelling of daily CLIMate sequences (GLIMCLIM) is a GLM based downscaling model (Chandler and Wheater, 2002). In GLM, the spatial dependence in precipitation occurrence determination using logistic

inter-site correlation and the empirical correlation between each pair of sites is used to model the precipitation amount. The residual is fitted with extreme value distribution to simulate downscaled ensembles. Thus the model parameters, the residual parameters and the spatial correlation need to be computed separately. The disadvantage of fitting the model and residual parameters separately is shown in the Chapters 2 and Chapter 3. Thus, a model that combines the model parameter and the residuals is needed for multi-site statistical downscaling. This can be achieved by using probabilistic downscaling model.

The probabilistic approach is developed to downscale variables at multiple sites namely expanded downscaling (Bürger, 1996). In probabilistic regression approaches, the distribution parameters of the predictand are computed to generate the predictive precipitation distribution directly. This eliminates the need to add the residuals to reproduce the observed variability in the prediction model. However, the disadvantage of placing constraint for preserving covariance in expanded downscaling led to the development of Probabilistic Gaussian Copula Regression (PGCR) for downscaling precipitation and temperature (Alaya *et al.*, 2014). In this framework, the probabilistic regression is used for downscaling precipitation and temperature. The dependence between the climate variables at multiple sites is described by Gaussian Copula in the climate variable simulation step. The PGCR is further improved by utilizing advanced Quantile Regression with Gaussian Copula (QRGC) to overcome the disadvantages of the traditional regression techniques (Alaya *et al.*, 2015). This methodology is implemented to downscale temperature and precipitation in the province of Quebec, Canada. The results of QRGC were compared with the traditional MMSDM and shown that QRGC performed well. In PGCR and QRGC, probabilistic regression and quantile regression are used to obtain the precipitation distribution at each site respectively. The dependence is incorporated using Gaussian Copula in both of these methods. In a previous research study, it is noticed that GLM produced heavy tail at some of the sites causing difficulty in capturing all types of behavior of rainfall amount (Alaya *et al.*, 2015). Alaya *et al.* (2015) proposed a multi-site downscaling model named Bernoulli–

generalized Pareto multivariate autoregressive (BMAR) model in which Bernoulli-generalized Pareto distribution is used to capture all types of variability in prediction. In their work, they considered several distributions for rainfall amount estimation such as gamma, mixed exponential, generalized Pareto and Weibull (WEI) distribution. Based on the test for applicability of different distributions, they found that generalized Pareto distribution was suited for modelling precipitation amount at all stations. The distribution of the precipitation obtained at each site by using probabilistic regression; the spatial dependence structure is estimated using the conditional sampling using the latent variable. The model output gave the simulation of precipitation amount cumulative probabilities ranging from 0 to 1. The cumulative probability was then transformed using multivariate first-order autoregressive (MAR(1)) to capture the spatial dependence structure between the sites for each day. However, the dependency between each day was not considered in downscaling. Some of the earlier research studies compared the performance of multi-site statistical downscaling techniques (Frost *et al.*, 2011; Liu *et al.*, 2013); none of the model is reported as the best model for multi-site downscaling of precipitation at multiple sites. This is because it is noted that in all these approaches the spatial dependence was not included in model calibration for precipitation amount estimation. The between-site correlation is incorporated after computing the parameters for determining the precipitation occurrence and estimating the precipitation amount. A multi-site downscaling model that simultaneously estimates the spatial correlation and residual fitting with model calibration is needed.

Another challenge is that downscaling models do not have a principled way to quantify and propagate uncertainty in GCM scenarios and in the model structure. The uncertainty propagates from GCM to downscaling model and then eventually to hydrological models. There are limited studies on quantifying uncertainty in multi-site statistical downscaling model. Typically the uncertainty in multisite downscaling is handled by analyzing the sensitivity of the different multisite statistical downscaling models in simulating future precipitation. The regression-based multi-site statistical downscaling model also has epistemic uncertainty coming from the model function and

aleatory uncertainty emerging from the residuals (2.3) similar to the single site SDM. The contemporary multi-site statistical downscaling techniques have not been able to couple the uncertainty quantification tool with the multisite statistical downscaling model. A multisite statistical downscaling model that addresses aforementioned issues in model structure and uncertainty quantification is needed.

Multi-site SDM can be considered as a problem of downscaling correlated precipitation from multiple sites. In other fields, this type of problem is also referred as multi-output regression. In Geostatistics, this is known as *cokriging* (Isaaks and Srivastava, 1989). GP is a non-parametric Bayesian approach where the model function is represented by covariance and mean functions (2.6). Multi-output GP is particularly useful when data from several rainfall sites are available. Since the precipitation at all the sites will be different, so pooling all the data within the model may not be appropriate. The advantage of multi-output GP is that dry days need not to be included in the model calibration; the model framework of the multi-output GP can handle different lengths of data at multiple sites. The wet days at multiple sites can be given as inputs for model calibration. In multi-output GP, the between-site spatial correlation is modeled using cross-covariance function and the correlation within precipitation for each site is modelled using auto-covariance function. As explained in equation (2.6) in Chapter 2, the errors are dependent and assumed to follow a stochastic process following Gaussian distribution. The main challenge in extending the single site GP-SDM presented in the Chapter 2 to multi-site downscaling is the requirement to compute between site cross covariance. The advantage is that the prediction accuracy can be improved by using cross covariance between the sites (Boyle and Frean, 2004). Other than specifying the covariance matrix with the model, the residual parameter calibration is also coupled within the model framework using Bayesian framework. The multisite statistical downscaling model developed using multi output GP classification and regression is named as MGP-SDM.

The hydrological studies need high temporal resolution future precipitation data which can be obtained by integrating the downscaled precipitation at multiple sites with the

multisite disaggregation model. Segond *et al.* (2007) proposed to combined downscaling and disaggregation approach to increase the spatial and temporal resolution of the climate variables. They have integrated GLM, HYETOS (temporal disaggregation) and multisite transformation using an artificial profile to downscale climate variables at multiple sites. In another study, Mezghani and Hingray (2009) combined downscaling and disaggregation to downscale temperature and rainfall. They used GLM for downscaling and KNN for disaggregation. The results from both of the above mentioned framework showed good performances and reproduced the statistical properties of the climate variables. Lu and Qin (2014) performed inter-comparison of different disaggregation methods integrated with multi-site GLM downscaling model to generate high resolution precipitation in Singapore. They used master station based approach to generate high resolution rainfall. In master station approach, the downscaled rainfall is disaggregated to hourly scale at single site (master station). The rainfall at single site was disaggregated using two approaches including HYETOS and KNN. The hourly rainfall was then simulated at the remaining stations using two approaches such as MuDRain and KNN disaggregation models. The test results showed that the hourly rainfall generated using multi-site GLM integrated with KNN for single site disaggregation and MuDRain for multisite disaggregation reproduced observed statistics and spatial correlation well at multiple sites. In this study, KNN is chosen for disaggregation for ease of implementation based on previous literature study results.

The objective of this study is to develop a coupled uncertainty quantification tool with multi-site SDM for downscaling precipitation in Singapore (an urban area with tropical climate). The novel contributions of this chapter work are as follows:

1) Development of multi-site regression based statistical downscaling model (MGP-SDM) using stochastic process where the spatial dependence, residual and model function parameters are estimated simultaneously using a Bayesian framework. The model structure couples model calibration, site correlation information, uncertainty quantification and prediction together. This Chapter uses a multi output Gaussian process (GP) to couple the uncertainty

quantification tool with the multisite downscaling model to predict the precipitation at all sites jointly. MGP-SDM is implemented to downscale precipitation at multiple sites in Singapore. This framework is implemented for each month to capture monthly variations in precipitation.

2) The proposed multi-site SDM consists of generalized uncertainty quantification framework to give the posterior predictive distribution of precipitation at all sites as the model output.
3) MGP-SDM is integrated with KNN temporal disaggregation model to simulate the hourly precipitation for climate change impact assessment on hydrology.

4.3 Data and study area

The study area is Singapore as in Figure 4-1 and is located between 1°N and 2°N latitudes and 103.8° E and 104°E longitudes. The three rainfall stations for multi-site downscaling are presented in Figure 4-1. Singapore has a land area of 716.1 km^2 with an average annual precipitation of about 2340 mm. There are two monsoon periods in Singapore which starts from December to March and from June to September (NEA, 2016). The observed rainfall data is available at all the stations shown**.** The observed precipitation data for Singapore for the period of 1980-2000 are obtained from S46, S55 and S69 rain gauge station for this study and the records for the stations considered are complete (NEA, 2016). The observed precipitation greater than 0.1 mm is considered as wet day for precipitation occurrence determination and precipitation amount estimation (Liu *et al.*, 2011; Lindau and Simmer, 2013; Taye and Willems, 2013; Pervez and Henebry, 2014).

HadCM3 represents the IPCC 4^{th} Assessment Report (AR4) scenarios data. There are eight predictors available for future climate scenarios in Singapore; they are mean sea level pressure (*mslp*), 500 hPa geopotential (*p500*), 850 hPa geopotential (*p850*), near surface relative humidity (*rhum*), relative humidity at 500 hPa height (*r500*), relative humidity at 850 hPa height (*r850*), and near surface specific humidity (*shum*) and temperature at 2 m (temp) (Pope *et al.*, 2000; Collins *et al.*, 2001).The HadCM3 data

freely available in http://ccds-dscc.ec.gc.ca/?page=pred-canesm2 is obtained for downscaling. The grid that corresponds to Singapore was chosen. The NCEP reanalysis data (Kalnay *et al.*, 1996) are chosen as atmospheric large scale predictors for model calibration. As the grid size is different for NCEP and HadCM3 predictors, the NCEP predictors are interpolated on to the same grid of $2.5^{\circ} \times 3.75^{\circ}$ latitude and longitude as the GCM. The NCEP and HadCM3 dataset are normalized using the historic data ranging from 1961-1990 period data (Dibike *et al.*, 2008). The period from 1980-1994 is chosen as the calibration period and the period from 1995-2000 is chosen as the validation period. As the main aim is to present a new methodology for multi-site SDM, HadCM3 A2 scenarios are chosen for future prediction.

Figure 4-1 Location of rain gauge stations at Singapore used in the study

Selection of predictors is an important step in establishing a statistical downscaling model since the characteristics of the downscaled scenarios largely depend on the predictors for downscaling (Wilby *et al.*, 2004). Generally, the basic requirements for predictor selection are 1) the predictors must be strongly correlated with the predictand, 2) captures the multiyear variability and 3) also physically sensible (Wilby *et al.*, 2004). GCMs simulate circulation predictors with some skill and they are commonly

used as predictors for statistical downscaling (Cavazos and Hewitson, 2005). In order to capture the precipitation driving mechanisms such as thermodynamics and moisture content, humidity predictors have been increasingly used in statistical downscaling (Karl *et al.*, 1990; Wilby and Wigley, 1997). Along with the above predictors selection, the predictors are also chosen based on the Two-sample Kolmogorov-Smirnov test and stepwise regression as explained in the Chapter.2

4.4 Multi-site precipitation occurrence determination using multi-output GPC

The *multi-output GPC* is also referred as *multi-task GPC* or *dependent GPC* (Boyle and Frean, 2004; Bonilla *et al.*, 2008). The relationship between outputs $\mathbf{y}_{mc}$ and the auxiliary variable g_{mc} is deterministic which is given by (4.3):

$$p(\mathbf{y}_{mcj} \mid g_{mcj}) = \begin{cases} \delta(g_{mcj})\delta(\mathbf{y}_{mcj}) \; if \; \mathbf{y}_{mcj} = +1 \\ \delta(-g_{mcj})\delta(-\mathbf{y}_{mcj}) \; if \; \mathbf{y}_{mcj} = -1 \end{cases} \tag{4.3}$$

where δ is 1, if the argument of δ is positive; zero otherwise. The g_{mcj} is a normal distribution with mean $\mathbf{f}_{mc}$ and variance 1. This is a probit model for classification. The posterior distribution $p(\mathbf{f}_{mc} \mid \mathbf{y}_{mc})$ is a non-Gaussian as the probit likelihood function is coupled with the GP prior placed over the latent function. This makes the inference not analytically tractable. Similar to the single site GPC, the analytical marginalization is not possible in multi-site GPC since the non-Gaussian likelihood is used. Multi-task Gaussian process classification using Expectation Propagation (EP) is adopted for implementing the model (Skolidis and Sanguinetti, 2011).

The joint likelihood is given by Bayes' theorem (4.4):

$$\begin{aligned} &p(\mathbf{y}_{mc}, \mathbf{g}_{mc}, \mathbf{f}_{mc}, K_{mc}^{y_{mc}}, \boldsymbol{\alpha}_{mc} \mid \boldsymbol{\theta}_{mc}^{y_{mc}}, \boldsymbol{\theta}_{mc}^{y_{mc}}) = \\ &p(\mathbf{y}_{mc} \mid \mathbf{g}_{mc}) p(\mathbf{g}_{mc} \mid \mathbf{f}_{mc}) p(\mathbf{f}_{mc} \mid K_{mc}^{y_{mc}}, \boldsymbol{\alpha}_{mc}) p(\boldsymbol{\alpha}_{mc} \mid \boldsymbol{\theta}_{mc}^{x_{mc}}) p(K_{mc}^{y_{mc}} \mid \boldsymbol{\theta}_{mc}^{y_{mc}}) \end{aligned} \tag{4.4}$$

where $\mathbf{g}_{mc}$ is auxiliary latent variable following normal distribution with the mean given by $\mathbf{f}_{mc}$ and variance 1, $\mathbf{f}_{mc}$ is a latent function which combines the information from the data and the tasks; $\mathbf{f}_{mc}$ is assumed to follow normal distribution with zero mean and covariance matrix given by $\mathrm{K}_{mc}^{y_{mc}} \otimes K_{mc}^{x_{mc}}$, $\mathrm{K}_{mc}^{y_{mc}}$ represents the between-site spatial dependence covariance matrix with dimension $s \times s$, $K_{mc}^{x_{mc}}$ is the within-site covariance matrix with dimension $n \times n$, $\boldsymbol{\theta}_{mc}^{y_{mc}}$ is the prior distribution parameters for dependence covariance matrix covariance matrix, $\boldsymbol{\theta}_{mc}^{x_{mc}}$ is the prior distribution of the within-site covariance matrix and $\boldsymbol{\alpha}_{mc}$ is the hyperparameters of the within-site covariance function.

The spatial dependence structure is learnt through optimization of hyperparameters using EP approximation. The covariance can be obtained through a covariance function or using free form covariance matrix. In this work, a free covariance matrix is chosen (Skolidis and Sanguinetti, 2011). The correlation function $\mathrm{K}_{mc}^{y_{mc}}$ captures the spatial dependence between the sites s. The matrix $\mathrm{K}_{mc}^{y_{mc}}$ must be positive semidefinite. The detailed description of spatial correlation function implementation is presented in (Skolidis and Sanguinetti, 2011). There are several other within site covariance functions available as well. The commonly used within site covariance function is squared exponential ARD covariance function (2.21) which is similar to the covariance function used in single site GP-SDM. In multi-site occurrence determination, only one covariance function and the corresponding hyperparameters are shared for all the sites to capture the at-site residual dependence.

4.4.1 Expectation Propagation (EP) approximation for inference

The exact inference for non-Gaussian posterior distribution obtained by coupling the probit likelihood with a GP prior is impossible. There are several approximate inference approaches available such as Gibbs sampling, EP approximation (Opper and Winther, 2000; Minka, 2001; Rasmussen and Williams, 2006), variational expectation-

maximization (EM) algorithm (Girolami and Rogers, 2006). In this Chapter, EP is followed and the hyperparameters are estimated using type II maximum likelihood (ML). The detailed explanation for EP approximation for multitask classification can be found in Skolidis and Sanguinetti (2011).

The product of the prior and the factorized non-Gaussian likelihoods is proportional to the posterior $p(\mathbf{f}_{mc} \mid \mathbf{y}_{mc})$ over the latent variables. The non-Gaussian likelihoods are approximated by a product of unnormalized Gaussians $t_i(\mathbf{y}_{mci})$ using EP and is given by (4.5):

$$\prod_{i=1}^{n} p(y_{mci} \mid f_{mci}) \sim \prod_{i=1}^{n} t_i(f_{mci} \mid \tilde{Z}_{mci}, \tilde{\mu}_{mci}, \tilde{\sigma}^2_{mci}) = N(\tilde{\boldsymbol{\mu}}_{mc}, \tilde{\Sigma}_{mc}) \prod_{i=1}^{n} \tilde{Z}_{mci} \tag{4.5}$$

where $\tilde{\boldsymbol{\mu}}_{mc}$ is avector of $\tilde{\mu}_{mci}$ and $\tilde{\Sigma}_{mc}$ is diagonal with $\tilde{\Sigma}_{mcii} = \tilde{\sigma}^2_{mci}$. The approximated distribution of the latent variables $q(f_{mc} \mid y_{mc})$ is (4.6):

$$q(f_{mc} \mid y_{mc}) = \frac{1}{Z_{EP}} p(f_{mc}) \prod_{i=1}^{n} t_i(f_{mc} \mid \tilde{Z}_{mci}, \tilde{\mu}_{mci}, \tilde{\sigma}^2_{mci}) \tag{4.6}$$

where $\mu_{mc} = \sum \tilde{\Sigma}^{-1} \tilde{\mu}_{mc}$ and $\sum_{mc} = [(\mathrm{K}_{mc}^{y_{mc}} \otimes K_{mc}^{x_{mc}})^{-1} + \tilde{\Sigma}_{mc}^{-1}]^{-1}$

In EP, the likelihood is updated termwise where the current t_i is removed and the cavity distribution $q_{-i}(t_i)$ is combined with the true likelihood $p(y_{mci} \mid f_{mci})$ to obtain a non-Gaussian distribution.

The parameters of t_i are then computed by matching the moments between the Gaussian approximations of the non-Gaussian from the previous step. The log marginal likelihood is used to estimate the hyperparameters and the gradient with respect to the hyperparameters $\boldsymbol{\theta}_{mc}^{x_{mc}}$ and $\boldsymbol{\theta}_{mc}^{y_{mc}}$ is given by (4.7):

$$\frac{\partial \log Z_{EP}}{\partial \theta_j} = \frac{1}{2}\tilde{\mu}_{mc}^{\mathrm{T}}(\mathrm{K}_{mc}^{y_{mc}} \otimes K_{mc}^{x_{mc}} + \tilde{\Sigma}_{mc})\mathbf{\Omega}(\mathrm{K}_{mc}^{y_{mc}} \otimes K_{mc}^{x_{mc}} + \tilde{\Sigma}_{mc})^{-1}\tilde{\mu}_{mc} - \frac{1}{2}tr((\mathrm{K}_{mc}^{y_{mc}} \otimes K_{mc}^{x_{mc}} + \tilde{\Sigma}_{mc})^{-1}\mathbf{\Omega}) \tag{4.7}$$

where $\Omega = \begin{cases} \mathrm{K}_{mc}^{y_{mc}} \otimes \dfrac{\partial K_{mc}^{x_{mc}}}{\partial \theta^j} & if\ \theta^j = \theta^{x_{mc}} \\ \dfrac{\partial \mathrm{K}_{mc}^{y_{mc}}}{\partial \theta^j} \otimes K_{mc}^{x_{mc}} & if\ \theta^j = \theta^{y_{mc}} \end{cases}$

The task covariance matrix helps to transfer/infer the knowledge between the sites (Bonilla *et al.*, 2008).

4.4.2 Predictive distribution

The predictive distribution of the j^{th} site is given by (4.8):

$$E[p(y_{mci*} \mid \mathbf{y}_{mc})] = (\mathbf{k}_{mcj}^{y_{mc}} \otimes \mathbf{k}_{\mathbf{x}_{mc}, x_{mc*}}^{x_{mc}})^{\mathrm{T}} \sum_{mc}^{-1} \mathbf{y}_{mc} \tag{4.8}$$

The observations from the site i are weighed by the i^{th} element of $\mathbf{k}_{mcj}^{y_{mc}}$. Therefore the variance of the predictive distribution is given by (4.9):

$$\mathrm{var}(y_{mci*} \mid \mathbf{y}_{mc}) = (\mathbf{k}_{mcjj}^{y_{mc}} \mathbf{k}_{x_{mc*}, x_{mc*}}^{x_{mc}}) - (\mathbf{k}_{mcj}^{y_{mc}} \otimes \mathbf{k}_{\mathbf{x}_{mc}, x_{mc*}}^{x_{mc}})^{\mathrm{T}} \sum_{mc}^{-1} (\mathbf{k}_{mcj}^{y_{mc}} \otimes \mathbf{k}_{\mathbf{x}_{mc}, x_{mc*}}^{x_{mc}}) \tag{4.9}$$

4.5 Multisite Precipitation amount estimation using multi-output GPR

The precipitation measured at multiple sites is correlated and the challenge is to downscale precipitation jointly. Multi output Gaussian process regression has been implemented in many fields to predict the multiple outputs jointly. This method is also referred as multi-task Gaussian process regression in machine learning or Co-Krigingin Geostatistics. The Gaussian process regression is represented by the mean function and the covariance function. When the Gaussian process is extended for multiple outputs, the covariance matrix also consists of cross-covariance between the outputs. Ensuring

that the covariance matrix as positive definite is a challenging task; thus, the property of the convolution is used to obtain the positive definite matrix in the machine learning literature. Dependent Gaussian process was proposed to predict multiple outputs using a single model (Boyle and Frean, 2004). The GP requires positive definite symmetric covariance matrix for modelling. However, in multiple output modelling, there is no direct function that computes positive semi definite symmetric cross-covariance matrix. This problem is solved by employing a convolution process where each output is represented as the convolution between a smoothing kernel function and a Gaussian white noise (Bilionis and Zabaras, 2012). This approach will yield positive definite covariance and cross-covariance.

GP is a non-parametric Bayesian inference which provides efficient way to quantify uncertainty, model calibration and prediction simultaneously. In GPR, the Gaussian probability distribution is placed on the modelling functions instead of placing on the parameters. GP is represented by a mean function and the covariance function. The mean function can be linear, polynomial or any non-linear function. The covariance function defines the relationship between the inputs and measures the correlation between the predictions. It is assumed that similar inputs give similar outputs. The covariance structure is modified from the ordinary GP to take into account the auto- and cross- correlation between the input predictors and the different outputs.

Multi-task GP is viewed as learning multiple outputs given the inputs and outputs. GP provides a principled framework from specific priors over functions. The multi-task GP differs from single GP by construction of covariance function. The crucial step to encode the dependence structure of multiple outputs is by constructing the covariance function. The covariance matrix consists of auto- and cross-covariance matrix of different outputs and inputs. The number of hyperparameters to learn is increased. The constraint is that the covariance matrix should be positive semidefinite. This is achieved by convolution process.

This Chapter presents the derivation for rainfall amount estimation for multi-site SGP-SDM model. This type of modelling improves the prediction accuracy of one output using the correlation between the other outputs. Multi-output Gaussian process is a non-parametric Bayesian framework to predict multiple correlated outputs simultaneously. The advantage of this method is that the spatial correlation of all the outputs is considered in model calibration and the uncertainty quantification tool is coupled with the multi-site downscaling model.

The MGP-SDM modelling considers the spatial correlation between the datasets at different locations using the cross-covariances and the spatial correlation within the datasets using the auto-covariances. The cross-covariances are represented by convolving GP with a smoothing kernel by assuming GP being a white noise (Boyle and Frean, 2004; Vasudevan *et al.*, 2011; Robin, 2012). In this Chapter, the Gaussian process based *multivator* framework developed by (Robin, 2012) is adopted for implementing multisite downscaling. *Multivator* is a combined term used for multivariate emulator. Complex computer models need to be used in many scientific disciplines. Emulators are softwares developed to replace the complex computer models to generate similar outputs as the complex computer models. The mean function and the covariance matrix for the multi-output GP can then be specified by (4.10):

$$\mathbf{y}_{ma} \sim GP(\boldsymbol{\mu}_{ma}, \Sigma_{ma}) \tag{4.10}$$

where $\boldsymbol{\mu}_{ma} = \begin{bmatrix} \mathbf{g}_{ma1}(\mathbf{X}_{ma}^{1})^{\mathrm{T}} & 0 & \cdots & 0 \\ 0 & \mathbf{g}_{ma2}(\mathbf{X}_{ma}^{2})^{\mathrm{T}} & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & \mathbf{g}_{ma}(\mathbf{X}_{ma}^{s})^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_{ma1} \\ \vdots \\ \boldsymbol{\beta}_{mas} \end{bmatrix}$ and

$$\Sigma_{ma} = \begin{bmatrix} \Sigma_{ma}^{11} & \Sigma_{ma}^{12} & \cdots & \Sigma_{ma}^{1s} \\ \Sigma_{ma}^{21} & \Sigma_{ma}^{22} & \cdots & \Sigma_{ma}^{1s} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{ma}^{s1} & \Sigma_{ma}^{s2} & \cdots & \Sigma_{ma}^{ss} \end{bmatrix}$$

where $\mathbf{g}_{ma}(\mathbf{X}_{ma}^{s})^{\mathrm{T}}$ is the downscaling model mean function for the site m which can be either linear or non-linear, $\boldsymbol{\beta}_{ma}$ are the coefficients of the mean function corresponding to each site s, $\sum_{ma}^{pq}$ is the cross-covariance between the dataset X_{ma}^{s} corresponding to the site p and q, the diagonal entries of $\sum_{ma}^{ss}$ refers to the auto-covariance. The derivation of cross-covariance using convolution was presented in the work by Robin (2012). They used non-separable covariance matrix to derive the cross covariance for multi-output GP. The covariance function is given in equation (4.11):

$$\mathrm{K}_{mapq}(\mathbf{d}) = \begin{cases} \exp\{-\mathbf{d}^{\mathbf{T}}\mathbf{B}_p\mathbf{d}\} & \textit{if } p = q \\ \dfrac{\exp\{-\mathbf{d}^{\mathbf{T}}(\frac{1}{2}\mathbf{B}_{\mathbf{p}}^{-1} + \frac{1}{2}\mathbf{B}_{\mathbf{q}}^{-1})^{-1}\mathbf{d}\}}{|(\frac{1}{2}\mathbf{B}_{\mathbf{p}} + \frac{1}{2}\mathbf{B}_{\mathbf{q}})(\frac{1}{2}\mathbf{B}_{\mathbf{p}}^{-1} + \frac{1}{2}\mathbf{B}_{\mathbf{q}}^{-1})|^{1/4}} & \textit{otherwise} \end{cases} \tag{4.11}$$

where $\mathbf{B}_p = \dfrac{\mathbf{l}_p^{-1}}{2}$ and l_p is the correlation length corresponding to each dimension of the output and $\mathbf{d}$ is the distance $\| \mathbf{x}_{mai}^{\mathbf{p}} - \mathbf{x}_{maj}^{\mathbf{q}} \|^2$. The covariance function is then multiplied with the matrix M_{mapq} ($s \times s$) that comprises of covariance between the observations at multiple sites.

$$\sum_{ma}(d) = M_{mapq}K_{mapq}(d) \tag{4.12}$$

where $\sum_{ma}$ is a positive definite function.

The marginal likelihood equation is represented by (4.13):

$$L(M_{ma}, \mathbf{B}_{ma1}, ..., \mathbf{B}_{mas}) = \frac{|\sum_{ma}^{-1}|^{1/2}}{|\mathbf{G}_{ma}^{\mathrm{T}} \Sigma_{ma}^{-1} \mathbf{G}_{ma}|^{1/2}} \exp(-\frac{1}{2}(\mathbf{y}_{ma} - \mathbf{G}_{ma}\boldsymbol{\beta}_{ma})^{T} \sum_{ma}^{-1} (\mathbf{y}_{ma} - \mathbf{G}_{ma}\boldsymbol{\beta}_{ma})) \quad (4.13)$$

However, the likelihood cannot be optimized directly. Thus multi-stage approach is proposed to estimate the optimal parameter values. The steps for multi-stage approach are as follows:

1) The single task GP methodology is used to estimate the correlation length for each of the observations.
2) The posterior mode is used calculate the diagonal elements of the marginal variance terms M_{ma}
3) The off-diagonal elements of M_{ma} are determined numerically using the posterior mode.

With the optimal value of the hyperparameters, the predictive mean and the predictive covariance for the future data $\mathbf{x}_{ma}^{*}$ can be estimated using the conditional Gaussian distribution property (4.14). The predictions are conditional on the historic data.

$$\begin{bmatrix} \mathbf{y}_{ma} \\ \mathbf{y}_{ma}^{*} \end{bmatrix} \sim N\left(\begin{pmatrix} f_{ma}(\mathbf{x}_{ma}, \boldsymbol{\theta}_{ma}) \\ f_{ma}(\mathbf{x}_{ma}^{*}, \boldsymbol{\theta}_{ma}) \end{pmatrix}, \begin{pmatrix} \sum_{ma\, x_{ma}, x_{ma}} & \sum_{ma\, x_{ma}^{*}, x_{ma}} \\ \sum_{ma\, x_{ma}, x_{ma}^{*}} & \sum_{ma\, x_{ma}^{*}, x_{ma}^{*}} \end{pmatrix} \right) \quad (4.14)$$

$$m(\mathbf{x}_{ma}^{*}) = f_{ma}(\mathbf{x}_{ma}^{*}, \boldsymbol{\theta}_{ma}) + \sum_{ma\, x_{ma}, x_{ma}} \sum_{ma\, x_{ma}, x_{ma}}^{-1} (\mathbf{y}_{ma} - f_{ma}(\mathbf{x}_{ma}, \boldsymbol{\theta}_{ma}))$$

$$Cov(\mathbf{x}_{ma}^{*}) = \sum_{ma\, x_{ma}^{*}, x_{ma}^{*}} - \sum_{ma\, x_{ma}, x_{ma}^{*}}^{\mathbf{T}} \sum_{ma\, x_{ma}, x_{ma}}^{-1} \sum_{ma\, x_{ma}, x_{ma}^{*}}$$

where $m(\mathbf{x}_{ma}^{*})$ and $Cov(\mathbf{x}_{ma}^{*})$ are predictive mean and predictive covariance respectively.

4.6 KNN Disaggregation

K-Nearest Neighbor (KNN) resampling is a non-parametric approach to disaggregate rainfall into the required time scale. KNN resampling approach has been implemented (Lall and Sharma, 1996; Prairie *et al.*, 2007; Nowak *et al.*, 2010) to disaggregate annual river flow into monthly flow. The KNN disaggregation algorithm described in this section is based on the algorithm proposed by Nowak *et al.* (2010).

Each of the ensembles obtained from MGP-SDM is disaggregated into hourly precipitation at multiple sites using KNN resampling. Thus, the uncertainty ranges for the disaggregation of daily rainfall to hourly rainfall at multiple sites simultaneously. The algorithm is as follows:

Let Q_n is a $n \times 24$ matrix where $_n$ is number of days and 24 is the hours/day. The elements in the matrix Q_n are divided by the total daily flow and thus the each row sums to unity. Let Q_d^* be the daily precipitation which needs to be disaggregated. The K-nearest neighbors of Q_d^* is identified from the historical daily precipitation Q_d.

The K neighbors are computed using the Euclidean distance E_{dist} between Q_d^* and all the historical daily precipitation. A weight is assigned to the K-nearest neighbors using the weight scheme proposed by Lall and Sharma (1996) as (4.15):

$$W_{dist}(i) = \frac{1/i}{\sum_{i=1}^{K} 1/i} \tag{4.15}$$

where K is the number of nearest neighbors, i is the index of the neighbor and $i = 1$ is the nearest neighbor. The weight is used to pick one of the K-nearest neighbors which is ranked high. The corresponding hourly precipitation, $\mathbf{Q}_h^k$ for the chosen daily

precipitation using k-nearest neighbors is selected. The disaggregation for $\mathbf{Q}_d^*$ is obtained by multiplying $\mathbf{Q}_d^*$ with $\mathbf{Q}_h^k$ hourly proportion (4.16):

$$\mathbf{Q}_h^* = \mathbf{Q}_d^* \times \mathbf{Q}_h^k \tag{4.16}$$

The disaggregated $\mathbf{Q}_h^k$ vector sums to daily precipitation. The above steps are repeated for all the days which need to be disaggregated. For multisite disaggregation, the daily values are summation of the rainfall observed at all the stations considered and the hourly matrix consists of hourly rainfall from all the stations concatenated. The selection of best K is obtained by the optimization steps presented in Lu and Qin (2014). The first 10 closest neighbors are selected using the weight schemes and 10 selected neighbors are considered as ensembles. The optimal neighbor is estimated by minimizing the object function K_{obj}(4.17):

$$K_{obj} = \frac{\sum_{i=1}^{N_k} MAPE_i}{N_k} \tag{4.17}$$

where $MAPE$ is Mean Absolute Percentage Error, N_k is number of statistical properties used for objective function, and i represents the statistical properties such as mean, standard deviation, lag-1 autocorrelation, lag-2 autocorrelation, probability of wet hour, skewness and cross-correlation for multi-site disaggregation.

