4. Methodology

4.1 General

This chapter shows the methodology adopted for the individual objectives. The chapter is divided into six sections, each section contains the specific methodology for the specific objective.

4.2 Methodology

1. **Objective:** To demonstrate a comparative assessment of discrepancy in the hydrological behaviour of the DEMs in terms of terrain representation at the catchment scale.

   To evaluate the sensitivity of data sources and their vertical accuracies, two hydrologic applications, watershed boundary and river network extraction, are used along with various statistical measures. Hydrologic applications are selected because they heavily rely on DEM data. The workflow is divided into following three steps:

   Datum Transformation: Datum transformation is carried out to bring the DEMs to common horizontal datum and vertical datum. SRTM and ASTER data are referenced to WGS84 horizontal datum and EGM96 vertical datum. But, the ellipsoidal height of terrain (in meters), with WGS84 ellipsoid as a horizontal and a vertical datum, in Geographic Projection System (i.e., X and Y in terms of latitude and longitude) is provided by Cartosat DEM. So, the Cartosat DEM has been reprojected by using the Vdatum transformation tool provided by NOAA's National Ocean Service in a Geographic (lat./long.) projection, to WGS84 as a horizontal datum and EGM96 as a vertical datum.

   Visual Comparison: The aim of visual comparison was to detect changes between the results, such as streams and watershed derived from the different DEMs by using the shaded relief map and the high-resolution satellite imagery. The Vishwamitri watershed was selected for heterologous comparison of slope maps, ridge lines and streams generated by ASTER, SRTM, and Cartosat DEMs. The maximum rate of change of the elevation of the plane (the angle that the plane makes with a horizontal surface) is called the slope gradient. A declivity map with a pixel size of 30 m was created for analyzing the influence of the terrain slope on the models. Watershed delineation was performed by GIS software by importing DEMs. A pixel or a set of spatially connected pixels whose flow direction cannot be assigned to one of the eight valid values in a raster of the flow direction is called a sink. In order to remove small imperfections in the data, the Fill Sink tool was used. Sinks must be filled to ensure a proper delineation of basins and streams. A derived drainage network may be discontinuous if the sinks are not filled. A raster of the flow direction from each pixel to its downslope neighbours is created by the flow-direction tool. The accumulated flow as the accumulated weight of all pixels flowing into each

downslope pixel in the output raster is calculated by the flow accumulation tool. Pixels with a high flow accumulation are termed as areas of concentrated flow, which may be used for identifying stream channels. Similarly, pixels with a flow accumulation of 0 are termed as local topographic highs, which may be used for identifying ridges. A stream network can be delineated by applying a threshold value to the flow accumulation raster. A user-defined and important parameter, which is known as the stream threshold, directly affects the drainage network and basin boundaries that would be obtained by hydrological analysis. In this study, the stream threshold has been considered as 1% of the maximum flow accumulation value (Paul et al., (2015)). The point on the surface at which water flows out of an area is called the outlet or the pour point. The outlet is the lowest point along the boundary of a watershed. Figure 4.1 shows the methodology adopted for watershed delineation. Map algebra that determines where the Fill tool had filled the sinks was used to investigate the cause of the errors in the streams network.

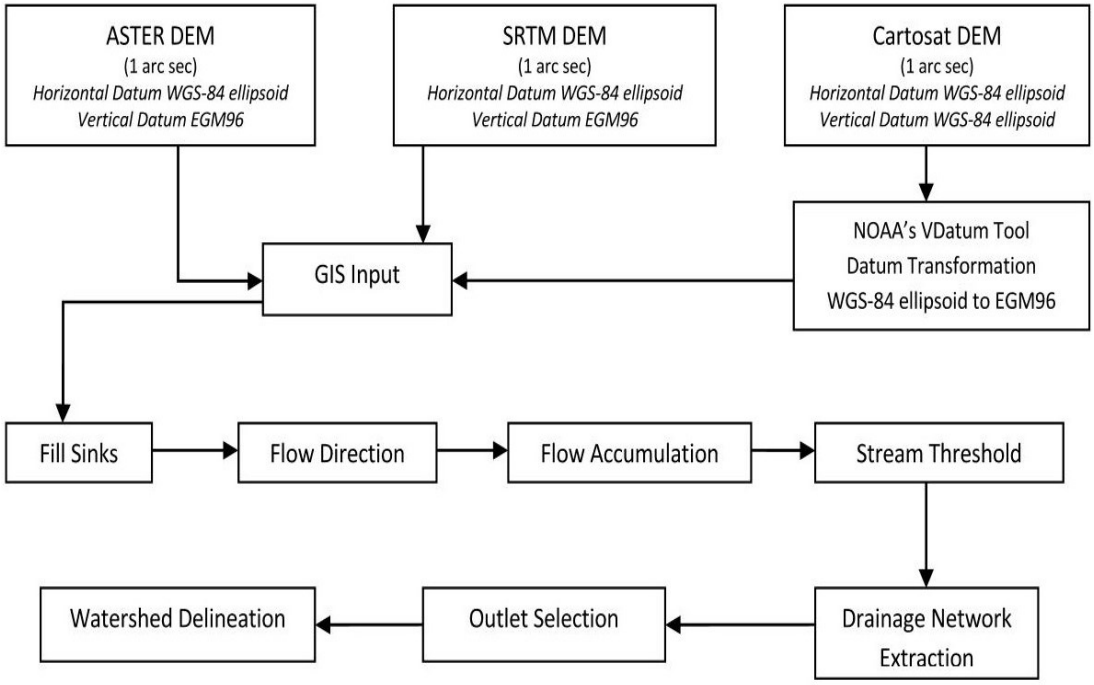

Figure 4.1: Methodology adopted for watershed delineation.

