BUASCSDSEC — Uncertainty Assessment of Coupled Classification and Statistical Downscaling Using Gaussian Process Error Coupling

Queen Suraajini Rajendran and Sai Hung Cheung

***Abstract*—The statistical downscaling models which are used as a bridging model to connect the global climate model output and the local weather variables have uncertainty associated with it. The uncertainty present in the model as well as in the results should be quantified for reliable climate change impact studies. The sources of uncertainty include natural variability, uncertainty in the climate model(s), downscaling model, model inadequacy and in the predicted results. Uncertainty analysis and quantification in the models is a promising approach for climate change impact studies. In this paper, a new approach called BUASCSDSEC (Bayesian uncertainty analysis for stochastic classification and statistical downscaling with stochastic dependent error coupling) is proposed. It is a robust Bayesian uncertainty analysis methodology and tools for combined classification (to predict the occurrence of rainfall) and statistical downscaling. It is based on coupling dependent modelling error which is viewed as a function modelled as a stochastic process with classification and statistical downscaling models in a way that the dependency among modelling errors will impact the result of the classification and statistical downscaling model calibration and uncertainty analysis for future prediction. Gaussian Process is considered in the error modelling. Singapore data are used and the uncertainty and prediction results are obtained for the validation period (1995-2000). It is observed that the CDFs of the daily predicted samples are consistent with the observed CDF of precipitation. The uncertainty is smaller for the extreme rainfall and the uncertainty for smaller amount of rainfall is more compared to that for the extreme rainfall. From the results obtained, ongoing research for improvement is briefly presented.**

***Index Terms*—Stochastic process, gaussian process, stochastic classification, statistical downscaling, uncertainty quantification, model inadequacy.**

I. Introduction

Statistical downscaling models are being widely used in studying climate change impact on hydrology. There have been increasing natural hazards such as flooding due to climate change. These extreme flood events cause damage to the properties and affect daily life. The impact of climate change on hydrology needs to be assessed to assist in future planning and risk mitigation during extreme flood events. Planning the adaption measures due to climate change for future is very important. For more robust future planning, the uncertainties in the models and results should be modelled and quantified properly [1]. The quantification of uncertainty in the model structure, model parameters and results is very important before using it for impact studies. From the literature survey, it is understood that the development of a method for assessing the models' prediction ability needs more attention of researchers since there are only limited studies existing so far [1], Probabilistic framework for assessing uncertainty [2] and Multi-model approach [3] are some of the methods proposed for uncertainty quantification in statistical downscaling models. Probabilistic framework can be used to link all the information from different models based on its efficiency and contribution to the output. A Bayesian method to quantify uncertainty in multimodal ensemble Atmospheric-Ocean General Circulation Models (AOGCMs) and to predict future scenarios were developed by [4]. They considered the present and future precipitation values as uncertain parameters and it is modelled random variables. They showed the range of uncertainty for future predicted precipitation values. They have also showed that large inter-regional variability is one of the reasons for uncertainty in the results. Tebaldi, (2004) applied Bayesian multi-model approach for assessing uncertainty in statistically downscaled precipitation [5]. They have showed that the Bayesian framework is helpful in extracting information from simulated precipitation values. It is also suggested that Bayesian framework along with statistical model can be used to couple the climate model and the uncertainty quantification model together. Hashmi, Shamseldin *et al*., (2009) developed weighted multi-model ensemble using Bayesian framework for statistical downscaling models [6]. They have combined the statistical methods based on multiple linear regression, multiple nonlinear regression and stochastic weather generator in the framework. They have showed that Bayesian method is helpful in quantifying uncertainty in the downscaling model results. J. Chen *et al*., (2011) compared six downscaling methods are compared to quantify uncertainty for studying climate change impact on hydrology. It showed that the regression based downscaling method contributes more uncertainty to the results [7].

There is a need for an efficient tool to quantify the uncertainty in the models as well as in the model parameters. The classification and statistical downscaling models have inherent errors in them. The presence of errors reduces the reliability of the model for downscaling. The uncertainties in the model are due to incomplete data, the spatial and temporal variation, incomplete understanding of the real world processes especially about the extreme events and model structure [8].

Manuscript received March 1, 2014; revised April 25, 2014.

The authors are with the School of Civil and Environmental Engineering, Nanyang Technological University, Singapore, 639798 (e-mail: queen1@e.ntu.edu.sg, shcheung@ntu.edu.sg).