4.7 Results and Discussions

The climate change impact studies require the future precipitation to be projected accurately. In this regard, the performances of MGP-SDM and KNN disaggregation model in reproducing the observed precipitation statistical properties that capture the spatial and temporal dependence are analyzed. The downscaled daily precipitation and disaggregated hourly precipitation are compared with the observed daily and hourly precipitation respectively. The statistical properties that are compared are mean (daily

and hourly), standard deviation of rainfall (daily and hourly) (Hessami *et al.*, 2008; Fowler and Ekström, 2009; Maraun *et al.*, 2010), the 90th percentile of daily precipitation (PERC90) of the precipitation on wet days (Haylock *et al.*, 2006; Goodess *et al.*, 2007), maximum daily rainfall (Max) (Hessami *et al.*, 2008), lag-1 autocorrelation of hourly rainfall ($AC1_h$), probability of wet day (Pwet), probability of wet hour ($Pwet_h$) (Semenov *et al.*, 1998), skewness of hourly rainfall ($skewness_h$) and cross-correlation coefficients as in (4.18) (CC_d/CC_h) (Ying *et al.*, 2011).

$$CC_d / CC_h = \frac{\sum_{i=1}^{n}(c_{obs,i} - \overline{c}_{obs,i})(c_{sim,i} - \overline{c}_{sim,i})}{\sqrt{\sum_{i=1}^{n}(c_{obs,i} - \overline{c}_{obs,i})^2 \sum_{i=1}^{n}(c_{sim,i} - \overline{c}_{sim,i})^2}} \tag{4.18}$$

where $c_{obs,i}$ is the observed rainfall, $c_{sim,i}$ is the simulated rainfall data, $\overline{c}_{obs,i}$ is the mean of the observed data, $\overline{c}_{sim,i}$ is the mean of the simulated rainfall data. The accuracy (*acc*) of the wet and the dry day classification (Chen *et al.*, 2010) is given in (4.19):

$$acc = \frac{C_{dry} + C_{wet}}{TP_{dry} + TP_{wet}} \tag{4.19}$$

where C_{dry} is the total number of correctly classified dry days, C_{wet} is the total number of correctly classified wet days, TP_{dry} is the total number of dry days and TP_{wet} is the total number of wet days.

4.7.1 MGP-SDM precipitation occurrence determination

The multisite precipitation occurrence model is calibrated for the period of 1980 to 1987 using NCEP reanalysis data. HadCM3 predictors for 1980 to 2010 are used for validating the classification model. The mean of the 100 occurrence determination ensembles generated from the classification step is used to choose the wet days for the precipitation amount estimation. The precipitation occurrence is determined using

averaged predictive distribution. The prediction itself is averaged and thus there is no uncertainty range in precipitation occurrence. Three criteria including dry day proportion, accuracy and wet and dry transition probability are used to assess the performance of MGP-SDM in downscaling precipitation at the three sites for all the months (Jeong *et al.*, 2012).

Figure 4-2 shows the dry day proportion compared with the observed data at the three stations for all the months. At station S46, it is observed that for the months of February, May, September and October, the dry day proportion is underestimated; large deviations are seen in the month of May. For the months of March and June, the dry day proportion is overestimated. The dry day proportion for the remaining months is close to the observed proportion. Similar results are seen at the other stations (S55 and S69) as well. The purpose of this chapter is to show the ability of the multisite precipitation occurrence model in wet and dry determination at all the sites simultaneously. With the increase in the data size, the computation cost increases in the multisite classification model; large data sizes involve large size of covariance matrices which causes memory issues. Thus, large size of training data cannot be used for calibration which leads to poor results for some months (Skolidis and Sanguinetti, 2011). This issue can be solved by using the advanced approximation technique for classification model implementation. Due to computational complexity, research scope and time factor, the approximation technique is not applied in this thesis. This needs to be explored in the future study.

Table 4-1 shows the average of the accuracy of the simulated precipitation occurrence ensembles at the three stations S46, S55 and S69 for the validation period. The average accuracy is around 50% at all the stations for all the months. The table shows that the model's classification accuracy for all the stations does not vary significantly. As part of future works, the performance of the model needs to be further assessed by comparing the results from MGP-SDM with the other multi-site statistical downscaling models.

Figure 4-3 and Figure 4-2 show the dry and wet day transition probability for the simulated occurrence series compared with the observed wet and dry day transition probability for the validation period respectively. At S46 and S55 station, the dry day transition probability is overestimated for the month of June and April; for the remaining months, dry day transition probability is underestimated. At S69, the dry day transition probability of June is closer to the observed data; however, for all the other months, the dry day transition probability is underestimated. Compared to dry day transition probability, the wet day transition probability is predicted well by MGP-SDM at all the three stations for all the months. The wet day transition probability for the month of May is overestimated and for the month of March, the wet day transition probability is underestimated compared to all the other months. The proposed method of multi-site precipitation occurrence determination needs to be further refined to improve the prediction of dry day proportion and the transition probabilities.

4.7.2 MGP-SDM precipitation amount estimation

The MGP-SDM downscaling is calibrated for the period from 1980-2000 using NCEP reanalysis predictors. The HadCM3 predictors from 1980-2010 are used as the validation period for downscaling. Twenty ensembles are simulated for the prediction period to assess the performance statistics and uncertainty range in the model simulations. The precipitation is downscaled at the three stations simultaneously. The MGP-SDM model is calibrated with different initial values to avoid local minimum as the optimal points.

The comparison of cumulative distribution of the generated ensembles and the observed data for each month during the validation period at the three stations (S46, S55 and S69) is shown in Figure 4-5, Figure 4-6 and Figure 4-7 respectively. The cdf of the downscaled precipitation shows that the ensembles from the model predictions cover the observed data for some months while it is over/underestimated for most of the months. The figures also prove that the model performance is consistent for all the

months at all the stations. However, for the month of May the ensembles are slightly overestimated.

This is attributed to the inaccurate classification model results for most of the months. The evaluation statistics for the all the months at the three stations (S46, S55 and S69) are presented in Figure 4-8, Figure 4-9 and Figure 4-10 respectively. The MGP-SDM shows better performance in simulating the mean daily precipitation for each month. For the month of May, the prediction results are overestimated. It can also be seen that the uncertainty in the classification results is propagated to the downscaling results. For the other statistical properties, the observed and the mean of the simulated ensembles are closer to each other and are within the range of the ensembles. In the maximum evaluation statistics from GLM, the mean daily maximum for the month of the December and March are not within the simulated ensembles. However, in MGP-SDM the statistics for all the months are captured well. The results prove that the model is superior even with fewer number of stations used for downscaling. The performance of the model is also consistent with all other stations.

Table 4-2 provides the comparison of the spatial correlation coefficient of the downscaled and observed precipitation between the three stations during the validation period. The interstation correlations are computed between the pairs of the predicted and observed precipitation at the three rain gauge locations. The table shows that the interstation correlation is underestimated for all the station pairs. This may be because of the uncertainty from the precipitation occurrence determination step due to the less amount of calibration used.

4.7.3 KNN disaggregation

In this section, the observed hourly rainfall from 1980 to 2000 is used for calibrating the disaggregation model and the observed data from 1980 to 2010 (baseline period) is used for model validation. The validation period is the same as the downscaling model since the precipitation from downscaling model is used for disaggregation for the future

period. Figure 4-11, Figure 4-12 and Figure 4-13 show the statistical properties of the disaggregated hourly data from the daily precipitation downscaled from MGP-SDM at the three stations S46, S55 and S69 respectively. At the station S44, KNN can keep track of the statistical properties for a few months and there is a deviation from the observed statistical properties for the remaining months. There is a notable underestimation of AC1 for the months of February, May, June, July and September; for the months of January, October and November, AC1 is overestimated. Similarly for $Pwet_{h,}$ all the months are underestimated except for October and December. The $skewness_h$ is estimated well for the months of January, February, April, June and August to October. The standard deviation is also not predicted well for at S44 station. At S55 station, the statistical properties are predicted well compared to S46; however, there are several months for which the statistical properties are underestimated or overestimated. The deviations of the statistical properties from the observed data are less at S69 station. The underestimation of precipitation can be because the seasonal effects are not considered in KNN. Also only limited number of data is available for each month and the seasonal effects are not considered. Even though KNN predicts some of the properties for some of the months, still there is a notable underestimation. Table 4-3 provides the comparison of the spatial correlation coefficient of the disaggregated and observed precipitation between the three stations during the validation period. The interstation correlation between the pairs of the disaggregation rainfall at the three locations is captured well compared to downscaled rainfall. However, the rainfall correlations are still underestimated as KNN does not take spatial correlation into account.

4.7.4 HadCM3 A2 scenarios future precipitation projection (2011-2099)

The daily precipitation downscaled using MGP-SDM is disaggregated using KNN to the hourly timescale for the future periods. The hourly precipitation is then used for analyzing the change in the intensity of future precipitation due to climate change. The projected precipitation for the next century is presented by using the HadCM3 predictors for 2011-2099. The HadCM3 scenarios are used for assessing the change in

the future periods. Figure 4-14 shows the comparison of the average, minimum and maximum of the mean hourly precipitation downscaled from HadCM3 A2 scenarios for the three periods 2011-2040, 2041-2070 and 2071-2099 with the baseline period. Figure 4-15 shows the comparison of the average, minimum and maximum of the maximum hourly precipitation downscaled from HadCM3 A2 scenarios for the three periods 2011-2040, 2041-2070 and 2071-2099 with the baseline period. The prediction results show that the mean hourly rainfall is increasing compared to the baseline period in 2011-2040 and the precipitation trend is increasing for the period from 2041-2099. The results show the highest increase in rainfall for the month February, March, May, July and December from 2011-2099.

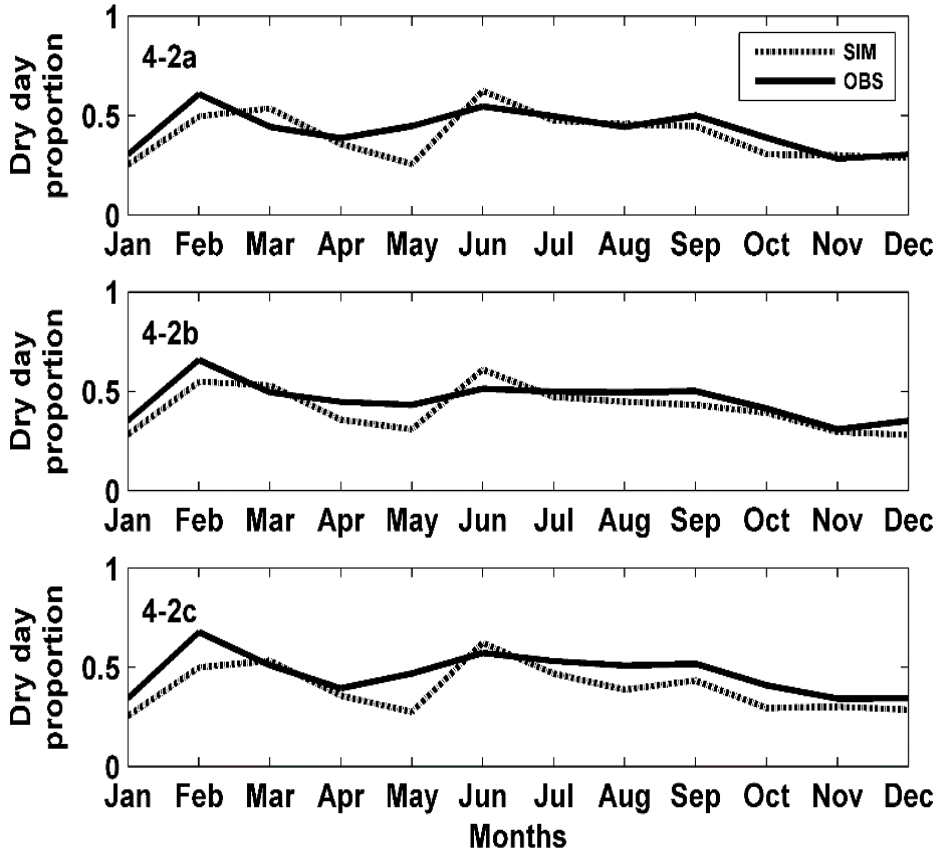


Figure 4-2 Comparison of observed and simulated dry day proportion by MGP-SDM at three stations a) S46, b) S55 and c) S69 for the validation period (1980-2010)

Table 4-1 Accuracy of the MGP-SDM classification ensembles at three stations

Accuracy (%)	S46	S55	S69
January	59.14	56.13	56.77
February	47.10	49.03	47.85
March	50.75	47.74	50.97
April	52.26	51.61	53.12
May	52.04	52.80	50.75
June	48.71	48. 28	50.75
July	47.85	49.46	50.65
August	50.97	50.75	48.60
September	50.40	52.47	49.25
October	52.80	51.29	55.16
November	58.39	58.17	56.77
December	59.68	57.63	58.92

Table 4-2 Cross-correlation of the downscaled rainfall between the sites

	MGP-SDM	OBS
S46-S55	0.19	0.58
S55-S69	0.17	0.58
S69-S46	0.2	0.7

Table 4-3 Cross-correlation of the disaggregated rainfall between the sites

	KNN	OBS
S46 –S55	0.25	0.36
S55 –S69	0.25	0.34
S69-S46	0.46	0.55

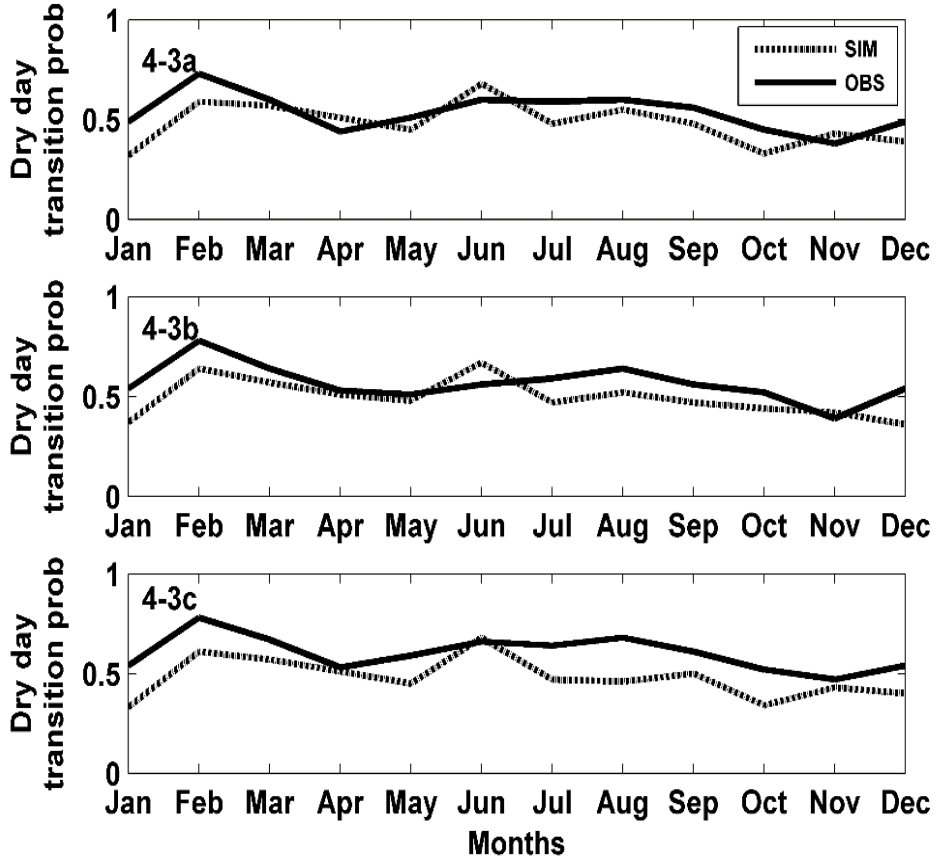


Figure 4-3 Comparison of observed and simulated dry day transition probability at three stations a) S46, b) S55 and c) S69 for the validation period (1980-2010)

The model shows decrease in mean hourly rainfall in the months of January and September. The maximum hourly prediction for the month of September shows the highest decreasing trend for all the three periods considered. The overall results show that the average precipitation prediction is decreasing for most of the months. The highest increase in mean hourly precipitation is seen in the month May. The fourth IPCC assessment report is also in agreement with the prediction results presented in this chapter (Solomon, 2007). However, the maximum hourly rainfall for each month shows an increasing trend. The maximum precipitation is predicted to increase in February. In summary, the mean hourly precipitation is predicted to increase in the next century for Singapore. The integrated multisite downscaling and disaggregation model with uncertainty quantification tool presents the first step to simulate future

precipitation for urban hydrological analysis. The projected future scenarios are based on the predictions only from one GCM scenario. It is necessary to run the proposed framework with various model results and with various emission scenarios for decision making and planning adaption measures in the future. The future precipitation results are obtained based on the stationarity relationship assessment between the current and the future period. In order to reduce uncertainty due to stationarity assumptions, it is necessary to develop non-stationary models for downscaling.

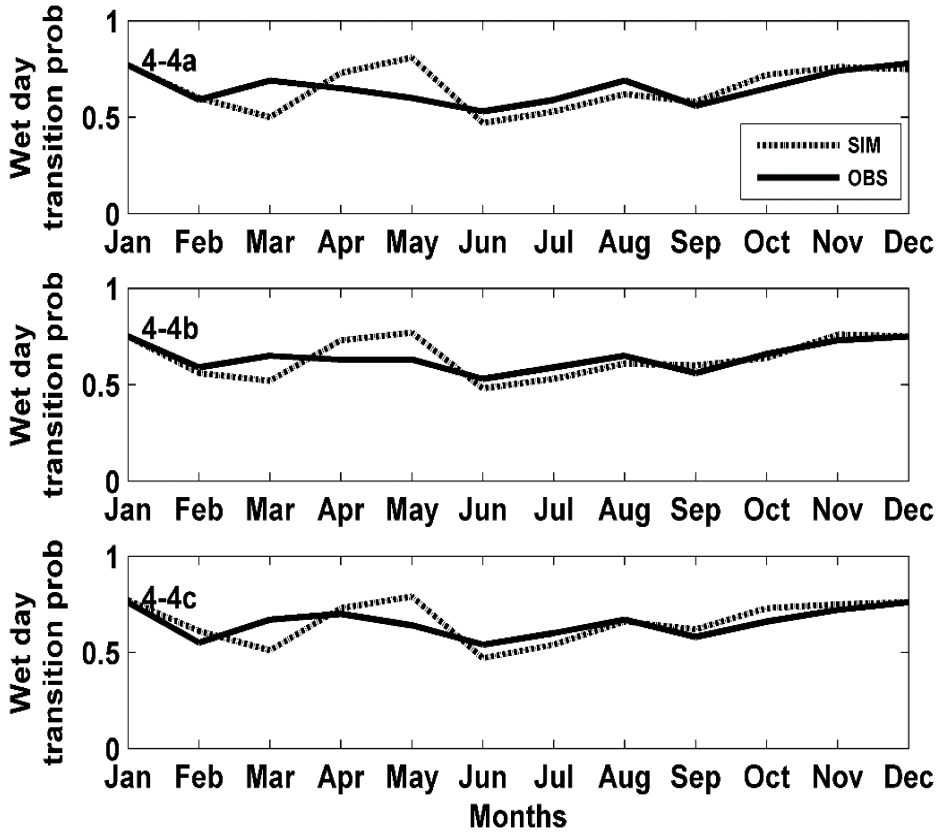
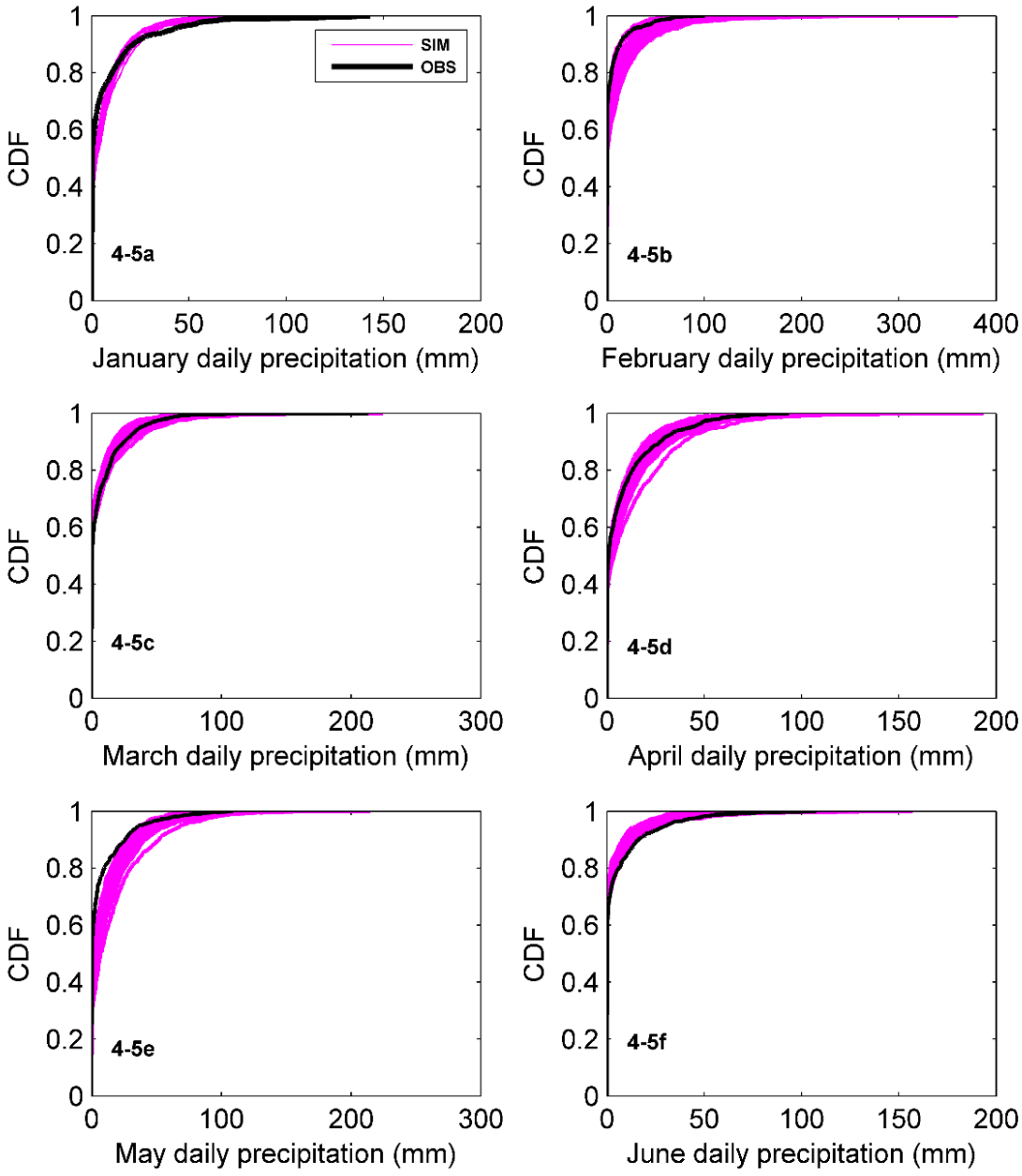


Figure 4-4 Comparison of observed and simulated wet-day transition probability at three stations a) S46, b) S55 and c) S69 for the validation period (1980-2010)


SIM
OBS
CDF
4-5a
January daily precipitation (mm)
4-5b
February daily precipitation (mm)
4-5c
March daily precipitation (mm)
4-5d
April daily precipitation (mm)
4-5e
May daily precipitation (mm)
4-5f
June daily precipitation (mm)

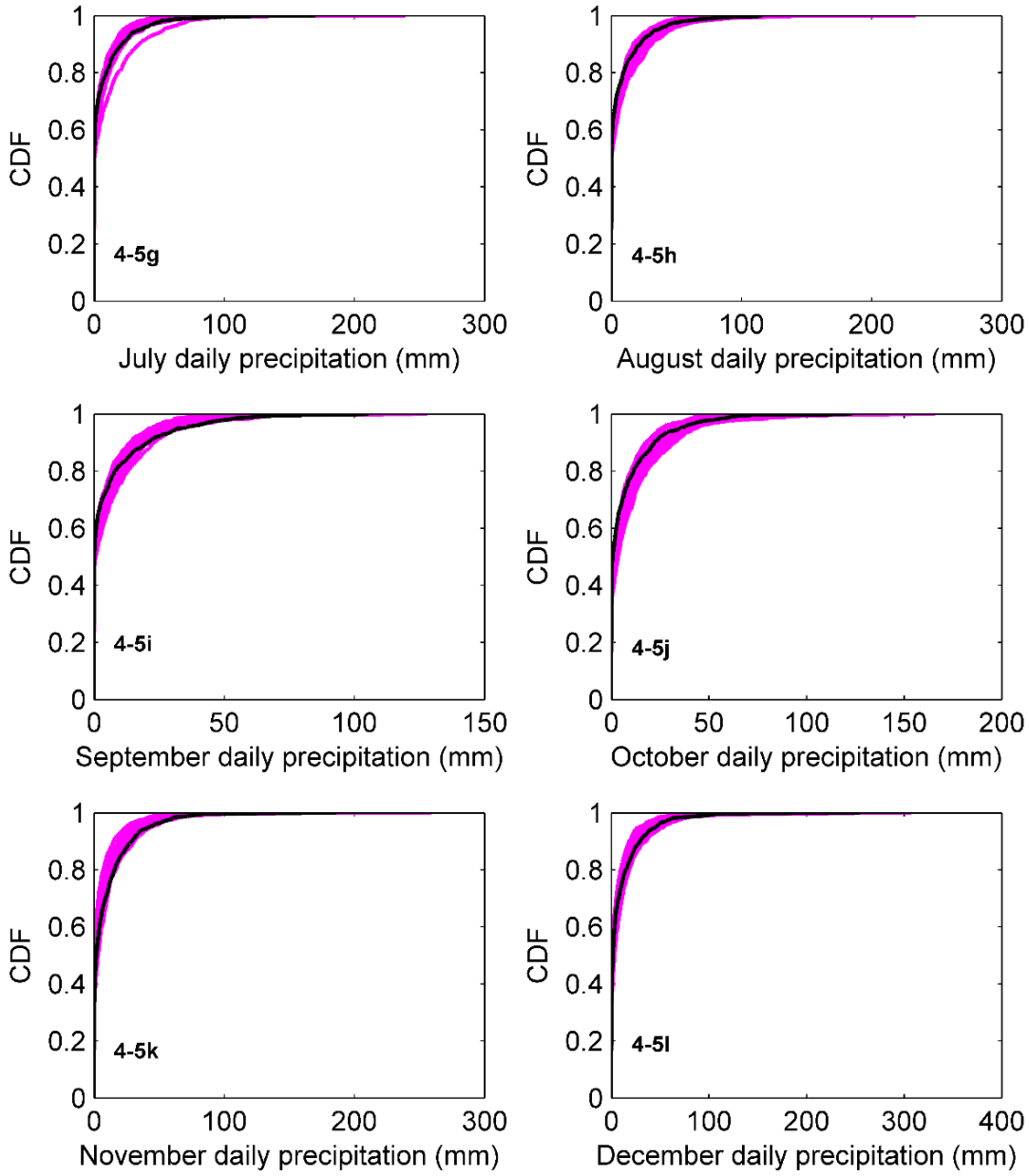
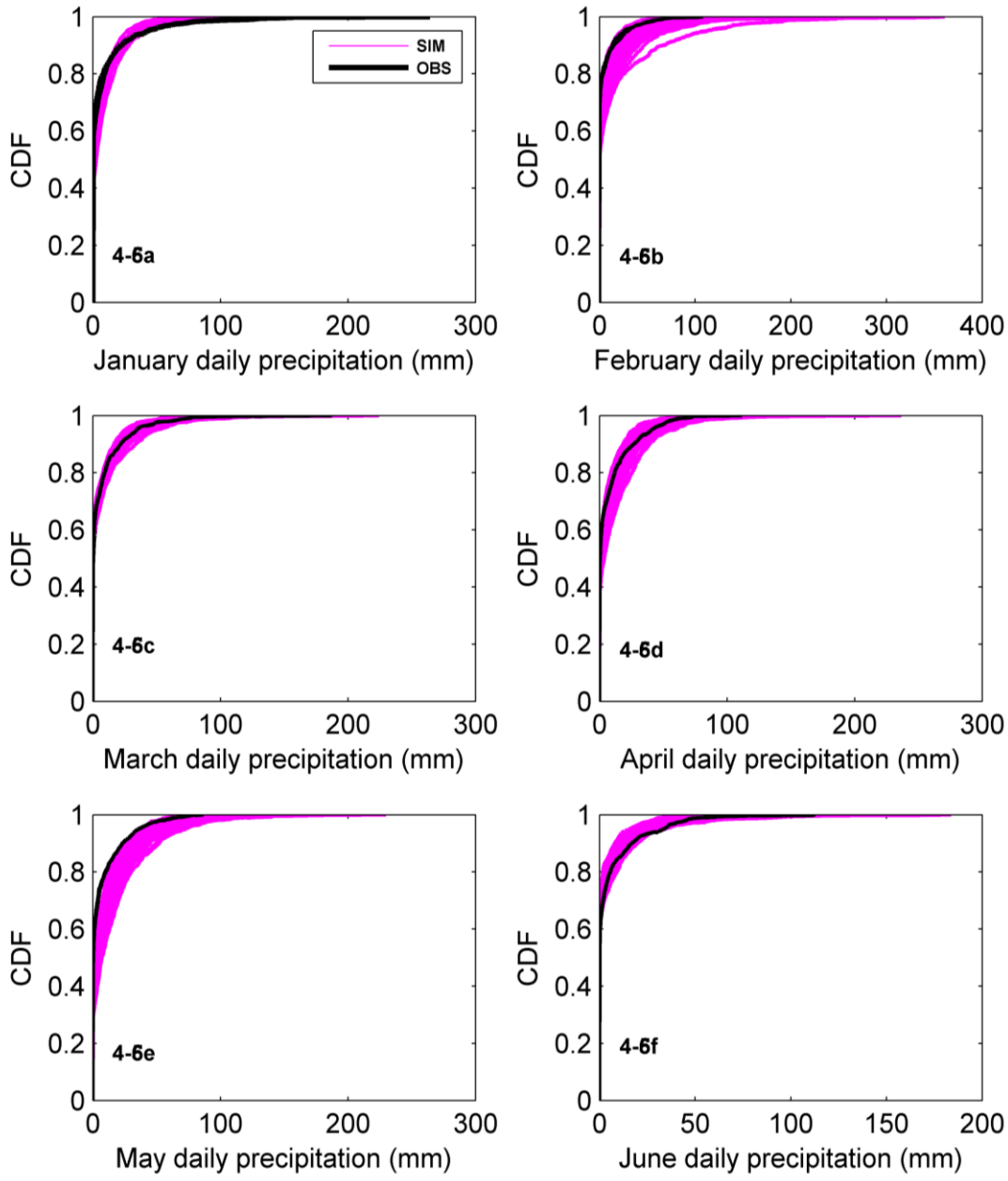


Figure 4-5 Comparison of cdf of the downscaled precipitation ensembles with observed cdf for all the months at Station S46


