Universität Stuttgart

Auslandsorientierter Studiengang
Wasserwirtschaft
Master of Science Program
Water Resources Engineering and
Management - WAREM

Master's Thesis:

Uncertainty Estimation of Precipitation Covariance Functions and its Effect on Discharge Simulation using Random Mixing

submitted by :
Mustafa Ahmed Mustafa

Matriculation number:
399374

Date : **May 28, 2018**

Supervisor : **Prof. rer.nat. Dr.-Ing. Andras Bardossy**

Institut für Wasser- und Umweltsystemmodellierung
Lehrstuhl für Hydrologie und Geohydrologie
Prof. Dr. rer.nat. Dr.-Ing. Andras Bardossy
Pfaffenwaldring 61 D-70569 Stuttgart, Germany 70550 Stuttgart

Universität Stuttgart

Water Resources Engineering and Management

akkreditiert durch ASIIN
im Auftrag des Akkreditierungsrats (KMK / HRK)

Universität Stuttgart · WAREM · Pfaffenwaldring 7a · 70569 Stuttgart

Pfaffenwaldring 7
70569 Stuttgart
Telefon: (0711) 685 - 66615 / 66616
Telefax: (0711) 685 – 66600
warem@iws.uni-stuttgart.de
http://www.warem.uni-stuttgart.de/

Anne Weiss M.A., M.Sc.
(Durchwahl: - 66616)

Master Thesis Abstract

of Mustafa Ahmed Mustafa

Uncertainty Estimation of Precipitation Covariance Functions and its Effect on Discharge Simulation using Random Mixing

The covariance or the spatial correlation function plays a vital role in the Random Mixing technique for precipitation simulation, the uncertainties of the covariance function is assessed in this study, The Study was performed on the upper Neckar catchment, located in southwest Germany, specifically in the sub-catchment Horb, the catchment’s rainfall data were collected through rain gauge observations for a period of 15 years (2001 to 2015).Multiple covariance functions were fitted for the same data set, then the precipitation was simulated using Random Mixing, the data is then used as input for the Hydrologiska Byrans Vattenbalansavdelning (HBV) model that is calibrated using differential evolution The input data is created using a mix of external drift kriging and IDW.

The results of this study demonstrate that using multiple covariance functions with the same dataset can be helpful in estimating the discharge flow, the simulated values matched and even exceeded the recorded values, since this method doesn’t underestimate the precipitation as much as other interpolation methods, however, the covariance function can cause a relatively high degree of error when used with low precipitation dates, and using multiple covariance functions for a big dataset requires more computation time and power.

Acknowledgements

I would first like to thank my thesis advisor Faizan Anwer for providing the scripts I needed for this work. the door to his office was always open whenever I ran into a trouble spot or had a question about my research or writing. He consistently allowed this thesis to be my own work, but steered me in the right the direction whenever he thought I needed it.

Prof. Andras Bardossy for coming up with both the modified maximum likelihood method and the Random mixing technique, for without him this thesis wouldn't exist, and Dr. Sebastian Hörning for writing the Random mixing scripts.

Finally, I must express my very profound gratitude to my parents and to my friends and family for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them.

Thank you.

Mustafa, Mustafa

Contents

List of Figures

Chapter 1

Introduction

1.1 Hydrological Modeling

According to Devia et al. (2015), a Hydrological model is a simplified representation of real world system which provides results as close to reality as possible by using a set of equations that help in the estimation of runoff as a function of other watershed parameters.

The two most important inputs that are required for a hydrological model are precipitation data, temperature data, and drainage area. Along with these are other watershed characteristics like vegetation, temperature, and evapo-transpiration rates, topography, soil moisture content, characteristics of ground water aquifer are also considered.

Throughout the years, a large number of hydrological models that ranges from small catchment models up to global models have been developed, these models are mainly used for predicting system behaviour and understanding various hydrological processes.

There are many advantages of using Hydrological models, these models help in flood forecasting, proper water resource management and evaluation of water quality, erosion and sedimentation, nutrient and pesticide circulation, land use and climate change etc. but there are also some drawbacks to using these models, like lack of user friendliness, large data requirements, absence of clear statements of their limitations (Devia et al., 2015)

1.1.1 Why Modeling?

There are many reasons as to why one would need to model hydrological processes, and due to the complexity of real world hydrological systems, our hydrological measurement techniques are very limited, therefore a more sophisticated method of extrapolation of information from those measurements are needed, particularly in catchments where measurements are either unavailable or inaccurate, or to assess the impact of future hydrological changes where measurements are impossible.

Generally, hydrological models are used for research purposes to formulate knowledge about hydrological systems in the real world, (Beven, 2012) argues that we learn most when a model or theory is shown to be in conflict with reliable data so that some modification of the understanding on which the model is based must be sought. and that due to increasing global demands on water resources, the ultimate aim of prediction using models must be to improve the decision making process about a hydrological problem.

1.1.2 Types of Models

according to (Devia et al., 2015), Hydrological Models can be classified as either event based or continuous, where event based models can only produce outputs for a specific time period.

The aforementioned models can also be sub-classified as lumped or distributed models, where lumped models treats the entire basin as a single unit without any regard to spatial variability, while distributed models divides the catchment or basin into cells or polygons so the parameters can vary spatially.

These models can also be sub-sub-classified as stochastic and deterministic models, where deterministic models will generate the same output for a single set of inputs, while stochastic models can generate different values of output from a single set of inputs.

Some of the most important classifications are empirical model, conceptual models and physically based models.

Empirical models

Empirical models are observation oriented models and represent real systems with mathematical explanation derived from input and output time series without any consideration to the physical processes of the catchment, hence, the models are called data driven models. Empirical models are only valid within the boundaries of the catchment (Devia et al., 2015).

Empirical models use non-linear statistical relationships between inputs and outputs, Unit Hydrograph is an example of this method. The general equation for the empirical models is a function of inputs:

$$Q = f(x) \tag{1.1}$$

where:

- Q: runoff output

- x: input datasets of rainfall and historic runoff.

The function used to transform rainfall to runoff is either an unknown procedure (as in machine learning) or without any reference to the physical processes.

Empirical models are best used when other outputs are not needed, also due to lack of information about un-gauged watersheds, in many cases, empirical models can provide accurate simulations including long time steps and recreating past run-off values (EPA, 2017)

Conceptual models

Conceptual models connect simplified components of hydrological processes, it consists of a number of connected reservoirs to provide a conceptual idea of the catchments behaviour.

Conceptual models represent the water balance equation with the conversion of rainfall to runoff, evapotranspiration, and groundwater. Each component in the water balance equation is estimated by mathematical equations that distributes the precipitation input data. The general governing equations for conceptual models are versions of the water balance equation which control surface water and storage fluctuations shown below : (EPA, 2017).

$$\frac{dS}{dt} = P - ET - Q_s \pm GW \tag{1.2}$$

Conceptual rainfall-runoff models are mostly used for resource planning, and the models can vary in complexity, but conceptual they provide good estimates of flows in gauged and un-gauged catchments provided that the needed data is available (eWater CRC, 2011)

Physically based models

This is a mathematically idealized representation of the real phenomenon. These are also called mechanistic models, they are based on the physics related to hydrological processes. physically based models are used to represent real hydrological responses in the catchment (Devia et al., 2015).

Many physics laws are used in this model, these include: water balance equations, conservation of mass and energy, momentum, and kinematics, St. Venant, Boussinesq's, Darcy and Richard's...etc. Therefore, the models require a huge amount of data to be available, such as: Soil moisture content, initial water depth, topography, topology, dimensions of river network etc. (EPA, 2017).

Physically based models can provide more information (even outside the boundary) than the two other models due to the paramaters having physical interpretation, and can also be applied to a wide range of situations. (example: SHE / MIKE SHE model).