DOI: 10.7763/IJESD.2015.V6.592

The variables that are used to describe the atmospheric process are not constant over time. They vary continuously with time which makes it difficult to predict the parameters with certainty. The known information about the parameters in the model is very limited. Incomplete knowledge causes uncertainty in making decisions. The error in the model is due to imperfect representation of the real world by the models, parameterization, measurement error and natural variability. The observations cannot be reproduced exactly by these models because of the presence of the aforementioned errors. Uncertainty is inherent in any numerical models and quantification of uncertainty in the output of the model is very important for future impact analysis.

There are two major problems which need to be addressed. The first problem is that the uncertainty quantification methods that are being used in the literature do not consider model calibration, prediction and uncertainty assessment simultaneously. Hence a framework that couples both the model calibration, prediction and the uncertainty quantification and propagation process and model inadequacy quantification for more reliable and robust predictions is necessary.

Bayesian framework provides a mechanism for model calibration which allows for uncertainty quantification of the model parameters and modelling errors. The development of uncertainty quantification tools using Bayesian statistical framework gives more accurate and robust predictions and will be helpful in assessing the accuracy of the models and the results [9]. Bayes’ theorem is used to combine the prior information and the likelihood function of the uncertain variables to get their updated posterior probability distribution. This posterior probability distribution can be propagated to obtain the updated (posterior) uncertainty in the model predictions. This process referred to as uncertainty propagation with uncertainty quantification using Bayesian model updating. In a lot of cases, the posterior distribution cannot be obtained analytically. Laplace asymptotic approximation method [9] can be 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 encapsulate its probabilistic information and to solve the multi-dimensional integrals. State of the art Markov chain Monte Carlo (MCMC) sampling methods such as those in Cheung and Beck, (2010) and Ching and Chen, (2007) can be used for stochastic sample simulation to reduce the computation in evaluating likelihood function many times [10], [11]. Statistical averaging of functions of Markov chain samples is used to estimate the integral. Bayesian model class averaging can then be used combine the predictions by all the candidate model classes together to make a more robust prediction [10].

The second problem is the assumption of probabilistic independency among the residuals in the model in most applications of Bayesian analysis. This assumption is not valid for the problem of interest here. Residuals should be viewed as uncertain function of input variables corresponding to different time instants and the dependent structure associated with the residuals [12] should also be considered for a better representation of a real-world system.

The objective of this study is to develop a new Bayesian uncertainty tool for combined stochastic classification and statistical downscaling with stochastic dependent error coupling. In this paper, Gaussian process [13] is considered to model the dependent errors.

II. Study Area and Data

The study area and rainfall stations for this research are Singapore which is shown in Fig. 1. It is located at 1 °20’N, 103 °50’E. The land area of Singapore is 716.1 km^2. The average annual precipitation is about 2340 mm. Singapore climate has no distinct seasons and is classified as Tropical rainforest climate. There are two monsoon periods in Singapore; northeast monsoon from December to March and southwest monsoon from June to September [14].

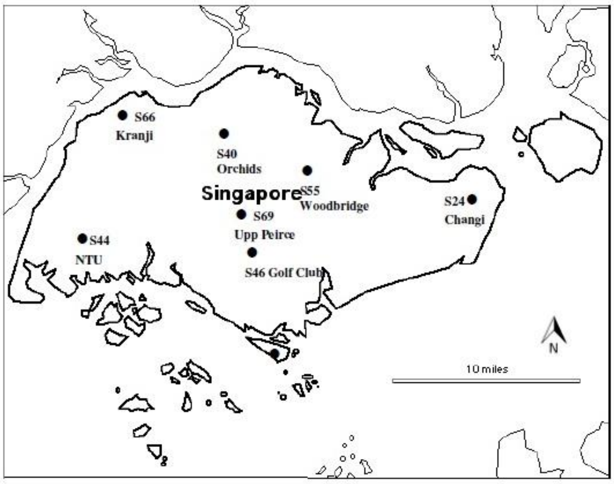

Fig. 1. Study area and location of rainguage stations [15].

The observed precipitation data used in this study were obtained from rain gauge stations in Singapore. The data from HadCM3 (Hadley Centre Coupled Model, version 3) are used as predictors in downscaling. HadCM3 is a coupled Atmospheric and Oceanic model and is extensively used for climate prediction. The levels in atmospheric component are 19 and the horizontal resolution of this Global Climate Model (GCM) is 2.5 degrees of latitude by 3.75 degrees of longitude. The total global grid cells are 96 × 73. The levels in oceanic component are 20 and the horizontal resolution is 1.25 × 1.25 degrees. The GCM data are available from 1961-2099. NCEP (National Centers for Environmental Protection) reanalysis data are used for calibrating and validating the model. The NCEP data have 41 years of daily observed data from 1961 to 2001. The calibration period for this study is from 1980 to 1994 and the validation period is from 1995 to 2000.