SIM
OBS
CDF
4-6a
January daily precipitation (mm)
4-6b
February daily precipitation (mm)
4-6c
March daily precipitation (mm)
4-6d
April daily precipitation (mm)
4-6e
May daily precipitation (mm)
4-6f
June daily precipitation (mm)

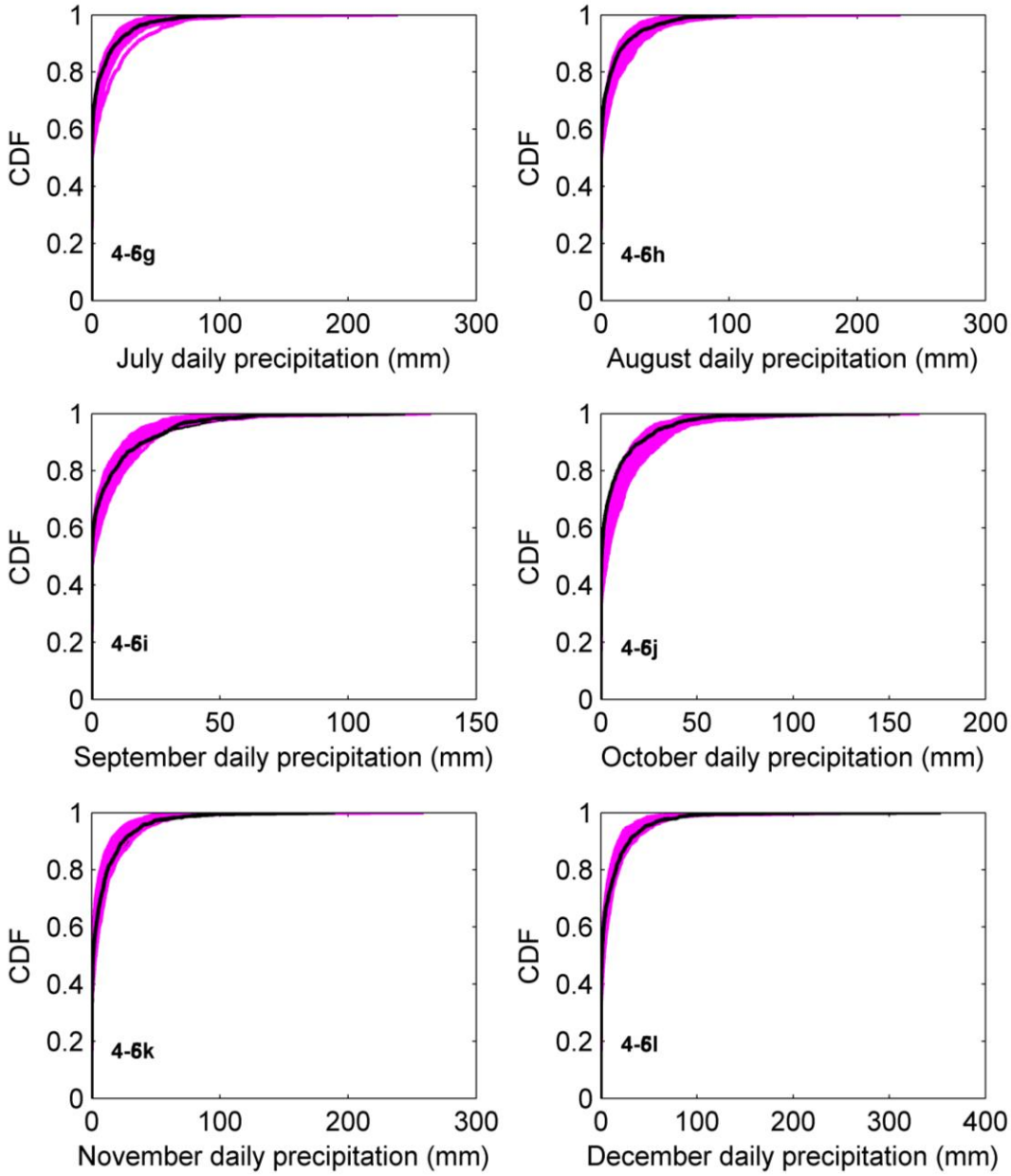
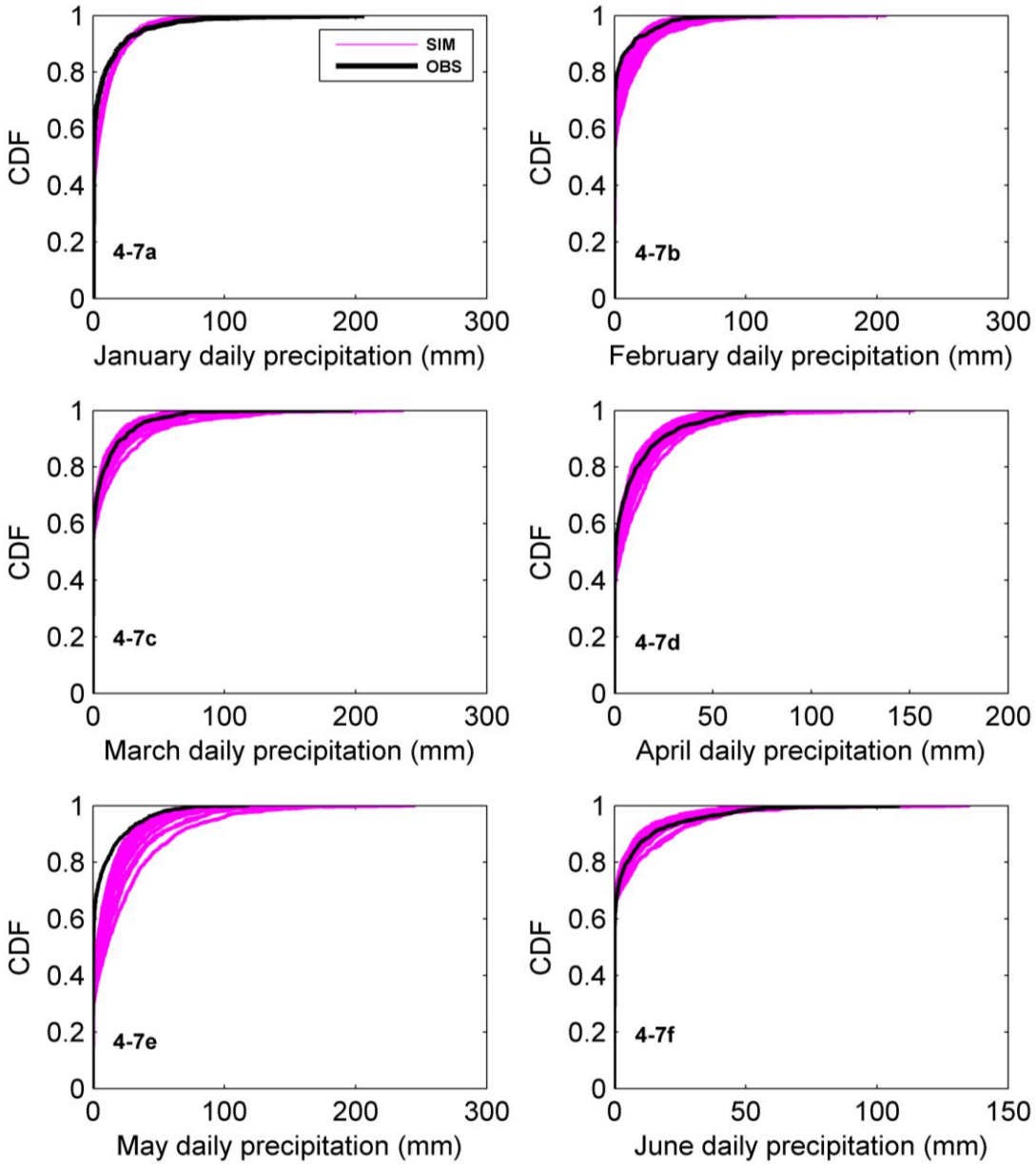


Figure 4-6 Comparison of cdf of the downscaled precipitation ensembles with observed cdf for all the months at station S55


SIM
OBS
CDF
4-7a
January daily precipitation (mm)
4-7b
February daily precipitation (mm)
4-7c
March daily precipitation (mm)
4-7d
April daily precipitation (mm)
4-7e
May daily precipitation (mm)
4-7f
June daily precipitation (mm)

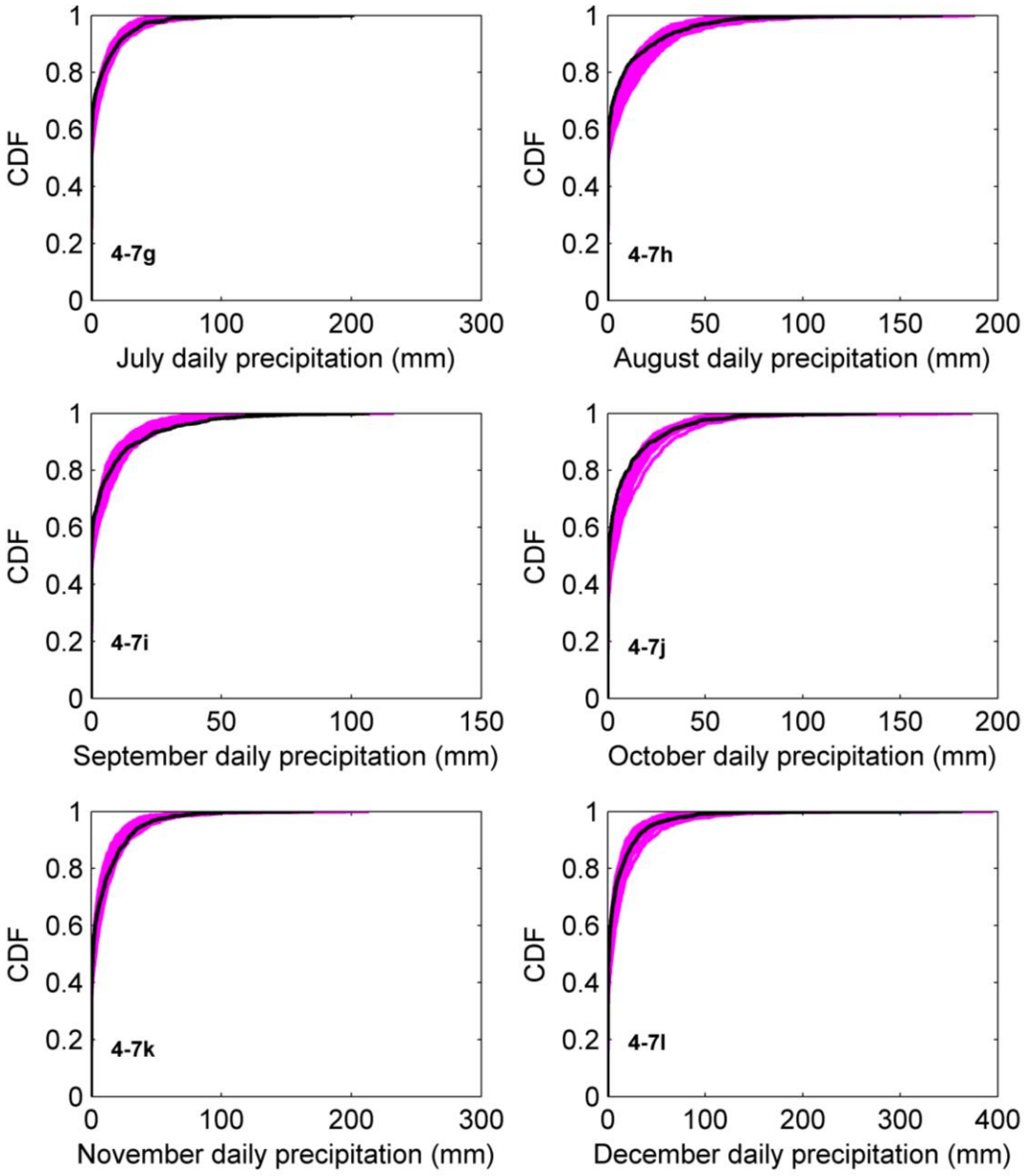


Figure 4-7 Comparison of cdf of the downscaled precipitation ensembles with observed cdf for all the months at station S69

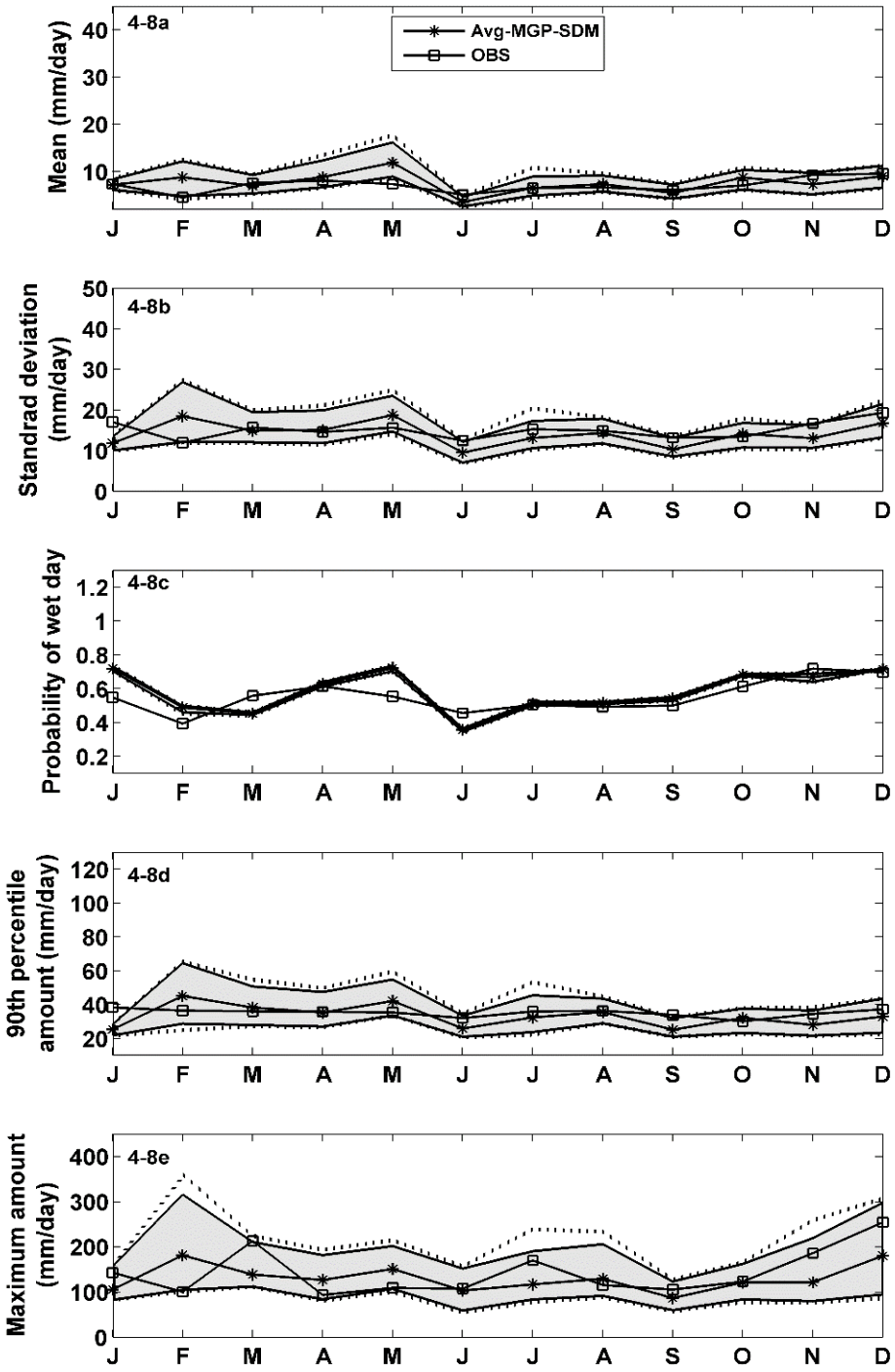


Figure 4-8 Evaluation Statistics for the station S46. The line with square represents the observed data. The shaded region shows the 5th and 95th percentile of the prediction ensembles. The dotted lines represent the maximum and minimum range.

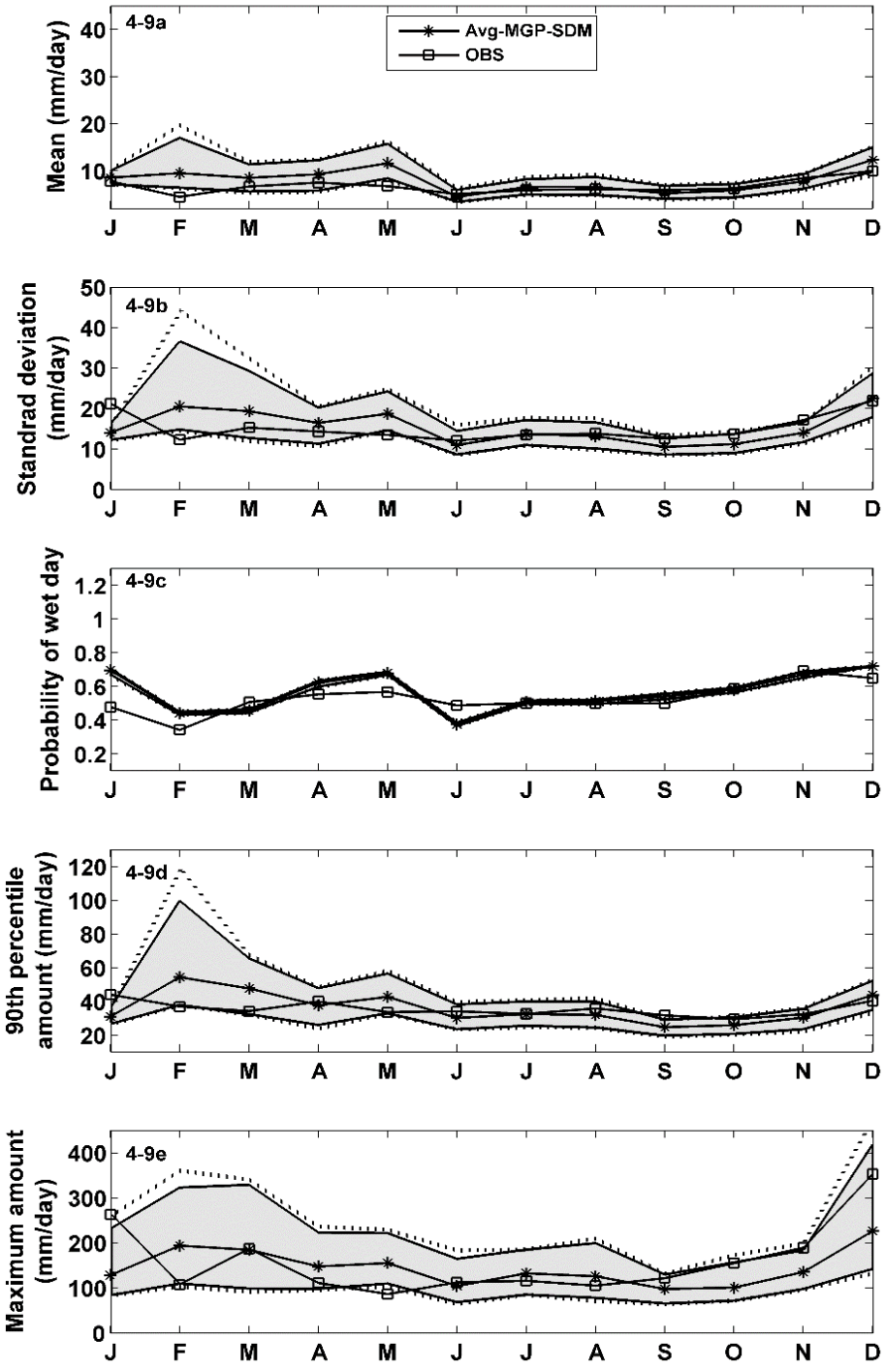


Figure 4-9 Evaluation Statistics for the station S55. The line with square represents the observed data. The shaded region shows the 5th and 95th percentile of the prediction ensembles. The dotted lines represent the maximum and minimum range.

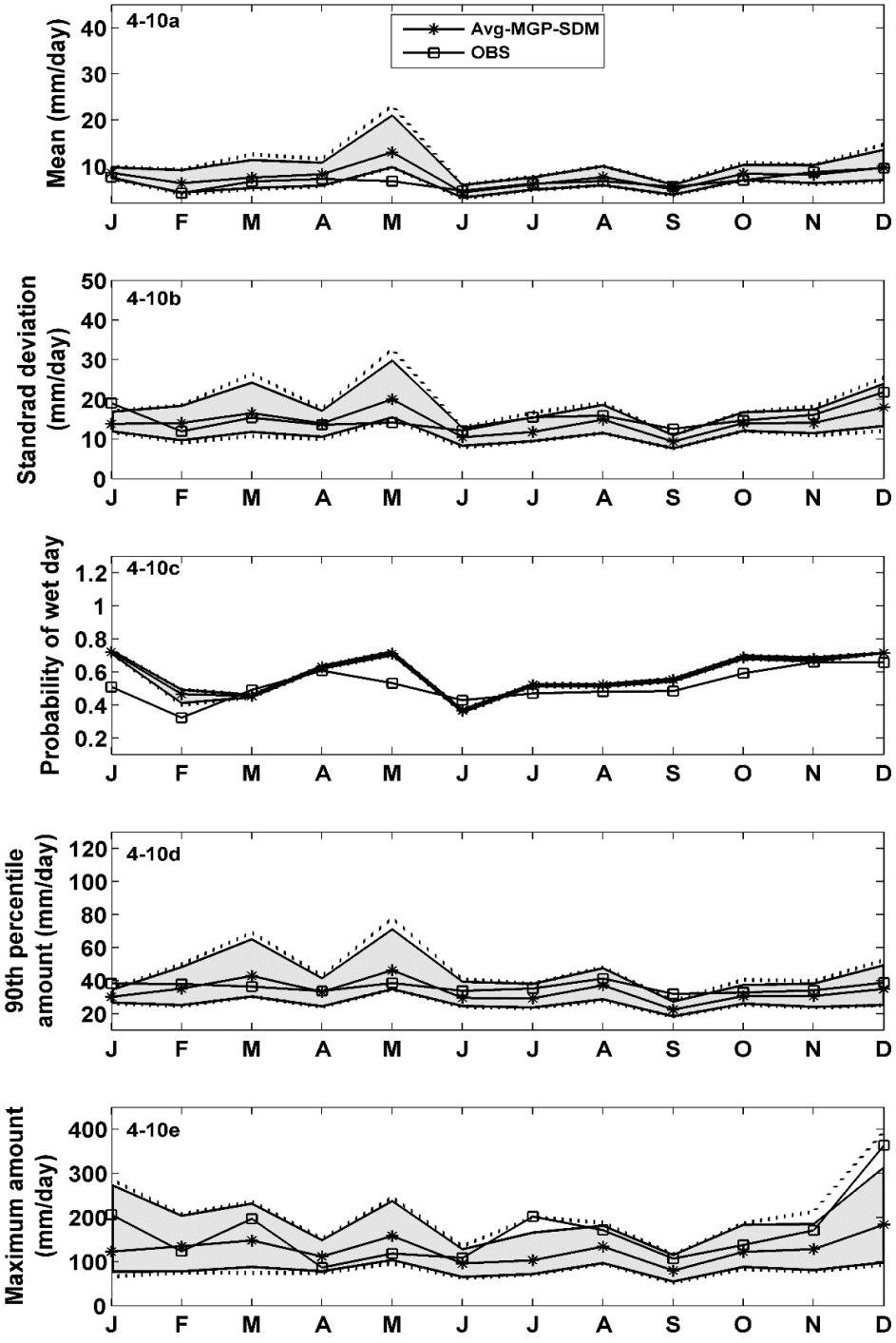


Figure 4-10 Evaluation Statistics for the station S69. The line with square represents the observed data. The shaded region shows the 5th and 95th percentile of the prediction ensembles. The dotted lines represent the maximum and minimum range.

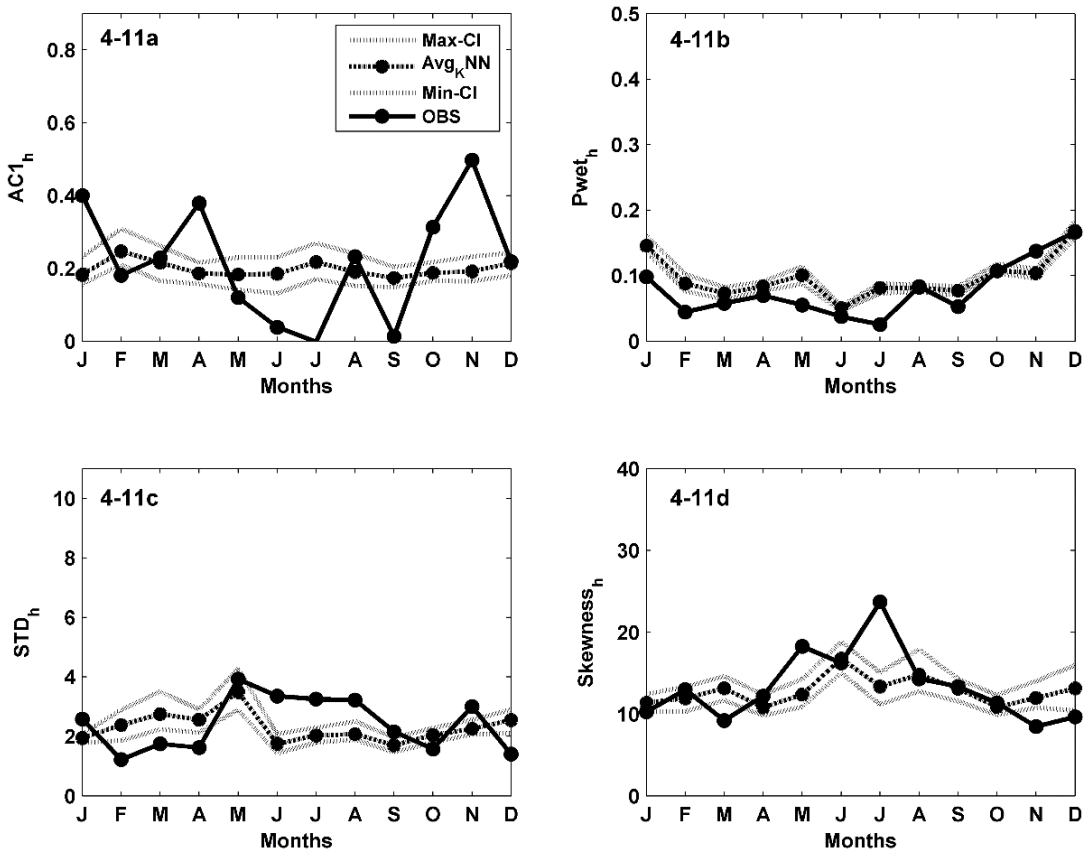


Figure 4-11 Disaggregated precipitation projection at station S46 for the validation period 1980 - 2010. The average hourly disaggregated data for the baseline period calculated from 20 ensembles

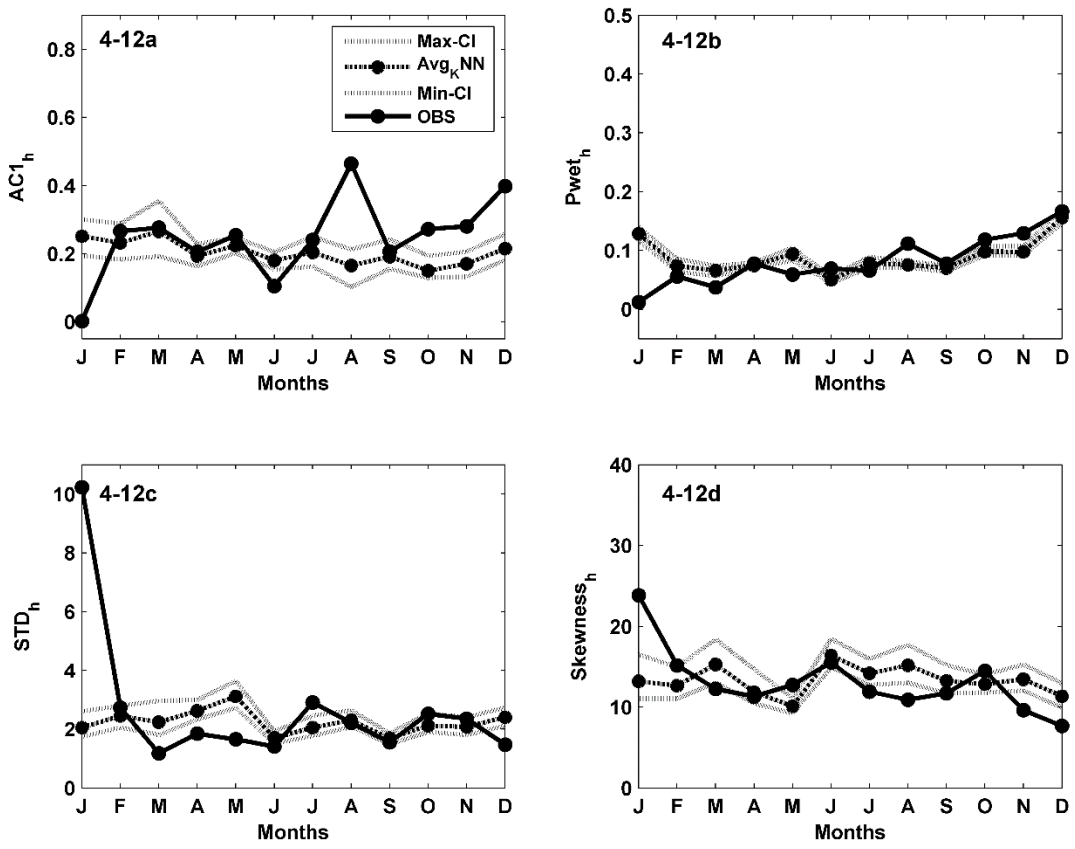


Figure 4-12 Disaggregated precipitation projection at station S55 for the validation period 1980 - 2010. The average hourly disaggregated data for the baseline period calculated from 20 ensembles.

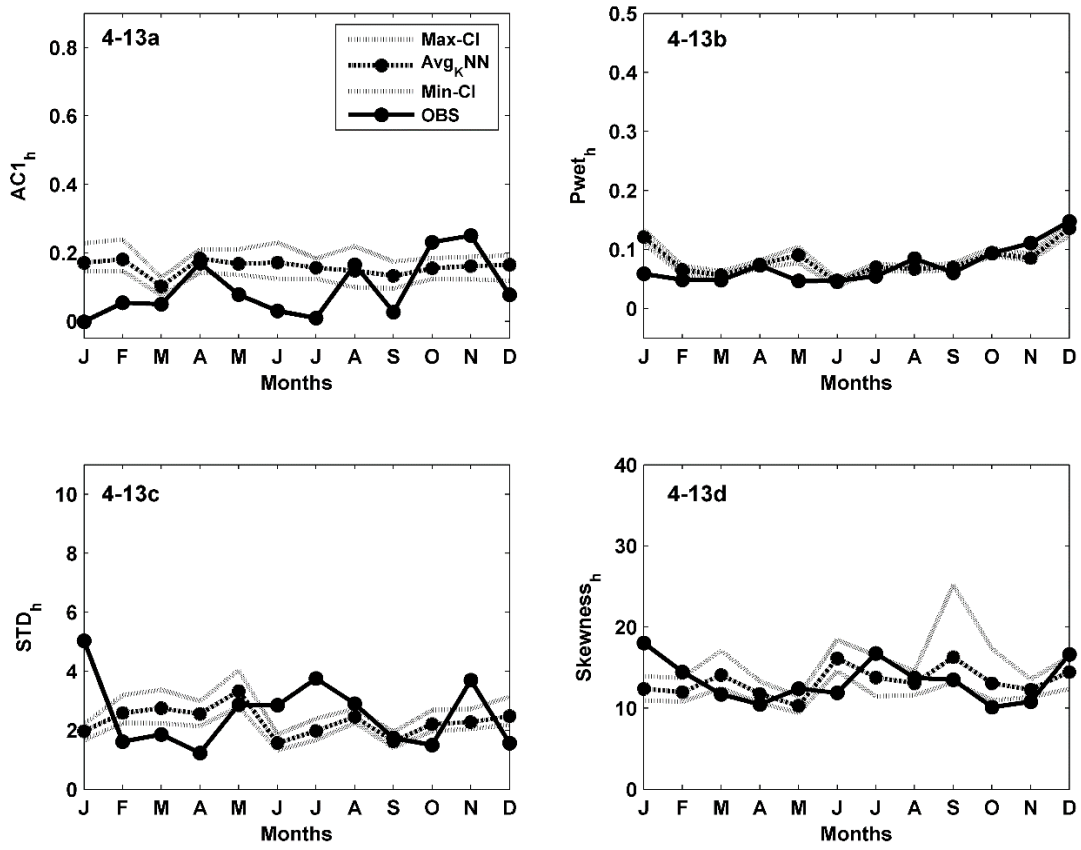


Figure 4-13 Disaggregated precipitation projection at station S69 (4-13a-4-13d) for the validation period 1980 -2010. The average hourly disaggregated data for the baseline period calculated from 20 ensembles.