1.2 Motivation

> “ The future in rainfall–runoff modelling is therefore one of uncertainty: but this then implies a further question as to how best to constrain that uncertainty. The obvious answer is by conditioning on data, making special measurements where time, money and the importance of a particular application allow. (Beven, 2012)”

Due to the complexity of hydrological systems, all models of these systems are prone to a certain degree of uncertainty, these uncertainties can arise from:

1. Input uncertainty
2. Parameter uncertainty
3. Conceptual uncertainty

input uncertainties can be caused by: the density of measurements, transmission errors, temporal resolution errors, as well as precipitation which is one of the main sources of input uncertainty in hydrological modelling.

there are many different factors that contributes to parameter uncertainties, a few examples:

- the degree of spatial differentiation of the investigated area
- the chosen parameter estimation approach
- the applied calibration procedure
- the type and number of verification points

while conceptual uncertainties are generally caused by the modeller himself, mainly due to the lack of information regarding the catchment or basin, or even the approach chosen by the modeller for process description. (Pluntke et al., 2014)

With the introduction of digital computer models, and the use of more complex spatial analysis methods, the spatial correlation or the covariance function emerged as one of the major sources of uncertainty, this is due to the fact that the covariance function gives the statistical correlation between random variables depending on the spatial distance between those variables, the following chapters the uncertainties introduced by the covariance function and the effects of said uncertainties on the rainfall run-off models are investigated.

1.3 Scope and structure of this thesis

This thesis is based on the techniques and methods described in Bárdossy and Li (2008), and Hörning (2016), Chapter 1 introduces the basics of hydrological modelling, while Chapter 2 discusses the basic methods and interpolation methods used to prepare the input data for Random mixing.

Chapter 3 describes the methodology of the work, the data selection process, grid sampling, and the kernel density estimation method, as well as the the variogram fitting process and the discharge simulation. The results of the simulations are analysed and discussed in Chapter 4

Chapter 2

Literature Review

2.1 HBV Model

The HBV model is a semi distributed conceptual model where catchments are divided into sub-basins as primary hydrological units, and within these an area-elevation distribution and a crude classification of land use (forest, open, lakes) are made.

The model explained in detail in (Bergström, 1992) uses daily and monthly precipitation data, evaporation and air temperature. the latter is used to calculate snow accumulation. If potential evaporation data are lacking, monthly evaporation estimates based on, forinstance, the Thornthwaite equation can be used (Rusli et al., 2015).

The snow, snow-melt, and snow accumulation are calculated by a degree-day method, while the groundwater recharge and evaporation are calculated as a function of actual water storage, and runoff is calculated as a function of water storage. Precipitation in HBV is modelled as either snow or rain depending on the temperature threshold TT, snowfall is multiplied by a correction factor C_{SF} to compensate for errors in measurements and evaporation, while snow-melt $MELT$ is calculated using the degree-day factor (see equation 2.1).

$$MELT = C_{MELT}.(T(t) - TT) \tag{2.1}$$

where:

- $MELT$ = snow-melt (mm/day)

- C_{MELT} = degree day factor (mm/C.day)
- TT = Threshold temperature (C)

The second routine of the HBV model is the soil moisture or evapo-transpiration routine, which computes an index of wetness for the basin and integrates interception and soil moisture storage which is controlled by 3 parameters, the first parameter FC is the maximum soil moisture storage, and the relative contribution to runoff from precipitation or snow melt at a given soil moisture deficit is controlled by parameter Beta, while parameter LP controls the potential evaporation curve and the evapo-transpiration rates.

the soil moisture routine uses a correction factor to account for temperature anomalies as shown in equation 2.2

$$PE_A = (1 + C.(T - T_M)).PE_M \tag{2.2}$$

where:

- PE_A = adjusted potential evapo-transpiration
- C = empirical model parameter
- T = daily mean air temperature
- T_M = monthly long term average temperature
- PE_M = monthly long term average potential evapo-transpiration

The final routine is a runoff response routine which transforms excess water (δQ) from the soil moisture routine to discharge it in each sub-basin, the model consists of 2 reservoirs, three recession parameters, K_0, K_1, K_2, Threshold L, Percolation rate, and a triangular filter (see equations 2.3 2.4 2.5 2.6)

$$Q_0 = K_0(S_1 - L_1) \tag{2.3}$$

$$Q_1 = K_1 S_1 \tag{2.4}$$

$$Q_2 = K_2 S_2 \tag{2.5}$$

$$Q = T(Q_0 + Q_1 + Q_2) \tag{2.6}$$

Although the HBV model was developed in Europe, it has also produced relatively accurate results in different climatic conditions. There are several other strengths of the HBV model, its physically based parameters, which are useful due to the simplicity of linking them to physical attributes; the unexcessive number of free parameters as compared with other models (the HBV model has only eight parameters, while the Sacramento model, Xinanjiang model, NedbØr-AfstrØmnings (NAM) model, and Pitman model have 21, 15, 15, and 16 parameters, respectively, simple data demands; user-friendliness; ease of operation; and high level of performance (Rusli et al., 2015)

2.2 Inverse distance weighting

Inverse distance weighting (IDW) is a widely used deterministic method, it is based on Tobler's first law "everything is related to everything else, but near things are more related than distant things", and it applies to geographical space for interpolation. in hydrology, it means that the attribute value of an ungauged site is the weighted average of the known attribute values within the neighbourhood, and weights are associated with the horizontal distances between the gauged and un-gauged sites (Waseem et al., 2016)

Due to the IDW being a univariate method with only the horizontal distance as an influence factor, there are some drawbacks associated with the performance of this method, the main drawback being that a direct application of the interpolation method in geographical space might cause unrealistic results.

Another use for IDW is in GIS software packages, it is based on the assumption that the attribute value of an unsampled point is the weighted average of the known values within the neighborhood. The method involves the process of assigning values to unknown points using values from a scattered set of known points. The value of an unknown point is a weighted sum of the values of the known points.

Waseem et al. (2016) explains that from a hydrological point of view, if m source sites (i.e., known points) transfer information to an unknown point (i.e., ungauged site), the required streamflow value at the ungauged site can be computed as the weighted average of the estimates of the m source sites. The computation of the required streamflow value for an ungauged site can be obtained using the IDW equations as follows:

$$Q_P(x) = \sum_{k=1}^{m} \frac{w_k}{\sum_{k=1}^{m} w_k} Q_P(w_k) \tag{2.7}$$

$$w_k = \frac{1}{(D_{IDW(x,x_k})^C} \tag{2.8}$$

$$D_{IDW(x,x_k} = d_{DWS} = \sqrt{(x_k - x)^2 + (y_k - y)^2} \tag{2.9}$$

- $D_{IDW(x,x_k)}$: distance between two sites.
- IDW subscript in $D_{IDW(x,x_k)}$: the method used.
- d_{DWS}: is the distance based on the geographical distance weighted scheme DWS.
- $Q_p(x)$: the hydrological variable at the ungagued site located at point (x,y).
- $Q_p(x_k)$: the hydrological variable at the neighboring donor site k at point (x_k, y_k) in the region.
- m: the total number of donor site.
- c the power parameter.
- w_k the interpolation weight assigned to the k^{th} donor site.

2.3 External Drift Kriging

External drift kriging is a case of normal kriging which allows the prediction of the variable Z (known only partially in the study area) modelled with a random function Z(x) through a deterministic variable S(x) where variable s is known in that area, with the assumptions:

- The two quantities Z and S are assumed to be linearly related
- Z(x) on avaerage is equal to s(b) up to a constant a_0 and a coefficient b_1

$$E[Z(x)] = a_0 + b_1 s(x) \tag{2.10}$$

the function $s(x)$ provides a finer detail than the sample spacing of $Z(x)$, and the predictor is a linear combination of the sample values at location $x_i (i = 1, ..., n)$ with unit sum weight w_i

$$Z^*(x_0) = \sum_{i=1}^{n} w_i Z(x_i) \tag{2.11}$$

where:

$$\sum_{i=1}^{n} w_i = 1$$

it is also assumed that the expectation of the predictor is zero so that

$$E[Z(x_0)] = E[Z^*(x_0)] \tag{2.12}$$

then the equality can be:

$$E[Z^*(x_0)] = \sum_{i=1}^{n} w_i E[Z(x_i)] = a_0 + b_1 s(x_0) \tag{2.13}$$

which means that the weights should be constant with the interpolation of $s(x)$

$$s(x_0) = \sum_{i=1}^{n} w_i s(x_i) \tag{2.14}$$

the function O is minimised

$$O = \sigma_E^2 - \mu_1 (\sum_{i=1}^{n} w_i - 1) - \mu_2 (\sum_{i=1}^{n} w_i s(x_i) - s(x_0)) \tag{2.15}$$

where the prediction variance σ_E^2:

$$\sigma_E^2 = var[Z^* - Z] = \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j C(x_i - x_j) - n\sum_{i=1}^{n} w_i C(x_i - x_0) + C(0) \tag{2.16}$$

where C is the Covariance function

The partial derivatives of the function $O(w_i, \mu_1, \mu_2)$ is set to zero to find the minimum and the kriging equations are:

$$\begin{cases} \sum_{j=1}^{n} w_j C(x_i - x_j) - \mu_1 - \mu_2 s(x_i) = C(x_i - x_0) for i = 1, 2, ..., n \\ \sum_{j=1}^{n} w_j = 1 \\ \sum_{j=1}^{n} w_j s(x_j) = s(x_0) \end{cases}$$

with the prediction variance:

$$\sigma_E^2 = C(0) - \sum_{i=1}^{n} w_i C(x_i - x_0) + \mu_1 + \mu_2 s(x_0) \tag{2.17}$$

The external drift kriging means incorporating additional conditions about one or more drift variables $s_i(x), i = 1..., M$ where the function $s_i(x)$ must be known at all locations x_i of $Z(x_i)$

this method can only be applied when the two variables are linearly related, and the function should be used to transform the data of the auxiliary variable which can then be used as an external drift (Bourennane et al., 2000).