Statistical Comparison: Several descriptive statistic measures are employed to describe and compare the elevation distributions in each DEM. The root-mean-square error (RMSE), a typical proportion of measuring vertical exactness in DEMs, is computed for DEMs. The elevation of each ASTER and SRTM DEM pixel is compared with that of the respective Cartosat DEM pixel. In addition, skewness and kurtosis are determined for DEMs. The degree of asymmetry of a distribution around its mean is measured by skewness. The range of skewness is considered to be from minus infinity ($-\infty$) to positive infinity ($+\infty$). A distribution with a tail extending out to

the right is called positively skewed distribution, whereas a distribution with an asymmetric tail extending out to the left is called negatively skewed (see Figure 4.2). The degree to which a distribution is more or less peaked than a normal distribution is measured by excess kurtosis. Kurtosis is a unitless measure that indicates how sharp the data peak is. A kurtosis value of >0 indicates a peaked distribution, whereas a kurtosis value of <0 indicates a flat distribution (see Figure 4.3). A measure of linear association between quantitative variables is called the correlation coefficient (r). By analyzing the respective scatter plots, the correlation between Cartosat-, SRTM-, and ASTER-derived elevation values are estimated. The value of the correlation coefficient ranges from 1 (indicating the perfect positive correlation) to −1 (indicating the perfect negative correlation). Pearson's correlation between variables a and b is determined by Eq. (4.1).

$$r = \frac{\sum_i (a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_i (a_i - \bar{a})^2 (b_i - \bar{b})^2}} \qquad 4.1$$

r = *correlation coefficient*
a_i = *values of the a-variable in a sample*
$\bar{a}$ = *mean of the values of the a-variable*
b_i = *values of the b-variable in a sample*
$\bar{b}$ = *mean of the values of the b-variable*

Moreover, the normality tests have also been performed over the datasets. For instance, The Kolmogorov–Smirnov (K–S) sample test is a nonparametric test with null hypothesis indicating that the data have been derived from a normal distribution. Similarly, the Wilcoxon signed-rank test is a nonparametric statistical hypothesis test that is used to compare two related samples in order to assess whether their population mean ranks differ (i.e., it is a paired difference test). When the population data are not normally distributed, the Wilcoxon test can be used as an alternative to the paired Student's t-test (also known as "t-test for matched pairs" or "t-test for dependent samples"). The related-sample Wilcoxon signed-rank test conducted with the null hypothesis indicates that the median of differences between the data equals 0. In order to compare the mean and standard deviation and to verify whether the data are underestimated or overestimated, an error map was developed for ASTER and SRTM by subtracting their elevation values from the respective Cartosat values.

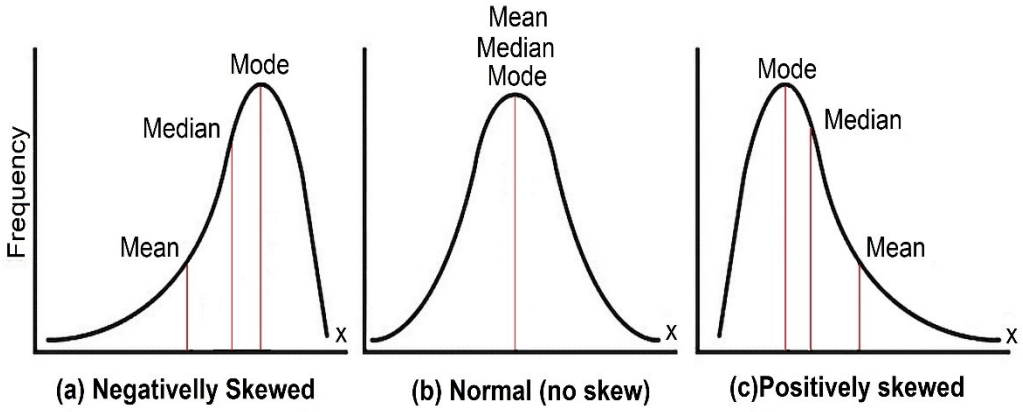

Figure 4.2: Types of Skewness.

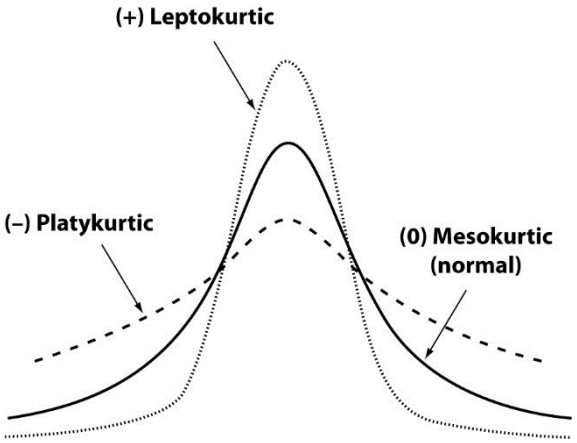

Figure 4.3: Types of Kurtosis.

2. **Objective:** To develop an approach to analyze Sentinel–2 satellite images using traditional and principal component analysis based approaches to create land use and land cover map, which is a prerequisite for developing the curve number.

   The Sentinel–2 cloud-free Level 1C data product (L1C_T43QCE_A008039_20180920T054434) acquired on 20 September 2018 was downloaded from the Sentinel Hub developed by European Space Agency. Sentinel–2 Level 1C data were processed from Top-Of-Atmosphere Level 1C to Bottom-Of-Atmosphere Level 2A. QGIS desktop 3.6.1 is a free and open-source cross-platform desktop geographic information system application that supports viewing, editing, and analysis of geospatial data. QGIS desktop 3.6.1 interface was used with Semi-Automatic Classification

Plugin (SCP), to convert the Sentinel–2 MSI data to reflectance values and for dark object subtraction atmospheric correction (DOS1) of the data.