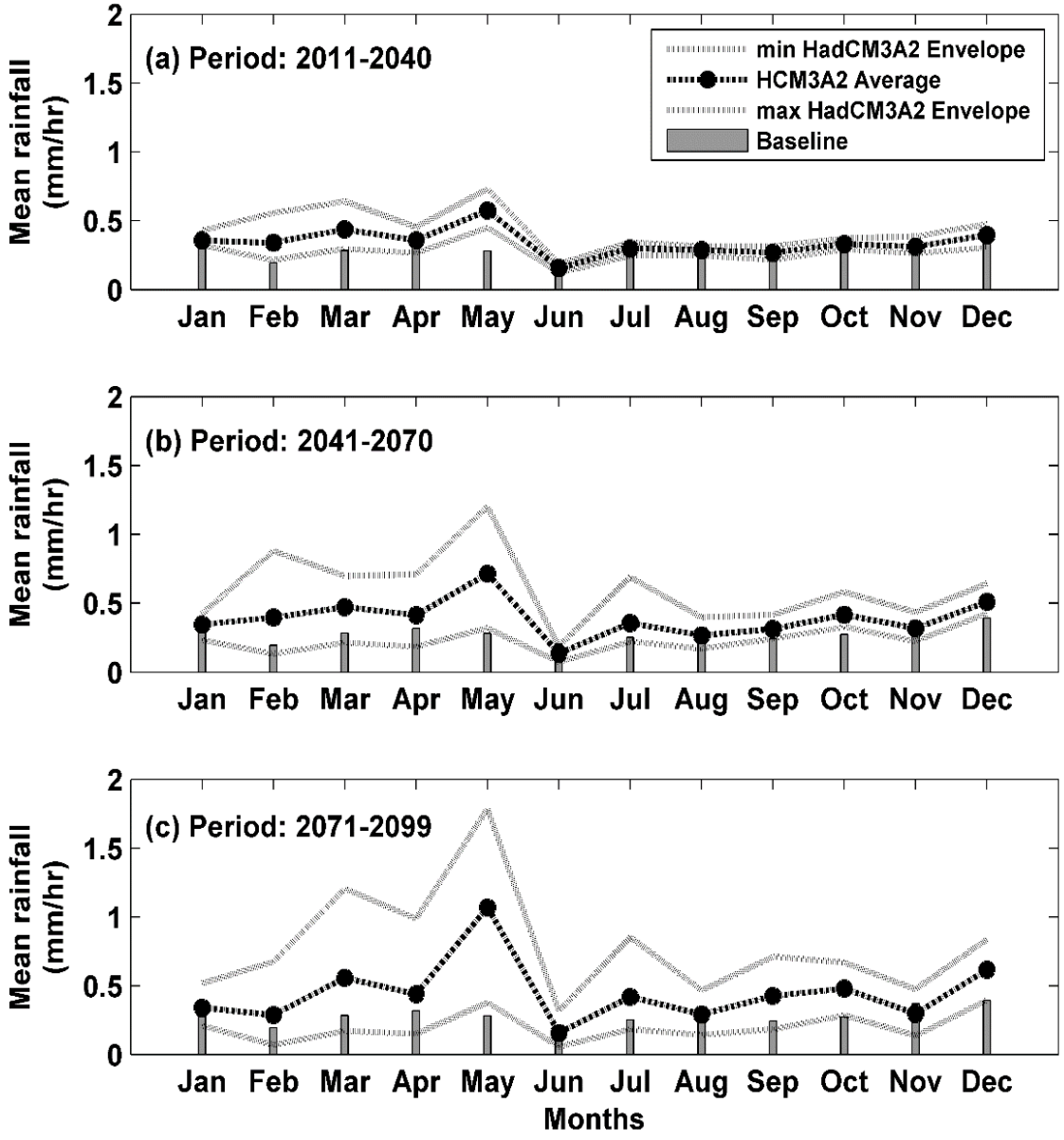


Figure 4-14 Mean hourly precipitation projection for future periods (2011-2099)

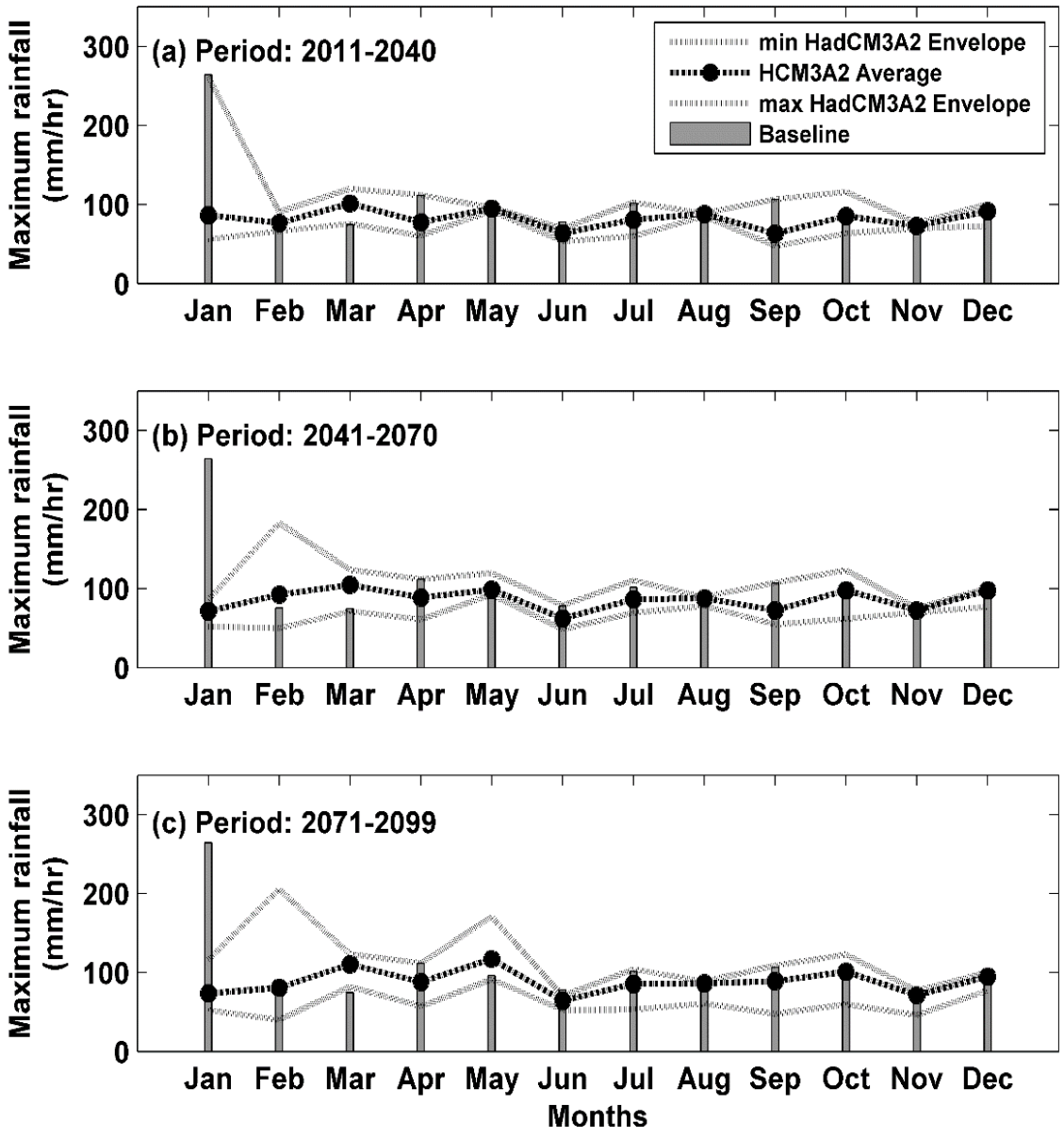


Figure 4-15 Maximum hourly precipitation projection for future periods (2011-2099)

CHAPTER 5 Integrated MGP-SDM and BUQ-SDDHM for uncertainty quantification to study the impact of climate change on future flood events

5.1 Introduction

The content of this Chapter is extracted from a manuscript that is to be submitted to a journal. There has been an increase in the frequency and intensity of flood events in many parts of the world due to the change in climate conditions (Solomatine *et al.*, 2008). This is evident in Quebec where it witnessed frequent flood events in the past (Gagnon *et al.*, 2005; Jones, 2008). Thus, it is important to understand the impact of climate change on the future flood events for robust planning and effective decision making. The GCM scenarios do not give accurate predictions at local level due to coarse spatial resolution. Accurate predictions for finer spatial resolution are needed for studying the impact of climate change on water resources (Tisseuil *et al.*, 2010). The hydrological models require high resolution precipitation than the downscaled climate variables for simulating stream flow. The multisite downscaling technique has to be implemented to simulate climate scenarios that represent the local conditions as the river flow depends on the rainfall occurrences at multiple sites in the watershed. Several downscaling models based on statistical relationship between the GCM predictors and local climate variables have been developed previously as it is computationally less intensive and easy to implement (Wilby and Wigley, 1997; Fowler *et al.*, 2007). The integration of multisite downscaling techniques and the multisite disaggregation model is proposed in the recent studies to simulate high spatial and temporal resolution climate variables at a local level (Mezghani, 2009). These downscaled and disaggregated climate variables are used as input in hydrological models to simulate high resolution runoff which is very useful in urban planning especially urban drainage design planning. It is found that there are very limited studies on integrated downscaling model, hydrological model and disaggregation model approach to assess the impact of climate change on hydrology.

The uncertainty in studying the impact of climate change on hydrology (rainfall, runoff and water resources) has gained interest recently (Xu, 1999). It is important to have the knowledge about the uncertainty of the change in future climate variables for decision making and planning. The uncertainty is encapsulated in each stage of future hydrological flow simulation (GCM predictors $\rightarrow$ downscaling model $\rightarrow$ disaggregation model $\rightarrow$ hydrological model) as all of them are numerical models which are generalized representation of reality. The uncertainty in the predictions is also due to factors such as anthropogenic greenhouse gas emissions and natural climate processes (Foley, 2010). The sources of uncertainty in GCM are dynamic climate system, generalization of climate in the numerical model, the parameterizations used for representing the natural process, initial boundary conditions of the climate model and incomplete knowledge about the system being modelled. The uncertainty in the SDM is introduced by the GCM predictions which are used as inputs, model structure, parameters, random variations and numerical errors (Rajendran and Cheung, 2015). It is also noted that no generalized uncertainty quantification framework exists for quantifying uncertainty in the above mentioned integrated approach. This study focuses on developing an integrated downscaling, disaggregation and hydrological model coupled with Bayesian uncertainty quantification framework. Development of a stochastic data-driven hydrological model is also considered in this study.

5.1.1 Integrated multi-site SDM and DDHM for runoff simulation

Tisseuil *et al.* (2010) used the statistical downscaling model directly to simulate river flows instead of downscaling climate variables. In another study, SDSM was used to simulate climate data for streamflow modelling in Quebec (Gagnon *et al.*, 2005). It was shown that the use of downscaling prior to the use in hydrological models improved the flow prediction. Grouillet *et al.* (2016) analyzed the sensitivity of a hydrological model to different statistical downscaling models that downscaled precipitation and temperature. The outputs from the three statistical downscaling models such as an analog method (ANALOG), a Stochastic Weather Generator (SWG) and a cumulative distribution function-transform approach (CDFt) were used as inputs to the GR4j

conceptual model to simulate a streamflow. Their results showed that there was improvement in simulation of streamflow when a high resolution downscaled climate variables were used to simulate streamflow compared to using low resolution climate model outputs. Their results also concluded that the ANALOG and CDFt models performed better than SWG model. Chen *et al.* (2010) compared Smooth Support Vector Machine (SSVM), SDSM for statistical downscaling of climate variables; the downscaled future scenarios were used as input in Xin-anjiang and HBV hydrological models for runoff simulation. The results showed that the SDSM performed well in simulating the rainfall and the runoff was simulated well by SSVM downscaled scenarios. The suitability of statistical downscaling models for runoff simulation was studied by Samadi *et al.* (2013). They integrated downscaling techniques such as SDSM, ANN (Artificial Neural Network) with a hybrid conceptual hydrological model to generate future runoff. Their results showed that the choice of the model used for downscaling had a major impact of the simulated streamflow ensembles. Fowler *et al.* (2007) also stated that downscaling model was one of the sources of uncertainty and it played a critical role in hydrological studies. Liu *et al.* (2016) proposed Bayesian model averaging to combine the downscaled precipitation from three downscaling models such as SVM, BCC/RCG-Weather Generators (BCC/RCC-WG) and SDSM. The ensembles combined using BMA were then used as input in Soil and Water Assessment Tool (SWAT) for runoff modelling. Their result showed that ensemble downscaling method performed well in runoff simulation compared to runoff simulation using separate downscaling methods. Lu *et al.* (2016) developed an integrated statistical and data-driven (ISD) framework to simulate river flow for analyzing flood frequencies under climate change in the Duhe River, China. In ISD, ASD and KNN was integrated to downscale rainfall and Conditional Density Estimate Network (CDEN) was used to downscale minimum temperature and relative humidity at multiple local weather stations. In the next step, a data-driven Bayesian Neural Network (BNN) was used to generate monthly future streamflows using downscaled climate variables. KNN were then used to disaggregate the monthly streamflow to daily streamflow for flood frequency analysis. ISD approach gave a generalized framework

for analyzing future flood frequency analysis. The drawback of their framework is that several models need to be run to simulate high resolution flow time series for assessing the impact. Another drawback is that it does not have uncertainty framework for ISD. There are several other studies that analyzed the major sources of uncertainty in streamflow simulation from GCM scenarios (Khan *et al.*, 2006). Their results concluded that GCM and emission scenarios were the major sources of uncertainty, statistical downscaling models were second major sources of uncertainty whereas the hydrological model and its parameters contributed less uncertainty to the output. These studies provided a generalized framework to compare the performance of the model. However, there requires a generalized uncertainty quantification framework that can propagate uncertainty from GCM to hydrological simulations.

5.1.2 Hydrological models

The commonly used streamflow simulation models can be divided into categories such as statistics or stochastic model data-driven model and conceptual or physics based model (Price, 2000; Refsgaard *et al.*, 2013). The statistics or stochastic data-driven model is based on precipitation, temperature and relative humidity (Solomatine and Price, 2004). Several stochastic time series model such as Autoregressive (AR), Moving Average (MA), Autoregressive Moving Average (ARMA) and Autoregressive Integrated Moving Average (ARIMA) are used as data-driven hydrological models. These models do not consider the physical processes and are based on the linear relationship between the input and output (Tokar and Johnson, 1999; Riad *et al.*, 2004). In a study by Tisseuil *et al.* (2010), Generalized Linear Model (GLM) and Generalized Additive Model (GAM), Aggregated Boosted Trees (ABT) and multi-layer preceptor neural networks (ANN) were used to downscale the river flow directly. Their study results showed that the non-linear models performed better in flow simulation compared to GLM downscaling models. Thus, artificial intelligence based Artificial Neural Network (ANN) for streamflow prediction was proposed to capture the nonlinearity (Govindaraju and Rao, 2000). For hydrological studies, ANN has been in use to simulate river flows (Wei *et al.*, 2012). The ANN models have also been applied

in various hydrological applications and the results showed that the ANN model had better prediction capability (Govindaraju and Rao, 2000). In the research work of Khan *et al.* (2006), they compared the ability of BNN and ANN to simulate streamflows. Their results showed that BNN performed better than ANN model in simulating mean and extreme flows. The applicability of BNN have been explored by Lu *et al.* (2016). Oyebode *et al.* (2014) compared the performance of Genetic Programming and differential equation-trained artificial neural networks for streamflow prediction. Their results showed that the Genetic Programming performance was superior to ANNs especially for the non-linear variations of the hydro-meteorological parameters. The drawback of ANN in streamflow simulation such as overfitting and instability of the model when the training data was less was noticed by Hsieh and Tang (1998). The kernel based SVM method was introduced to address the overfitting issues in ANN models. The advantage of SVM is that the non-linear relationship between the predictor and the predictand can be represented using a kernel implicitly. This is referred as 'kernel trick' (Bishop, 2006). However, all these data-driven models are deterministic and the prediction uncertainty cannot be obtained directly.

5.1.3 Uncertainty analysis in hydrological models

In case of hydrological models, the uncertainty arises from the numerical hydrological models and the incomplete knowledge about the hydrological processes (Mirzaei *et al.*, 2015). In hydrological models, a statistical generalized likelihood uncertainty estimation (GLUE) has been commonly used to characterize the uncertainty in the future predictions using the posterior distribution of the parameters (Beven and Binley, 1992; Freer *et al.*, 1996). The plausibility of the possible outcomes of the hydrological model is obtained by Monte Carlo simulations. The effect of uncertainties in the physical process based hydrological models was studied by many researchers (Engeland *et al.*, 2005; Kavetski *et al.*, 2006; Chowdhury and Sharma, 2007; Marshall *et al.*, 2007; Srivastav *et al.*, 2007; Jin *et al.*, 2010). The uncertainty quantification methodologies for hydrologic modelling consider only the parameter uncertainty. The uncertainty quantification techniques is modified to consider all types of uncertainty in

several studies (Schaefli *et al.*, 2007; Yang *et al.*, 2008). In addition, these studies were based on the assumption that the model errors were independent. Let the historic data for simulating streamflow be D_h of n_h observations, $D_h = \{(\mathbf{x}_{hi}, \mathbf{y}_{hi}) \mid i = 1, ..., n_h\}$ where $\mathbf{x}_h$ represents the climate variables such as precipitation, temperature and relative humidity with dimension m_h and $\mathbf{y}_h$ is the binary classification output (wet/dry day). In vector form, the predictor data can be represented as a matrix X_h with dimension $m_h \times n_h$ and the wet/dry day classification output vector is denoted as $\mathbf{y}_h$. Let the data-driven hydrological model be represented as equation (5.1):

$$\mathbf{y}_h = \underbrace{f_h(\mathbf{x}_h)}_{epistemic} + \underset{aleatory}{\boldsymbol{\varepsilon}_h} \tag{5.1}$$

where $\mathbf{y}_h$ is the flow data, $f_h(\mathbf{x}_h)$ is the hydrological modeling function which is a source of epistemic uncertainty in the predictions, $\boldsymbol{\theta}_h$ is the model parameters and $\boldsymbol{\varepsilon}_h$ is the residual/error of the data-driven hydrological model and is also a source of aleatory uncertainty in the predictions. Generally, ANN and SVM are deterministic models; they predict only the $f_h(\mathbf{x}_h)$. The model error, $\boldsymbol{\varepsilon}_h$ is calculated using the training and the validation data. The error is then fitted separately with normal or lognormal-3 distribution to estimate the confidence interval of the prediction (Salas *et al.*, 2000). This is a drawback since the model function and the error are fitted separately with different distribution assumption. RVM was developed to generate probabilistic predictions where the posterior distributions of the runoff can be simulated directly instead of deterministic predictions. RVM uses Bayesian framework to determine the posterior distribution of the model weights; the model weights are assumed to be random variables. RVM was used for long-term forecasting of streamflow by Liu *et al.* (2017). However, the applicability of RVM in hydrological applications is very limited. RVM has been successfully implemented in statistical downscaling of the precipitation (Ghosh and Mujumdar, 2008). The disadvantages of RVM are that if a future data point is located far from the relevance vectors in RVM, the predictive results are unreliable;

in this case, the predictive distribution follows a Gaussian distribution with mean and variance zero (Rasmussen and Williams, 2006). This issue was overcome by using GPR. Gaussian Process (GP) is a stochastic process models and it proves to be efficient in quantifying uncertainty in the models by capturing the dependency between the errors (Rasmussen and Williams, 2005). GP is non-parametric models in which the prior is placed over the model function rather than over the parameters; thus, GP captures all types of uncertainty. GP enables a principled way to quantify the uncertainty by coupling the uncertainty tool with the model calibration. GP uses Bayesian updating framework to calibrate the model and the uncertainty parameters simultaneously. Thus, the Bayesian updating framework gives both the posterior distribution of the model parameters along with the predictive mean and the predictive variance. Gaussian process models have been used in several other fields such as Geostatistics (Isaaks and Srivastava, 1989) and structural mechanics (Cheung and Beck, 2009). GPR was used by Sun *et al.* (2014) to simulate monthly streamflow probabilistically. Their method was not implemented for daily streamflow. In addition, their GPR framework assumed zero mean. GPR with mean function (linear/non-linear) needs to be developed for simulating daily streamflow. In the recent study, the uncertainty in SDM is the Bayesian updating model framework using GP for single site statistical downscaling named SGP-SDM (Rajendran and Cheung, 2015). In their work, the uncertainty quantification tool was coupled with the downscaling model framework to obtain the predictive mean and variance instead of point estimates. This method is shown to be efficient over the classical methods such ASD and GLM for single site downscaling in the Chapter 2 and Chapter 3 of this thesis. GP has also been used to downscale precipitation at multiple-site respectively in the Chapter 4 of this thesis.

The methodology developed by Lu *et al.* (2016), the precipitation was first downscaled using ASD and KNN; the Conditional Density Estimation Network (CDEN) was employed to simulate other climate variables such as minimum and maximum temperature and relative humidity conditioned on the downscaled precipitation. The runoff was generated using the monthly meteorological data as input in BNN data-

driven hydrological model. The drawback of their framework is that several models need to be run to simulate high resolution flow time series for assessing the impact. The proposed framework can downscale climate variables at multiple sites simultaneously using MGP-SDM thus eliminating the need to use CDEN and KNN for downscaling other climate variables and spatial disaggregation respectively. The river flow is then simulated using the BUQ-SDDHM and KNN is used for disaggregation of river flow to a finer time scale. The precipitation occurrence determination is not used since the monthly data is chosen for downscaling.

The main objective of this Chapter is twofold. An integrated multi-site statistical downscaling and data-driven hydrological model is proposed to simulate river flow in future from climate model output. The integrated framework implicitly couples the Bayesian uncertainty quantification tool with the downscaling model and the hydrological model. First, multi-site MGP-SDM developed in Chapter 4 is applied to downscale monthly precipitation, temperature at multiple sites simultaneously. In order to generate future runoff using downscaled climate variables, a data-driven hydrological model using GP is developed as the second step. The downscaled high resolution climate variables are used as inputs in data-driven hydrological models to simulate posterior predictive distribution of the runoff. Flood frequency analysis of the simulated runoff is performed to estimate the return time in future periods 2011-2040, 2041-2070 and 2071-2099.

5.2 Data and Study area

Vermillon river basin located in the province of Quebec, Canada is chosen for studying the impact of climate change on hydrology shown in Figure 5-1. The impact of climate change on water resources in province of Quebec has gained attention recently. As there is hydroelectric power production in Vermillon river basin, the prediction of the impact of climate change on water resources is important (Robinson, 1997). The predictors such as the precipitation, temperature and relative humidity are obtained from the Environment Canada climate database. The large-scale atmospheric variables

from HaDCM3 A2 GCM scenarios are used for validating and projecting the future climate scenarios. The HadCM3 predictors are downloaded from Canadian Climate Data and Sceanrios (http://www.cccsn.ec.gc.ca/?page=pred-hadcm3). The NCEP reanalysis climate data re-gridded to HaDCM3 A2 scenario data are used for model calibration (Kalnay *et al.*, 1996). The available predictors for Quebec province are mean sea level pressure (mslp), mean temperature at 2m (temp), relative humidity (rhum), specific humidity (shum), geopotential height at 500m (p500), geopotential height at 850 (p850), airflow strength at 500m (s500), airflow strength at 850m (p850), zonal velocity, meridional velocity, vorticity, wind direction and divergence. The hydrologic data of the observed streamflow data are obtained from the hydrometric stations in the HYDAT database given by Environment Canada. The hydrometric stations are located at the outlet of all the river basin. The locations of the climate stations and hydrometric station are shown in Figure 5-1. Table 5-1 and Table 5-2 describe the Vermillon climate station and hydrometric station.

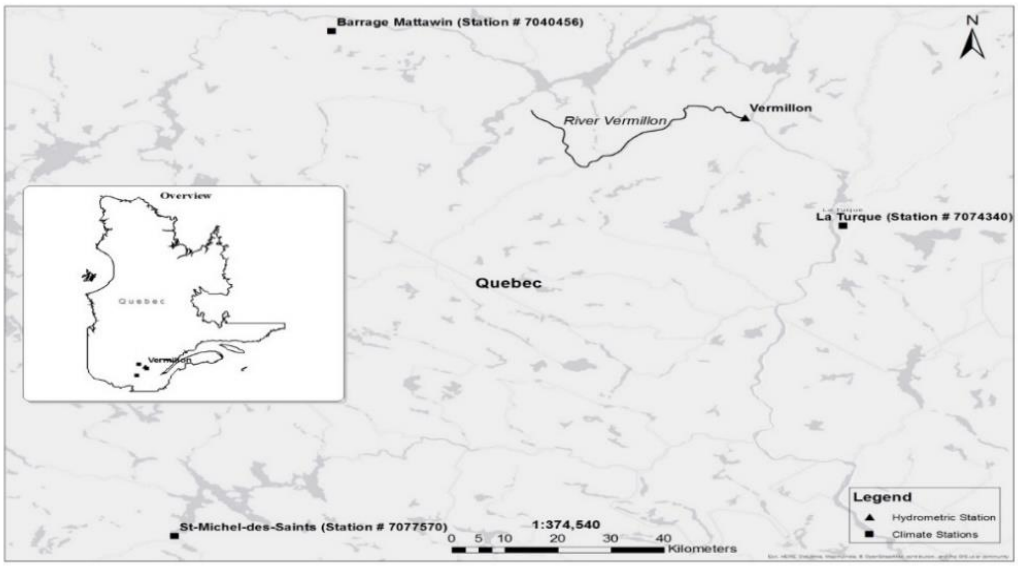


Figure 5-1 Climate station and hydrometric data location for Vermillon watershed, Quebec, Canada

The predictors need to be standardized using (5.2) before using it for precipitation amount estimation and precipitation occurrence determination:

$$z_i = \frac{r_i - r_{\min}}{r_{\max} - r_{\min}} \tag{5.2}$$

where r_i is the i^{th} input variable; $r_{\min}$ and $r_{\max}$ are the minimum and maximum values of the input variables, respectively.

Table 5-1 Vermillon climate station information

Vermillon Climate Station	Station Number	Latitude	Longitude
La Turque	7074340	47°24'N	72°47'W
Barrage Mattawin	7040456	46°51'N	73°39'W
St-Michel-des-Saints	7077570	46°41'N	73°55'W

Table 5-2 Vermillon hydrometric station information

Vermillon Hydrometric Station	Station Number	Latitude	Longitude
Vermillon	02ND001	47°39'N	72°57'W

5.3 Methodology

The integrated multisite SDM and data-driven hydrological model framework is proposed for river flow simulation and flood frequency analysis under climate change. The workflow is represented in Figure 5-2.

5.3.1 MGP-SDM

The MGP-SDM proposed in Chapter 4 for downscaling precipitation at multiple sites jointly is adopted for downscaling the climate variables such as precipitation, temperature and relative humidity. The MGP-SDM, the spatial correlation and the dependency between the residuals are coupled with the model calibration, thus enabling estimation of the model parameters, the spatial correlation parameters and the residual parameters simultaneously. The model is formulated based on the assumption that the residuals are stochastic processes following Gaussian distribution as the dependencies

between the residuals affect the model predictions. The output from the model itself is the predictive mean and the predictive variance as the residual are coupled with the model. The difference between the single site downscaling

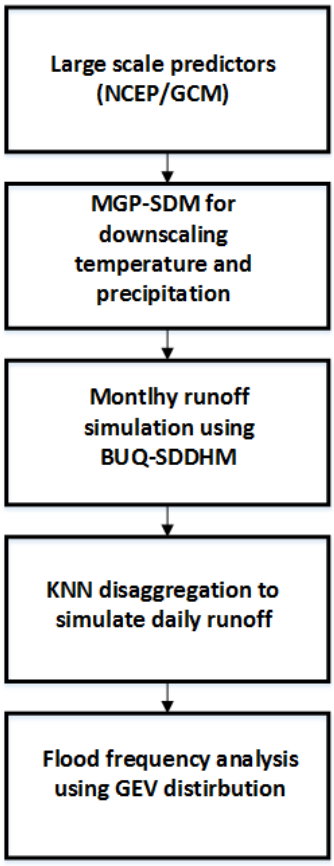


Figure 5-2 Integrated MGP-SDM and BUQ-SDDHM workflow framework for hydrological impact studies

and the multisite downscaling is that there is a change in the mean function matrix and covariance function matrix representation. Let the historic data be D_{mh} for the precipitation amount estimation model of n_{mh} observations represented as $D_{mh} = \{(\mathbf{x}_{mhi}, y_{mhi}) \mid i = 1,...,n_h\}$, where $\mathbf{x}_{mh}$ is the GCM predictor with dimension m_{mh} and $\mathbf{y}_{mh}$ represents the rainfall amount. The number of sites is represented as s_h. The mean function and the covariance matrix for the multi-output GP are then specified by equation (5.3):

$$\mathbf{y}_{mh} \sim GP(\boldsymbol{\mu}_{mh}, \Sigma_{mh}) \tag{5.3}$$

$$\text{where } \boldsymbol{\mu}_{mh} = \begin{bmatrix} \mathbf{g}_{mh1}(\mathbf{X}_{mh}^{1})^{\mathrm{T}} & 0 & \ldots & 0 \\ 0 & \mathbf{g}_{mh2}(\mathbf{X}_{mh}^{2})^{\mathrm{T}} & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & \mathbf{g}_{mh}(\mathbf{X}_{mh}^{s_h})^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_{mh1} \\ \vdots \\ \boldsymbol{\beta}_{mhs_h} \end{bmatrix} \text{ and}$$

$$\Sigma_{mh} = \begin{bmatrix} \Sigma_{mh}^{11} & \Sigma_{mh}^{12} & \cdots & \Sigma_{mh}^{1s} \\ \Sigma_{mh}^{21} & \Sigma_{mh}^{22} & \cdots & \Sigma_{mh}^{1s} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{mh}^{s_h 1} & \Sigma_{mh}^{s_h 2} & \cdots & \Sigma_{mh}^{s_h s_h} \end{bmatrix}$$

where $\mathbf{g}_{mh}(\mathbf{X}_{mh}^{s_h})^{\mathrm{T}}$ is the downscaling model mean function for the site s_h which can be either linear or non-linear, $\boldsymbol{\beta}_{mh}$ represents the coefficients of the mean function corresponding to each site s_h, $\sum_{mh}^{pq}$ is the cross-covariance between the datasets $\mathbf{x}_{mh}$ corresponding to the sites p and q, the diagonal entries of $\sum_{mh}^{ss}$ refer to the auto-covariance. The derivation of cross-covariance using convolution was presented in the work by Robin (2012). They used non-separable covariance matrix to derive the cross-covariance for multi-output GP (Higdon, 2002). The challenging part is to define a positive definite covariance matrix since there are no direct functions for defining the positive-definite cross-covariance. Auto-and cross-covariance function described by Boyle and Frean (2004) using the convolution property for handling multiple outputs is given in (5.4).

$$\mathrm{K}_{mhpq}(\mathbf{d}_h) = \begin{cases} \exp\{-\mathbf{d}_h^{\mathrm{T}}\mathbf{B}_{hp}\mathbf{d}_h\} & \textit{if } p = q \\ \dfrac{\exp\{-\mathbf{d}_h^{\mathrm{T}}(\frac{1}{2}\mathbf{B}_{hp}^{-1} + \frac{1}{2}\mathbf{B}_{hq}^{-1})^{-1}\mathbf{d}_h\}}{|(\frac{1}{2}\mathbf{B}_{hp} + \frac{1}{2}\mathbf{B}_{hq})(\frac{1}{2}\mathbf{B}_{hp}^{-1} + \frac{1}{2}\mathbf{B}_{hq}^{-1})|^{1/4}} & \textit{otherwise} \end{cases} \tag{5.4}$$

where $\mathbf{B}_{hp} = \frac{\mathbf{l}_p^{-1}}{2}$ and l_p is the correlation length corresponding to each dimension of the output, $\mathbf{d}_h$ is the distance and $\| \mathbf{x}_{mhi}^{p} - \mathbf{x}_{mhj}^{q} \|^2 . \sum_{mh}(d_h) = M_{mhpq} K_{mhpq}(d_h)$ is the covariance matrix that accounts for the covariance between the observations at the multiple sites whereas M_{mhpq} is $s_h \times s_h$ the positive-definite matrix between the sites.