The predictors for model calibration are chosen from NCEP reanalysis data. They are 500 hPa geopotential (*p*500), 850 hPa geopotential (*p*850), near surface relative humidity (*rhum*), relative humidity at 500 hPa height (*r*500), relative humidity at 850 hPa height (*r*850) and near surface specific humidity (*shum*). Due to space limitations, stochastic models are calibrated and validated only for the month of December. The daily observed rainfall is used for downscaling and comparing the downscaled results. In this study, the observed rainfall from a Singapore golf station is used.

III. Methodology

A. Proposed Bayesian Stochastic Classification and Statistical Downscaling with Gaussian Process Error Coupling

A robust Bayesian uncertainty analysis methodology and tools have been developed for classification and statistical downscaling. It is based on coupling dependent modelling error which is viewed as a function modelled as a stochastic process with classification and statistical downscaling models in a way that the dependency among modelling errors will impact the result of stochastic classification and statistical downscaling model calibration and uncertainty analysis for future prediction. This proposed method for classification and statistical downscaling is called BUASCSDSEC (Bayesian uncertainty analysis for stochastic classification and statistical downscaling with stochastic dependent error coupling). In this proposed stochastically coupled classification -downscaling model, the quantified uncertainty from model parameters and in model structures (with inter-dependence and error quantified during the process) are propagated from the classification process to the output of downscaling model stochastically. The uncertainties are encapsulated in each successive prediction and models can be propagated forward to the other parts of model predictions. In this study, a special case of Gaussian process error coupling is considered. The main components of the proposed method are presented in sections *B* and *C* in this Section III.

The workflow of this framework is as follows. There are two stages in the model. The predictors and observed precipitation data are given as the inputs and labels of dry or wet days as the outputs to the binary classification model for model calibration. A Gaussian process is used to model the latent function. The optimal model parameters and hyperparameters values are obtained from the model calibration. The estimated values are then used to predict wet and dry day classification for the future days (validation period). From the proposed stochastic classification model, several samples (about 10000) of time histories of wet day or dry day label are simulated for the future days. The simulated samples are used as the input to decide which future days the wet days which require statistical precipitation downscaling with error coupling. The statistical downscaling model with Gaussian error coupling is calibrated using the data corresponding to the wet days in the past. The calibrated stochastic model is used for simulate samples of precipitation amount time history prediction for predicted future wet days. These samples are then used to analyze uncertainty in the predicted results.

B. Proposed Gaussian Process Binary Classification Model

A robust Bayesian binary (two classes) Gaussian Process rainfall classification model is developed. This model allows simulation of samples of discrete random vectors indicating which days are wet (i.e., having rainfall larger than some small threshold) and dry (i.e., having no rainfall or rainfall being smaller than some threshold) in the future based on data in the past. In this model, the event of a certain future day being wet (or dry) is dependent of that of the other days in the future and the past. The latent function $h(\boldsymbol{x})$, maps GCM predictors $\boldsymbol{x}$ to the probability of a day being wet (or dry) through a sigmoid function (e.g., logistic function, cumulative Gaussian). $h(\boldsymbol{x})$ is viewed as a stochastic process in $\boldsymbol{x}$ with dependent error coupling; such dependency among the input values $\boldsymbol{x}$ should influence the calibration of the stochastic model and future prediction for a more robust and realistic prediction and uncertainty analysis..

Given a data set $D_c = \{(x_i, y_i) \mid i = 1, ..., n\}$, $\mathbf{X} = [x_1 x_2 ... x_n]^T$ and $\mathbf{y} = [y_1 y_2 ... y_n]^T$ where $\mathbf{X}$ is the GCM predictors for all the days during the calibration period (n is the number of days in the calibration period); $\mathbf{y}$ = Observed Rainfall (wet/dry) for the calibration period, with binary class labels y_i = [0, 1], the proposed stochastic process classifier infers class label probabilities for the days in the future.