2.4 Random Mixing

Random mixing is a copula based simulation approach first introduced in Bárdossy and Hörning (2016) which is an extension of Gradual deformation by (Hu 2000), where spatial fields are generated from a linear combination of independent random fields, where weights are selected to satisfy the required linear constraints, then an infinite number of solutions are generated inside a vector space, where non linear constraints can be incorporated via optimization.

2.4.1 Simulating daily precipitation with Random Mixing

Hörning (2016) investigated the use of Random mixing for daily precipitation simulation in a high spatial resolution in order to assess the uncertainty of rainfall runoff simulations for a time period of 62 years in the upper Neckar catchment in south-west Germany, and used a distribution of precipitation amounts $F_{t_j}(z)$ as shown below:

$$F_{t_j}(z) = \begin{cases} p_0 & z = 0 \\ p_0 + (1 - exp(-\lambda z)) & 0 < z \leq L \\ p_L + (1 - P_L)G_{t_j}(z) & L < z \end{cases}$$

where:

- $F_{t_j}(z)$: the daily distribution of precipitation amount
- p_0 : the discrete probability of zero
- $p_0 + (1 - exp(-\lambda z))$: exponentially distributed precipitation amount for the wet days with precipitation amounts below a selected threshold L
- $p_L + (1 - P_L)G_{t_j}(z)$: distribution of the precipitation amounts exceeding the threshold L

Hörning (2016) explains that there are several reasons for splitting the function into three parts, first, the probability of 0 precipitation has to be treated separately, second, the precipitation amounts (lower than the threshold L are very frequent and they area measured with a relatively high degree of observation errors, they can distort the estimation of the whole distribution and must be treated separately, lastly, the precipitation amounts above the threshold L have skewed distributions, hence, they have to be treated separately as well.

in conclusion, Hörning (2016) discovered that Random mixing provides a reasonable representation of the uncertainty in daily precipitation simulation since 84% of the investigated days were not rejected by a Kolmogorov-Smirnov test on the 90% level, which is acceptable, however, it was pointed out that this procedure can only be performed on *wet days* since the probability of zero observations cannot be identified.

2.4.2 Using Random Mixing with incomplete records

Hörning (2016) also shows that random mixing can be used in situations where observation records are incomplete, or suffer from a high degree of inaccuracy, the standard approach is to remove the affected stations, but random mixing is believed to provide a solution for this problem, and in order to prove this theory, an experiment was conducted where 311 precipitation stations in the state of Baden-Württemberg were selected, and 30 of those stations were randomly selected, and 5 days of precipitation data were removed, and 500 conditional simulations are performed as follows:

- first, monthly precipitation is calculated using daily values.
- the spatial field must be transformed to a multi-normal field Z prior to the Random Mixing procedure using the equation:

  $$Z(s) = \phi^{-1}(F(W(s))) \tag{2.18}$$

  where:

  - F(W): the univariate marginal distribution of the field W applied for each location s
  - ϕ^{-1}: the inverse univariate. standard normal distribution
- the spatial correlation function of the normalized values is estimated using the modified maximum likelihood method described in more detail in section 3.4
- the simulations are then performed and the values are back transformed into precipitation.

- the mean of the simulations can be considered as interpolated monthly precipitation.

Hörning (2016) created a second model based on the common approach (by removing the stations with missing data entirely) and compared it to the results of the above simulations and also to the results of the original data, and it was shown that the simulation model has more similarities to the original model than the model with the omitted stations, hence proving that random mixing is a valuable tool when simulating catchments with missing or inaccurate data.

2.5 The Covariance Function

The estimation of the covariance function using a modified maximum likelihood method which is based on spatial copulas was first introduced in Bárdossy (2011) where it was used to spatially interpolate observation values with concentrations below the sensitivity of measuring devices.

The method can be viewed in detail in Bárdossy (2011), however, the covariance function estimation equations are summarized below:

$$L(\beta) = \prod_{(j,k)\in I_1} \phi_2\Big(y_j, y_k, R(h_{j,k},\beta)\Big) \prod_{(j,k)\in I_2} \Phi_1\left(\frac{y_j^d - y_k R(h_{j,k},\beta),\beta)}{\sqrt{1-R(h_{j,k},\beta)^2}}\right) \prod_{(j,k)\in I_3} \Phi_2\Big(y_j^d, y_k^d, R(h_{j,k},\beta)\Big) \tag{2.19}$$

where:

- $R(h_{j,k},..)$ comes from the correlation matrix Γ:

$$\Gamma = \Big((\rho_{i,j})_{l,l}^{n,n}\Big)$$

  where $\rho_{i,j}$ depends on the vector h separating the points x_i and x_j:

$$\rho_{i_j} = R(x_i - x_j) = R(h_{i,j})$$

- $R(..,\beta)$: the correlation function is assumed to have a parametric form with the parameter vector β

- y_k, y_j^d: the observed values transformed to the standard normal distribution using:

$$y_k = \Phi_1^{-1}(G(z(x_k))) \qquad k = 1, ..., n_z$$

$$y_j^d = \Phi_1^{-1}(G(d(x_j))) \qquad j = 1, ..., n_d$$

- Φ_1: the distribution function of the standard normal distribution N(0,1).
- $\Phi_2(x, y, r)$: the distribution function of the 2 dimensional normal distribution with correlation r and standard normal marginal distribution N(0,1).
- $\phi_2(x, y, r)$ is the density function of Φ_2
- I_1: a set that contains pairs of locations with both variables (j,k) from measured values.
- I_2 a set that contains pairs of one measured ppt value and one non-detect value.
- I_3 a set that contains pairs of both non-detect values.

Bárdossy (2011) used this method to interpolate groundwater quality parameters with values below the detection limit, or non detects as described in the paper, and compared it with the more commonly used methods of interpolation such as ordinary kriging and indicator kriging and found that the copula based interpolation is exact at the observation locations since the interpolated value was equal to the observed value, and that it outperformed the other interpolation methods, where non-detect values were set to below the detection limit, equal to the detection limit, or zero, where both methods lead to systematic errors.

Chapter 3

Methodology

Simulating precipitation using random mixing is a multi-step process that starts by selecting the data, then the data is adapted to a map, distributions are fitted to the input data, random mixing process is then executed and the results are then modelled in HBV, the steps in this chapter were performed using python scripts that were collected from Anwar (2016) and Hörning (2016), the scripts were to meet the requirements of this thesis, and the results are mentioned in Chapter 4

3.1 Data Selection

The Study was performed on the upper Neckar catchment, located in southwest germany, specifically in the subcatchment horb (see figure 3.1), with a mean elevation of 550 m.a.s.l., a long term average daily temperature of 8.1 (C^o), and an annual precipitation of approximately 908 mm, the area of the subcatchment is 420.18 km^2.
The catchment rainfall data were collected through rain gauge observations for a period of 15 years (2001 to 2015) , Station numbers, date and time, and the precipitation values were extracted from its original files along with their coordinates.

3.2 Grid Sampling

In grid sampling the catchment is divided into cells and only one station in the cell is selected, which is usually the station closest to the centre of this cell, and if other stations are present in the cell, it is selected as a validation station so that it would minimise the uncertainties in the simulation as much as possible.