The estimation of joint likelihood of M_{mhpq} and $\mathbf{B}_{hp}$ simultaneously by optimizing the following marginal likelihood is given in equation (5.5):

$$L(M_{mha}, \mathbf{B}_{ha1}, ..., \mathbf{B}_{has}) = \frac{|\sum_{mha}^{-1}|^{1/2}}{|\mathbf{G}_{mha}^{\mathrm{T}} \boldsymbol{\Sigma}_{mha}^{-1} \mathbf{G}_{mha}|^{1/2}} \exp(-\frac{1}{2}(\mathbf{y}_{mha} - \mathbf{G}_{mha}\boldsymbol{\beta}_{mha})^T \sum_{mha}^{-1} (\mathbf{y}_{mha} - \mathbf{G}_{mha}\boldsymbol{\beta}_{mha})) \quad (5.5)$$

Estimation of this marginal likelihood function is impractical (Boyle and Frean, 2004). Thus, a multi-stage optimization procedure is used for finding the optimal hyperparameter values. The procedure for estimation of the hyperparameter was explained in detail by Robin (2012) and followed in multivariate emulator package in R. The steps for multi-stage optimization are as follows:

1) Estimate the correlation length for each observation using the single task GP procedure.
2) Calculate diagonal elements of the marginal variance terms M_{mha} using the posterior mode.
3) Determine off-diagonal elements of M_{mha} numerically using the posterior mode.

Once the hyperparameters are optimized, the predictive mean and the predictive covariance for the future data $\mathbf{x}_{mh}^{*}$ can be estimated using the conditional Gaussian distribution property (5.6):

$$\begin{bmatrix} \mathbf{y}_{mh} \\ \mathbf{y}^{*}_{mh} \end{bmatrix} \sim N\left(\begin{pmatrix} f_{mh}(\mathbf{x}_{mh}, \boldsymbol{\theta}_{mh}) \\ f_{mh}(\mathbf{x}^{*}_{mh}, \boldsymbol{\theta}_{mh}) \end{pmatrix}, \begin{pmatrix} \sum_{mhx_{mh}, x_{mh}} & \sum_{mhx^{*}_{mh}, x_{mh}} \\ \sum_{mhx_{mh}, x^{*}_{mh}} & \sum_{mhx^{*}_{mh}, x^{*}_{mh}} \end{pmatrix} \right)$$

$$m(\mathbf{x}^{*}_{mh}) = f_{mh}(\mathbf{x}^{*}_{mh}, \boldsymbol{\theta}_{mh}) + \sum_{mhx_{mh}, x_{mh}} \sum^{-1}_{mhx_{mh}, x_{mh}} (\mathbf{y}_{mh} - f_{mh}(\mathbf{x}_{mh}, \boldsymbol{\theta}_{mh})) \quad (5.6)$$

$$Cov(\mathbf{x}^{*}_{mh}) = \sum_{mhx^{*}_{mh}, x^{*}_{mh}} - \sum^{\mathbf{T}}_{mhx_{mh}, x^{*}_{mh}} \sum^{-1}_{mhx_{mh}, x_{mh}} \sum_{mhx_{mh}, x^{*}_{mh}}$$

where $m(\mathbf{x}^{*}_{mh})$ and $Cov(\mathbf{x}^{*}_{mh})$ are predictive mean and covariance respectively.

5.3.2 Bayesian Uncertainty Quantification framework for Stochastic Data-Driven Hydrological Model (BUQ-SDDHM)

BUQ-SDDHM is the stochastic data-driven hydrological model. Bayesian framework has gained recent attention to quantify all types of uncertainty in the numerical models. The GPR can be used to approximate the statistical downscaling function by taking the system and parameter uncertainties into account while calibrating the model, thus representing the model stochastically (Williams, 1999). The advantage of GPR is that it provides Bayesian framework where the posterior distribution of the parameters consists of the information about the observed data. This framework is also referred to Bayesian model updating in which the prior assumption about the parameters is updated using the information in the observed data. The posterior distribution of GPR can be solved analytically as it also follows Gaussian distribution.

In contrast to other methods, the uncertainties including both the model uncertainty and the residual uncertainty in the input data, are propagated to the output. The predictive mean and the predictive variance can be obtained directly from the posterior distribution of the model. GPR was used for rainfall amount estimation in a statistical downscaling proposed by Rajendran and Cheung (2015) and their results have showed the superiority of GP models over classical downscaling techniques to generate posterior predictive distribution of river flow to quantify uncertainty in the predictions. This GP framework is similar to statistical downscaling methodology presented in Chapter 2 for precipitation amount estimation.

The prediction of the flow and the uncertainty in the predicted flow can be estimated as the output from the GPR model. Since the posterior of the parameters itself is Gaussian, the maximum likelihood-II parameter estimation techniques can be used to find the optimal estimates of the parameter. The proposed methodology for data-driven hydrological approach follows the same method of model framework, optimization and prediction used in precipitation amount estimation in SGP-SDM. Each element of model function is assumed to be a realization of stochastic process as in (5.7):

$$\mathbf{y}_h \sim GP(\boldsymbol{\mu}_h, K_h) \tag{5.7}$$

where $\boldsymbol{\mu}_h = f_h(\mathbf{x}_h, \boldsymbol{\theta}_h)^{\mathrm{T}} \mathbf{B}_h$ is the mean function, $\mathbf{B}_h$ is the mean function coefficient vector and K_h is the covariance function to represent the mean and covariance of the predictions respectively. The uncertainty of the predictions can be obtained by using the simulations of the predictive mean and covariance. The performance of the GPR is affected by the choice of the mean function and the covariance function used for model formulation (Rasmussen and Williams, 2006). For the development of BUQ-SDDHM, the linear/quadratic mean function and the squared exponential covariance function is used. The mean and the covariance functions are estimated using the input data points $\mathbf{x}_h$. The computational costs increase with the increase in the covariance function. When a flat prior is assumed, the posterior distribution of the parameters given the data is proportional to the marginal likelihood $p(\boldsymbol{\theta}_h \mid X_h, \mathbf{y}_h) \propto p(\mathbf{y}_h \mid X_h, \boldsymbol{\theta}_h)$. Based on Bayes' theorem and the Total theorem of probability, the log marginal likelihood is represented by the equation (5.8):

$$\begin{aligned} \log p(\mathbf{y}_h \mid X_h, \boldsymbol{\theta}_h) = & -\frac{1}{2}(\mathbf{y}_h - \boldsymbol{\mu}_h)^{\mathrm{T}} (K_h + \sigma_{hn}^2 \mathbf{I})^{-1} (\mathbf{y}_h - \boldsymbol{\mu}_h) \\ & -\frac{1}{2}\log | K_h + \sigma_{hn}^2 \mathbf{I} | - \frac{N}{2}\log 2\pi \end{aligned} \tag{5.8}$$

The marginal likelihood is optimized by using maximum likelihood II estimates to obtain mean and covariance function hyperparameters. Once the hyperparameters are

obtained, the predictive mean and variance can be computed using the conditional Gaussian distribution property. The predictive mean $\mathbf{M}_h^*$ in (5.10) and the predictive variance, $\mathbf{C}_h^*$ in (5.11) for the future data, X_h^* conditioned on the historic data X_h are shown in (5.9):

$$\begin{bmatrix} \mathbf{f}_h \\ \mathbf{f}_{h^*} \end{bmatrix} \sim N\left(\begin{pmatrix} \mu_h(\mathbf{x}_h) \\ \mu_h(\mathbf{x}_{h^*}) \end{pmatrix}, \begin{pmatrix} K_h & K_{h^*} \\ K_{h^*}^T & K_{h^{**}} \end{pmatrix} \right) \tag{5.9}$$

$$M(\mathbf{x}_h^*) = f_h(\mathbf{x}_h^*, \boldsymbol{\theta}_h) + K_{hx^*,x} K_{hx,x}^{-1} (\mathbf{y}_h - f_h(\mathbf{x}_h, \boldsymbol{\theta}_h)) \tag{5.10}$$

$$Cov(\mathbf{x}_h^*) = K_{hx^*,x^*} - K_{hx^*,x}^{\mathrm{T}} K_{hx,x}^{-1} K_{hx^*,x} \tag{5.11}$$

5.3.3 KNN runoff disaggregation model

KNN disaggregation model is a non-parametric model used to generate precipitation time series at a finer time scale at a site by preserving the statistics such as standard deviation, skewness, minimum and maximum value, lag-1 and lag-2 autocorrelation. The model can be easily thought of as the disaggregation model which can be easily implemented. The KNN disaggregation model can be thought of as a disaggregation precipitation from a conditional probability distribution $p(\mathbf{Q}_d \mid \mathbf{Q}_m)$ where $\mathbf{Q}_m$ is the monthly rainfall and $\mathbf{Q}_d$ is the disaggregated daily rainfall which can be added up to $\mathbf{Q}_m$ (Prairie *et al.*, 2007; Nowak *et al.*, 2010) and they have shown that KNN methodology is efficient in capturing nonlinear and non-Normal features.

Let the matrix $\mathbf{Z}_{N\times 24}$ represent the proportion of observed hourly rainfall to the observed daily rainfall where N is the total number of observed data and 24 is the number of hours. Each row of the matrix $\mathbf{Z}$ sums to unity. $\mathbf{Q}_d$ is the monthly flow which needs to be disaggregated to daily scale. The K-nearest neighbor of $\mathbf{Q}_d$ is identified from the

historical month flow vector $\mathbf{Q}_m$ by utilizing the distance between $\mathbf{Q}_d$ and each element in the $\mathbf{Q}_m$ vector.

A weight is assigned for each of the selected neighbors using the equation (5.12).

$$W_k(j) = \frac{\frac{1}{j}}{\sum_{j=1}^{K} \frac{1}{j}} \tag{5.12}$$

where K is the nearest neighbors and j refers to the index of the neighbor in which $j = 1$ represents the nearest neighbors. The weights are normalized.

Based on the weight, one of the K-nearest neighbors (one of the observed monthly data) is chosen. The proportion vector corresponding to the selected monthly observed data of the K-nearest neighbor is multiplied by the streamflow which needs to be disaggregated.

$$\mathbf{Q}_d = \mathbf{Q}_m \times z_m \tag{5.13}$$

where the disaggregated $\mathbf{Q}_d$ adds up to unity. To generate ensembles, the different K-nearest neighbors are chosen and the steps are repeated. In this study, $\sqrt{N}$ nearest neighbors are chosen. Thus, for each of the monthly ensembles of flow from BUQ-SDDHM, daily flow time series is obtained using KNN disaggregation.

5.4 Results and Discussion

5.4.1 Multi-site downscaling using MGP-SDM

The hydrological impact studies require accurate future precipitation projection along with spatial and temporal variability for the prediction of runoff in the river basin. The downscaling model is analyzed in terms of its ability to simulate the climate variables at multiple sites. In this section, the results of rainfall amount and minimum

temperature at three locations downscaled simultaneously by MGP-SDM at monthly timescale is presented. The MGP-SDM is calibrated using NCEP reanalysis predictors and observed climate variables (precipitation and minimum temperature) at the three climate stations considered in this study. The data availability at the three stations has different time period. Thus, the calibration data period is different for each station. MGP-SDM allows having different predictand length for downscaling. The calibration period for the station La Turque and the station Barrage Mattawin is from 1961 to 1975 and for the station St-Michel-des-Saints the calibration period is from 1976 to 1982. The validation period for all the three stations is from 1976 to 1982. The posterior predictive distribution of the precipitation and the minimum temperature at the three stations are downscaled simultaneously. The uncertainty is represented by generating 20 ensembles of precipitation and minimum temperature. The 20 ensembles are chosen based on the previous literature studies (Lu *et al.*, 2016); their studies have shown that the uncertainty range is large when a large number of ensembles are simulated. In this study also 20 ensembles of climate variables are enough to cover the observed data. The downscaled precipitation and minimum temperature are compared with the observed climate variables at the three stations during the validation period to assess the performance of MGP-SDM in downscaling multiple climate variables at multiple sites simultaneously. The downscaled precipitation and temperature for all the three stations for the validation period are shown in Figure 5-3, Figure 5-4 and Figure 5-5. It can also be seen that the dark line is not continuous because of the missing observed data. The results indicate that the downscaled ensembles of precipitation can well cover the observed data for most of the years at La Turque. However, there are a few months where the downscaled ensembles are slightly overestimated. The extreme data for few years are not predicted well by BUQ-SDDHM model for a couple of months in 1978 at La Turque. However, for the other two stations, the downscaled ensembles of precipitation can well cover the observed data for all the months except for few months in 1981 at Barrage Mattawin and at St-Michel-des-Saints. With respect to minimum temperature, at La Turque station, the observed minimum temperature is well covered by downscaled ensembles for most of the months in all the years. However, for the

months from January to March in the years 1976-79 and 1981-82, the downscaled minimum temperature is overestimated. At the Barrage Mattawin, the months from January to March in the year 1982 is overestimated. At St-Michel-des-Saints, for the months from January to March in the year 1976, 1978, 1980 and 1981, the temperature is overestimated; from September to December in the year 1982, the temperature prediction is underestimated.

Table 5-3 Comparison of simulated minimum, average and maximum monthly rainfall with the observed rainfall at three stations during the validation period (1976-1980)

Station	Minimum	Average	Maximum	Observed
La Turque	49.3134	75.9879	139.7532	67.8440
Barrage Mattawin	45.8102	72.7902	103.0413	63.3536
St-Michel-des-Saints	5.6314	69.0322	97.0147	75.2810

Table 5-3 compares the mean statistics for the observed rainfall with the simulated ensembles of rainfall by MGP-SDM. In Table 5-3, the average of 20 rainfall ensembles at all the three stations is overestimated. This can be due to fewer monthly training data at all the three stations. The observed rainfall lies within the minimum and maximum average rainfall. It can be seen in the table that the predicted minimum rainfall is very small at St-Michel-des-Saints compared to the other two stations. The number of minimum rainfall months with less than 30 mm/month is only one at St-Michel-des-Saints and that with less than 40 mm/month rainfall is three. Thus, a major percentage of rainfall in the simulated ensembles is above 40 mm/month. The main reason for this can be the data availability for the calibration period at St-Michel-des-Saints.

Table 5-4 compares the mean statistics for the minimum temperature with the simulated ensembles of minimum temperature by MGP-SDM. In Table 5-4, the average of 20 minimum temperature ensembles at all the three stations is closer to the observed temperature. The observed minimum temperature can be well covered by the minimum and maximum average rainfall. It can be seen in the table that the predicted range of minimum and maximum monthly minimum temperature is wide. With the availability of more data, the prediction can be improved. The downscaled results

demonstrate that the future predictions simulated from MGP-SDM is closer to the observed climate variables. The results also show that the simultaneous downscaling of the climate variables removes the need to use conditional downscaling of climate variables as in the classical MGP-SDM techniques. There is no need to consider the statistical correlations between the climate variables at different sites separately.

5.4.2 Future Climate Projections for the period 2011-2099

The validation results show that the proposed GP based multisite downscaling is efficient in downscaling the precipitation and minimum temperature at multiple sites simultaneously. For the future scenarios projection (2011-2099), the predictors from HadCM3 A2 scenarios are used. The climate variables are also downscaled for the future period using HadCM3 A2 future scenarios form 2011 to 2099. The downscaled monthly precipitation (a) and monthly minimum temperature (b) corresponding to future scenarios at the three climate stations are shown in the Figure 5-6, Figure 5-7, Figure 5-8, Figure 5-9, Figure 5-10, Figure 5-11, Figure 5-12, Figure 5-13 and Figure 5-14. The variations in the climate variables under future conditions can be assessed using the downscaled scenarios. The minimum temperature for the future scenarios shows decreasing tendencies. Overall, the results show both increasing and decreasing trends.

Table 5-4 Comparison of simulated minimum, average and maximum monthly minimum temperature with the observed minimum temperature at three stations during the validation period (1976-1980)

Station	Minimum	Average	Maximum	Observed
La Turque	-22.1146	-2.9460	14.3486	-1.8524
Barrage Mattawin	-21.3362	-3.6091	12.8555	-3.2238
St-Michel-des-Saints	-28.6187	-4.5407	14.1136	-3.3571

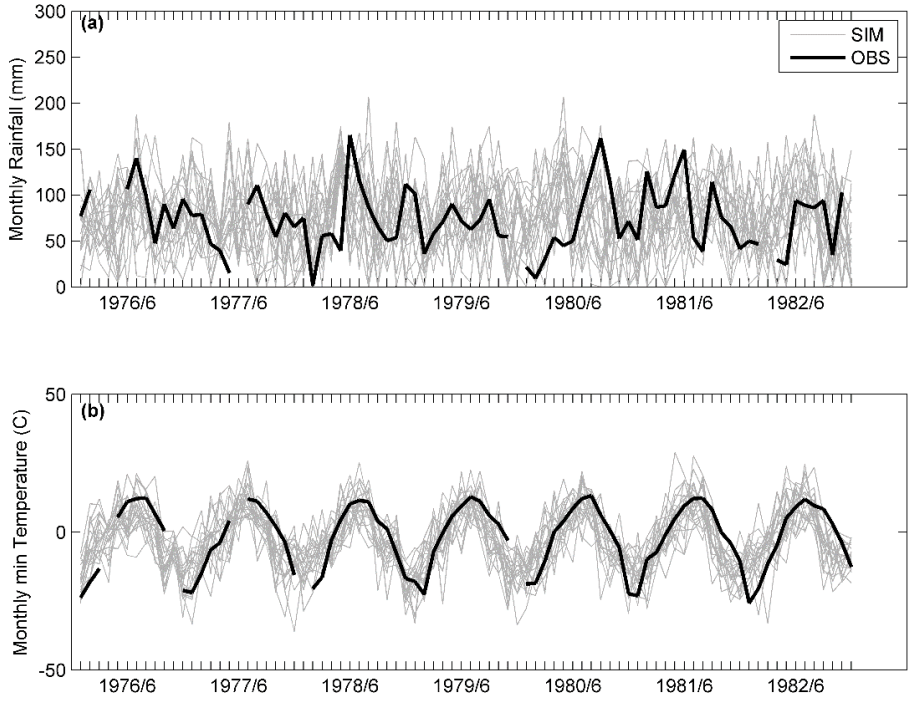


Figure 5-3 Comparison of (a) observed and downscaled precipitation ensembles and (b) observed and downscaled minimum temperature ensembles at La Turque

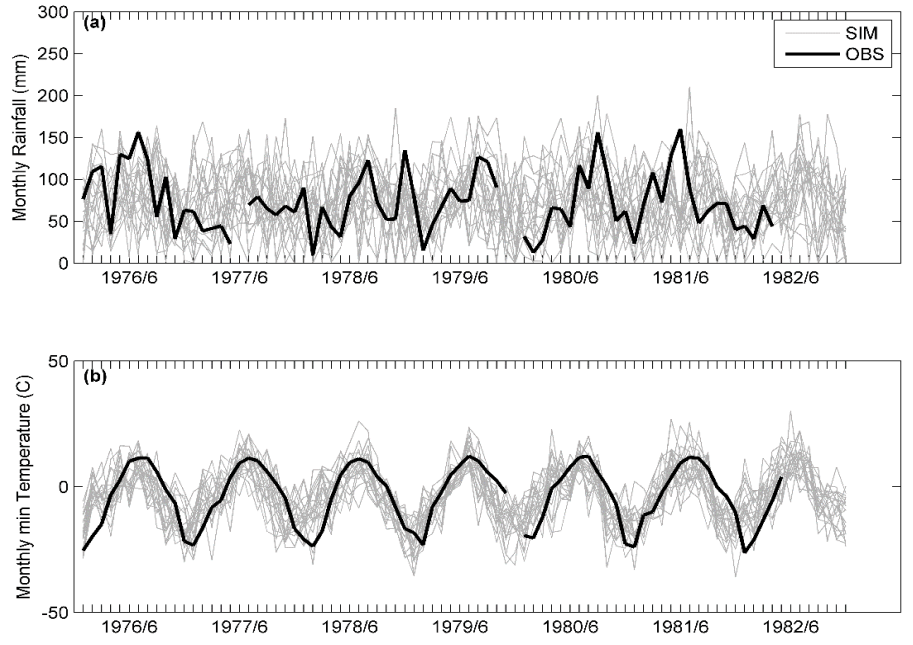


Figure 5-4 Comparison of (a) observed and downscaled precipitation ensembles and (b) observed and downscaled minimum temperature ensembles at Barrage Mattawin

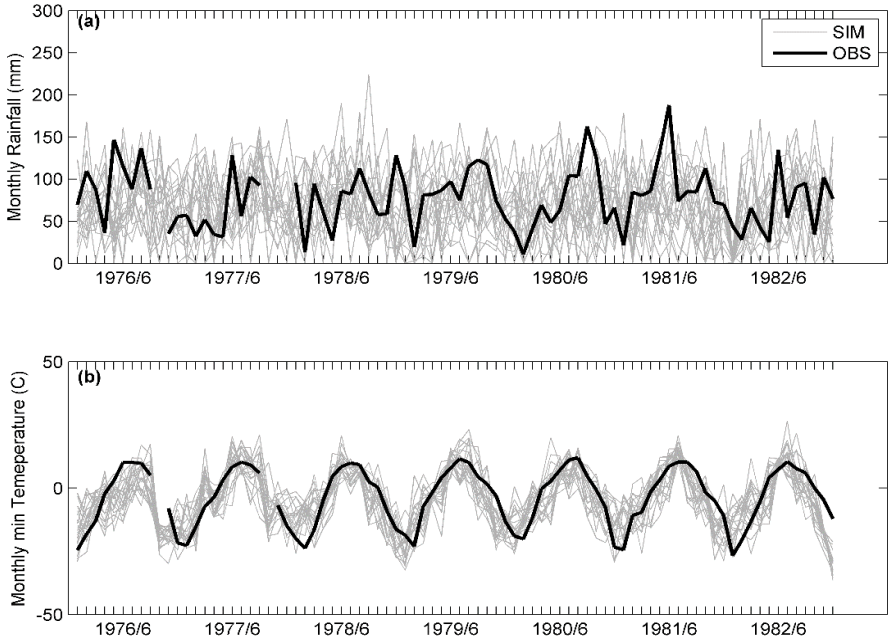


Figure 5-5 Comparison of (a) observed and downscaled precipitation ensembles and (b) observed and downscaled minimum temperature ensembles at St-Michel-des-Saints

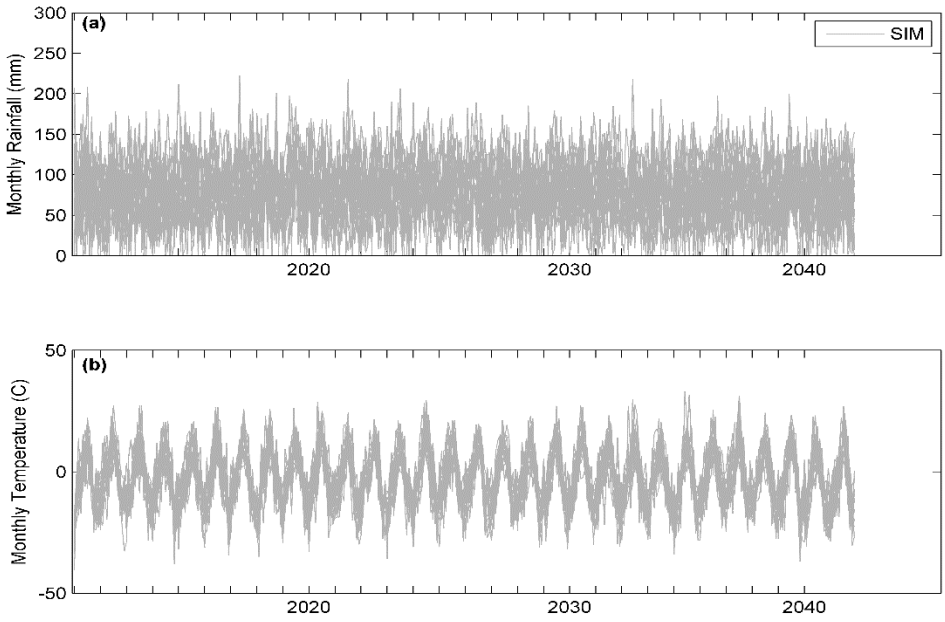


Figure 5-6 Downscaled precipitation ensembles of downscaled (a) monthly rainfall and (b) minimum temperature ensembles for future HadCM3 scenarios (2011-2040) at La Turque

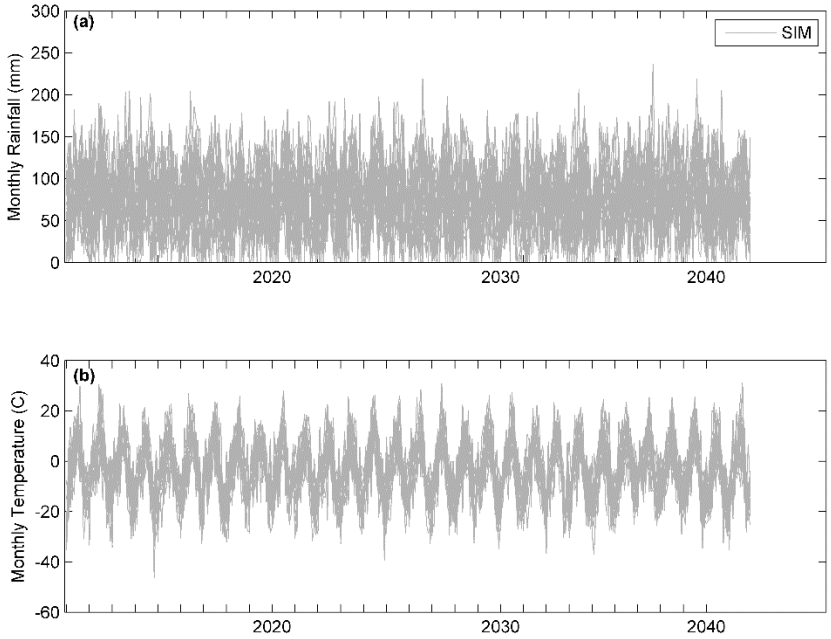


Figure 5-7 Downscaled precipitation ensembles of downscaled (a) monthly rainfall and (b) minimum temperature ensembles for future HadCM3 scenarios (2011-2040) at Barrage Mattawin

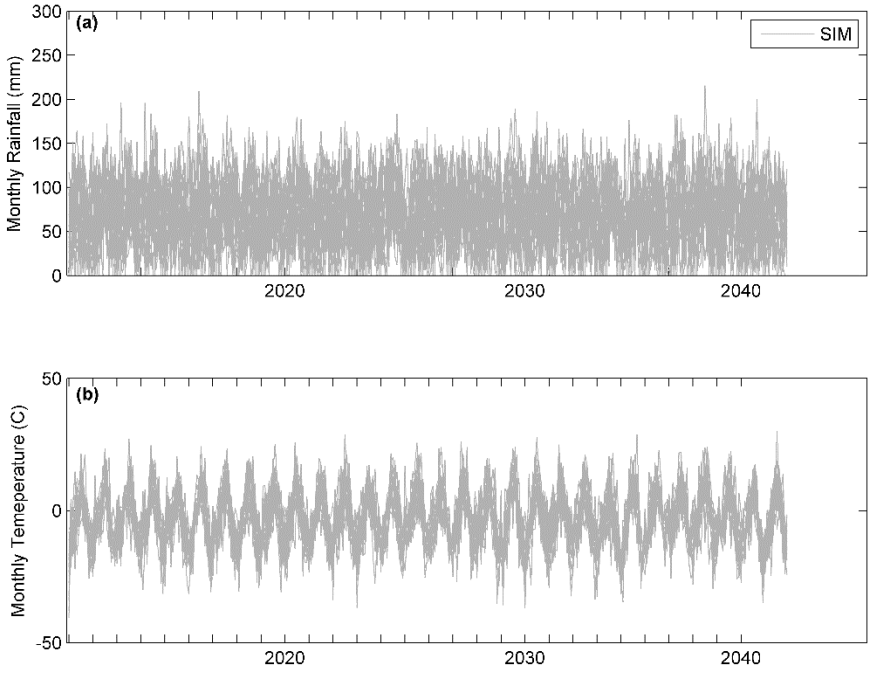


Figure 5-8 Downscaled precipitation ensembles of downscaled (a) monthly rainfall and (b) minimum temperature ensembles for future HadCM3 scenarios (2011-2040) at St-Michel-des-Saints

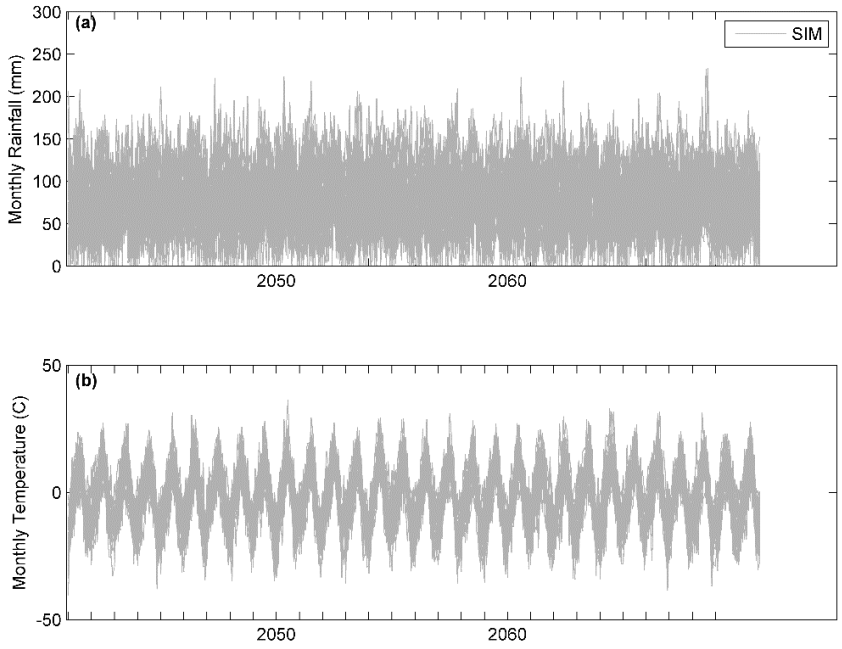


Figure 5-9 Downscaled precipitation ensembles of downscaled (a) monthly rainfall and (b) minimum temperature ensembles for future HadCM3 scenarios (2041-2070) at La Turque