The latent function $h(\boldsymbol{x})$ is assumed to the form $g(\boldsymbol{x}, \boldsymbol{w})+\varepsilon(\boldsymbol{x})$ with the error function $\varepsilon(\boldsymbol{x})$ modeled as a Gaussian process in this paper. The covariance function of $\varepsilon(\boldsymbol{x})$ is assumed to follow a squared exponential covariance function with correlation length l and variance of $\varepsilon(\boldsymbol{x})$ being σ^2:

$$\mathrm{cov}(\varepsilon(x_i), \varepsilon(x_j)) = \sigma^2 exp(-\frac{\| x_i - x_j \|_2}{l^2}) \quad (1)$$

where ||.|| represents the norm of the vector inside it. In this model, the error in the latent function $h(\boldsymbol{x})$ is the stochastic dependent error $\varepsilon(\boldsymbol{x})$. A covariance function is used to capture the similarity or nearness of the input data, $\boldsymbol{x}$. It is based on the assumption that similar input values are likely to have similar target values.

By Bayes' Theorem, the posterior distribution of $\boldsymbol{\theta}$, $p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y})$ can be obtained: $p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y}) = p(\mathbf{y}| \mathbf{X}, \boldsymbol{\theta})p(\boldsymbol{\theta})/p(\mathbf{y}|\mathbf{X})$ with the prior distribution of $\boldsymbol{\theta}$ and the likelihood function given by the following:

$$p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta}) = \int p(\mathbf{y} \mid \mathbf{h}) p(\mathbf{h} \mid \mathbf{X}, \boldsymbol{\theta}) d\mathbf{h} \quad (2)$$

where $p(\mathbf{y} \mid \mathbf{h})$ is the class probability given the latent function (product of sigmoid functions) where $\mathbf{h} = [h_1 ... h_n]^T$; $p(\mathbf{h} \mid \mathbf{X}, \boldsymbol{\theta})$ is due to the prior placed over $h(\boldsymbol{x})$ which is assumed to follow Gaussian process in this paper and $g(\boldsymbol{x}, \boldsymbol{w})$ is assumed to be zero; the vector $\boldsymbol{\theta}$ includes the parameters and hyper-parameters of the latent function $h(\boldsymbol{x})$ including those appearing in the covariance function of $\varepsilon(\boldsymbol{x})$. In this paper, a logistic function is considered as the sigmoid function which is required to obtain the likelihood function. The posterior predictive distribution of $\mathbf{h}_*$ corresponding to future days given in (3) can be approximated by Laplace's asymptotic approximation given in (4):

$$p(\mathbf{h}_* \mid \mathbf{X}, \mathbf{y}, \mathbf{X}_*) = \int p(\mathbf{h}_* \mid \mathbf{X}, \mathbf{y}, \mathbf{X}_*, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{y}) d\boldsymbol{\theta} \quad (3)$$

$$q(\mathbf{h}_* | \mathbf{X}, \mathbf{y}, \mathbf{X}_*) \sim N(\mathbf{K}_*^T \nabla \log p(\mathbf{y} \mid \hat{\mathbf{h}}), \mathbf{K}(x_*, x_*) - \mathbf{K}_*^T (\mathbf{K}^{-1} + \mathbf{W})^{-1} \mathbf{K}_*) \quad (4)$$

where $\mathbf{W} = -\nabla\nabla \log p(\mathbf{y} \mid \mathbf{h})$ is the Hessian matrix of the negative log $p(\mathbf{y} \mid \mathbf{h})$, $\mathbf{K}$ is the covariance matrix of $\varepsilon(\boldsymbol{x})$ with $\boldsymbol{x}$ corresponding to those days in the calibration period; $\mathbf{K}_*$ is the covariance matrix between $\varepsilon(\boldsymbol{x})$ corresponding to the calibration days and $\varepsilon(\boldsymbol{x}_*)$ corresponding to the future prediction days; $\mathbf{K}(\boldsymbol{x}_*, \boldsymbol{x}_*)$ is the covariance matrix of $\varepsilon(\boldsymbol{x}_*)$ corresponding to the future prediction days. The estimates of the parameters and hyperparameters cannot be obtained analytically. Optimization such as Conjugate gradient method is used to obtain the estimates of parameters.

Posterior predictive joint class probability distribution for *N* future days given past data *D* including **X** and **y** and GCM predictors $\mathbf{X}_*$ for future days is given by (5).

$$p(\mathbf{y}_* \mid \mathbf{X}, \mathbf{y}, \mathbf{X}_*) = \int p(\mathbf{y}_* \mid \mathbf{h}_*)\, p(\mathbf{h}_* \mid \mathbf{X}, \mathbf{y}, \mathbf{X}_*)\, d\mathbf{h}_* \qquad (5)$$

where $p(\mathbf{y}_* \mid \mathbf{h}_*) = \prod_{j=1}^{N} p(y_{*,\mathrm{j}} \mid h_{*,\mathrm{j}})$

A lot of technical details for all the above are left out here due to space limitations and can be found in the journal version of this paper.