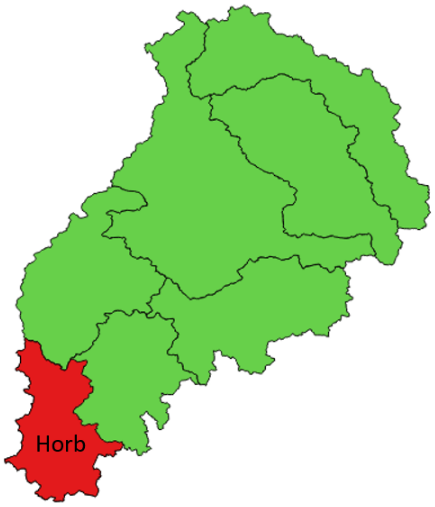

Figure 3.1: Subcatchment Horb in the upper Neckar catchment

Data such as precipitation and temperature from measuring stations are usually represented by a single point on the map, the data can be considered constant within the cell, however, cell size can vary from model to model, a larger celled model may experience significant and abrupt changes in properties between cells.

In this work, the catchment shape file was processed with GDAL Python API, and it was divided into cells of approximately 2 kilometres, and a maximum cell threshold of 0.9 is assigned (where a station farther from the centre are not considered for grid sampling), a coordinates matrix is then created with the threshold distance adapted into the model and is used to either include or exclude stations in grid sampling.

3.3 KDE Fitting

KDE fitting or Kernal density estimation, is a method of estimations of curves, a technique to estimate the unknown probability distribution of a random variable, based on a sample of points taken from that distribution. The target is estimating the probability density function of the variable using kernels, hence the name. it is also used in statistical applications to show the basic properties of a dataset such as skewness, dispersion...etc.

the general equation for the kernal density estimation is:

$$f_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - X_i}{h}) \tag{3.1}$$

where:

- f: a probability density function
- K: the kernel (or the chosen density function)
- h: the smoothing parameter (bandwidth)
- n: the total sample size

however, before going through with the kernal density estimation, it is important to remember the distribution of precipitation amounts equation mentioned in section 2.4.1, where precipitation is divided into 3 different parts, one distribution for the discreet probabilty of zero, a second distribution for precipitation below the selected threshold, and the last distribution for precipitation amounts exceeding the threshold.

the KDE method used in this case was the multivariate KDE method from the statsmodel API in python, the multivariate KDE differs slightly from equation 3.1 and is explained in more detail in Gramacki (2018)

In this step, since we are only interested in days with high precipitation values, a minimum precipitation threshold value is introduced and a list of dates with precipitation below the threshold value which are not used for Random Mixing is generated, the values for those dates are taken from a previously interpolated values using kriging with external drift, and if those values were not available then they are taken from IDWs interpolations.

3.4 Variogram fitting

The covariance function (or the spatial correlation function) is estimated using the modified maximum likelihood method detailed in Bárdossy (2011), The method is based on spatial copulas, the marginal distributions of parameters is first estimated using a mixed maximum likelihood approach, then the parameters of spatial dependence are estimated using the maximum likelihood estimation method

Precipitation data are usually highly skewed due to the high presence of 0 precipitation days (or non-detects) or precipitation with very low values (below the threshold), which makes their interpolation difficult for hydrologists, and often lead to problems with covariance function estimation, these effects can usually be reduced by data transformations, however, sometimes these transformations cannot be used in a straightforward manner, and the empirical distribution function of observations must be calculated, however, sometimes the mean and standard deviation cannot be calculated directly and the estimation of parameters θ of a selected parametric distribution via method of moments is not possible, and the maximum likelihood method is required. (Bárdossy, 2011)

the main part in this step is identifying the spatial structure from the available data, in cases with missing or inaccurate data, spatial structure identification can be difficult, as it may cause a reduction in variance, neglecting the data on the other hand can lead to an over estimation of the variance, which in turn leads to an underestimation of the spatial dependence.

The uncertainties in this step come from the fact that the selected pairs for the sets I_1, I_2, I_3 are chosen randomly, so that every time the method is executed a different covariance function will be generated, which could drastically alter the model outputs, to demonstrate, a sample is selected as inputs for the modified maximum likelihood method detailed in section 2.5 and the method was executed 10 times and produced the covariance functions shown in table 3.2.

0.550276	1.414677	-1.38702	-0.22126	-1.56328	1.74E+00	1.29E+00	2.66E-01
0.569114	-0.19786	-0.21073	-0.05583	-1.03474	1.16E+00	1.27E+00	6.08E-01
0.913154	0.704211	0.101199	0.396723	-0.02546	8.10E-01	1.07E+00	4.95E-01
1.387913	1.172007	0.013695	-0.25453	-0.20782	2.24E+00	7.95E-01	8.11E-01
0.12916	0.567426	-0.20613	-1.13397	-0.94693	9.58E-01	8.69E-01	7.80E-01
0.869099	1.377639	-1.24891	0.655947	-0.52958	8.98E-01	1.54E+00	2.92E-01
-0.20284	1.580683	0.436823	-0.12494	-0.72897	1.88E+00	4.17E-01	-4.74E-01
0.357916	1.049939	0.268917	-0.25639	-1.24891	1.24E+00	1.39E+00	2.14E-01
1.096648	0.436918	0.336577	-0.39746	-0.12701	1.05E+00	1.61E+00	-2.62E-01
2.00531	0.706083	-1.56328	-0.43073	-1.24891	1.11E+00	1.88E+00	-2.96E-01

Table 3.1: Transformed Data Sample

0.78890 Sph(34.59843)
0.54462 Exp(39.85648)
0.71240 Sph(31.59250)
0.49765 Sph(93.86109)
0.58260 Exp(37.31967)
0.77250 Sph(63.76907)
0.74940 Sph(49.80326)
0.76765 Sph(72.61200)
0.45713 Sph(51.91685)
0.70547 Sph(66.81781)

Table 3.2: Fitted Variograms

the above procedure is used to estimate the covariance function, and is executed multiple times to estimate 20 different covariance functions for each day for a period of 15 years (2000 - 2015), the covariance functions are then fitted.

3.5 Random mixing and HBV modeling

Up until this step, everything that has been done was to prepare the data for the random mixing process, Random Mixing is based on linear combinations of unconditional spatial random fields where the corresponding weights of the linear combination have to be selected such that certain predefined linear constraints are fulfilled. It uses spatial copulas as spatial random function; thus Gaussian as well as non-Gaussian spatial dependence structures and arbitrary marginal distributions can be considered. where unconditional fields are used to solve 10 conditional fields per covariance function per day, which produces 200 realisations for each day for a period of 15 years, since it is a lumped model, the realisations are converted into a single value and a time series of those values is created.

The data from the time series are then looped into an HBV model that is calibrated with external drift kriging data in order to simulate the discharge of station 411, the results of those simulations are shown and discussed in chapter 4

It is also worth mentioning that a buffer distance of 20 km was introduced to the model in order to include the stations that are adjacent to the catchment since they can have significant effect on the outputs of this model.

Chapter 4

Results and Discussion

As mentioned in section 1.2, all rainfall-runoff models are prone to a certain degree of uncertainty, which in this case can either be input uncertainty from infilling, or parameter uncertainty like the chosen covariance function, or conceptual uncertainty like uncertainties in the chosen random fields, in this chapter, these uncertainties are evaluated. the year 2007 was chosen as a case study year (see figure 4.1) since precipitation events with varying intensities were observed, ranging from low precipitation events (below 5mm) to high precipitation events (above 35mm).