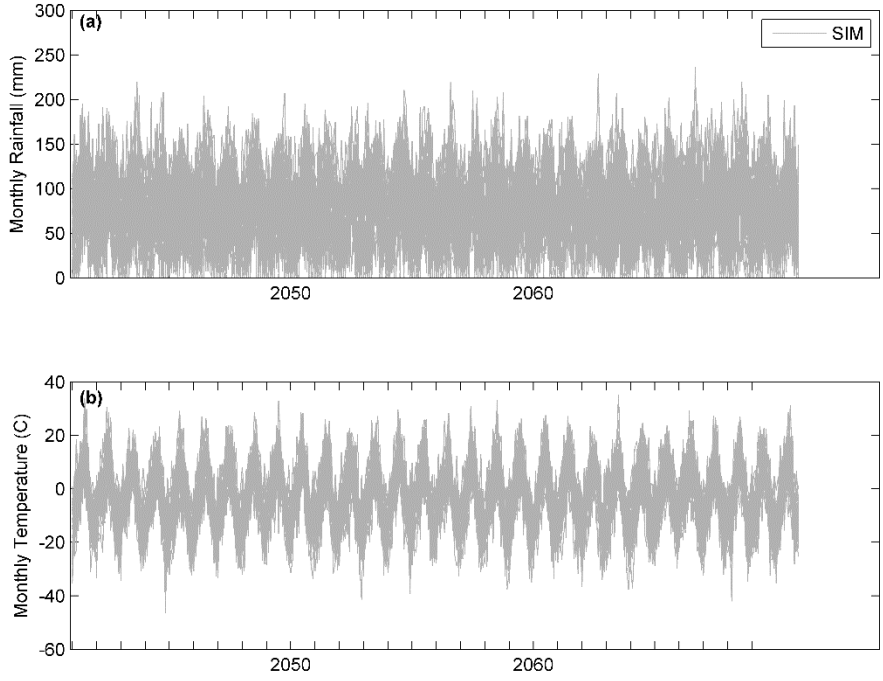


Figure 5-10 Downscaled precipitation ensembles of downscaled (a) monthly rainfall and (b) minimum temperature ensembles for future HadCM3 scenarios (2041-2070) at Barrage Mattawin

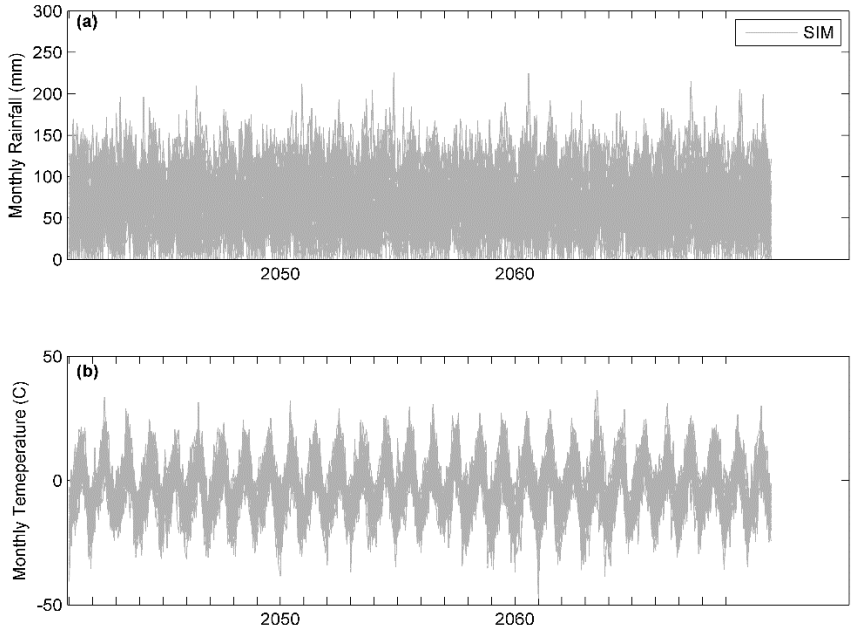


Figure 5-11 Downscaled precipitation ensembles of downscaled (a) monthly rainfall and (b) minimum temperature ensembles for future HadCM3 scenarios (2041-2070) at St-Michel-des-Saints

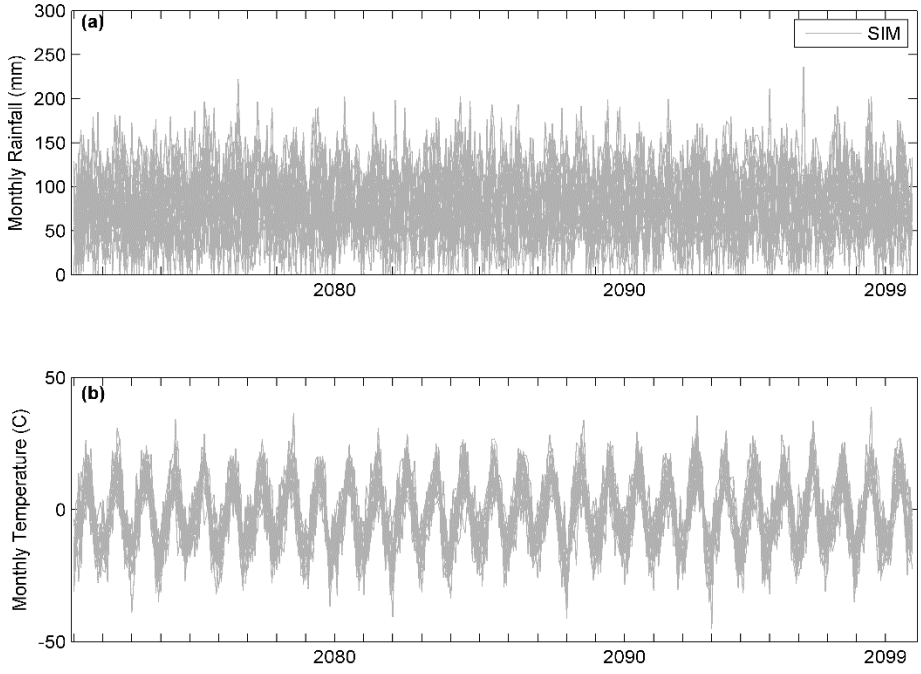


Figure 5-12 Downscaled precipitation ensembles of downscaled (a) monthly rainfall and (b) minimum temperature ensembles for future HadCM3 scenarios (2071-2099) at La Turque

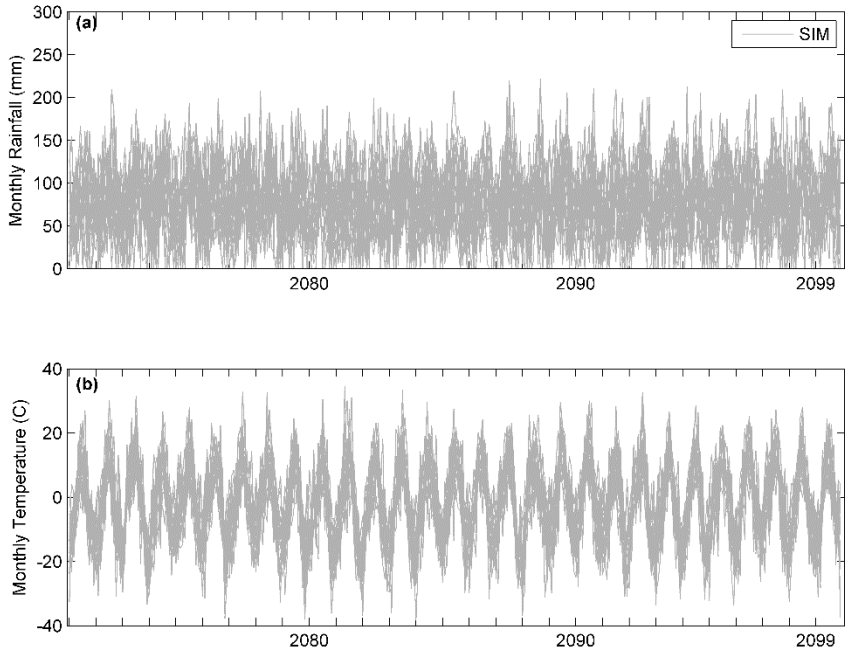


Figure 5-13 Downscaled precipitation ensembles of downscaled (a) monthly rainfall and (b) minimum temperature ensembles for future HadCM3 scenarios (2071-2099) at Barrage Mattawin

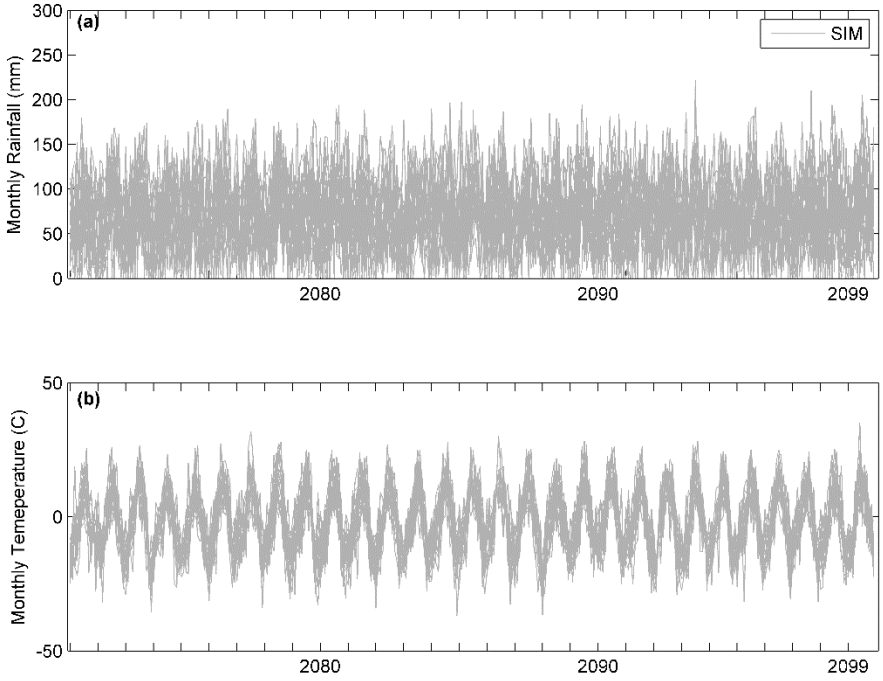


Figure 5-14 Downscaled precipitation ensembles of downscaled (a) monthly rainfall and (b) minimum temperature ensembles for future HadCM3 scenarios (2071-2099) at St-Michel-des-Saints

5.4.3 BUQ-SDDHM flow simulation

Since only two variables are downscaled, the inputs for the data driven model BUQ-SDDHM are not enough for the river flow simulation. The meteorological data that will affect the river flow of the hydrological processes are chosen as inputs for implementing the proposed hydrological models. Thus, the high resolution gridded meteorological data are needed other than the downscaled variables for the river flow prediction. In this chapter, the relative humidity from the GCM is directly used as one of the inputs for river flow prediction using BUQ-SDDHM for testing the proposed hydrological model's ability in simulating the flow. The meteorological variables such as rainfall, relative humidity, temperature and their time series at different lag times are used as inputs for data-driven hydrological models in the previous literatures (Gao *et al.*, 2010). As the usage of both the maximum temperature and minimum temperature would not have a significant impact in the prediction of the river flows, only the minimum temperature is chosen in this study. The downscaled precipitation, minimum temperature and the monthly relative humidity extracted from the NCEP reanalysis data are used for calibrating, validating and predicting the river flow for the study area. Twenty ensembles are generated from the model to represent the uncertainty in the prediction. However, from the downscaling model, there are already 20 ensembles of monthly precipitation and temperature. For hydrological simulation, 20 ensembles of runoff are simulated for each of the 20 ensembles of climate variables from downscaling. Thus, there are 400 ensembles of runoff to represent the uncertainty from downscaling model as well as the hydrological model. The monthly runoff is simulated by using the monthly downscaled climate variables from SGP-SDM. The monthly runoff is then disaggregated to daily timescale for flood frequency analysis.

The BUQ-SDDHM model is calibrated for the period from 1966 to 1975 on monthly runoff data for the study area. The observed data from 1976 to 1982 (similar to downscaling model) are used as validation data for BUQ-SDDHM. Figure 5-15 shows the comparison of the observed and mean and maximum of simulated river flows for the validation period 1976 to 1982 using the proposed BUQ-SDDHM model. The blue

lines represent the mean of the 400 ensembles of runoff and the grey lines represent the maximum values of the 400 ensembles. It can be observed from the figure that the simulated ensembles for the river flow are close to the observed river flow. The simulated ensembles can well cover the observed data including the most of the extreme values. There is a slight overestimation of river flow during the low flow months. Overall, the proposed BUQ-SDDHM simulated ensembles match the observed data reasonably well. Figure 5-16, Figure 5-17 and Figure 5-18 show the mean and the maximum of the simulated monthly runoff under HadCM3 A2 scenarios for the three future periods including 2011-2041, 2040-2071 and 2071-2099. The blue lines represent the average of runoff ensembles and the grey lines represent the maximum of the runoff ensembles. The average of the observed runoff for the verification period is 40.68 m^3/s and the simulated runoff is 45.13 m^3/s. The average runoff during the validation period from BUQ-SDDHM is overestimated. The average runoff simulated during 2011-2040 is 46.47 m^3/s, during 2041-2070 is 45.4 m^3/s and during 2071-2099 is 45.38 m^3/s. It is observed that for future conditions, HadCM3 A2 scenarios results show an increasing trend from 2011 to 2070 and from 2071 to 2099, the rainfall trend is decreasing.

Figure 5-19 shows the predicted flow by BUQ-SDDHM. Only one random sequence from the total of 400 ensembles in comparison with the observed runoff for the validation period is presented. Figure 5-19 analyzes the ability of BUQ-SDDHM to reproduce the extreme values of runoff. The threshold for extreme runoff is set as 100 m^3/s. The number of extreme runoff events in the observed data during the validation period is 9. The number of extreme runoff events simulated in the predicted flow shown in Figure 5-19 is 10. The average number of extreme events of 400 ensembles simulated by BUQ-SDDHM for the validation period is 7.7. The maximum extreme runoff value for the observed runoff is 192 m^3/s and the simulated runoff sample shown in the figure is 196.15 m^3/s. The average maximum runoff values calculated from 400 ensembles is 195.39 m^3/s. The statistics show that the simulated runoff from

the proposed BUQ-SDDHM is close to the observed runoff. Thus, the performance of BUQ-SDDHM in simulating extreme flow events is appreciable.

The distribution of the observed and downscaled monthly runoff is investigated for the validation period. The cumulative distributions of the observed and downscaled monthly runoff for all the 20 ensembles of runoff of BUQ-SDDHM generated from each of the 20 ensembles from MGP-SDM are shown in Figure 5-20. The graphical representation of the ensemble distribution from the proposed hydrological model captures the observed distribution of the runoff well for the study area. However, it can be seen that there are fewer number of distribution samples with the runoff at 50 m^3/s even though it is covered well by the ensembles. It can be seen that all the 400 ensembles capture the uncertainty from the multi-site downscaling model and the data-driven hydrological model. It is also observed that the uncertainty range is wider for the runoff above 40 m^3/s and less than 60 m^3/s. In the calibration period, only 11% of data is available between 40 m^3/s and 60 m^3/s which is the reason for wider uncertainty range.

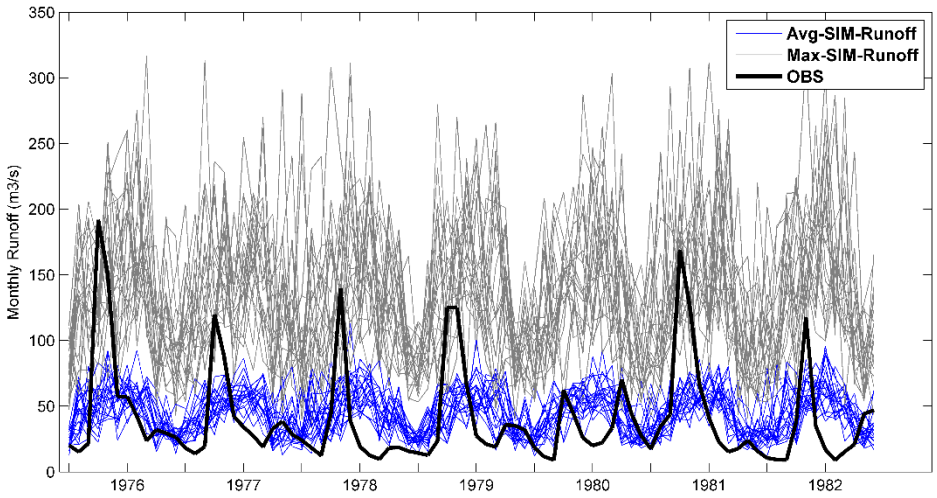


Figure 5-15 Comparison of observed and simulated monthly river flows for the validation period 1976-1982 using BUQ-SDDHM

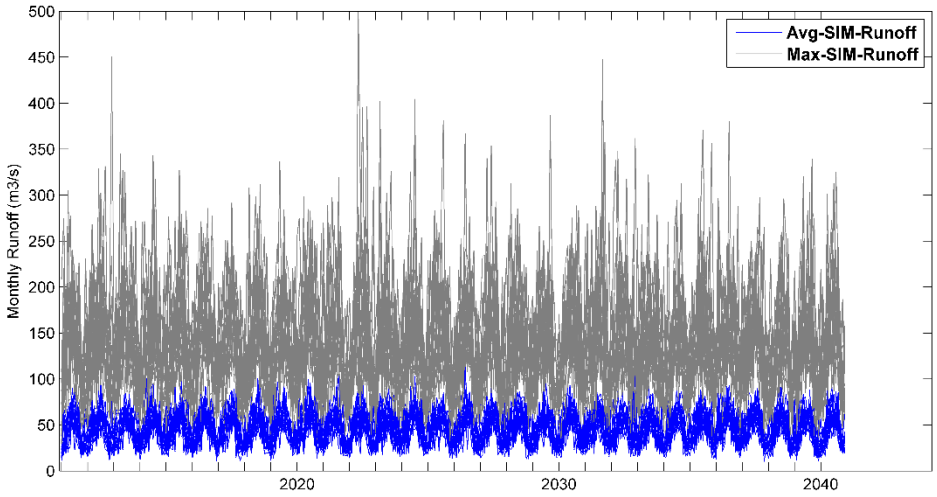


Figure 5-16 Simulated monthly river flows for future HadCM3 A2 scenarios 2011-2040 using BUQ-SDDHM

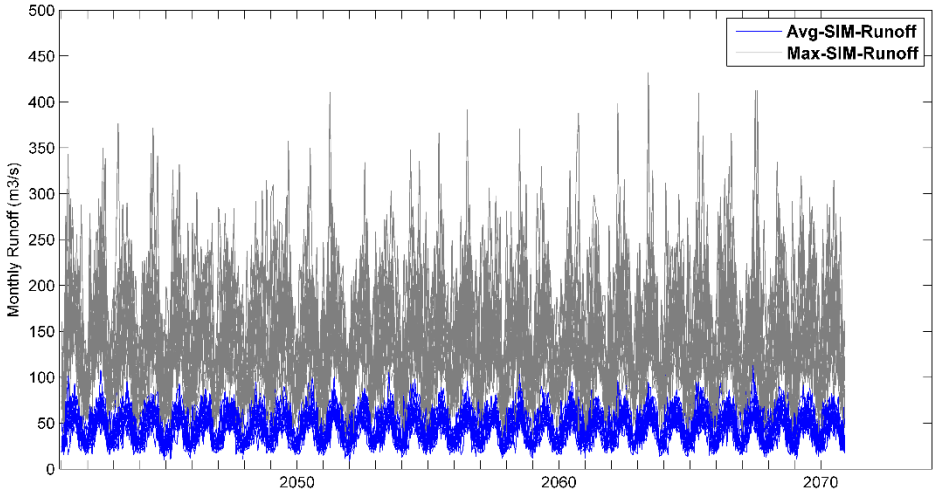


Figure 5-17 Simulated monthly river flows for future HadCM3 A2 scenarios 2041-2070 using BUQ-SDDHM

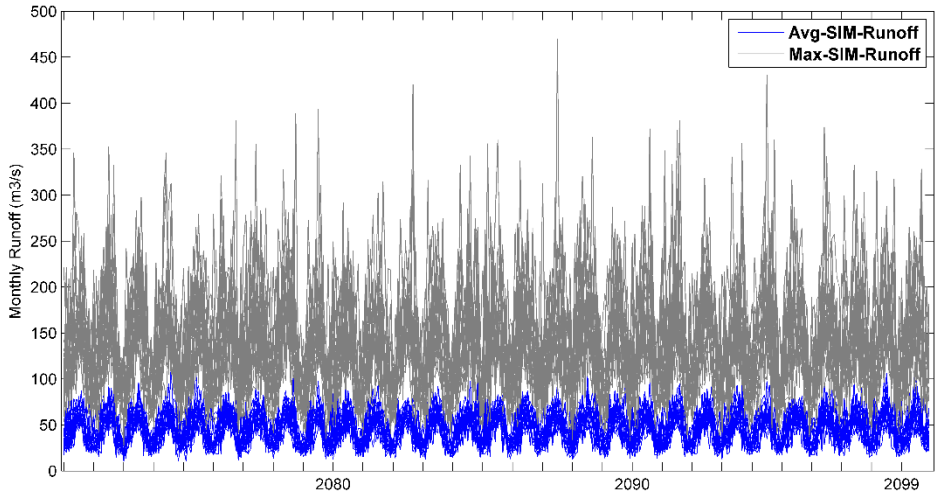


Figure 5-18 Simulated monthly river flows for future HadCM3 A2 scenarios 2071-2099 using BUQ-SDDHM

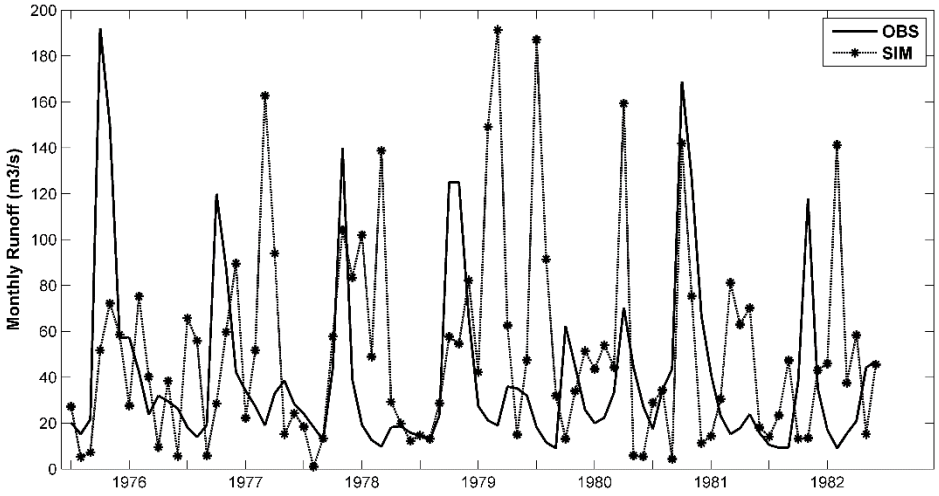


Figure 5-19 Comparison of observed runoff with one of the simulated runoffs from SGP-SDM

5.4.4 Flood frequency analysis

In order to analyze the daily runoff, the monthly runoff needs to be disaggregated to daily timescale. KNN disaggregation method is used for simulating the daily runoff from the monthly runoff. In hydrology, the normal or extreme value distribution is used

to predict the return periods of the extreme flood events in the future. It is necessary to simulate flood frequency estimates to minimize the damage costs due to flood events. In this study, the Gumbel distribution is fitted to the annual maximum of the daily runoff simulated from the disaggregation model for the validation period (1976-1982), 2011-2040, 2041-2070 and 2071-2099.

Figure 5-21 presents the box plot for the observed versus simulated flood frequencies ensembles of the validation period and future conditions. The flood frequency estimated for the validation period using observed runoff data is used as benchmark for comparison. For the validation period, Figure 5-21 shows that the return period for the flood frequencies is underestimated even though the frequencies fall within the maximum and minimum range. The return period for the future years 2011-2099 shows

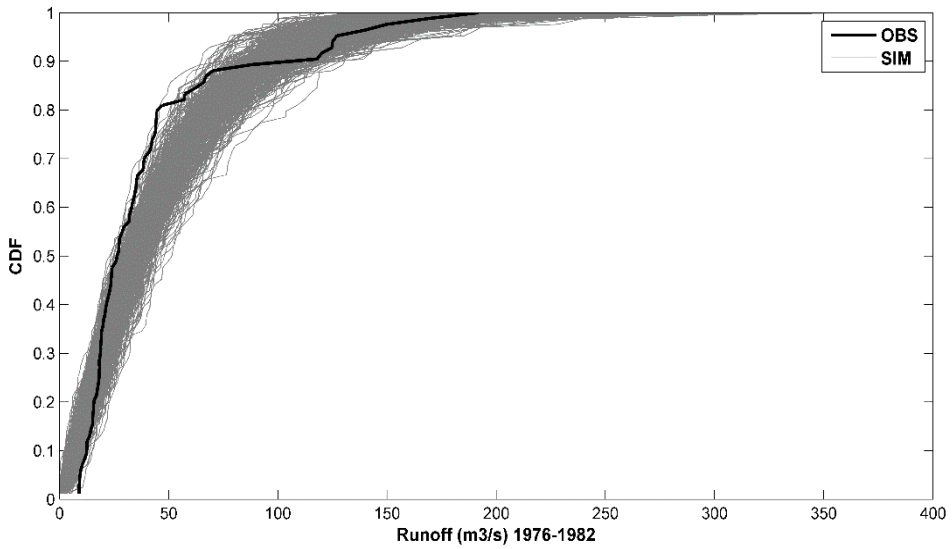


Figure 5-20 Comparison of CDF of observed runoff with simulated runoff from BUQ-SDDHM

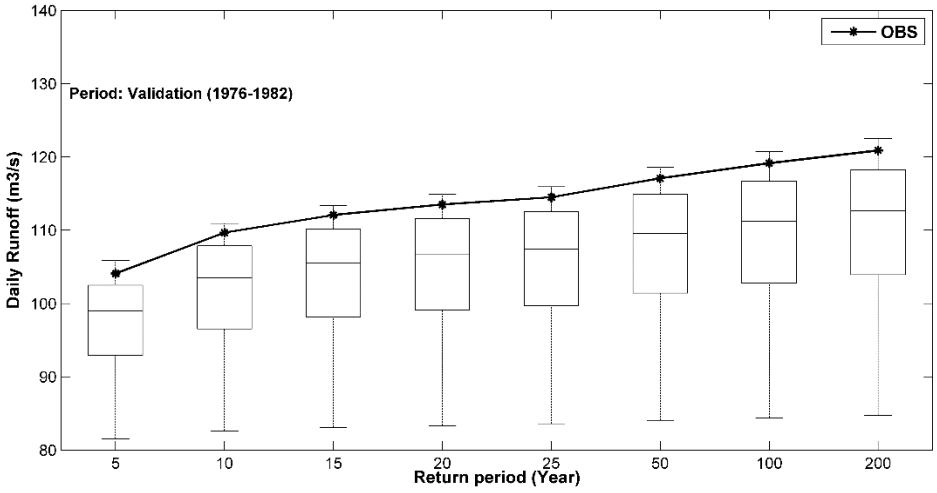


Figure 5-21 Flood frequency analysis of the flows predicted using BUQ-SDDHM for the validation period (1976-1982). The median of the results is represented as middle line of the box, the 25th and 75th percentile is presented at the top and bottom lines and the whiskers are represented as the bars at the top and the bottom

a decreasing trend in comparison with the return period for the baseline validation period. However, the presented results cannot be used directly for future planning as only one scenario from GCM is used for analysis. The prediction results from the other GCM scenarios need to be compared with the HadCM3 A2 scenarios simulation results. Figure 5-22 presents the box plot for the flood frequencies estimated of the daily runoff for the 2011-2040, 2041-2070 and 2071-2099 compared with the baseline period. The line with star represents the return period for the observed data. The median of the results is represented as the middle line of the box, the 25th and 75th percentile is presented at the top and bottom lines and the whiskers are represented as the bars at the top and the bottom. The results also show that the uncertainty increases with the increased in the return period years. However, the uncertainty range for the validation period is very high as there are only 7 years in the validation period. Less number of years is used for validation period because of the unavailability of the flow data for a longer period. The uncertainty range is less for the periods from 2011-2099 since many years are considered for prediction. The uncertainty in the validation period can be reduced when the number of years is increased. The results also show the

decrease in the return period of all the flood events in the future especially in the extreme flow amount.

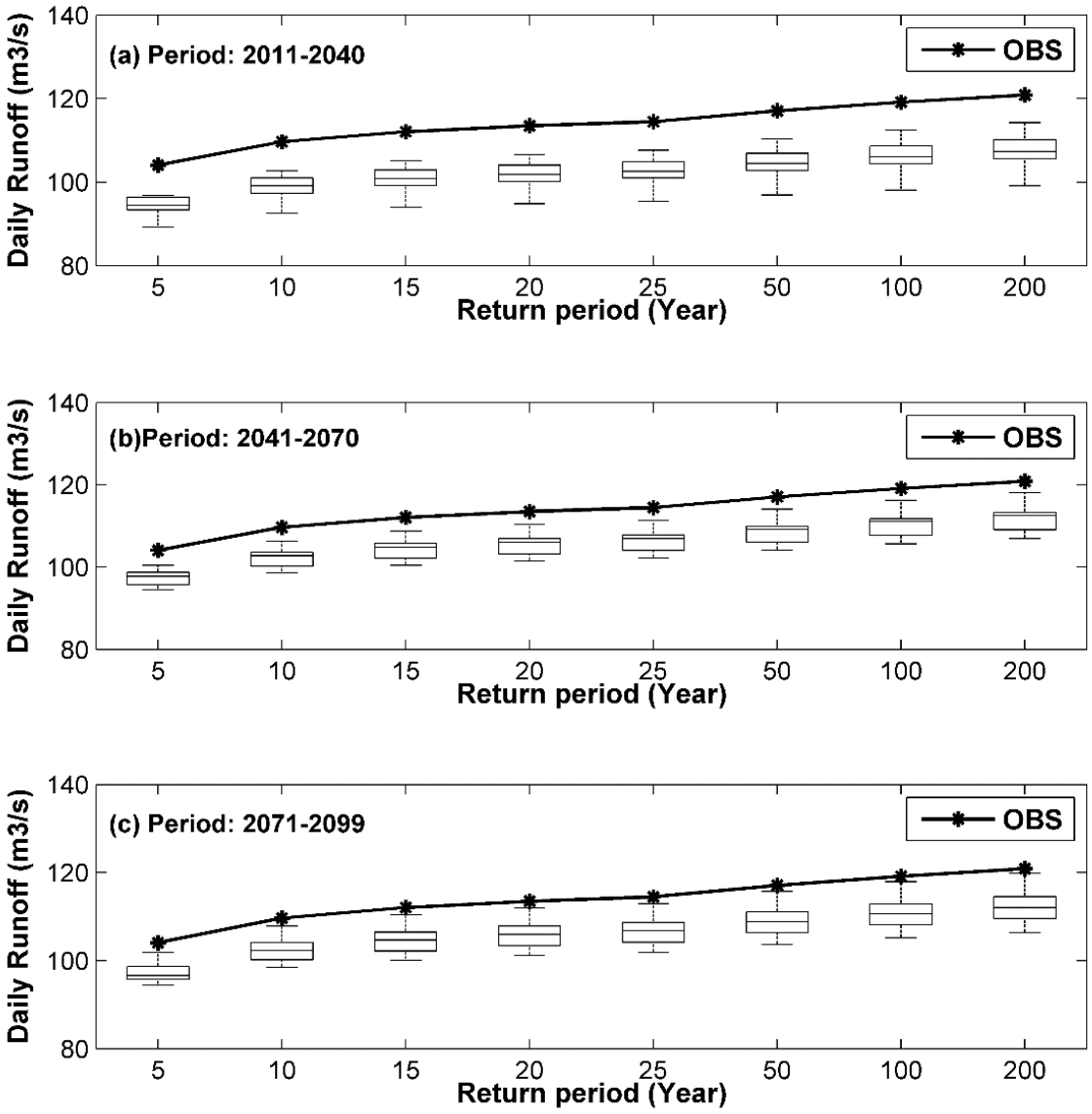


Figure 5-22 Flood frequency analysis of the flows predicted using BUQ-SDDHM for 2011-2040. The median of the results is represented as middle line of the box, the 25th and 75th percentile is presented at the top and bottom lines and the whiskers are represented as the bars at the top and the bottom

CHAPTER 6 **Conclusions and future directions**

This thesis focuses on developing the uncertainty quantification tool for single site statistical downscaling model, combined multi-site statistical downscaling model and disaggregation model, data-driven hydrological model for studying the climate change impact on streamflow. The thesis pays special attention to the coupling of the uncertainty quantification tool with the SDM and data-driven hydrological model to propagate and characterize the important types of uncertainty using Bayesian updating model framework. The conclusions of the main contributions of all the chapters in this thesis are presented as follows.