After atmospheric correction, ten bands (2–8, 8A, 11 and 12) were composited and clipped to the study area. The processed data were georeferenced to the WGS 84 UTM 43N projected coordinate system. In order to test the effectiveness of PCA, two stacks were created for the classification in ESRI's ArcGIS Desktop 10.5 software. Stack 1 (Figure 5.14 (a)) contained atmospherically corrected bands (2–8, 8A, 11 and 12) and Stack 2 (Figure 5.15 (a)) contained 3 major PCA bands accounting for the 97.96% of eigenvalues. The PCA technique was used to reduce the number of bands or dimensions necessary for classification. Dimension reduction leads to a reduction in the computation costs without compromising the desired variability in the data. According to Mather, (2010) the process of PCA can be divided into three steps. The first step is to calculate the covariance or correlation matrix of multiband images. The covariance matrix is calculated by Eq. (4.2).

$$C_{XY}=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{n-1} \quad 4.2$$

C_{XY} = *Covariance between Band X and Band Y*
n = *The number of pixels*
X_i = *Individual pixel value vectors of Band X*
$\bar{X}$ = *Mean of Band X*
Y_i = *Individual pixel value vectors of Band Y*
$\bar{Y}$ = *Mean of Band Y*

The diagonal elements of the covariance matrix are the band variances, and the off-diagonals are band covariances. If a correlation matrix is used instead of a covariance matrix, each entry in C should be further divided by the product of the standard deviations of the features represented by the corresponding row and column.

$$R_{XY}=\frac{C_{XY}}{\sigma_X \sigma_Y} \quad 4.3$$

R_{XY} = *Correlation between Band X and Band Y*
C_{XY} = *Covariance between Band X and Band Y*
σ_X = *Standard deviations of Band X*
σ_Y = *Standard deviations of Band Y*

The second step is to calculate the eigenvectors of the covariance matrix. Following equation is used for the calculation:

$$(C-\lambda_i I)A_i = 0 \quad 4.4$$

C = *Covariance matrix*
λ_i = *Eigenvector*
I = *Identity matrix*
A_i = *Eigenvalue*

The normalized eigenvectors of the covariance or correlation matrix form the new coordinate system. The mapping location f_i of each pixel X=(x_1, x_2, …, x_k) on the i^{th} principal component is given by:

$$f_i = XA_i = x_1 a_1, x_2 a_2, \ldots, x_k a_k \quad 4.5$$

which shows the rotation of the axes of the feature space.

The traditional approach and PCA based approach used Stack 1 and Stack 2, respectively, as inputs for land use and land cover classification. The training data were collected based on the manual interpretation of the original Sentinel–2 data and DigitalGlobe's WorldView-4 high-resolution imagery and was kept the same for all the three classifiers to avoid the optimistic bias in classification. The training sample size was kept below 1000 pixels per class to evaluate the influence of the training sample size, as well as the performance of classification algorithms. Training data for each land use and land cover class were collected as a group of pixels. The input data and corresponding ground truth data (training sample) were used to train the classifiers. The classifiers learn the complex relationships between the input and ground truth data (training sample). To determine the accuracy of each classification and class, thematic accuracy assessment was performed. For this purpose, firstly a reference data set including a total of 100 points was created. Stratified random sampling was used with 100 points to obtain the ground truth data from the manual interpretation of the original 10 m resolution Sentinel–2 data (Band 2, 3 and 4) and DigitalGlobe's WorldView-4 data (Product Id: 1ba34688-3ee0-41e4-9187-de68fdb075df-inv) acquired on 25-10-2018 at 5:30 am with 31 cm resolution.The results of the classifications were not post-processed (e.g., filtered). The classification maps were evaluated in terms of their overall accuracy (OA), producer's accuracy (PA), user's accuracy (UA) and the Kappa index of agreement (k) or Kappa coefficient and a Confusion matrix was created. The key element of a quantitative accuracy assessment is the creation of a confusion matrix. The confusion matrix is represented by a table that shows correspondence between the classification result and a reference image assigned to a particular category, which is relative to the actual category as indicated by the reference data. Producer's accuracy is the probability that value in a given class was correctly classified.

$$\textit{Producer's accuracy} = \frac{t\text{otal number of correct pixels in a class}}{\text{total pixels in that class as derived from the reference data}} \quad 4.6$$

User's accuracy is the probability that a value predicted to be in a certain class is really in that class.

$$\textit{User's accuracy} = \frac{t\text{otal number of correct pixels in a class}}{\text{total pixels that were classified in that class}} \quad 4.7$$

The kappa coefficient measures the agreement between classification and truth-values. A kappa value of 1 represents perfect agreement, while a value of 0 represents no agreement.

$$\textit{Kappa coefficient} = \frac{Observed\ accuracy\ -\ Expected\ agreement}{1\ -\ Expected\ agreement} \quad 4.8$$

The Overall accuracy is given by the ratio of the proportion of the correctly classified pixels to the total number of pixels in the confusion matrix.

The land use and land cover maps were later used in HEC-GeoHMS for the integration of land use land cover and soil data for Curve Number grid preparation. A logical condition was defined in ArcGIS to generate the curve number raster file from the raster files of hydrologic soil group and land use and land cover using TR-55 table (Feldman, (2000)). The tables provide estimates of the Curve Number as a function of hydrologic soil group, cover type, treatment, hydrologic condition, antecedent runoff condition, and impervious area in the catchment. Selected Curve Number values for the study area are given in Table 4.1. The sub classes have been created using interpretive overlays based on elements of image interpretation such as texture, tone, association and pattern.