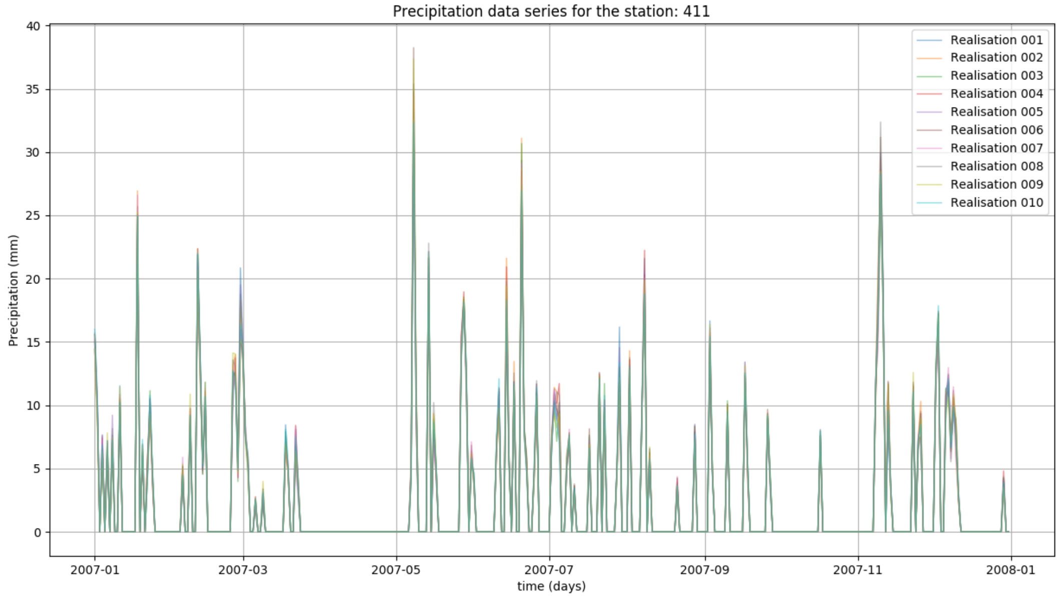

Figure 4.1: Precipitation data for the year 2007

4.1 Uncertainties in Random Mixing

In order to evaluate the uncertainties in Random Mixing, the data should be isolated as much as possible from other sources of uncertainties, and due to the fact that a different covariance function is fitted in each iteration, the realisations of each iteration must be evaluated separately, in addition, low precipitation, and high precipitation events should also be evaluated separately in order to further understand the behaviour of the random fields under different circumstances.

4.1.1 Low Precipitation Events

Low precipitation events were observed throughout different times in the year 2007, the lowest simulated event occurred on the 6^{th} of March, 2007, all 20 iterations were analysed, and while realisations may vary a little in each iteration, it was in iteration 3 where the realisations were closer together with a difference of $0.49mm$ between the maximum and minimum realisation values (see figure 4.2) while the maximum difference in realisation values occurred in Iteration 11 (see figure 4.3) with a difference of $1.7mm$.

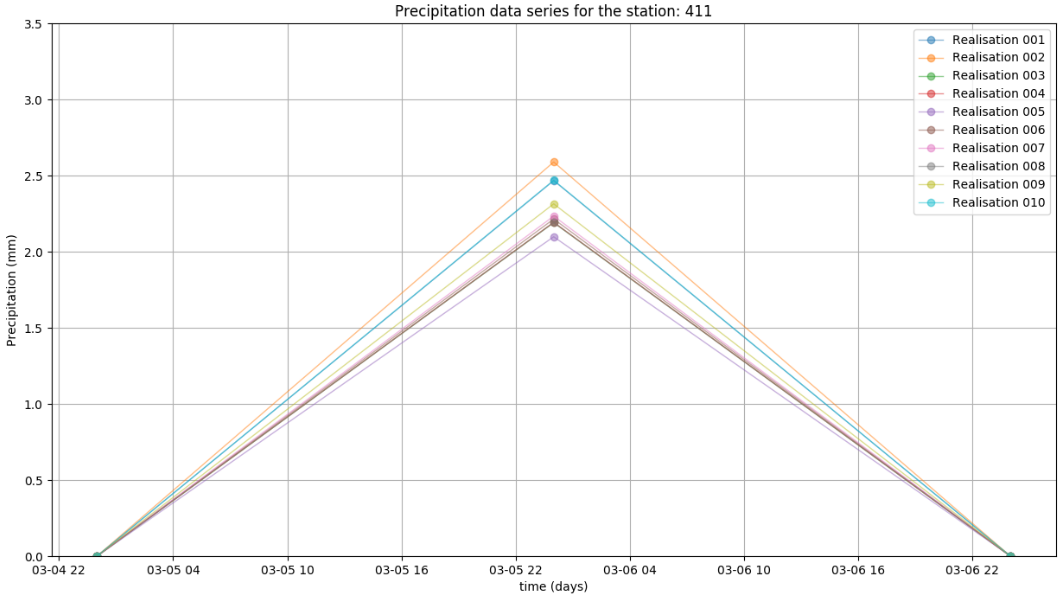

Figure 4.2: Realisations March 6, 2007 - Iteration 3

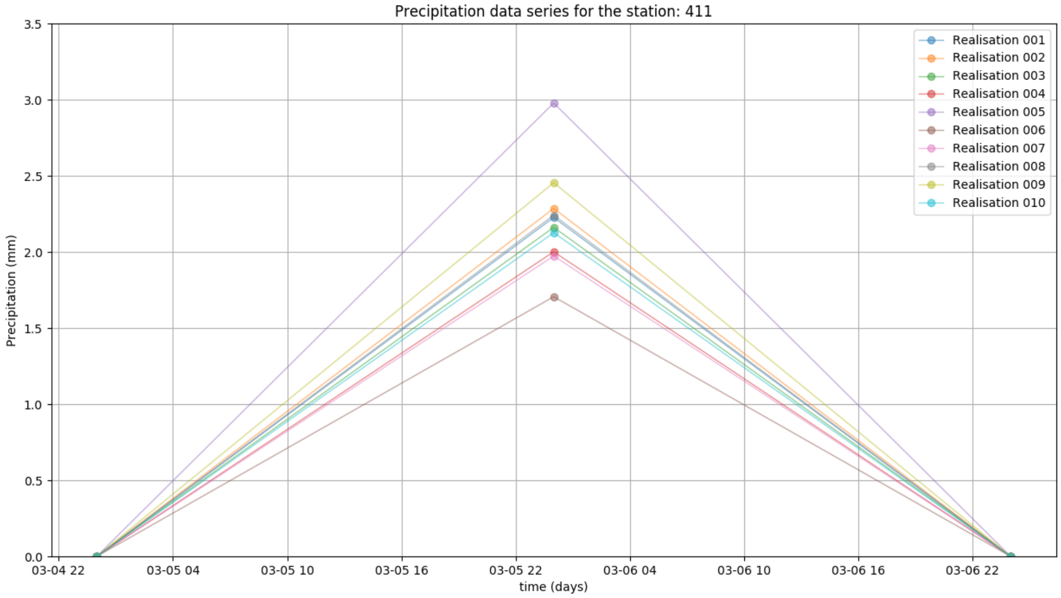

Figure 4.3: Realisations for March 6, 2007 - Iteration 11

4.1.2 High Precipitation Events

The highest precipitation event in 2007 peaked on the 8^{th} of May 2007, with precipitation above $30mm$, the data for all iterations were analysed, and iteration 20 that witnessed the highest range in values with $12.75mm$ difference between the highest and lowest realisation value, while the lowest difference in realisation values was observed in iteration 7 with only $5.25mm$

4.2 Uncertainties in the Covariance Function

The covariance functions or the spatial correlation functions cannot be directly compared with each other, averaging the realisation and plotting them leads to a reduction in variance and it does not provide much information, however, in figures 4.6, and 4.7 the first realisation of each iteration is plotted and compared for both, the low precipitation and the high precipitation events. While the CDFs for those dates are also shown in figures 4.8 and 4.9.