6.1 Chapter 2 conclusions

This chapter has shown that KNN-BUQSDM can be an alternate tool for statistical downscaling of precipitation and quantifying uncertainty in the model structure (epistemic uncertainty) and residuals (aleatory uncertainty) simultaneously. The proposed model can be considered as a hybrid model of weather typing and regression based SDM as KNN-BUQSDM is a combination of classification and regression. Another advantage of KNN-BUQSDM is that it can provide uncertainty information along with prediction in the form of error bars or confidence interval; this is achieved by assuming the residuals are dependent and the covariance function is used to capture dependency. This information is needed to make drainage design planning in an urban area and decision making for adaption measures.

In KNN-BUQSDM, KNN is used for determination of occurrence of dry and wet days and for stratification of rainfall types. K-means is used for finding the threshold for classifying the data into rainfall types. GPR is used to estimate the precipitation amount for the wet days in each rainfall type and to quantify uncertainty. This is achieved by coupling the uncertainty quantification tool with the statistical downscaling model using GPR which is a Bayesian statistical and machine learning technique. A Gaussian prior with a mean and a covariance function is assumed over the model function; the

errors of the model are assumed to be dependent and are modelled as a stochastic process following Gaussian distribution. The posterior predictive distribution of SDM function that relates the predictors and the model parameters to the predictand is computed using Bayes' theorem. The posterior distribution of the SDM function represents the important types of uncertainty in the downscaling process in terms of the predictive variance. The predictive distribution constitutes of the model function parameters and data noise which makes it straightforward to predict the future data that lies within and outside the training data range. The predictive mean and the variance not only depend on the optimal parameters and the future data but also the training data. The conditional distribution of the training data and the prediction data is used to capture the correlation between them in the predictive distribution to improve the accuracy of the prediction.

The effectiveness of the KNN-BUQSDM is demonstrated by downscaling daily precipitation for each month to a rain gauge station scale in Singapore using CFSR reanalysis data (0.5 X 0.5 spatial resolution) as the large scale predictors. The results show that it is possible to capture all the uncertainty in modelling such as model structure uncertainty (or epistemic) and residual uncertainty (or aleatory uncertainty). In addition to providing confidence interval through full posterior distribution of the predictive distribution, it is shown that KNN-BUQSDM yields better predictive performance compared to ASD, GLM and KNN-BNN in terms of accuracy and evaluation statistics such as mean, standard deviation, proportion of wet days, 90^{th} percentile and maximum. However in Table 8, it can be seen that the uncertainty range of KNN-BUQSDM is slightly greater than that of KNN-BNN. This is because KNN-BNN uses eight or six rainfall types for downscaling and KNN-BUQSDM uses only two rainfall classes. KNN-BUQSDM outperforms KNN-BNN in other evaluation statistics even with less number of classes. It is also shown that there is lower misclassification rate by GPR and fewer negative values in the prediction by GPR. It confirms that the assumption of GP does not simulate more negative values.

This research study does not include the predictions for the future period as the scope of this work is to demonstrate and verify the KNN-BUQSDM for statistical downscaling. The application of the proposed method to downscale GCM scenarios integrated with advanced Bayesian classification model is an ongoing research work. Further research is needed to extend the proposed methodology to downscale other climate variables such as temperature and humidity for climate change impact studies. As explained in the discussion section that some of the features such as trend, periodicity and seasonality features in the GLM can be represented implicitly in Gaussian Process models using covariance function; further research is needed to assess the performance using different covariance functions.

6.2 Chapter 3 conclusions

This chapter has shown that SGP-SDM can be an alternate tool for downscaling the precipitation statistically; this method is also efficient in quantifying the important types of uncertainty (epistemic and aleatory) in the statistical downscaling model structure. The proposed method uses GPC and GPR for rainfall occurrence determination and rainfall amount estimation respectively. SGP-SDM couples the uncertainty quantification tool within a Bayesian framework by assuming the residuals of the model are dependent and are stochastic processes following Gaussian distribution. This enables simulation of posterior probabilistic prediction ensembles directly from the downscaling model for both occurrence determination and amount estimation instead of getting point estimates thus eliminating the need to add residuals to simulate the ensembles.

The effectiveness of the GPC model in predicting wet and dry days is evaluated using two datasets including CFSR and CanESM2. The GPC model is then integrated with the GPR model to estimate the precipitation amount. First, the theoretical formulation of Bayesian GPC and GPR is provided; then the Laplace's algorithm to approximate the analytically intractable GPC model is described; the Bayesian method to learn the hyperparameters of GPC is also presented. The practical implementation issues with

the GPC model have also been described. The results in this chapter are downscaled using CFSR reanalysis data. When compared to other SDMs for precipitation, the results from SGP-SDM show better results in terms of accuracy, monthly mean square error and the monthly evaluation statistics such as the mean, the standard deviation, the proportion of wet days, the 90^{th} percentile and the maximum. The results indicate that the SGP-SDM consistently compared to other models such as ASD, GLM, KNN-BNN and KNN-BUQSDM outperforms for two rain gauge stations.

SGP-SDM is a kernel based SDM. The advantage of SGP-SDM over other kernel methods such as RVM and SVM is their ability to infer the *latent model functio*n by placing GP prior using the natural Bayesian formulation to yield the downscaled output probabilistically. The *hyperparameters* are optimized by maximizing the marginal likelihood using the gradient-based optimizers. It is also possible to determine the relevant importance of large scale predictors automatically using ARD kernel.

Despite the advantages mentioned above, training of SGP-SDM requires large computation efforts especially when huge training historic data are used. This drawback is compensated since the GP models require fewer iterations in learning model. The computational cost can also be reduced by using sparse GP models (Lawrence *et al.*, 2003). In sparse approximation technique, the number of calibration data size can be reduced; sparse approximation selects important training data that are relevant to the predictand. With the computational advancement research works, the computational time difficulties can also be reduced. Future research works are also suggested to automatically select the appropriate models among the different combination of models (for example combination of covariance function and mean functions) for the given data similar to GLM (Cheung and Beck, 2010).

This study exploits the potential of SGP-SDM only for downscaling precipitation for future scenarios. The SGP-SDM is suggested to be useful for downscaling other climate variables such as temperature and humidity mainly due to the advantages offered by Bayesian model selection using log marginal likelihood, automatic selection

of predictors using ARD kernels and estimation of confidence interval for the prediction. Hence, it is recommended that future studies can analyze the use of GP in improving the uncertainty quantification methodology for downscaling.

6.3 Chapter 4 conclusions

For multi-site downscaling of daily precipitation along with uncertainty quantification tool, this research study develops a robust stochastic statistical downscaling framework named as MGP-SDM. The MGP-SDM is a Bayesian updating framework to capture the correlated information among the precipitation at multiple sites simultaneously with the model calibration. The proposed framework consists of two stages: multi-site rainfall occurrence and multi-site rainfall amount estimation using multi-task Gaussian process classification and multi-output GPR respectively. In MGP-SDM, the spatial cross-correlation between the sites and the residual fitting is captured in the model calibration simultaneously to enable simulation of future scenario ensembles at all the sites jointly. In summary,

(i) MGP-SDM provides a principled way to quantify all important types of uncertainty using Bayesian framework in determining precipitation occurrence and estimating precipitation amount at multiple sites.

(ii) The predictive mean and the predictive variance for all the sites are obtained jointly by considering the spatial cross-correlation and the dependency between the residuals. The predictions are conditioned on the historic data.

(iii) The integrated MGP-SDM with KNN disaggregation model yields high temporal resolution (e.g., hourly) precipitation at multiple sites. The advantage is that the uncertainty in disaggregation model can also be computed by using the ensembles generated using MGP-SDM. The disaggregated ensembles can be used for analysing the impact of climate change on the water resources.

The stochastic MGP-SDM is implemented for each month separately and the results show that the mean of the simulated monthly statistics such as mean, standard deviation, proportion of wet days and max are close to the observed monthly statistics at all three selected stations. The proposed Bayesian framework for multi-site downscaling can be used for studying the impact of the climate change such as extreme events in the tropical areas. In order to make effective decision making, different GCM scenarios need to be downscaled and compared. The accuracy of the model can also be improved by utilizing more advanced covariance function and mean function respectively.

6.4 Chapter 5 conclusions

A robust stochastic uncertainty quantification tool for the integrated SDM, disaggregation model and data-driven hydrological model for studying the impact of climate change on hydrology is proposed in this study. This study also proposes a data-driven hydrological model based on Bayesian updating framework and stochastic error coupling named BUQ-SDDHM. In this framework, 1) MGP-SDM is used for downscaling monthly climate variables such as precipitation, minimum and maximum temperature and relative humidity at multiple sites simultaneously along with uncertainty information; 2) BUQ-SDDHM is used for simulating monthly river flows for the future; 3) KNN disaggregation model is used for converting monthly to daily flow time series.

The combined MGP-SDM and BUQ-SDDHM framework provides the stochastic methodology to propagate and quantify uncertainty in each stage of climate change impact studies on water resources. This methodology helps to capture the major source of uncertainty in the GCM predictors, the SDM model structure and the data-driven hydrological model structure and the random noise. The proposed methodology reduces computational complexity of using several models for studying the climate change impact on the future events and is easy to implement. However, the BUQ-SDDHM can be tested with other high resolution gridded data in future studies to

assess the proposed model performance. The prediction results can be compared with the contemporary data-driven hydrological models for assessing the model's ability in simulating the river flows.

6.5 Suggested future works

The possible future extension works of this thesis are explained in this section.

1. This study focuses on application of the model for downscaling CFSR predictors and has not attempted to predict future scenarios and there is an on-going work to predict future scenarios using GCM predictors. The research study has opened potentially useful areas to apply the state-of-the-art Bayesian framework for uncertainty quantification of the current and future work related to the downscaling techniques. As the precipitation does not follow Gaussian distribution, there is a need to take cubic transform of the predictand before using it for regression. The results can further be improved by utilizing the non-Gaussian process in the model and also there is no need to transform the data.
2. While the works in Chapters 2-4 serves as beginning step in developing uncertainty quantification tool specifically for statistical downscaling, further research is needed for the investigation of downscaling climate variables such as minimum temperature, maximum temperature and relative humidity.
3. The proposed framework for single site and multi-site statistical downscaling model in Chapters 2-4 consists of two steps including precipitation occurrence determination and precipitation amount estimation. In occurrence determination using GP model, the mean function is assumed to be zero while the precipitation amount estimation uses linear and quadratic mean functions. The accuracy of the classification model can be further improved by implementing GP classification model with a non-zero mean function. The precipitation amount can be estimated more accurately by using more complicated non-linear mean function.

4. The proposed method of SDM uses the classical predictor selection methods such as stepwise regression for precipitation amount estimation and Two-sample Kolmogorov-Smirnov test (Chapter 2, Chapter 3 and Chapter 4) for precipitation occurrence determination. This research study also uses the predictors that have strong relationship with the climate variables that are being downscaled. A SDM framework that automatically selects the relevant predictors is needed. While Automated Statistical Downscaling (ASD) incorporated this idea, the predictors were chosen based on the statistical tests. In most of the cases, the statistical test based results are misleading since there is no principle way to select the predictors. Thus, another area of research based on Bayesian updating framework to choose the relevant predictors automatically needs to be developed for SGP-SDM, MGP-SDM and BUQ-SDDHM.
5. The efficiency of the proposed approximation with large size dataset needs to be assessed in future works. The SGP-SDM and MGP-SDM are complicated and computationally intensive models as the large size of covariance matrix is used in the model. Thus, a sparse approximation technique should be adopted if the large dataset is used for downscaling.
6. Even though there are many recent studies on data-driven hydrological models, the hydrological researchers are sceptical about the efficiency of the data-driven models in simulating the real world scenarios compared to the physical hydrological models. In Chapter 5, BUQ-SDDHM is proposed to predict river flow along with uncertainty quantification. There needs to be a study which compares the BUQ-SDDHM with the physics-based hydrological models.
7. All the statistical downscaling models are developed based on the assumption that the present and the future scenarios are stationary. A statistical downscaling model that can capture non-stationary relationship between the present and the future climate conditions can be developed.
8. It is also important to develop a statistical downscaling model that automatically incorporates the model selection within the framework.

REFERENCES

Abbott MB. 1991. Hydroinformatics: information technology and the aquatic environment. Avebury Technical.

Abrahart RJ, See L. 2000. Comparing neural network and autoregressive moving average techniques for the provision of continuous river flow forecasts in two contrasting catchments. Hydrological processes, **14**: 2157-2172.

Abrahart RJ, See LM. 2007. Neural network modelling of non-linear hydrological relationships. Hydrol. Earth Syst. Sci., **11**: 1563-1579. DOI: 10.5194/hess-11-1563-2007.

Alaya B, Ali M, Chebana F, Ouarda TB. 2014. Probabilistic Gaussian Copula Regression Model for Multisite and Multivariable Downscaling. Journal of Climate, **27**: 3331-3347.

Alaya MAB, Chebana F, Ouarda TBMJ. 2015. Probabilistic Multisite Statistical Downscaling for Daily Precipitation Using a Bernoulli–Generalized Pareto Multivariate Autoregressive Model. Journal of Climate, **28**: 2349-2364. DOI: 10.1175/jcli-d-14-00237.1.

Alaya MB, Chebana F, Ouarda T. 2015. Multisite and multivariable statistical downscaling using a Gaussian copula quantile regression model. Climate Dynamics: 1-15.

Altman NS. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician, **46**: 175-185.

Anzai Y. 2012. Pattern recognition and machine learning. Elsevier.

Armstrong JS, Collopy F. 1992. Error measures for generalizing about forecasting methods: Empirical comparisons. International journal of forecasting, **8**: 69-80.

Arnbjerg-Nielsen K, Willems P, Olsson J, Beecham S, Pathirana A, Gregersen IB, Madsen H, Nguyen V-T-V. 2013. Impacts of climate change on rainfall extremes and urban drainage systems: a review. Water Science and Technology, **68**: 16-28.

Bali N, Gupta P, Gandhi C. 2007. A Textbook of Quantitative Techniques. Firewall Media.

Barber D. 2012. Bayesian reasoning and machine learning. Cambridge University Press.

Bates BC, Campbell EP. 2001. A Markov Chain Monte Carlo Scheme for parameter estimation and inference in conceptual rainfall-runoff modeling. Water Resources Research, **37**: 937-947. DOI: 10.1029/2000WR900363.

Beck JL, Cheung SH. 2009. Probability logic, model uncertainty and robust predictive system analysis. In: Proceedings of the 10th International Conference on Structural Safety and Reliability, CRC Press.

Beck JL, Katafygiotis LS. 1991. Updating of a Model and its Uncertainties Utilizing Dynamic Test Data. In: Computational Stochastic Mechanics, Spanos PD, Brebbia CA (eds.) Springer Netherlands, pp: 125-136.

Beck JL, Katafygiotis LS. 1998. Updating Models and Their Uncertainties. I: Bayesian Statistical Framework. Journal of Engineering Mechanics, **124**: 455-461. DOI: doi:10.1061/(ASCE)0733-9399(1998)124:4(455).

Benestad RE, Hanssen-Bauer I, Chen D. 2008. Empirical-statistical downscaling. New Jersey : World Scientific Pub Co Inc., c2008.

Berger JO. 1985. Statistical Decision Theory and Bayesian Analysis. [electronic resource]. New York, NY : Springer New York : Imprint: Springer, 1985. Second Edition.

Bergström S, Carlsson B, Gardelin M, Lindström G, Pettersson A, Rummukainen M. 2001. Climate change impacts on runoff in Sweden assessments by global climate models, dynamical downscaling and hydrological modelling. Climate research, **16**: 101-112.

Bernardo J, Berger J, Dawid A, Smith A. 1998. Regression and classification using Gaussian process priors. Bayesian statistics, **6**: 475.

Beven K, Binley A. 1992. The future of distributed models: Model calibration and uncertainty prediction. Hydrological Processes, **6**: 279-298. DOI: 10.1002/hyp.3360060305.

Biegler L, Biros G, Ghattas O, Heinkenschloss M, Keyes D, Mallick B, Tenorio L, van Bloemen Waanders B, Willcox K, Marzouk Y. 2011. Large-scale inverse problems and quantification of uncertainty. John Wiley & Sons.

Bilionis I, Zabaras N. 2012. Multi-output local Gaussian process regression: Applications to uncertainty quantification. J. Comput. Phys., **231**: 5718-5746. DOI: 10.1016/j.jcp.2012.04.047.

Bishop CM. 2006. Pattern recognition and machine learning. New York : Springer.

Bonilla E, Chai K, Williams C. 2008. Multi-task {G}aussian Process Prediction. In: Advances in Neural Information Processing Systems 20, Platt JC, Koller D, Singer Y, Roweis S (eds.) MIT Press.

Booij M. 2005. Impact of climate change on river flooding assessed with different spatial model resolutions. Journal of hydrology, **303**: 176-198.

Boyle P, Frean M. 2004. Dependent Gaussian Processes. In: Advances in Neural Information Processing Systems 17, Saul LK, Weiss Y, Bottou L (eds.).

Bürger G. 1996. Expanded downscaling for generating local weather scenarios. Climate Research: 111-128.

Buser CM, Künsch H, Lüthi D, Wild M, Schär C. 2009. Bayesian multi-model projection of climate: bias assumptions and interannual variability. Climate Dynamics, **33**: 849-868.

Cavazos T, Hewitson BC. 2005. Performance of NCEP–NCAR reanalysis variables in statistical downscaling of daily precipitation. Climate Research, **28**: 95-107.

Celisse A, Mary-Huard T. 2012. Exact Cross-Validation for kNN: application to passive and active learning in classification. Journal de la Société Française de Statistique, **152**: 83-97.

Challis E, Hurley P, Serra L, Bozzali M, Oliver S, Cercignani M. 2015. Gaussian process classification of Alzheimer's disease and mild cognitive impairment from resting-state fMRI. NeuroImage, **112**: 232-243. DOI: http://dx.doi.org/10.1016/j.neuroimage.2015.02.037.

Chandler RE, Wheater HS. 2002. Analysis of rainfall variability using generalized linear models: A case study from the west of Ireland. Water Resources Research, **38**: 10-11-10-11. DOI: 10.1029/2001WR000906.

Chen H, Xu C-Y, Guo S. 2012. Comparison and evaluation of multiple GCMs, statistical downscaling and hydrological models in the study of climate change impacts on runoff. Journal of Hydrology, **434–435**: 36-45. DOI: http://dx.doi.org/10.1016/j.jhydrol.2012.02.040.

Chen J, Brissette FP, Leconte R. 2011. Uncertainty of downscaling method in quantifying the impact of climate change on hydrology. Journal of Hydrology, **401**: 190-202.

Chen J, Brissette FP, Poulin A, Leconte R. 2011. Overall uncertainty study of the hydrological impacts of climate change for a Canadian watershed. Water Resources Research, **47(12)**: 1-16.

Chen S-T, Yu P-S, Tang Y-H. 2010. Statistical downscaling of daily precipitation using support vector machines and multivariate analysis. Journal of Hydrology, **385**: 13-22.

Cheung S, Beck J. 2008. Near real-time loss estimation of structures subjected to strong seismic excitation. In: Inaugural International Conference of the Engineering Mechanics Institute (EM08), University of Minnesota, Minneapolis, Minnesota, USA.

Cheung S, Beck J. 2008. On using posterior samples for model selection for structural identification. In: Proceedings of the Asian-Pacific Symposium on Structural Reliability and its Applications, Hong Kong University of Science and Technology, pp: 125-130.

Cheung S, Beck J. 2008. Updating reliability of monitored nonlinear structural dynamic systems using real-time data. In: Proceedings of the Inaugural International Conference of Engineering Mechanics Institute.

Cheung S, Beck J. 2009. New Bayesian updating methodology for model validation and robust predictions of a target system based on hierarchical subsystem tests. In: Computer Methods in Applied Mechanics and Engineering, Doctoral Dissertation California Institute of Technology.

Cheung SH, Bansal S. 2013. A new gibbs-sampling based algorithm for Bayesian model updating of linear dynamic systems with incomplete complex modal data. In: Proceedings of the international multi conference of engineers and computer scientists.

Cheung SH, Beck JL. 2007. Algorithms for Bayesian model class selection of higher-dimensional dynamic systems. In: ASME 2007 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers, pp: 1549-1558.

Cheung SH, Beck JL. 2007. New stochastic simulation method for updating robust reliability of dynamic systems. In: Proceedings of the 8th Engineering Mechanics Division Conference.

Cheung SH, Beck JL. 2009. Bayesian Model Updating Using Hybrid Monte Carlo Simulation with Application to Structural Dynamic Models with Many Uncertain Parameters. Journal of Engineering Mechanics, **135**: 243-255. DOI: doi:10.1061/(ASCE)0733-9399(2009)135:4(243).

Cheung SH, Beck JL. 2009. Comparison of different model classes for Bayesian updating and robust predictions using stochastic state-space system models. In:

Proceedings of the 10th International Conference on Structural Safety and Reliability, CRC Press, pp: 474-474.

Cheung SH, Beck JL. 2010. Calculation of posterior probabilities for Bayesian model class assessment and averaging from posterior samples based on dynamic system data. Computer-Aided Civil and Infrastructure Engineering, **25**: 304-321.

Cheung SH, Beck JL. 2010. Calculation of Posterior Probabilities for Bayesian Model Class Assessment and Averaging from Posterior Samples Based on Dynamic System Data. Computer-Aided Civil and Infrastructure Engineering, **25**: 304-321. DOI: 10.1111/j.1467-8667.2009.00642.x.

Cheung SH, Beck, J.L. 2007. Bayesian Model Class Selection of Higher-Dimensional Dynamic Systems Using Posterior Samples. In: 18th Engineering Mechanics Division Conference for the American Society of Civil Engineers (EMD2007).

Cheung SH, Oliver TA, Prudencio EE, Prudhomme S, Moser RD. 2011. Bayesian uncertainty analysis with applications to turbulence modeling. Reliability Engineering & System Safety, **96**: 1137-1149. DOI: http://dx.doi.org/10.1016/j.ress.2010.09.013.

Chowdhury S, Sharma A. 2007. Mitigating parameter bias in hydrological modelling due to uncertainty in covariates. Journal of Hydrology, **340**: 197-204.

Chylek P, Li J, Dubey M, Wang M, Lesins G. 2011. Observed and model simulated 20th century Arctic temperature variability: Canadian earth system model CanESM2. Atmospheric Chemistry and Physics Discussions, **11**: 22893-22907.

Clark MP, Wilby RL, Gutmann ED, Vano JA, Gangopadhyay S, Wood AW, Fowler HJ, Prudhomme C, Arnold JR, Brekke LD. 2016. Characterizing Uncertainty of the Hydrologic Impacts of Climate Change. Current Climate Change Reports, **2**: 55-64. DOI: 10.1007/s40641-016-0034-x.

Coe R, Stern RD. 1982. Fitting Models to Daily Rainfall Data. Journal of Applied Meteorology, **21**: 1024-1031. DOI: 10.1175/1520-0450(1982)021<1024:fmtdrd>2.0.co;2.

Collins M, Tett BSF, Cooper C. 2001. The internal climate variability of HadCM3, a version of the Hadley Centre coupled model without flux adjustments. Climate Dynamics, **17**: 61-81. DOI: 10.1007/s003820000094.

Das K, Srivastava AN. 2011. Sparse Inverse Gaussian Process Regression with Application to Climate Network Discovery. In: CIDU, pp: 233-247.

Davies DL, Bouldin DW. 1979. A cluster separation measure. IEEE transactions on pattern analysis and machine intelligence: 224-227.

De Toffol S, Laghari AN, Rauch W. 2009. Are extreme rainfall intensities more frequent? Analysis of trends in rainfall patterns relevant to urban drainage systems. Water Science and Technology, **59**: 1769-1776. DOI: 10.2166/wst.2009.182.

Dibike BY, Gachon P, St-Hilaire A, Ouarda JTBM, Nguyen T-VV. 2008. Uncertainty analysis of statistically downscaled temperature and precipitation regimes in Northern Canada. Theoretical and Applied Climatology, **91**: 149-170. DOI: 10.1007/s00704-007-0299-z.

Dile YT, Srinivasan R. 2014. Evaluation of CFSR climate data for hydrologic prediction in data-scarce watersheds: an application in the Blue Nile River Basin. JAWRA Journal of the American Water Resources Association, **50**: 1226-1241.

Dixon KW, Lanzante JR, Nath MJ, Hayhoe K, Stoner A, Radhakrishnan A, Balaji V, Gaitán CF. 2016. Evaluating the stationarity assumption in statistically downscaled climate projections: is past performance an indicator of future results? Climatic Change, **135**: 395-408.

Dobler C, Hagemann S, Wilby RL, Stötter J. 2012. Quantifying different sources of uncertainty in hydrological projections in an Alpine watershed. Hydrol. Earth Syst. Sci., **16**: 4343-4360. DOI: 10.5194/hess-16-4343-2012.

Dobson A. 2001. An introduction to generalized linear models. CRC press.

Dunn JC. 1973. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. 32-57.

Engeland K, Xu C-Y, Gottschalk L. 2005. Assessing uncertainties in a conceptual water balance model using Bayesian methodology/Estimation bayésienne des incertitudes au sein d'une modélisation conceptuelle de bilan hydrologique. Hydrological Sciences Journal, **50**: 45-63.

Environmnet and Heritage E. 2016. Certainty and Uncertainty In: Climate projections for NSW, NSW Australia.

Foley AM. 2010. Uncertainty in regional climate modelling: A review. Progress in Physical Geography, **34**: 647-670. DOI: 10.1177/0309133310375654.

Fowler H, Ekström M. 2009. Multi-model ensemble estimates of climate change impacts on UK seasonal precipitation extremes. International Journal of Climatology, **29**: 385-416.

Fowler HJ, Blenkinsop S, Tebaldi C. 2007. Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling. International Journal of Climatology, **27**: 1547-1578. DOI: 10.1002/joc.1556.

Freer J, Beven K, Ambroise B. 1996. Bayesian Estimation of Uncertainty in Runoff Prediction and the Value of Data: An Application of the GLUE Approach. Water Resources Research, **32**: 2161-2173. DOI: 10.1029/95WR03723.

Frost A. 2007. Australian application of a statistical downscaling technique for multi-site daily rainfall: GLIMCLIM. In: MODSIM 2007: International Congress on Modelling and Simulation, Modelling and Simulation Society of Australia and New Zealand, pp: 553-559.

Frost AJ, Charles SP, Timbal B, Chiew FH, Mehrotra R, Nguyen KC, Chandler RE, McGregor JL, Fu G, Kirono DG. 2011. A comparison of multi-site daily rainfall downscaling techniques under Australian conditions. Journal of Hydrology, **408**: 1-18.

Fung CF, Lopez A, New M. 2011. Modelling the impact of climate change on water resources. John Wiley & Sons.

Gagnon S, Singh B, Rousselle J, Roy L. 2005. An Application of the Statistical DownScaling Model (SDSM) to Simulate Climatic Data for Streamflow Modelling in Québec. Canadian Water Resources Journal / Revue canadienne des ressources hydriques, **30**: 297-314. DOI: 10.4296/cwrj3004297.

Gal Y, Ghahramani Z. 2015. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. arXiv preprint arXiv:1506.02142, **2**.

Gao C, Gemmer M, Zeng X, Liu B, Su B, Wen Y. 2010. Projected streamflow in the Huaihe River Basin (2010–2100) using artificial neural network. Stochastic Environmental Research and Risk Assessment, **24**: 685-697.

Ghosh S, Katkar S. 2012. Modeling Uncertainty Resulting from Multiple Downscaling Methods in Assessing Hydrological Impacts of Climate Change. Water Resources Management, **26**: 3559-3579. DOI: 10.1007/s11269-012-0090-5.

Ghosh S, Mujumdar P. 2009. Climate change impact assessment: Uncertainty modeling with imprecise probability. Journal of Geophysical Research: Atmospheres, **114**: 1-17.

Ghosh S, Mujumdar PP. 2008. Statistical downscaling of GCM simulations to streamflow using relevance vector machine. Advances in Water Resources, **31**: 132-146. DOI: http://dx.doi.org/10.1016/j.advwatres.2007.07.005.

Gibbs MN. 1998. Bayesian Gaussian processes for regression and classification. Doctoral dissertation University of Cambridge.

Gibson N, Aigrain S, Roberts S, Evans T, Osborne M, Pont F. 2012. A Gaussian process framework for modelling instrumental systematics: application to transmission spectroscopy. Monthly Notices of the Royal Astronomical Society, **419**: 2683-2694.

Giorgi F, Mearns LO. 1991. Approaches to the simulation of regional climate change: A review. Reviews of Geophysics, **29**: 191-216. DOI: 10.1029/90RG02636.

Giorgi F, Mearns LO. 2002. Calculation of Average, Uncertainty Range, and Reliability of Regional Climate Changes from AOGCM Simulations via the "Reliability Ensemble Averaging" (REA) Method. Journal of Climate, **15**: 1141-1158. DOI: 10.1175/1520-0442(2002)015<1141:coaura>2.0.co;2.

Girolami M, Rogers S. 2006. Variational Bayesian multinomial probit regression with Gaussian process priors. Neural Computation, **18**: 1790-1817.

Goodess C. 2003. Statistical and regional dynamical downscaling of extremes for European regions: STARDEX. European geophysical union information newsletter, **6**.

Goodess C, Anagnostopoulou C, Bárdossy A, Frei C, Harpham C, Haylock M, Hundecha Y, Maheras P, Ribalaygua J, Schmidli J. 2007. An intercomparison of statistical downscaling methods for Europe and European regions–assessing their performance with respect to extreme temperature and precipitation events. Climatic Change.

Govindaraju RS, Rao AR. 2000. Artificial Neural Networks in Hydrology. [electronic resource]. Dordrecht : Springer Netherlands : Imprint: Springer.

Graham LP, Andersson L, Horan M, Kunz R, Lumsden T, Schulze R, Warburton M, Wilk J, Yang W. 2011. Using multiple climate projections for assessing hydrological response to climate change in the Thukela River Basin, South Africa. Physics and Chemistry of the Earth, Parts A/B/C, **36**: 727-735. DOI: http://dx.doi.org/10.1016/j.pce.2011.07.084.

Graham LP, Andréasson J, Carlsson B. 2007. Assessing climate change impacts on hydrology from an ensemble of regional climate models, model scales and linking methods – a case study on the Lule River basin. Climatic Change, **81**: 293-307. DOI: 10.1007/s10584-006-9215-2.

Grouillet B, Ruelland D, Vaittinada Ayar P, Vrac M. 2016. Sensitivity analysis of runoff modeling to statistical downscaling models in the western

Mediterranean. Hydrol. Earth Syst. Sci., **20**: 1031-1047. DOI: 10.5194/hess-20-1031-2016.

Grum M, Jørgensen AT, Johansen RM, Linde JJ. 2006. The effect of climate change on urban drainage: an evaluation based on regional climate model simulations. Water Science and Technology, **54**: 9-15. DOI: 10.2166/wst.2006.592.

Gurley KR. 1997. Modelling and simulation of non-Gaussian processes. Doctoral dissertation University of Notre Dame.

Hawkins E, Sutton R. 2009. The Potential to Narrow Uncertainty in Regional Climate Predictions. Bulletin of the American Meteorological Society, **90**: 1095-1107. DOI: 10.1175/2009BAMS2607.1.

Haylock MR, Cawley GC, Harpham C, Wilby RL, Goodess CM. 2006. Downscaling heavy precipitation over the United Kingdom: a comparison of dynamical and statistical methods and their future scenarios. International Journal of Climatology, **26**: 1397-1415.