C. Proposed Statistical Downscaling Model with Gaussian Process Error Coupling

A statistical downscaling model is a function of predictors of GCM. Common examples of this function include Multiple linear/nonlinear function, Generalized linear/nonlinear model, Artificial neural network and Support vector machines. The errors are added at the same time when the statistical downscaling is being done. The model is represented by the following equation:

$$Y = f_d(x, \boldsymbol{\theta}_d) + \varepsilon_d \qquad (6)$$

Where *Y* is the dependent variable (i.e. precipitation), $f_d(\boldsymbol{x}, \boldsymbol{\theta}_d)$ is the function depending on statistical downscaling model (e.g. Multiple linear/nonlinear function; Generalized linear/nonlinear model, Artificial neural network, Support vector machines), **x** is the vector of independent variables (inputs), $\boldsymbol{\theta}_d$ includes the parameters of the function $f_d(\boldsymbol{x}, \boldsymbol{\theta}_d)$ and in our proposed method, ε_d is viewed as a function of $\boldsymbol{x}$ modelled as a stochastic process (or random field) with extra uncertain model parameters $\boldsymbol{\theta}_\varepsilon$.

In this work, we start with a simple case where the stochastic process is Gaussian with some form of covariance function between the errors corresponding to two different *x*'s which depends on the "distance" between the two *x* vector-valued variables. In the dependent error model, the errors are dependent with covariance function in the form of equation (1). The model parameters include both $\boldsymbol{\theta}_d$ in the function $f_d(\boldsymbol{x}, \boldsymbol{\theta}_d)$ and covariance function with hyper parameters (variance σ^2, and correlation length, *l*). The collection of random variables in Gaussian process follows multivariate normal distribution. It is assumed that if the "distance" between two input variable vectors is less, the correlation between the corresponding dependent variables (outputs) is more. If the distance between the input variable vectors is more, between the corresponding dependent variables (outputs) is less.

The posterior PDF of $\boldsymbol{\theta}_d, \boldsymbol{\theta}_\varepsilon$ given data $D_d = \hat{Y}$ (By Bayes' Theorem) is given in (7).

$$p(\boldsymbol{\theta}_d, \boldsymbol{\theta}_\varepsilon \mid D_d = \hat{Y}) = \frac{p(D_d \mid \boldsymbol{\theta}_d, \boldsymbol{\theta}_\varepsilon)\, p(\boldsymbol{\theta}_d, \boldsymbol{\theta}_\varepsilon)}{p(D_d)} \qquad (7)$$

where $p(D_d \mid \boldsymbol{\theta}_d, \boldsymbol{\theta}_\varepsilon)$ is the likelihood function, $p(\boldsymbol{\theta}_d, \boldsymbol{\theta}_\varepsilon)$ is the prior PDF of $\boldsymbol{\theta}_d, \boldsymbol{\theta}_\varepsilon$ and $p(D_d)$ is the evidence (marginal likelihood).

With sufficiently large sample size, Laplace's asymptotic method [8] can be used to approximate the posterior probability distribution of $\boldsymbol{\theta}_d, \boldsymbol{\theta}_\varepsilon$. Optimal parameters are the most probable model parameters of $p(\boldsymbol{\theta}_d, \boldsymbol{\theta}_\varepsilon)$, which maximizes $p(\boldsymbol{\theta}_d, \boldsymbol{\theta}_\varepsilon)$. With a huge amount of data, using the sufficiently diffuse prior PDF will not influence the final results. The likelihood function for $\boldsymbol{\theta}_d, \boldsymbol{\theta}_\varepsilon$ is given by (8):

$$p(D_d \mid \boldsymbol{\theta}_d, \boldsymbol{\theta}_\varepsilon) = \frac{1}{(2\pi)^{\frac{N}{2}} \mid \Sigma \mid^{\frac{1}{2}}} \exp(-\frac{1}{2} \begin{bmatrix} \hat{Y}_1 - f(\mathbf{x}_1, \boldsymbol{\theta}_d) \\ \vdots \\ \hat{Y}_N - f(\mathbf{x}_N, \boldsymbol{\theta}_d) \end{bmatrix}^T$$

$$\Sigma^{-1} \begin{bmatrix} \hat{Y}_1 - f(\mathbf{x}_1, \boldsymbol{\theta}_d) \\ \vdots \\ \hat{Y}_N - f(\mathbf{x}_N, \boldsymbol{\theta}_d) \end{bmatrix}) \qquad (8)$$

where $\Sigma(\sigma^2, l, p) = [\Sigma_{ij}] = \sigma^2 \exp(-\left\| \frac{\mathbf{x}_i - \mathbf{x}_j}{l} \right\|_p), 0 < p \le 2;$