Table 4.1: Selected Curve Number values for the study area using TR-55 table.

LULC	Sub classes of LULC	HSG-A	HSG-B	HSG-C	HSG-D
Water	Water	100	100	100	100
Cultivated land	Cultivated land crop1	64	75	82	85
	Cultivated land crop 2	71	80	87	90
	Sparsely vegetated	74	83	88	90
Barren land	Barren land	77	86	91	94
Fallow land (Vertisols dominance)	Fallow land 1 Vertisols dominated	76	85	90	93
	Fallow land 2 Vertisols dominated	76	85	90	93
Fallow land (Inseptisol dominance)	Fallow land 1 inseptisol dominated	74	83	88	90
	Fallow land 2 inseptisol dominated	74	83	88	90
Mixed forest	Mixed forest	36	60	73	79
Builtup	Builtup	89	92	94	95
	Mixed builtup 1	83	89	92	93
	Mixed builtup 2	89	92	98	98

To estimate the runoff from Curve Number, the first step to be followed is to delineate and measure the drainage area tributary to the point of analysis (delineation methodology shown

$$S = \frac{25400}{CN} - 254 \qquad 4.9$$

in objective 1). The potential maximum soil retention is then calculated using following formula: where S is in mm, and CN is the curve number (dimensionless).

The assumption of SCS-CN is that, for a single storm event, potential maximum soil retention is equal to the ratio of direct run-off to available rainfall. This relationship, after algebraic manipulation and inclusion of simplifying assumptions, results the following expression:

$$Q = \frac{(P - I_a)^2}{(P + S - I_a)} \quad 4.10$$

Q = *direct run-off depth*

P = *total rainfall*

I_a = *initial abstraction*

I_a and S can be related using the following equation:

$$I_a = \lambda S \quad 4.11$$

λ = 0.2 was assumed in original SCS-CN model

3. **Objective:** To perform Morphometrical analysis of Vishwamitri watershed and prioritization of sub-watersheds for assessing the flood influencing characteristics of sub-watersheds of the Vishwamitri river.

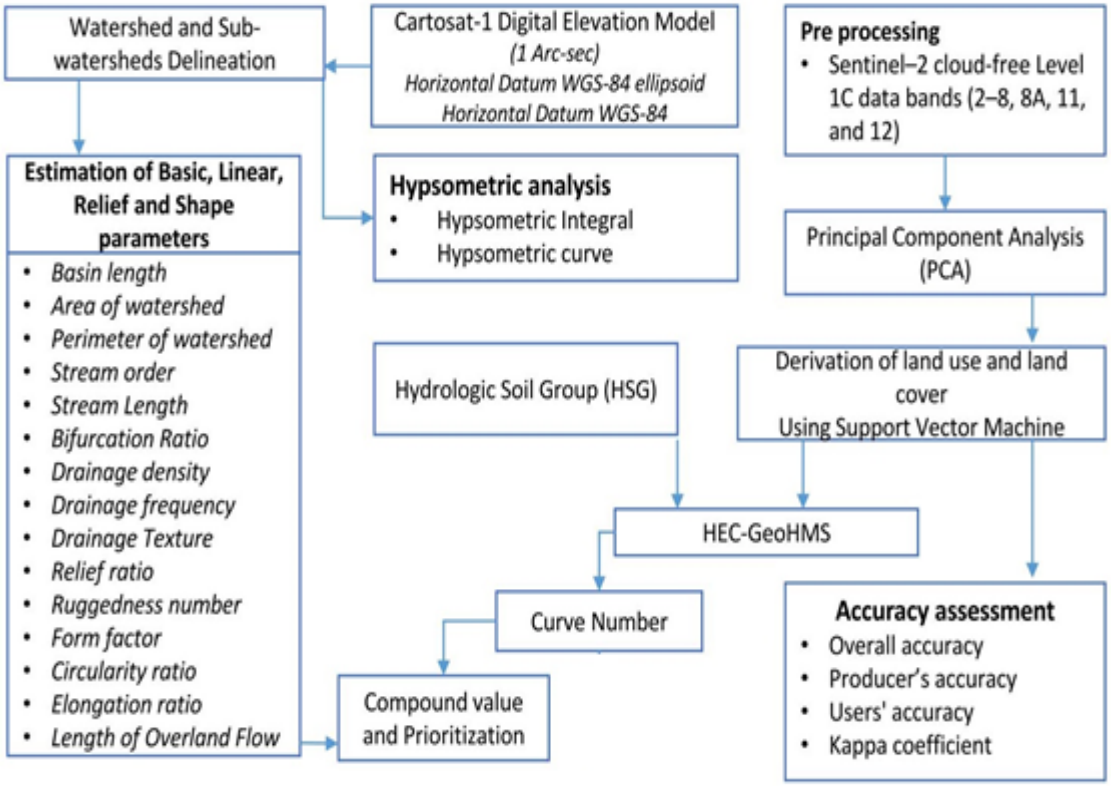

Figure 4.4: Methodology adopted for prioritization of sub-watersheds for assessing the flood influencing characteristics of sub-watersheds.