When comparing figures 4.5 and 4.7, where realisations of one iteration and the first realisation of each iteration are plotted respectively, it can be noticed that the plots are

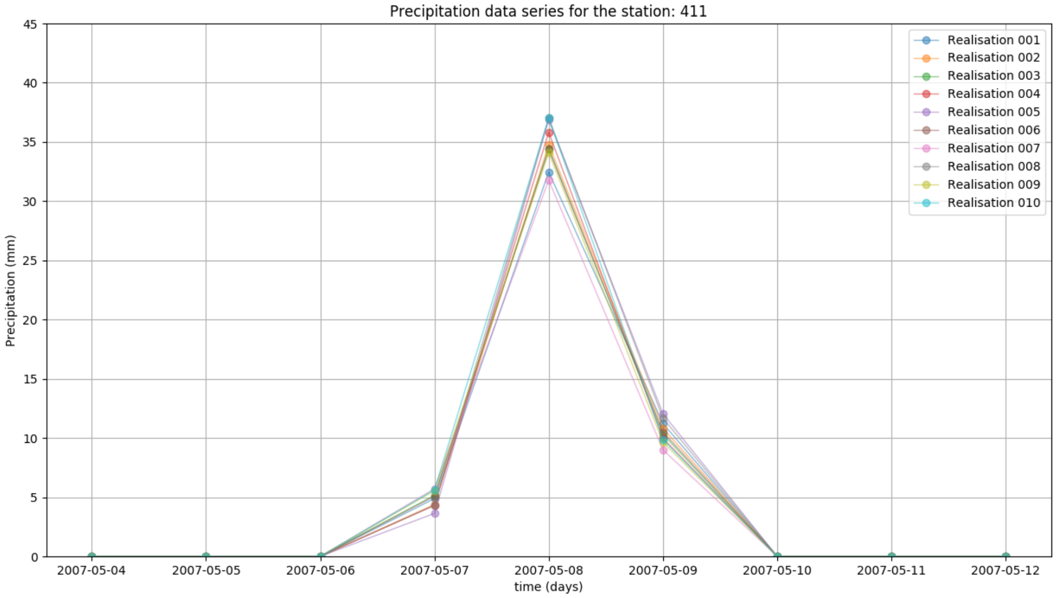

Figure 4.4: Realisations for May 8, 2017 - Iteration 07

very similar, with the only difference being that in the latter, the points are more or less equally spaced while they look more random in the realisations plot, which proves that a better method of comparison.

The cumulative distribution function figures 4.9 and 4.8 provides a more detailed view on the realisations, in those figures, all realisations from all the iterations are plotted, and upon examining the figures, it's clear that the largest concentration of values are more towards the centre of the plot (example: between 30 and 38 millimetres per day on the 8^{th} of may, and between 1.8 and 2.8 millimetres per day for the 6^{th} of march) however, due to the randomness of the covariance function, one could end up with higher concentration of values above or below the aforementioned numbers if only one covariance function was used.