$\boldsymbol{\theta}_\varepsilon = [\sigma^2 \ l \ p]^T$.

An algorithm has been developed to obtain the most probable parameters. Interested readers can refer to the journal version of this paper.

The non-informative prior for all uncertain parameters is used in this paper. Given the proposed Gaussian process error function considered in this paper, correlation is considered between the past and present outputs *Y* and predicted future output, Y_{f_d}. The predicted future outputs depend on all the past and present conditions and GCM predictors in the past explicitly not just the model parameters and GCM predictors in the future. Another essential aspect is outputs for all days for the past, present, future are dependent. Prediction for all future days needs to be simultaneously calculated (not day by day independently). In this paper, the precipitation prediction follows a multivariate normal distribution with mean μ_{f_d} and variance Σ_{f_d}. The precipitation prediction for future wet days is obtained by sampling from the posterior PDF of the predicted future precipitation Y_{f_d} in *Q* future wet days given past data D_d, precipitation classification data D_c and GCM predictors $[\mathbf{x}_{f_d,1}\ \mathbf{x}_{f_d,2} ... \mathbf{x}_{f_d,Q}]$ for the future wet days given $\boldsymbol{\theta}_d, \boldsymbol{\theta}_\varepsilon$ is given by (9):

$$p(Y_{f_d} \mid \boldsymbol{\theta}_d, \boldsymbol{\theta}_\varepsilon, D_d, D_c) \sim N(\mu_{f_d}, \Sigma_{f_d}) \tag{9}$$

$$\mu_{f_d} = \begin{bmatrix} f_d(\mathbf{x}_{f_d,1}, \boldsymbol{\theta}_d) \\ \vdots \\ f_d(\mathbf{x}_{f_d,Q}, \boldsymbol{\theta}_d) \end{bmatrix} + \Sigma_{21}\Sigma_{11}{}^{-1} \begin{bmatrix} \hat{Y}_1 - f_d(\mathbf{x}_1, \boldsymbol{\theta}_d) \\ \vdots \\ \hat{Y}_N - f_d(\mathbf{x}_N, \boldsymbol{\theta}_d) \end{bmatrix};$$

$$\Sigma_{f_d} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}{}^{-1}\Sigma_{12}$$

where,

$$\Sigma_{11}(\boldsymbol{\theta}_\varepsilon) = [\sigma^2 \exp(-\left\| \frac{\mathbf{x}_i - \mathbf{x}_j}{l} \right\|_p)],\ 1 \le i, j \le N$$

$$\Sigma_{21}(\ \boldsymbol{\theta}_\varepsilon) = [\sigma^2 \exp(-\left\| \frac{\mathbf{x}_{f_d,i} - \mathbf{x}_j}{l} \right\|_p)],\ 1 \le i \le Q,\ 1 \le j \le N,$$

$$\Sigma_{12} = \Sigma_{21}{}^T$$

$$\Sigma_{22}(\ \boldsymbol{\theta}_\varepsilon) = [\sigma^2 \exp(-\left\| \frac{\mathbf{x}_{f_d,i} - \mathbf{x}_{f,j}}{l} \right\|_p)],\ 1 \le i, j \le Q$$

In the proposed method for Bayesian uncertainty analysis for stochastic classification and statistical downscaling with stochastic dependent error coupling (BUASCSDSEC), the posterior PDF of predicted future rainfall Y_{f_d} in future days given past statistical downscaling data, D_d, the past classification class data D_c and the future GCM predictors, $\mathbf{X}_*$ (By the Theorem of Total Probability) is given by (10):

$$p(\mathbf{Y}_{f_d} \mid D_d, D_c, \mathbf{X}_*) = \int p(\mathbf{Y}_{f_d} \mid \mathbf{y}_*, \mathbf{X}_*, D_d, D_c)\, p(\mathbf{y}_* \mid D_c, \mathbf{X}_*) d\mathbf{y}_* \tag{10}$$

where

$$\begin{aligned} &p(\mathbf{Y}_{f_d} \mid \mathbf{y}_*, \mathbf{X}_*, D_d, D_c) \\ &= \int p(\mathbf{Y}_{f_d} \mid \mathbf{y}_*, \mathbf{X}_*, \boldsymbol{\theta}_d, \boldsymbol{\theta}_\varepsilon, D_d, D_c)\, p(\boldsymbol{\theta}_d, \boldsymbol{\theta}_\varepsilon \mid D_d) d\boldsymbol{\theta}_d d\boldsymbol{\theta}_\varepsilon \end{aligned} \tag{11}$$