The methodology used in the study is presented in Figure 4.4. The area of the derived watershed is determined by calculating the geometry of the watershed polygons in GIS environment. Using the mathematical formulas (Table 4.2), morphometric analysis of the parameters, namely, stream order, stream length, bifurcation ratio, relief ratio, drainage density, drainage frequency, drainage texture, form factor, length of overland flow, ruggedness number, circulatory and elongation ratio, area, perimeter, and basin lengths of all 5 sub-watersheds is carried out. Each of the linear, areal, and relief morphometric parameters along with CN is taken into consideration for assessing the flood influencing characteristics of the five sub-watersheds of

the Vishwamitri watershed, as these parameters have a direct but variable relationship with flood runoff. Prioritization was achieved through the allocation of weights to the individual indicators contributing to flood runoff and a compound value (C_v) was calculated for final prioritization. C_v is derived by calculating the average of weights assigned to the individual parameters. The sub-watershed with highest C_v is having highest flood influencing characteristics as a result needs highest priority for flood mitigation measures, whereas sub-watersheds with lowest C_v is contributing least to flood runoff thereby is low priority.

Hypsometric analysis is useful to understand the geomorphometric stage of a river basin and to assess factors forcing basin evolution. By graphing the relative area along the abscissa and relative elevation along the ordinate, the hypsometric curve is obtained. The relative area is obtained as a ratio between the area above a particular contour and the total area of the watershed encompassing the outlet. The relative elevation is calculated as the ratio between the height of a given contour (h) from the base plane and the maximum basin elevation (H) (up to the remote point of the watershed from the outlet) (Subedi & Tamrakar, (2020)). The curve obtained provides a measure of the distribution of landmass volume remaining below or above a basal reference plane. The area under the hypsometric curve (Hypsometric integral (HI)) indicates the erosion process dynamics in a watershed (Zakerinejad, (2016)). Actually, the shape of the hypsometry curve shows the evolutionary stage of a basin. Curves with convex shape are related to young basin morphologies while basins with concave curved shapes are more mature basins (Figure 4.5). Hypsometric Integral (HI):

$$HI = \frac{[Elev_{\text{mean}} - Elev_{\text{min}}]}{[Elev_{max} - Elev_{\text{min}}]} \quad 4.12$$

Where,

$Elev_{mean}$ = *average elevation of the catchment*
$Elev_{min}$ = *minimum elevation within the catchment*
$Elev_{max}$ = *maximum elevation within the catchment*

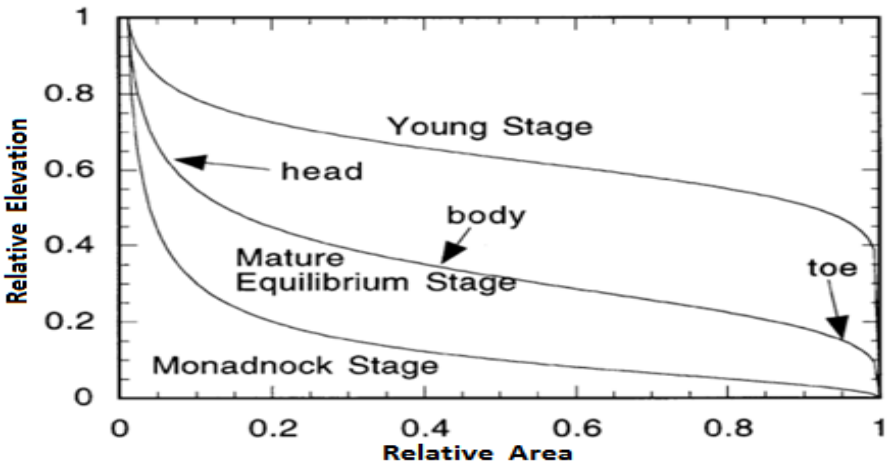

Figure 4.5: Hypsometric curves – young, mature and old stages – showing toe, head and body.

Hypsometric analysis is useful to understand the geomorphometric stage of a river basin and to assess factors forcing basin evolution (Markose & Jayapp, 2011). By graphing the relative area along the abscissa and relative elevation along the ordinate, the hypsometric curve is obtained. The relative area is obtained as a ratio between the area above a particular contour and the total area of the watershed encompassing the outlet. The relative elevation is calculated as the ratio between the height of a given contour (h) from the base plane and the maximum basin elevation (H) (up to the remote point of the watershed from the outlet) (Sarangi et al., 2001; lama & Maiti, 2019).

Table 4.2: Linear, relief and shape morphometric parameters.

Morphometric parameters	**Formulae**	**Units**
Basin length (L_b)	Maximum length of the watershed measured parallel to the main drainage line	Km
Area (A)	Area of watershed	Km^2
Perimeter (P)	Length of the watershed boundary	Km
Stream order (S_u)	Hierarchical rank (Strahler Scheme)	Dimensionless
Stream Length (L_u)	$L_u=L_1+L_2+…+L_n$; Length of the stream	Km
Stream number (N_u)	$N_u=N_1+N_2+…+N_n$;	Dimensionless
Bifurcation Ratio (R_b)	R_b = Nu/Nu + 1; R_b was computed as the ratio between the number of streams of any given order to the number of streams in the next higher order	Dimensionless
Mean Bifurcation Ratio (Rbm)	Rbm = Average of bifurcation ratios of all orders	Dimensionless
Drainage density (D_d):	Dd = ΣLu/A; The ratio between the total stream length of all orders to the area of the basin	(km/km^2)
Drainage frequency (F_s):	Fs = ΣNu/A; The ratio between total number of streams and area of the basin	($no./km^2$)
Drainage Texture (R_t):	T = ΣNu/P; Where, Rt = Drainage texture; ΣNu = Total no. of streams of all orders; P = Perimeter (km)	(no./km)
Relief ratio (R_r):	R_r = H/L; Where, R_r = Relief ratio; H = Total relief of the basin in Kilometre; L_b = Basin length	Dimensionless
Ruggedness number (R_n):	R_n = B_h × D_d; Where, B_h = Basin relief; D_d = Drainage density	Dimensionless
Form factor (F_f):	F_r = A/L_b^2; The ratio of the basin area to the square of the basin length	Dimensionless
Circularity ratio (R_c):	$Rc = 4 * \pi * A/p^2$; Where, Re = Circularity ratio; π = "Pi" value that is 3.14; A = Area of the basin (km2); P = Perimeter (km)	Dimensionless
Elongation ratio (R_e):	Re = $(2/L_b) * sqrt\ (A/\pi)$; Where, Re = Elongation ratio A = Area of the basin (km2); π = "Pi" value that is 3.14; L_b = Basin length	Dimensionless
Length of Overland Flow (L_g):	L_g = 1/($D_d * 2$); Where, L_g = Length of overland flow; D_d = Drainage density	Km