He QP, Wang J. 2007. Fault detection using the k-nearest neighbor rule for semiconductor manufacturing processes. IEEE Transactions on Semiconductor Manufacturing, **20**: 345-354.

Hermkes M, Kuehn NM, Riggelsen C. 2014. Simultaneous quantification of epistemic and aleatory uncertainty in GMPEs using Gaussian process regression. Bulletin of Earthquake Engineering, **12**: 449-466. DOI: 10.1007/s10518-013-9507-7.

Hessami M, Gachon P, Ouarda TBMJ, St-Hilaire A. 2008. Automated regression-based statistical downscaling tool. Environmental Modelling & Software, **23**: 813-834. DOI: http://dx.doi.org/10.1016/j.envsoft.2007.10.004.

Hewitson BC, Crane RG. 1996. Climate downscaling: techniques and application. Climate Research, **07**: 85-95.

Higdon D. 2002. Space and space-time modeling using process convolutions. In: Quantitative methods for current environmental issues, Springer, pp: 37-56.

Hocking RR. 1976. A Biometrics invited paper. The analysis and selection of variables in linear regression. Biometrics, **32**: 1-49.

Hsieh WW, Tang B. 1998. Applying Neural Network Models to Prediction and Data Analysis in Meteorology and Oceanography. Bulletin of the American Meteorological Society, **79**: 1855-1870. DOI: 10.1175/1520-0477(1998)079<1855:annmtp>2.0.co;2.

Isaaks EH, Srivastava RM. 1989. Applied geostatistics. [electronic resource]. New York : Oxford University Press, 1989.

Jenkins G, Lowe J. 2003. Handling uncertainties in the UKCIP02 scenarios of climate change. Hadley Centre, Technical note 44, Exeter, UK.

Jeong DI, St-Hilaire A, Ouarda TBMJ, Gachon P. 2012. Multisite statistical downscaling model for daily precipitation combined by multivariate multiple linear regression and stochastic weather generator. Climatic Change, **114**: 567-591. DOI: 10.1007/s10584-012-0451-3.

Jin X, Xu C-Y, Zhang Q, Singh V. 2010. Parameter and modeling uncertainty simulated by GLUE and a formal Bayesian method for a conceptual hydrological model. Journal of Hydrology, **383**: 147-155.

Jones NK. 2008. On the Impact of Recent Climate Change on Seasonal Floods—A Case Study from a River Basin in Southern Quebec. Canadian Water Resources Journal / Revue canadienne des ressources hydriques, **33**: 55-72. DOI: 10.4296/cwrj3301055.

Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, Iredell M, Saha S, White G, Woollen J. 1996. The NCEP/NCAR 40-year reanalysis project. Bulletin of the American meteorological Society, **77**: 437-471.

Kamp R, Savenije H. 2007. Hydrological model coupling with ANNs. Hydrology and Earth System Sciences Discussions, **11**: 1869-1881.

Kannan S, Ghosh S. 2011. Prediction of daily rainfall state in a river basin using statistical downscaling from GCM output. Stochastic Environmental Research & Risk Assessment, **25**: 457-474. DOI: 10.1007/s00477-010-0415-y.

Karl TR, Wang W-C, Schlesinger ME, Knight RW, Portman D. 1990. A Method of Relating General Circulation Model Simulated Climate to the Observed Local Climate. Part I: Seasonal Statistics. Journal of Climate, **3**: 1053-1079. DOI: doi:10.1175/1520-0442(1990)003<1053:AMORGC>2.0.CO;2.

Katz RW. 1977. Precipitation as a Chain-Dependent Process. Journal of Applied Meteorology, **16**: 671-676. DOI: doi:10.1175/1520-0450(1977)016<0671:PAACDP>2.0.CO;2.

Kavetski D, Kuczera G, Franks SW. 2006. Bayesian analysis of input uncertainty in hydrological modeling: 2. Application. Water Resources Research, **42**: 1-10.

Kay A, Davies H, Bell V, Jones R. 2009. Comparison of uncertainty sources for climate change impacts: flood frequency in England. Climatic Change, **92**: 41-63.

Khalili M, Van Nguyen VT, Gachon P. 2013. A statistical approach to multi-site multivariate downscaling of daily extreme temperature series. International Journal of Climatology, **33**: 15-32. DOI: 10.1002/joc.3402.

Khan MS, Coulibaly P, Dibike Y. 2006. Uncertainty analysis of statistical downscaling methods. Journal of Hydrology, **319**: 357-382. DOI: http://dx.doi.org/10.1016/j.jhydrol.2005.06.035.

Kigobe M, Van Griensven A. 2010. Assessing hydrological response to change in climate: Statistical downscaling and hydrological modelling within the upper Nile. In: 2010 International Congress on Environmental Modelling and Software,Modelling for Environment's Sake, Fifth Biennial Meeting, David A. Swayne WY, A. A. Voinov, A. Rizzoli, T. Filatov (eds.) (ed.) International Environmental Modelling and Software Society (iEMS), pp: n/a-n/a.

Kim U, Kaluarachchi JJ. 2009. Climate change impacts on water resources in the Upper Blue Nile River Basin, Ethiopia1. JAWRA Journal of the American Water Resources Association, **45**: 1361-1378. DOI: 10.1111/j.1752-1688.2009.00369.x.

Kiureghian AD, Ditlevsen O. 2009. Aleatory or epistemic? Does it matter? Structural Safety, **31**: 105-112. DOI: http://dx.doi.org/10.1016/j.strusafe.2008.06.020.

Krige D. 1951. A Statistical Approach to Some Basic Mine Valuation Problems on the Witwatersrand. Journal of the Chemical, Metallurgical and Mining Society of South Africa, **52**: 119-139. DOI: citeulike-article-id:3070804 doi: 10.2307/3006914.

Kuss M, Rasmussen CE. 2005. Assessing approximations for Gaussian process classification. In: Advances in Neural Information Processing Systems, pp: 699-706.

Kuss M, Rasmussen CE. 2006. Assessing approximations for Gaussian process classification. Advances in Neural Information Processing Systems, **18**: 699.

Lall U, Sharma A. 1996. A Nearest Neighbor Bootstrap For Resampling Hydrologic Time Series. Water Resources Research, **32**: 679-693. DOI: 10.1029/95WR02966.

Lawrence N, Seeger M, Herbrich R. 2003. Fast sparse Gaussian process methods: The informative vector machine. Advances in neural information processing systems: 625-632.

Liléo S, Petrik O. 2011. Investigation on the use of NCEP/NCAR, MERRA and NCEP/CFSR reanalysis data in wind resource analysis. In: European Wind

Energy Conference and Exhibition 2011, EWEC 2011, 14 March 2011 through 17 March 2011, Brussels, Belgium, pp: 181-185.

Lindau R, Simmer C. 2013. On correcting precipitation as simulated by the regional climate model COSMO-CLM with daily rain gauge observations. Meteorology and Atmospheric Physics, **119**: 31-42. DOI: 10.1007/s00703-012-0215-7.

Liu J, Yuan D, Zhang L, Zou X, Song X. 2016. Comparison of Three Statistical Downscaling Methods and Ensemble Downscaling Method Based on Bayesian Model Averaging in Upper Hanjiang River Basin, China. Advances in Meteorology, **2016**: 12. DOI: 10.1155/2016/7463963.

Liu W, Fu G, Liu C, Charles SP. 2013. A comparison of three multi-site statistical downscaling models for daily rainfall in the North China Plain. Theoretical and Applied Climatology, **111**: 585-600. DOI: 10.1007/s00704-012-0692-0.

Liu Y, Sang Y-F, Li X, Hu J, Liang K. 2017. Long-Term Streamflow Forecasting Based on Relevance Vector Machine Model. Water, **9**: 9.

Liu Z, Xu Z, Charles SP, Fu G, Liu L. 2011. Evaluation of two statistical downscaling models for daily precipitation over an arid basin in China. International Journal of Climatology, **31**: 2006-2020. DOI: 10.1002/joc.2211.

Lu Y, Qin XS. 2014. A coupled K-nearest neighbour and Bayesian neural network model for daily rainfall downscaling. International Journal of Climatology, **34**: 3221-3236. DOI: 10.1002/joc.3906.

Lu Y, Qin XS. 2014. Multisite rainfall downscaling and disaggregation in a tropical urban area. Journal of Hydrology, **509**: 55-65. DOI: http://dx.doi.org/10.1016/j.jhydrol.2013.11.027.

Lu Y, Qin XS, Xie YJ. 2016. An integrated statistical and data-driven framework for supporting flood risk analysis under climate change. Journal of Hydrology, **533**: 28-39. DOI: http://doi.org/10.1016/j.jhydrol.2015.11.041.

MacKay DJ. 1992. Bayesian interpolation. Neural computation, **4**: 415-447.

MacKay DJC. 1992. Bayesian Interpolation. In: Maximum Entropy and Bayesian Methods: Seattle, 1991, Smith CR, Erickson GJ, Neudorfer PO (eds.) Springer Netherlands, pp: 39-66.

MacQueen J. 1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Oakland, CA, USA., pp: 281-297.

Madsen H, Arnbjerg-Nielsen K, Mikkelsen PS. 2009. Update of regional intensity–duration–frequency curves in Denmark: Tendency towards increased storm intensities. Atmospheric Research, **92**: 343-349. DOI: http://dx.doi.org/10.1016/j.atmosres.2009.01.013.

Mani A, Tsai FT-C. 2017. Ensemble Averaging Methods for Quantifying Uncertainty Sources in Modeling Climate Change Impact on Runoff Projection. Journal of Hydrologic Engineering, **22**: 1.

Maraun D, Wetterhall F, Ireson AM, Chandler RE, Kendon EJ, Widmann M, Brienen S, Rust HW, Sauter T, Themeßl M, Venema VKC, Chun KP, Goodess CM, Jones RG, Onof C, Vrac M, Thiele-Eich I. 2010. Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. Reviews of Geophysics, **48**: 1-34. DOI: 10.1029/2009RG000314.

Marshall L, Nott D, Sharma A. 2007. Towards dynamic catchment modelling: a Bayesian hierarchical mixtures of experts framework. Hydrological Processes, **21**: 847-861.

Martinez WL, Martinez AR, Solka JL. 2011. Exploratory data analysis with MATLAB. [electronic resource]. Boca Raton, Fla. : CRC Press, 2nd ed.

Matthies HG. 2007. QUANTIFYING UNCERTAINTY: MODERN COMPUTATIONAL REPRESENTATION OF PROBABILITY AND APPLICATIONS. In: Extreme Man-Made and Natural Hazards in Dynamics of Structures, Ibrahimbegovic A, Kozar I (eds.) Springer Netherlands, pp: 105-135.

McCulloch WS, Pitts W. 1943. A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics, **5**: 115-133. DOI: 10.1007/bf02478259.

MEWR. 2012. Report on Key Conclusions and Recommendations of the Expert Panel on Drainage Design and Flood Protection Measures. Ministry of the Environment and Water Resources Singapore.

Mezghani A. 2009. A combined downscaling-disaggregation weather generator for stochastic generation of multisite hourly weather variables over complex terrain: Development and multi-scale validation for the Upper Rhone River basin. Journal of hydrology, **377**: 245.

Mezghani A, Hingray B. 2009. A combined downscaling-disaggregation weather generator for stochastic generation of multisite hourly weather variables over complex terrain: Development and multi-scale validation for the Upper Rhone

River basin. Journal of Hydrology, **377**: 245-260. DOI: http://dx.doi.org/10.1016/j.jhydrol.2009.08.033.

Mezghani A, Hingray B. 2009. A combined downscaling-disaggregation weather generator for stochastic generation of multisite hourly weather variables over complex terrain: Development and multi-scale validation for the Upper Rhone River basin. Journal of Hydrology, **377**: 245-260.

Minka TP. 2001. Expectation propagation for approximate Bayesian inference. Morgan Kaufmann Publishers Inc., pp: 362-369.

Minka TP. 2001. Expectation propagation for approximate Bayesian inference. In: Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc., pp: 362-369.

Mirzaei M, Huang YF, El-Shafie A, Shatirah A. 2015. Application of the generalized likelihood uncertainty estimation (GLUE) approach for assessing uncertainty in hydrological models: a review. Stochastic Environmental Research and Risk Assessment, **29**: 1265-1273. DOI: 10.1007/s00477-014-1000-6.

MSS. 2016. Climate of Singapore. In: Meteorological Service Singapore, Singapore Government.

Mujumdar PP, Nagesh Kumar D. 2012. Floods in a Changing Climate: Hydrologic Modeling. Cambridge University Press.

Nakićenović N. 2000. Greenhouse Gas Emissions Scenarios. Technological Forecasting and Social Change, **65**: 149-166. DOI: http://doi.org/10.1016/S0040-1625(00)00094-9.

NEA. 2016. Local Climatology Singapore. In: National Environmental Agency Singapore.

Neal RM. 1996. Bayesian learning for neural networks. Springer-Verlag New York, Inc.

Nieminen J, Ylinen J, Seppälä T, Alapaholuoma T, Loula P. 2012. A framework for classifying IPFIX flow data, case KNN classifier. In: Proceeding of ICNS 2012, The Eighth International Conference on Networking and Services, pp: 14-19.

Nowak K, Prairie J, Rajagopalan B, Lall U. 2010. A nonparametric stochastic approach for multisite disaggregation of annual to daily streamflow. Water Resources Research, **46**: n/a-n/a. DOI: 10.1029/2009WR008530.

O'Hagan A, Kingman JFC. 1978. Curve Fitting and Optimal Design for Prediction. Journal of the Royal Statistical Society. Series B (Methodological), **40**: 1-42.

Olsson J, Uvo CB, Jinno K. 2001. Statistical atmospheric downscaling of short-term extreme rainfall by neural networks. Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere, **26**: 695-700. DOI: http://dx.doi.org/10.1016/S1464-1909(01)00071-5.

Opper M, Winther O. 2000. Gaussian processes for classification: Mean-field algorithms. Neural computation, **12**: 2655-2684.

Osborne MA, Roberts SJ, Rogers A, Jennings NR. 2012. Real-time information processing of environmental sensor network data using bayesian gaussian processes. ACM Transactions on Sensor Networks (TOSN), **9**: 1.

Oyebode OK, Adeyemo J, Otieno FAO. 2014. Uncertainty sources in climate change impact modelling of water resource systems. Academic Journal of Science, **03**: 245–260.

Pachauri RK. 2004. Climate Change and its Implications for Development: The Role of IPCC Assessments. IDS Bulletin, **35**: 11-14.

Pelczer I, Cisneros-Iturbe H. 2008. Identification of rainfall patterns over the Valley of Mexico. In: 11th International Conference on Urban Drainage. Edinburgh, Scotland, UK, pp: 1-9.

Pervez MS, Henebry GM. 2014. Projections of the Ganges–Brahmaputra precipitation—Downscaled from GCM predictors. Journal of Hydrology, **517**: 120-134. DOI: http://doi.org/10.1016/j.jhydrol.2014.05.016.

Pope VD, Gallani ML, Rowntree PR, Stratton RA. 2000. The impact of new physical parametrizations in the Hadley Centre climate model: HadAM3. Climate Dynamics, **16**: 123-146. DOI: 10.1007/s003820050009.

Prairie J, Rajagopalan B, Lall U, Fulp T. 2007. A stochastic nonparametric technique for space-time disaggregation of streamflows. Water Resources Research, **43(3)**: n/a-n/a.

Price RK. 2000. Hydroinformatics and urban drainage: an agenda for the beginning of the 21st century. Journal of Hydroinformatics, **2**: 133-147.

Prudhomme C, Davies H. 2009. Assessing uncertainties in climate change impact analyses on the river flow regimes in the UK. Part 1: baseline climate. Climatic Change, **93**: 177-195.

Prudhomme C, Davies H. 2009. Assessing uncertainties in climate change impact analyses on the river flow regimes in the UK. Part 2: future climate. Climatic Change, **93**: 197-222.

Qian B, Corte-Real J, Xu H. 2002. Multisite stochastic weather models for impact studies. International Journal of Climatology, **22**: 1377-1397. DOI: 10.1002/joc.808.

Quintana Seguí P, Ribes A, Martin E, Habets F, Boé J. 2010. Comparison of three downscaling methods in simulating the impact of climate change on the hydrology of Mediterranean basins. Journal of Hydrology, **383**: 111-124. DOI: http://dx.doi.org/10.1016/j.jhydrol.2009.09.050.

Rajagopalan B, Lall U. 1999. A k-nearest-neighbor simulator for daily precipitation and other weather variables. Water resources research, **35**: 3089-3101.

Rajendran QS, Cheung SH. 2015. BUASCSDSEC--Uncertainty Assessment of Coupled Classification and Statistical Downscaling Using Gaussian Process Error Coupling. International Journal of Environmental Science and Development, **6**: 211.

Ramos J. 2012. MATLAB implementation of the Dunn Index.

Randall DA. 2000. General circulation model development. San Diego : Academic Press.

Rasmussen CE, Nickisch H. 2010. Gaussian processes for machine learning (GPML) toolbox. Journal of Machine Learning Research, **11**: 3011-3015.

Rasmussen CE, Williams CKI. 2005. Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press.

Rasmussen CE, Williams CKI. 2006. Gaussian processes for machine learning. Cambridge, Mass. : MIT Press.

Refsgaard JC, Arnbjerg-Nielsen K, Drews M, Halsnæs K, Jeppesen E, Madsen H, Markandya A, Olesen JE, Porter JR, Christensen JH. 2013. The role of uncertainty in climate change adaptation strategies—A Danish water management example. Mitigation and Adaptation Strategies for Global Change, **18**: 337-359. DOI: 10.1007/s11027-012-9366-6.

Reilly J, Stone PH, Forest CE, Webster MD, Jacoby HD, Prinn RG. 2001. Uncertainty and climate change assessments. Science, **293**: 430-433.

Riad S, Mania J, Bouchaou L, Najjar Y. 2004. Rainfall-runoff model usingan artificial neural network approach. Mathematical and Computer Modelling, **40**: 839-846. DOI: http://dx.doi.org/10.1016/j.mcm.2004.10.012.

Roberts S, Osborne M, Ebden M, Reece S, Gibson N, Aigrain S. 2013. Gaussian processes for time-series modelling. Phil. Trans. R. Soc. A, **371**: 20110550.

Robin KSH. 2012. Introducing multivator : A Multivariate Emulator. Journal of Statistical Software, **46**: 1-20.

Robinson PJ. 1997. Climate change and hydropower generation. International Journal of Climatology, **17**: 983-996.

Rousseeuw PJ. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, **20**: 53-65.

Rowell DP. 2006. A demonstration of the uncertainty in projections of UK climate change resulting from regional model formulation. Climatic Change, **79**: 243-257.

Saha S, Moorthi S, Pan H-L, Wu X, Wang J, Nadiga S, Tripp P, Kistler R, Woollen J, Behringer D, Liu H, Stokes D, Grumbine R, Gayno G, Wang J, Hou Y-T, Chuang H-Y, Juang H-MH, Sela J, Iredell M, Treadon R, Kleist D, Delst PV, Keyser D, Derber J, Ek M, Meng J, Wei H, Yang R, Lord S, Dool HVD, Kumar A, Wang W, Long C, Chelliah M, Xue Y, Huang B, Schemm J-K, Ebisuzaki W, Lin R, Xie P, Chen M, Zhou S, Higgins W, Zou C-Z, Liu Q, Chen Y, Han Y, Cucurull L, Reynolds RW, Rutledge G, Goldberg M. 2010. The NCEP Climate Forecast System Reanalysis. Bulletin of the American Meteorological Society, **91**: 1015-1057. DOI: doi:10.1175/2010BAMS3001.1.

Salas JD, Markus M, Tokar AS. 2000. Streamflow Forecasting Based on Artificial Neural Networks. In: Artificial Neural Networks in Hydrology, Govindaraju RS, Rao AR (eds.) Springer Netherlands, pp: 23-51.

Samadi S, Carbone GJ, Mahdavi M, Sharifi F, Bihamta MR. 2013. Statistical Downscaling of River Runoff in a Semi Arid Catchment. Water Resources Management, **27**: 117-136. DOI: 10.1007/s11269-012-0170-6.

Schaefli B, Talamba DB, Musy A. 2007. Quantifying hydrological modeling errors through a mixture of normal distributions. Journal of Hydrology, **332**: 303-315.

Schmocker-Fackel P, Naef F. 2010. More frequent flooding? Changes in flood frequency in Switzerland since 1850. Journal of Hydrology, **381**: 1-8. DOI: http://doi.org/10.1016/j.jhydrol.2009.09.022.

Schoof JT, Pryor S. 2001. Downscaling temperature and precipitation: A comparison of regression-based methods and artificial neural networks. International Journal of climatology, **21**: 773-790.

Seeger M. 2005. Expectation propagation for exponential families (No. EPFL-REPORT-161464).

Segond M-L, Neokleous N, Makropoulos C, Onof C, Maksimovic C. 2007. Simulation and spatio-temporal disaggregation of multi-site rainfall data for urban drainage applications. Hydrological Sciences Journal, **52**: 917-935.

Segond ML. 2006. Spatial–temporal disaggregation of daily rainfall from a generalized linear model. Journal of hydrology, **331**: 674.

Semenov MA, Barrow EM, Lars-Wg A. 2002. A stochastic weather generator for use in climate impact studies. User Manual, Hertfordshire, UK.

Semenov MA, Brooks RJ, Barrow EM, Richardson CW. 1998. Comparison of the WGEN and LARS-WG stochastic weather generators for diverse climates. Climate research, **10**: 95-107.

Shrestha DL, Solomatine DP. 2008. Data-driven approaches for estimating uncertainty in rainfall-runoff modelling. International Journal of River Basin Management, **6**: 109-122.

Skolidis G, Sanguinetti G. 2011. Bayesian Multitask Classification With Gaussian Process Priors. IEEE Transactions on Neural Networks, **22**: 2011-2021. DOI: 10.1109/TNN.2011.2168568.

Snelson EL. 2008. Flexible and efficient Gaussian process models for machine learning. University of London, University College London (United Kingdom).

Solomatine D, See LM, Abrahart RJ. 2008. Data-Driven Modelling: Concepts, Approaches and Experiences. In: Practical Hydroinformatics: Computational Intelligence and Technological Developments in Water Applications, Abrahart RJ, See LM, Solomatine DP (eds.) Springer Berlin Heidelberg, pp: 17-30.

Solomatine DP, Ostfeld A. 2008. Data-driven modelling: some past experiences and new approaches. Journal of hydroinformatics, **10**: 3-22.

Solomatine DP, Price RK. 2004. Innovative approaches to flood forecasting using data driven and hybrid modelling. In: Proc. 6th International Conference on Hydroinformatics.

Solomon S. 2007. Climate change 2007 : the physical science basis : contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge ; New York : Cambridge University Press.

Solomon S, Intergovernmental Panel on Climate Change I, Intergovernmental Panel on Climate Change I, Working Group I. 2007. Climate change 2007 : the physical science basis : contribution of Working Group I to the Fourth Assessment

Report of the Intergovernmental Panel on Climate Change. Cambridge University Press.

Srikanthan R, McMahon TA. 2001. Stochastic generation of annual, monthly and daily climate data: A review. Hydrol. Earth Syst. Sci., **5**: 653-670. DOI: 10.5194/hess-5-653-2001.

Srivastav R, Sudheer K, Chaubey I. 2007. A simplified approach to quantifying predictive and parametric uncertainty in artificial neural network hydrologic models. Water Resources Research, **43(10)**: n/a.

Stern R, Coe R. 1984. A model fitting analysis of daily rainfall data. Journal of the Royal Statistical Society. Series A (General): 1-34.

Sun AY, Wang D, Xu X. 2014. Monthly streamflow forecasting using Gaussian Process Regression. Journal of Hydrology, **511**: 72-81. DOI: https://doi.org/10.1016/j.jhydrol.2014.01.023.

Sunyer MA, Madsen H, Ang PH. 2012. A comparison of different regional climate models and statistical downscaling methods for extreme rainfall estimation under climate change. Atmospheric Research, **103**: 119-128. DOI: http://dx.doi.org/10.1016/j.atmosres.2011.06.011.

Sunyer Pinya MA, Hundecha Y, Lawrence D, Madsen H, Willems P, Martinkova M, Vormoor K, Bürger G, Hanel M, Kriaučiuniene J. 2015. Inter-comparison of statistical downscaling methods for projection of extreme precipitation in Europe. Hydrology and Earth System Sciences, **19**: 1827-1847.

Tang J, Niu X, Wang S, Gao H, Wang X, Wu J. 2016. Statistical downscaling and dynamical downscaling of regional climate in China: Present climate evaluations and future climate projections. Journal of Geophysical Research: Atmospheres, **121**: 2110-2129. DOI: 10.1002/2015JD023977.

Taye MT, Willems P. 2013. Influence of downscaling methods in projecting climate change impact on hydrological extremes of upper Blue Nile basin. Hydrol. Earth Syst. Sci. Discuss., **2013**: 7857-7896. DOI: 10.5194/hessd-10-7857-2013.

Tebaldi C, Smith RL, Nychka D, Mearns LO. 2005. Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multimodel ensembles. Journal of Climate, **18**: 1524-1540.

The MathWorks I. MATLAB and Statistics Toolbox Release 2012b Natick, Massachusetts, United States.

Thompson R, Green DN. 2004. Mediterranean precipitation and its relationship with sea level pressure patterns. Annals of Geophysics, Vol 47, Iss 5 DOI: 10.4401/ag-3364.

Tipping M. 2003. Relevance vector machine. U.S. Patent No. 6,633,857.

Tisseuil C, Vrac M, Lek S, Wade AJ. 2010. Statistical downscaling of river flows. Journal of Hydrology, **385**: 279-291. DOI: http://dx.doi.org/10.1016/j.jhydrol.2010.02.030.

Tokar AS, Johnson PA. 1999. Rainfall-runoff modeling using artificial neural networks. Journal of Hydrologic Engineering, **4**: 232-239.

Tripathi S, Srinivas VV, Nanjundiah RS. 2006. Downscaling of precipitation for climate change scenarios: A support vector machine approach. Journal of Hydrology, **330**: 621-640. DOI: http://doi.org/10.1016/j.jhydrol.2006.04.030.

Vanik MW, Beck JL, Au SK. 2000. Bayesian Probabilistic Approach to Structural Health Monitoring. Journal of Engineering Mechanics, **126**: 738-745. DOI: doi:10.1061/(ASCE)0733-9399(2000)126:7(738).

Vasudevan S, Ramos F, Nettleton E, Durrant-Whyte H. 2011. Non-stationary dependent Gaussian processes for data fusion in large-scale terrain modeling. In: Robotics and Automation (ICRA), 2011 IEEE International Conference on, pp: 1875-1882.

Vasudevan S, Ramos F, Nettleton E, Durrant-Whyte H. 2009. Gaussian process modeling of large-scale terrain. Journal of Field Robotics, **26**: 812-840.

von Storch H. 1999. The Global and Regional Climate System. In: Anthropogenic Climate Change, von Storch H, Flöser G (eds.) Springer Berlin Heidelberg, pp: 3-36.

Wang W, Xie P, Yoo S-H, Xue Y, Kumar A, Wu X. 2011. An assessment of the surface climate in the NCEP climate forecast system reanalysis. Climate Dynamics, **37**: 1601-1620. DOI: 10.1007/s00382-010-0935-7.

Wei S, Song J, Khan NI. 2012. Simulating and predicting river discharge time series using a wavelet-neural network hybrid modelling approach. Hydrological Processes, **26**: 281-296.

Wetterhall F, Bárdossy A, Chen D, Halldin S, Xu C-Y. 2006. Daily precipitation-downscaling techniques in three Chinese regions. Water Resources Research, **42**: n/a-n/a. DOI: 10.1029/2005WR004573.

Wilby R, Charles S, Zorita E, Timbal B, Whetton P, Mearns L. 2004. Guidelines for use of climate scenarios developed from statistical downscaling methods. Supporting material of the Intergovernmental Panel on Climate Change, available from the DDC of IPCC TGCIA, **27**.

Wilby R, Tomlinson O, Dawson C. 2003. Multi-site simulation of precipitation by conditional resampling. Climate Research, **23**: 183-194.

Wilby RL. 1998. Statistical downscaling of daily precipitation using daily airflow and seasonal teleconnection indices. Climate Research, **10**: 163-178.

Wilby RL, Dawson CW, Barrow EM. 2002. sdsm — a decision support tool for the assessment of regional climate change impacts. Environmental Modelling & Software, **17**: 145-157. DOI: http://dx.doi.org/10.1016/S1364-8152(01)00060-3.

Wilby RL, Harris I. 2006. A framework for assessing uncertainties in climate change impacts: Low-flow scenarios for the River Thames, UK. Water Resources Research, **42**: n/a-n/a. DOI: 10.1029/2005WR004065.

Wilby RL, Wigley T, Conway D, Jones P, Hewitson B, Main J, Wilks D. 1998. Statistical downscaling of general circulation model output: a comparison of methods. Water resources research, **34**: 2995-3008.

Wilby RL, Wigley TML. 1997. Downscaling general circulation model output: a review of methods and limitations. Progress in Physical Geography, **21**: 530-548. DOI: 10.1177/030913339702100403.

Willems P. 2012. Impacts of climate change on rainfall extremes and urban drainage systems. London : IWA.

Willems P, Olsson J, Arnbjerg-Nielsen K, Beecham S, Pathirana A, Gregersen IB, Madsen H, Nguyen V-T-V. Climate Change Impacts on Rainfall Extremes and Urban Drainage: a State-of-the-Art Review. In: World Environmental and Water Resources Congress 2013, pp: 1131-1135.

Williams CK, Barber D. 1998. Bayesian classification with Gaussian processes. IEEE Transactions on Pattern Analysis and Machine Intelligence, **20**: 1342-1351.

Williams CKI. 1999. Prediction with Gaussian processes: from linear regression to linear prediction and beyond. 599-621. DOI: citeulike-article-id:3109141.

Wilson AG, Adams RP. 2013. Gaussian Process Kernels for Pattern Discovery and Extrapolation. In: ICML (3), pp: 1067-1075.

Woldemeskel FM, Sharma A, Sivakumar B, Mehrotra R. 2014. A framework to quantify GCM uncertainties for use in impact assessment studies. Journal of Hydrology, **519, Part B**: 1453-1465. DOI: http://dx.doi.org/10.1016/j.jhydrol.2014.09.025.

Xoplaki E, Luterbacher J, Burkard R, Patrikas I, Maheras P. 2000. Connection between the large-scale 500 hPa geopotential height fields and precipitation over Greece during wintertime. Climate Research, **14**: 129-146.

Xu C-y. 1999. Climate Change and Hydrologic Models: A Review of Existing Gaps and Recent Research Developments. Water Resources Management, **13**: 369-382. DOI: 10.1023/a:1008190900459.

Yang C, Chandler RE, Isham VS, Wheater HS. 2005. Spatial-temporal rainfall simulation using generalized linear models. Water Resources Research, **41**: n/a-n/a. DOI: 10.1029/2004WR003739.

Yang J, Reichert P, Abbaspour K, Xia J, Yang H. 2008. Comparing uncertainty analysis techniques for a SWAT application to the Chaohe Basin in China. Journal of Hydrology, **358**: 1-23.

Yang MY, Liao W, Rosenhahn B, Zhang Z. 2015. Hyperspectral image classification using Gaussian process models. In: 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp: 1717-1720.

Ying L, Ke F, Hui-Jun W. 2011. Statistical Downscaling Prediction of Summer Precipitation in Southeastern China. Atmospheric and Oceanic Science Letters, **4**: 173-180. DOI: 10.1080/16742834.2011.11446925.

Zhao K, Popescu S, Zhang X. 2008. Bayesian learning with Gaussian processes for supervised classification of hyperspectral data. Photogrammetric Engineering & Remote Sensing, **74**: 1223-1234.

Zorita E, Storch Hv. 1999. The Analog Method as a Simple Statistical Downscaling Technique: Comparison with More Complicated Methods. Journal of Climate, **12**: 2474-2489. DOI: doi:10.1175/1520-0442(1999)012<2474:TAMAAS>2.0.CO;2.