$p(\mathbf{y}_* \mid D_c, \mathbf{X}_*) = p(\mathbf{y}_* \mid \mathbf{X}, \mathbf{y}, \mathbf{X}_*)$ is given in Section *B* from the proposed Bayesian Gaussian Process binary classification model; $p(\mathbf{Y}_{f_d} \mid \mathbf{y}_*, \mathbf{X}_*, \boldsymbol{\theta}_d, \boldsymbol{\theta}_\varepsilon, D_d, D_c)$ is given by (9) for $\mathbf{y}_*, \mathbf{X}_*$ correspond to the future wet days.

IV. Results and Discussions

The dataset is divided into two sets; calibration data and validation data. The validation set results are provided in this section. The model prediction results for the month of December for the validation period is shown here. In BUASCSDSEC model, the marginal likelihood is approximated using Laplace approximation. The uncertainty in the model parameters, model structure is quantified and propagated from classification model to the downscaling model.

The proposed Bayesian Gaussian Process binary classification model is implemented to generate time histories samples which indicate which future days are wet and which days are dry before the downscaling of the GCM variables is carried out. The dataset was divided into two sets; calibration data and validation data. The proposed model is calibrated using the calibration data and the model is validated using the validation data.

Fig. 2 shows the comparison of cumulative distribution function (cdf) of the real observations and predicted samples obtained from the proposed method. From the figure, it is observed that the CDFs of the daily predicted samples are consistent with the observed CDF of precipitation. The uncertainty is smaller for the extreme rainfall and the uncertainty for smaller amount of rainfall is more compared to that for the extreme rainfall. One possible reason for this could be that the uncertainty from the proposed classification model affects the results in the downscaling model. The selection of regression based method as statistical downscaling model function could be another reason.

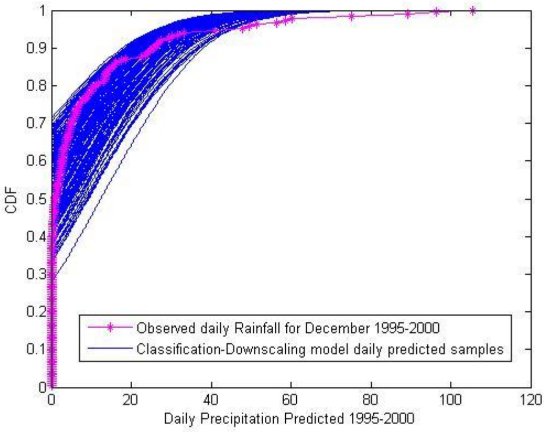

Fig. 2. Cumulative distribution function (cdf) of the real observations and predicted samples obtained from the proposed method.

Using the proposed BUASCSDSEC model, the different sources of uncertainty can be captured. Using Bayesian statistical framework, the likelihood for different models can be obtained. The results can be used for weighting different statistical downscaling models in multi model method for uncertainty quantification.

V. Conclusion

This paper presents a novel method for classification and statistical downscaling coupled with uncertainty quantification and propagation. Ongoing work includes the coupling of the proposed method with different forms of mean functions, for example, Bayesian Neural Network, support vector machine and generalized linear/nonlinear models etc. Multiple class classification instead of binary classification, different forms of sigmoid functions and covariance functions in the proposed methodology are currently under investigation. The results obtained from the proposed method will be compared with other commonly used models for downscaling and the comparative results will be provided in the journal version of this paper. The future scenario simulations will be provided. The uncertainty in GCM will also be quantified and propagated in the future work by a newly developed method by the authors for integrating multiple GCMs.

ACKNOWLEDGMENT

Earth Observatory of Singapore (EOS) Project (M4080707.B50, M4080891.B50 and M4430053.B50) supported this research work. The authors would like to thank National Environmental Agency (NEA) of Singapore for providing the data.