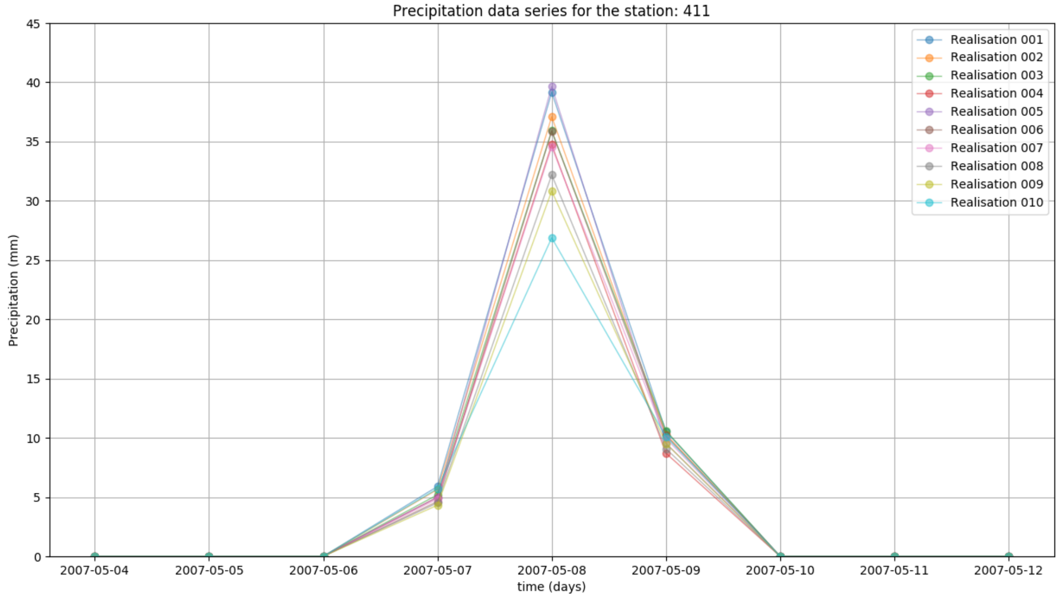

Figure 4.5: Realisations for May 8, 2017 - Iteration 20

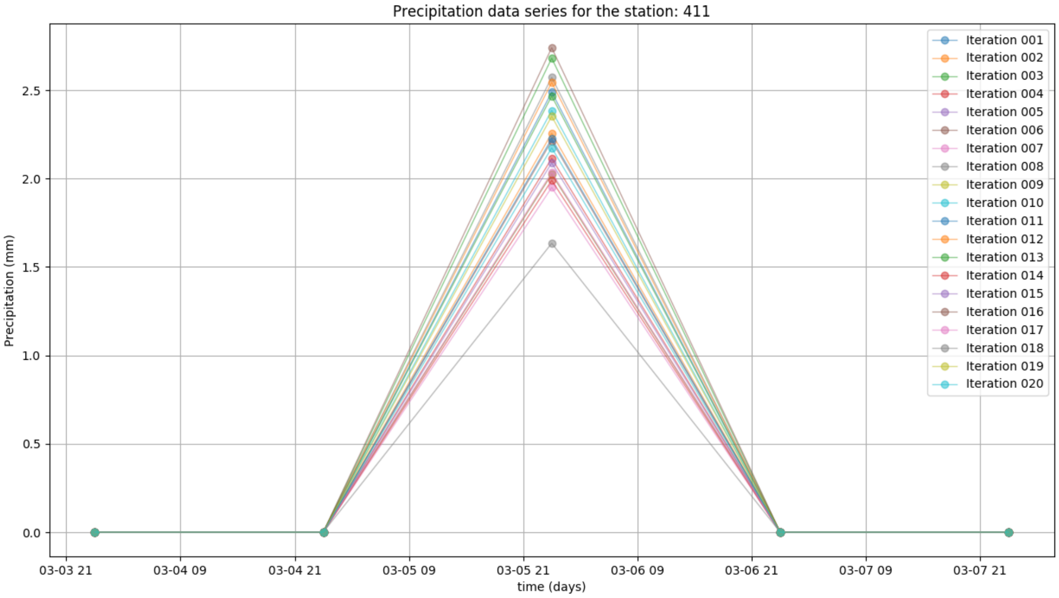

Figure 4.6: Simulated precipitation iterations, March 6, 2007

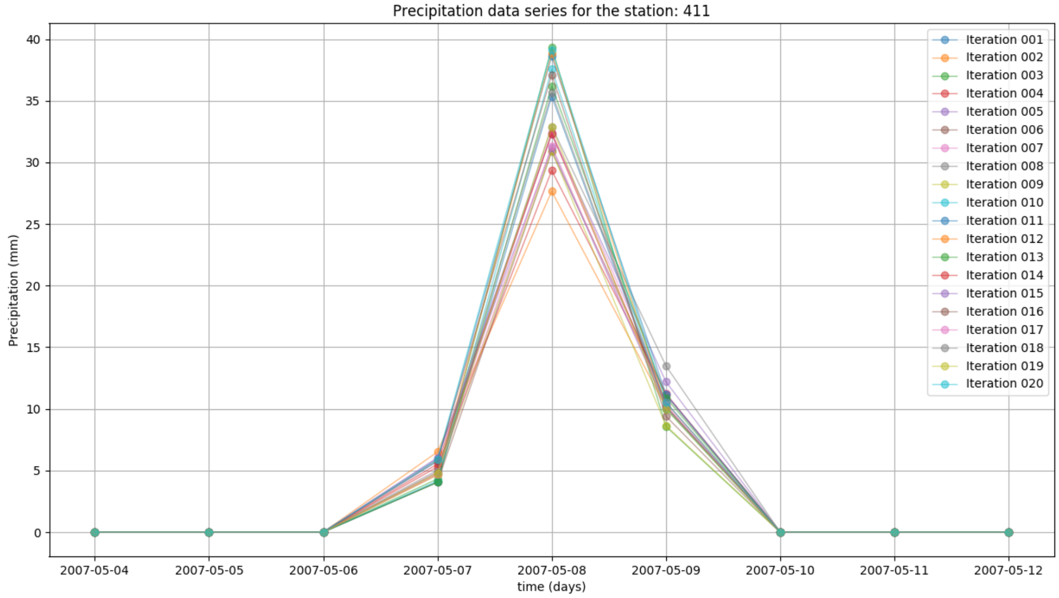

Figure 4.7: Simulation precipitation iterations, May 8, 2007

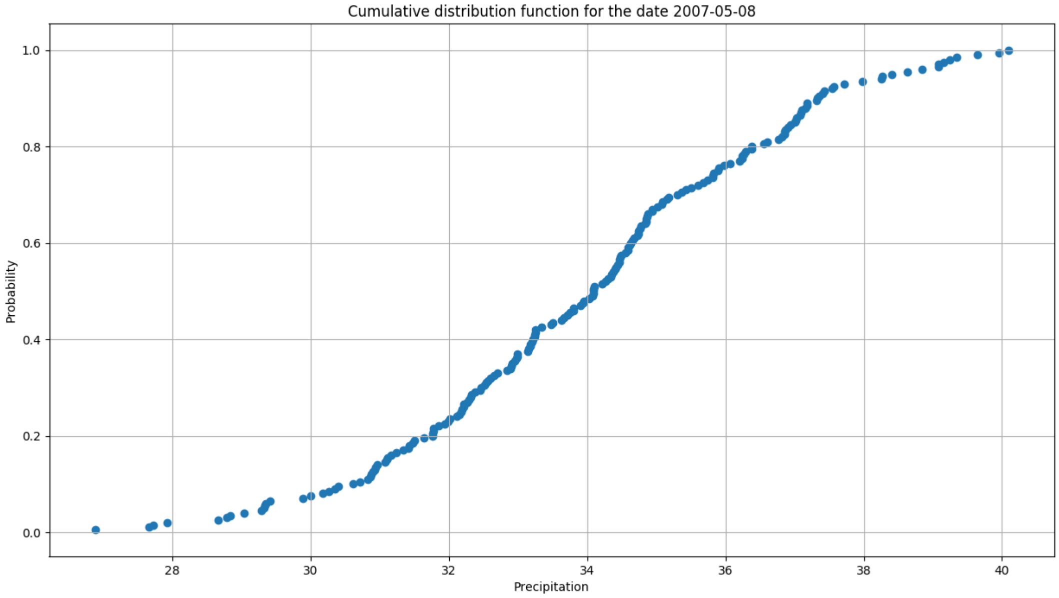

Figure 4.8: CDF for the date May 8th 2007

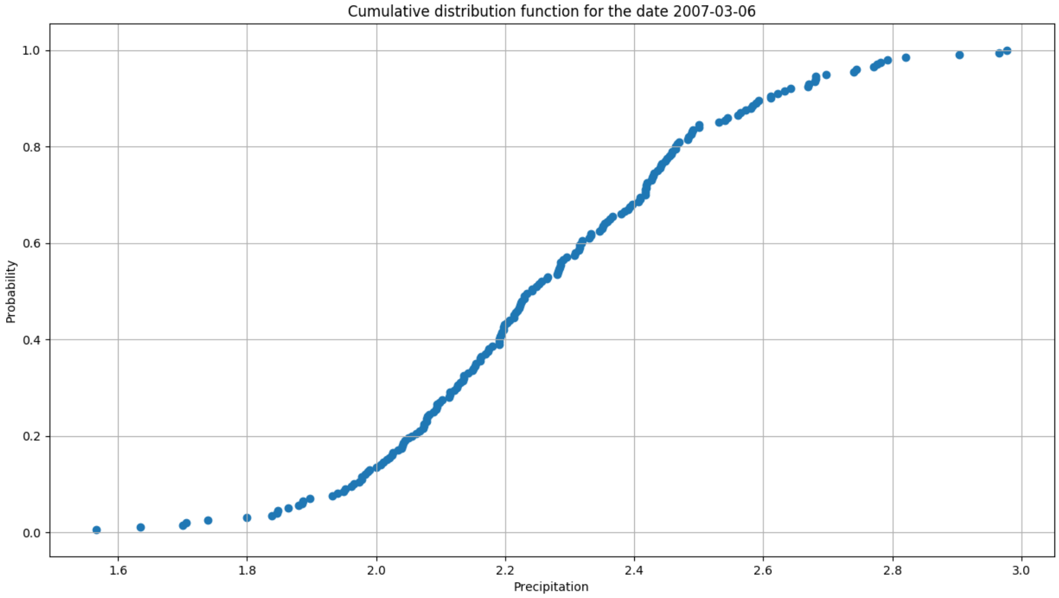

Figure 4.9: CDF for the date March 6th 2007

4.3 Discharge Comparisons

The HBV simulated discharge was found to be about 15% to 20% higher than the recorded discharge in station 411, figures 4.10 and 4.11 show the error ratio between the discharges, it was calculated by dividing the the simulated discharge values by the recorded discharge values. if the discharge values were perfectly similar, the ratio would be equal to 1.

The figures were selected according to their Nash-Sutcliffe values, with the intention of showing the ratios with low NS index value against high NS value, however, all the results have approximately the same NS value, with iteration number 10 - realisation number 05 being the least at 0.788 and iteration 16 realisation 07 with the NS value of 0.812. the ratios are very similar to each other with values reaching the being above the recorded discharge, this is mostly due to the fact that conditional simulation methods conserve the covariance function over a plain while IDWs or Kriging tend to the mean.

Figure 4.12 shows the Actual recorded discharge and the simulated discharge by the HBV model, as well as the snow accumulation, on most days the simulated flow matches or goes above the recorded discharge, however, on high peak days, the recorded flow goes well above the simulated flow, when compared to figure 4.13 these anomalies can be attributed to snow accumulation and the fact that the HBV model's snow simulation is a little simplistic.