4. **Objective**: To identify potential runoff storage zones based on the various physical characteristics of the Vishwamitri watershed using a GIS-based conceptual framework that combines through analytic hierarchy process using multi criteria decision-making method.

   To find the potential runoff storage zones, workflow was divided into 4 steps (Figure 4.6). Firstly, the rainfall analysis was carried out using SPI and annual rainfall. Secondly, processing of spatial data and creation of spatial data layers. Thirdly, criteria weights were determined using AHP. Lastly, executing weighted overlay process (WOP) within GIS.

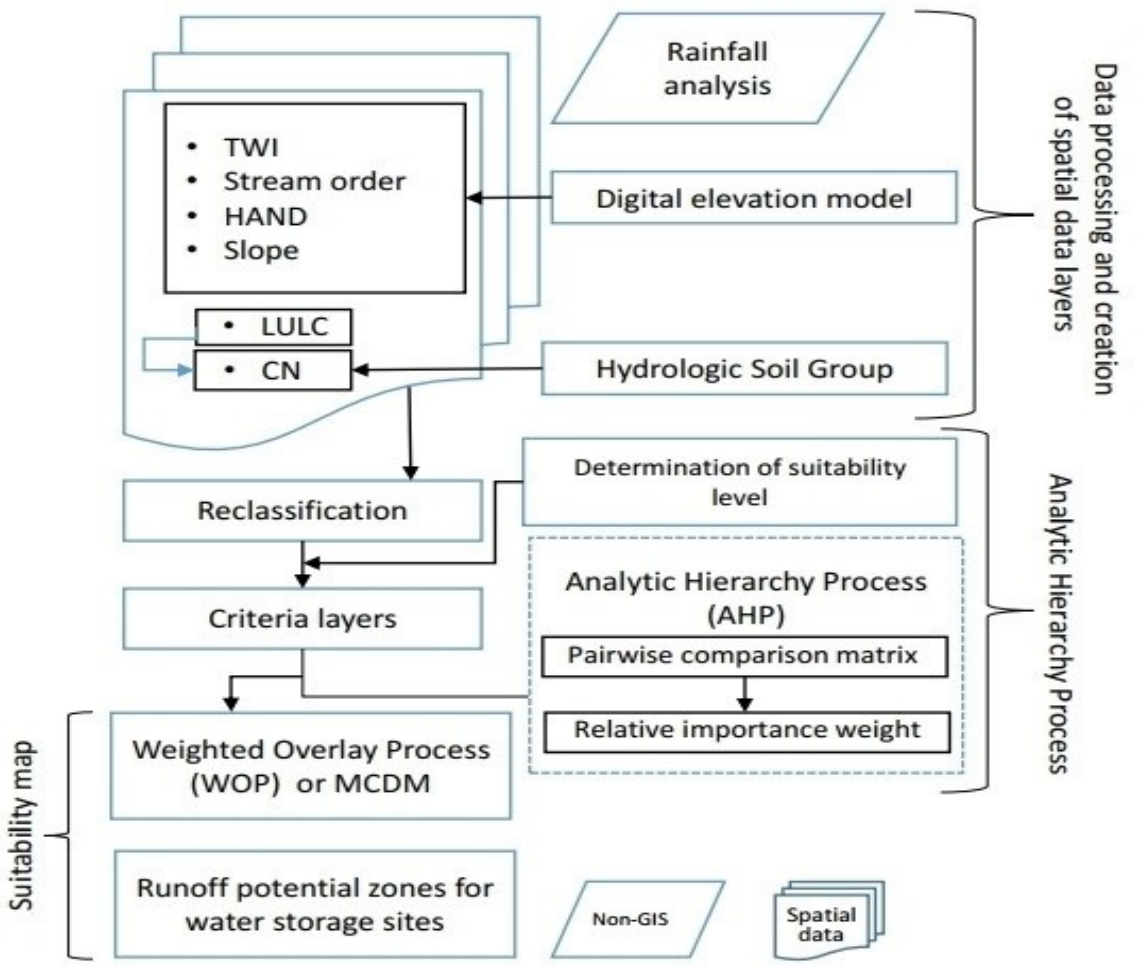

Figure 4.6: Multi criteria decision making (MCDM) technique workflow using AHP for identification of potential runoff storage zones for water storage.