REFERENCES

[1] H. J. Fowler, S. Blenkinsop, and C. Tebaldi, "Linking climate change modelling to impacts studies: recent advancs in downscaling techniques for hydrological modelling," *International Journal for Climatology,* vol. 27, pp. 1547-1578, 2007.

[2] R. L. Wilby and I. Harris, "A framework for assessing uncertainties in climate change impacts: Low-flow scenarios for the River Thames, UK," *Water Resources Research,* vol. 42, pp. 1-10, 2006.

[3] S. Ghosh and C. Misra, "Assessing Hydrological Impacts of Climate Change: Modelling Techniques and Challenges," *The Open Hydrology Journal,* vol. 4, pp. 115-121, 2010.

[4] C. Tebaldi, L. O. Mearns, D. Nychka, and L. Smith, "Regional probabilities of precipitation change: A Bayesian analysis of multimodel simulations," *Geophysical Research Letters*, vol. 31, 2004.

[5] C. Tebaldi, "Quantifying uncertainty in projections of regional climate change: A bayesian approach to the analysis multimodel ensembles," *Journal of Climate,* vol. 18, pp. 1524-1539, 2004.

[6] M. Z. Hashmi, A. Y. Shamseldin, and B. W. Melville, "Statistical downscaling of precipitation: state-of-the-art and application of bayesian multi-model approach for uncertainty assessment," *Hydrology and Earth System Sceinces,* vol. 6, pp. 6535-6579, 2009.

[7] J. Chen, F. B. Brisstte, and R. Leconte, "Uncertainty of downscaling method in quantifying the impact of climate change on hydrology," *Journal of Hydrology,* vol. 401, pp. 190-202, 2011.

[8] B. Axel and N. Danieland B. Gerd, "Effects of climate and land-use change on strom runoff generation: present knowledge and modelling capabilities," *Hydrological Processes,* vol. 16, pp. 509-529, 2002.

[9] Beck and Katafygiotis, "Updating Models and Their Uncertainties. I: Bayesian Statistical Framework," *Journal of Engineering Mechanics,* vol. 124, pp. 455-461, 1998.

[10] S. H. Cheung and J. L. Beck, "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,* vol. 25, pp. 304-321, 2010.

[11] J. Ching and Y. Chen, "Transitional Markov Chain Monte Carlo Method for Bayesian Model Updating, Model Class Selection, and Model Averaging," *Journal of Engineering Mechanics,* vol. 133, pp. 816-832, 2007.

[12] S. H. Cheung, T. A. Oliver, E. E. Prudencio, S. Prudhomme, and R. D. Moser, "Bayesian Uncertainty Analysis with Applications to Turbulence Modeling," *Reliability Engineering & System Safety*, vol. vol. 96, issue 9, pp. 1137-1149, 2011.

[13] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, the MIT press, 2006, ch. 2-5, pp. 7-128.

[14] NEA (National Environmetal Agency), *Weather Wise Singapore*, 2009.

[15] Outline map of Singapore. (March 6, 2014). [Online]. Available: http://www.enchantedlearning.com/asia/singapore/outlinemap/

Queen Suraajini Rajendran is a PhD student in the School of Civil and Environmental Engineering at Nanyang Technological University (NTU), Singapore. She earned her bachelor's degree with distinction in geo informatics at Anna University, Chennai, India in 2011. Her research areas include uncertainty modelling and analysis of climate models and climate change impact studies. She also contributes part of her research work to the Earth Observatory of Singapore, NTU, Singapore.

Sai Hung Cheung is an assistant professor in the School of Civil and Environmental Engineering at Nanyang Technological University, Singapore. He obtained his PhD in civil engineering from the California Institute of Technology (Caltech) with a GPA of 4.1 (A+) in 2009. He got his bachelor degree with First-class honors in civil and structural engineering with a minor in mathematics and his master degree in civil engineering from the Hong Kong University of Science and Technology (HKUST). During the period of March 2009-August 2010, he was a postdoctoral fellow at the Institute of Computational Engineering and Sciences (ICES) at the University of Texas at Austin (UT Austin) where he was involved in a multidisciplinary research project. His research areas include catastrophe risk modeling, analysis, mitigation and management due to natural disasters and man-made hazards; reliability, risk engineering and science; stochastic dynamics; complexity science; earthquake engineering, performance-based engineering; sustainable urban planning and development; climate change impact studies; optimal decision making, design and control under uncertainty; uncertainty quantification, system identification and Structural health monitoring.