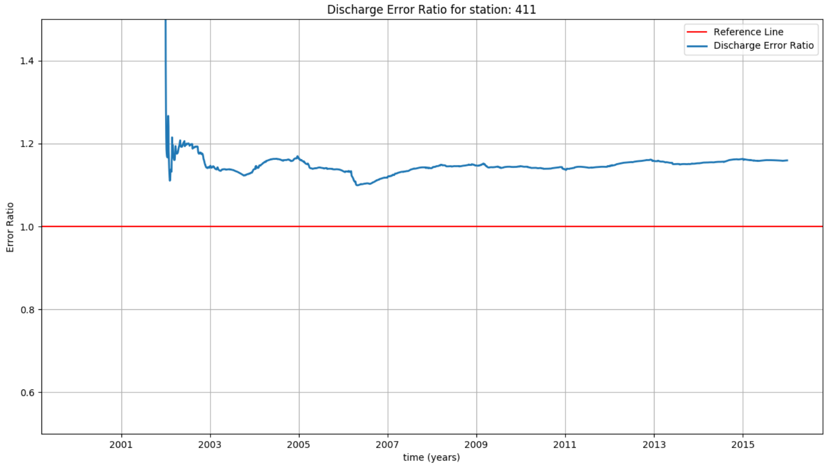

Figure 4.10: Mass balance for Iteration 10 realisation 05

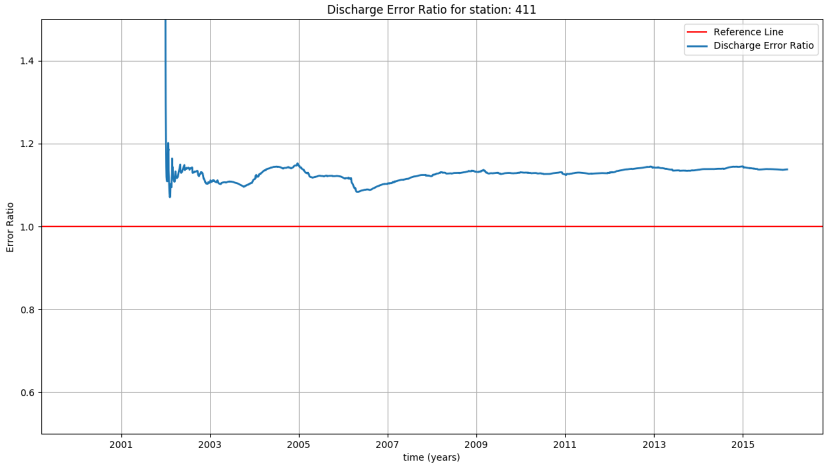

Figure 4.11: Mass balance for Iteration 16 realisation 07

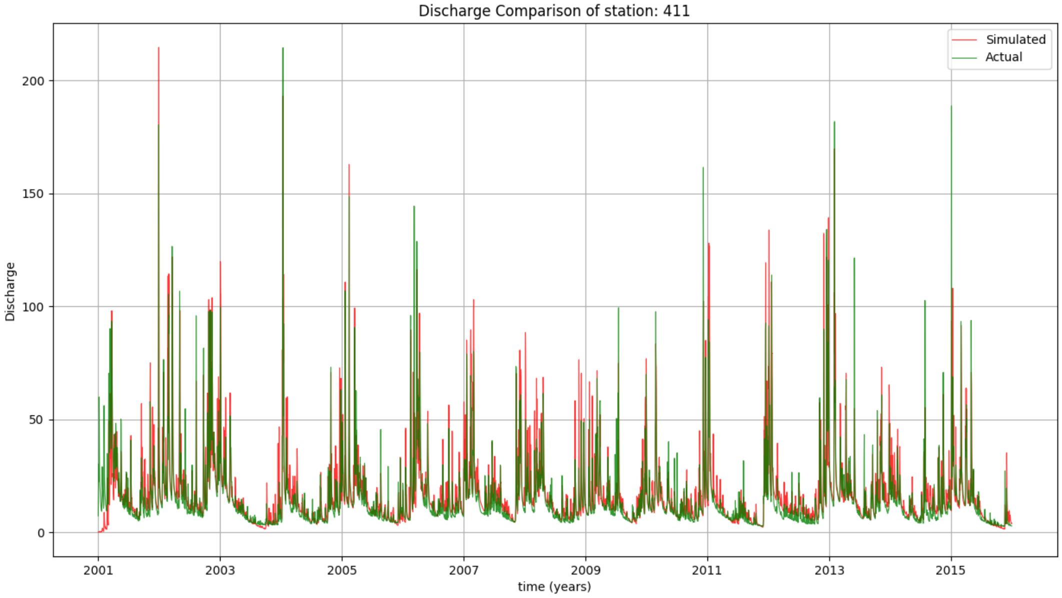

Figure 4.12: Simulated and Actual discharge comparison

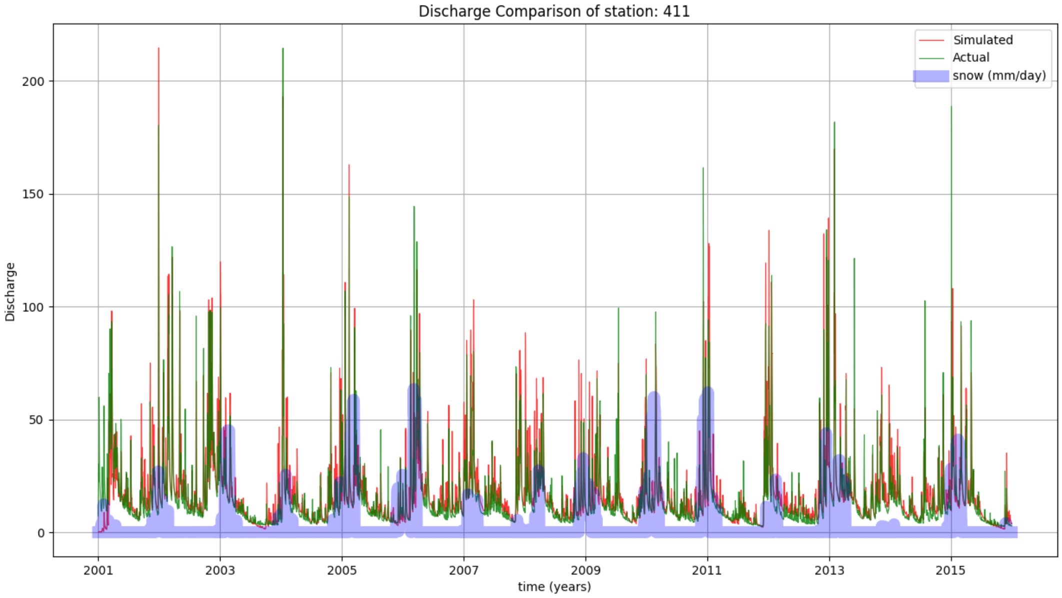

Figure 4.13: Simulated and Actual discharge comparison with snow

Chapter 5

Conclusion

While there are certain pros for using the covariance function, mainly since it proved valuable in estimating the precipitation, which leads to a relatively accurate estimation of the peak, specially when compared to other interpolation methods such as IDW or kriging which under estimate the precipitation.

There are many uncertainties in this model, as shown in chapter 4, uncertainties can arise from several sources, including the random mixing itself, since even different realisations of a single covariance function can be sporadic, adding to that the randomness of the covariance function, and the fact that a different function can be generated from the same dataset can lead to a different interpolation, however, the cumulative distribution for high precipitation events shows that 90% of the realisations of all the covariance functions for a single day all lean to the centre of the CDF, with some outliers on the edges. On the other hand, the difference in values in simulated precipitation observed in low precipitation events , is too high.

In conclusion, the results of this study demonstrate that using multiple covariance functions with the same dataset can be helpful in estimating the discharge flow, the simulated values matched and even exceeded the recorded values, since this method doesn't underestimate the precipitation as much as other interpolation methods, however, the covariance function can cause a relatively high degree of error when used with low precipitation dates, and using multiple covariance functions for a big dataset requires more computation time and power.

Bibliography

Anwar. *Estimation of a Historical Flood using Random Mixing*. Master's thesis, Institut für Wasser- und Umweltsystemmodellierung, Pfaffenwaldring 61, Stuttgart, Germany, 2016.

A. Bárdossy. Interpolation of groundwater quality parameters with some values below the detection limit. *Hydrology and Earth System Sciences*, 15(9):2763–2775, 2011. doi: 10.5194/hess-15-2763-2011.

A. Bárdossy and S. Hörning. Random mixing: An approach to inverse modeling for groundwater flow and transport problems. *Transport in Porous Media*, 114(2):241–259, 2016. ISSN 1573-1634. doi: 10.1007/s11242-015-0608-4. URL http://dx.doi.org/10.1007/s11242-015-0608-4.

A. Bárdossy and J. Li. Geostatistical interpolation using copulas. *Water Resources Research*, 44(7):357, 2008. ISSN 00431397. doi: 10.1029/2007WR006115.

Bergström. The hbv model - its structure and applications.

K. J. Beven. *Rainfall-runoff modelling: The primer / Keith Beven*. Wiley-Blackwell, Chichester, West Sussex and Hoboken, NJ, 2nd ed. edition, 2012. ISBN 978-0-470-71459-1.

H. Bourennane, D. King, and A. Couturier. Comparison of kriging with external drift and simple linear regression for predicting soil horizon thickness with different sample densities. *Geoderma*, 97(3-4):255–271, 2000. ISSN 00167061. doi: 10.1016/S0016-7061(00)00042-2.

G. K. Devia, B. P. Ganasri, and G. S. Dwarakish. A review on hydrological models. *Aquatic Procedia*, 4:1001–1007, 2015. ISSN 2214241X. doi: 10.1016/j.aqpro.2015.02.126.

EPA. An overview of rainfall-runoff model types. 2017.

eWater CRC. Guidelines for rainfall-runoff modelling: Towards best practice model application. 2011.

Gramacki. *Nonparametric Kernel Density Estimation and its Computational Aspects*. SPRINGER INTERNATIONAL PU, [S.l.], 2018. ISBN 978-3-319-71687-9.

Hörning. *Process-oriented modeling of spatial random fields using copulas*. PhD thesis, Institut für Wasser- und Umweltsystemmodellierung, Stuttgart, Germany, 2016. URL https://elib.uni-stuttgart.de/handle/11682/8888.

T. Pluntke, D. Pavlik, and C. Bernhofer. Reducing uncertainty in hydrological modelling in a data sparse region. *Environmental Earth Sciences*, 72(12):4801–4816, 2014. ISSN 1866-6280. doi: 10.1007/s12665-014-3252-3.

S. R. Rusli, D. Yudianto, and J.-t. Liu. Effects of temporal variability on hbv model calibration. *Water Science and Engineering*, 8(4):291–300, 2015. ISSN 16742370. doi: 10.1016/j.wse.2015.12.002.

M. Waseem, M. Ajmal, U. Kim, and T.-W. Kim. Development and evaluation of an extended inverse distance weighting method for streamflow estimation at an ungauged site. *Hydrology Research*, 47(2):333–343, 2016. ISSN 0029-1277. doi: 10.2166/nh.2015.117.