I. Rainfall analysis:

Study area falls under the areas of Gujarat Plains. Potential runoff storage zones or structures require considerable rainfall and hence it is important to analyse variability of rainfall within the watershed for the suitable locations of these storage zones or structures before execution. For rainfall variability analysis two indicators, viz., annual rainfall and Standardized Precipitation Index (SPI) have been used. Annual rainfall is highly influenced by the amount of the rainfall,

intensity of the rainfall, frequency of occurrence of the rainfall and distribution over area as well as time of the rainfall. The approach examines the annual rainfall along with the SPI drought index application for Vishwamitri watershed and it is calculated accordingly by historical precipitation data. Results of SPI and annual rainfall help in evaluating the area whether it is suitable for water storage structures or not. Favourable results qualifies the area for identification of suitable sites for water storage. The SPI calculation for any location is based on the long-term precipitation record for a desired period. Precipitation is normalized using a probability distribution function and allows for the estimation of both dry and wet periods. Daily rainfall data was collected from State Water Data Centre, Gandhinagar, Gujarat. A total of 56 years (1961 to 2016) rainfall data of 4 rain gauge stations namely Vadodara, Padra, Savli and Waghodia were used for computation of SPI and annual rainfall. Annual SPI classification system used by McKee et al., (1993) shown in the Table 4.3 and it was computed as described by Akinsanola & Ogunjobi, (2014) and Adegoke & Sojobi, (2015). Positive SPI values indicate greater than median precipitation and negative values indicate less than median precipitation. Drought starts when the SPI value is equal or below -1.0 and ends when the value becomes positive.

$$\text{SPI} = \frac{X - \bar{X}}{\sigma} \qquad 4.13$$

X = *rainfall in each particular year*
$\bar{X}$ = *mean rainfall in each particular year*
σ = *the standard deviation of rainfall in each particular year*

Table 4.3: Annual standardized precipitation index given by McKee et al., (1993).

Classification	SPI
Dry extreme	≤ −2.0
Severely dry years	−1.5 to −1.99
Moderately dry years	−1.0 to −1.49
Near normal	−0.99 to 0.99
Moderately wet years	1.0 to 1.49
Very wet	1.5–1.99
Wet extreme	≥ +2.0

II. Processing spatial data and creation of spatial data layers:

Topographic Wetness Index (TWI)

The topographic wetness index was first introduced by Beven & Kirkby, (1979). TWI is widely used topographically based soil wetness model that identifies wet areas. It is based on the assumption that local topography controls the movement of water in slopped terrain which quantifies the effect of the local topography on runoff generation. The index is represented as the natural logarithm of the ratio of upslope flow accumulation area and slope at the cell.

$$TWI = \frac{Ln(\text{flow accumulation } + 1)}{\tan(((\text{slope in degrees})3.14)/180)} \qquad 4.14$$

Topographic wetness at a particular point on the landscape is the ratio between the catchment area contributing to that point and the slope at that point (Wilson & Gallant, (2000)). Locations with a high TWI value have large upslope area and are expected to have higher water availability. On the other hand, locations with small TWI value have small upslope area that are assumed to have lower water availability. Also, Steep locations receive a small TWI value and are expected to be better drained than gently sloped locations, which receive a high TWI value (Sørensen & Seibert, (2007); Hojati & Mokarram, (2016); Bjelanovic, (2016); Ågren et al., (2014); Loritz et al., (2019)). The TWI calculation for this study was conducted with the use of Topography Toolbox for ArcGIS 10.1 (Dilts, (2015)). Firstly, the DEM was pre-processed in order to remove shallow sinks, thus an impact of model artefacts in further analysis could be reduced. The second step included calculation of prerequisites for further TWI computation slope and catchment area; the latter parameter was calculated using the multiple flow direction method (Quinn et al., (1991)).

Generation of slope map using Topographic Position Index (TPI)

The TPI is the basis of the landform classification system. Gallant & Wilson, (2000) defined TPI as the relative topographic position of the central point as the difference between the elevation at this point and the mean elevation within a predetermined neighbourhood. Using TPI, landscapes can be classified in slope position classes. Many researchers have used this index in the field of geomorphology (Tagil & Jenness, (2008); Liu et al., (2009); McGarigal et al., (2009)); geology (Mora-Vallejo et al., (2008); Deumlich et al., (2010); Illés et al., (2011)); hydrology (Lesschen et al., (2007); Francés & Lubczynski, (2011); Liu et al., (2011)); agricultural science (Pracilio et al., (2006)); archaeology (Patterson, (2008); Berking et al., (2010)). The TPI is the difference of a cell elevation in a digital elevation model from the mean elevation ($\bar{X}$) of a user specified neighborhood surrounding. Local mean elevation is subtracted from the elevation value at centre of the local window (Gallant, (2000)). The range of TPI depends not only on elevation differences but also on the adopted local window. Large local window values mainly reveal major landscape units, while smaller values highlight smaller features, such as minor valleys and ridges.

$$TPI_i = X_0 - \bar{X} \quad (4.15)$$

$$\bar{X} = \frac{\sum_{i=n} X_i}{n} \quad (4.16)$$

Where,

X_0 = *elevation at the central point*
X_i = *elevation of the ith cell*
$\bar{X}$ = *average elevation around the central point within the local window*
n = *total number of surrounding points employed in the evaluation*
TPI_i = *topographic position index of the ith cell*

TPI variation is shown in Figure 4.7 for arbitrary point elevations (A and B) in a DEM to the mean elevation of a specified neighbourhood around these point elevations. A small neighbourhood consisting of 33 × 33 cell units window was used in order to identify complex landscape features. The TPI provides a concise and effective technique of landscape classification in accordance with morphology. A higher degree of slope results in a higher run-off potential and low infiltration and a lower degree of slope favours the retention of water. There are a wide range of geomorphological methods and algorithms classify the landscape into morphological classes (Burrough et al., (2000); Deng, (2007); Iwahashi and Pike, (2007); Hengl and Reuter, (2009)). Weiss, (2001) and Jenness, (2006) recommend (Table 4.4) standard deviations (σ) away from the mean TPI raster as threshold values for classifying six slope positions:

Table 4.4: Recommended standard deviations (σ) away from the mean TPI raster as threshold values for classifying six slope positions.

Class	Description Breakpoints
Valley	$TPI \leq -1\ \sigma$
Lower Slope	$-1\ \sigma < TPI \leq -0.5\ \sigma$
Flat Slope	$-0.5\ \sigma < TPI < 0.5\ \sigma$, Slope $\leq 5°$
Middle Slope	$-0.5\ \sigma < TPI < 0.5\ \sigma$, Slope $> 5°$
Upper Slope	$0.5\ \sigma < TPI \leq 1\ \sigma$
Ridge	$TPI > 1\ \sigma$

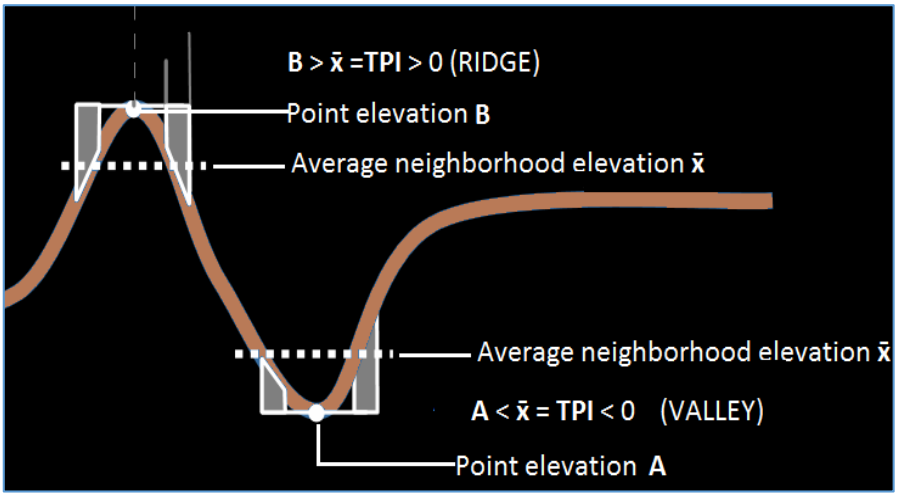

Figure 4.7: TPI variation is shown for arbitrary point elevations (A and B) in a DEM to the mean elevation of a specified neighbourhood around these point elevations.

Land use/land cover

Land use pattern of any watershed influences the runoff. To compute hydrological elements more accurately more accurate LULC map is required. By image processing techniques, image can be produced which depict some of the characteristics, notably the cover types such as areas with vegetation, water bodies, bare soils, etc (Durga & Bhaumik, (2003)). The LULC pattern and rainfall have a significant influence on the hydrological response of the watershed.

Soil texture

Soil texture refers to the relative proportion of clay, silt and sand. Soils containing large proportions of sand have relatively large pores through which water can drain freely. These soils produce less runoff. As the proportion of clay increases, the size of the pore space decreases. This restricts movement of water through the soil and increases the runoff. Soil data based on soil texture collected from National Bureau of Soil Survey and Land Use Planning (NBSS & LUP). The land use and land cover maps were later used in Hydrologic Engineering Center's Geospatial Hydrologic Modeling Extension (HEC-GeoHMS) for the integration of land use land cover and soil data for Curve Number grid preparation. The HEC-GeoHMS is extension to ESRI's ArcGIS software that compute the Curve Number and other loss rate parameters based on various soil and land use and land cover databases.

Curve number (CN)

The CN is most commonly used reliable and conceptual technique for estimating surface runoff. CN is basically a dimensionless number that reduces the rainfall to runoff. It depends upon two parameters LULC and Hydrologic Soil Group (HSG). HSG is one of the important parameters for assigning curve numbers and is generated by reclassifying the soil textural map considering their runoff potential into account (Singh et al., (2017); Hameed et al., (2019); Rizeei et al., (2018); Tripathi, (2018)).

Stream order

The availability of the total quantity of surface water is proportional to the stream order and some particular structure are suitable at a particular drainage order only, for example, check dams should be constructed at lower order streams only (IMSD, (1995) and Durga & Bhaumik, (2003)). The stream order of the Vishwamitri watershed was assigned using the Strahler, (1957) method. In the Strahler method, all streams without any tributaries are assigned an order of 1 and are referred to as first order. The stream segments starting from the confluence of two streams of the first order are called streams of second order and so on. The tail point of each stream is defined as the point from where a stream of higher order starts. Flow accumulation and flow direction rasters were used to generate stream network using hydrology toolset of ArcGIS. Stream ordering was done for proper planning of conservation measures in terms of storage and capacity.

Height Above Nearest Drainage (HAND)

The Height above the nearest drainage is a digital elevation model normalized using the nearest drainage. It normalizes topography according to the local relative heights found along the drainage network and in this way presents the topology of the relative soil gravitational potential, or local draining potentials. HAND allows for the calculation of the elevation of each

point in the catchment above the nearest stream it drains to, following the flow direction (Hamdani & Baali, (2019); Rennó et al., (2008); Nobre et al., (2011)). HAND raster was prepared for the 4th and 5th order streams of Vishwamitri watershed as they are highly susceptible to flooding. The first step in generating HAND raster is to remove small imperfection by filling sinks in the Cartosat-1 DEM. Sinks must be filled to ensure a proper delineation of streams. A derived drainage network may be discontinuous if the sinks are not filled (Rana & Suryanarayana, (2019)). Second step is to create flow-direction raster, it is computed from the DEM using the D8 method (Jenson & Domingue, (1988)) to determine the flow from each cell to its steepest downslope neighbour. An erroneous flow-direction raster may be resulted in the presence of sinks. Next, the accumulated flow direction is used to find the nearest stream cell for each cell. At last, the elevation of the nearest stream cell is deducted from the elevation of each cell to normalize the terrain and to get its corresponding HAND value (Figure 4.8).

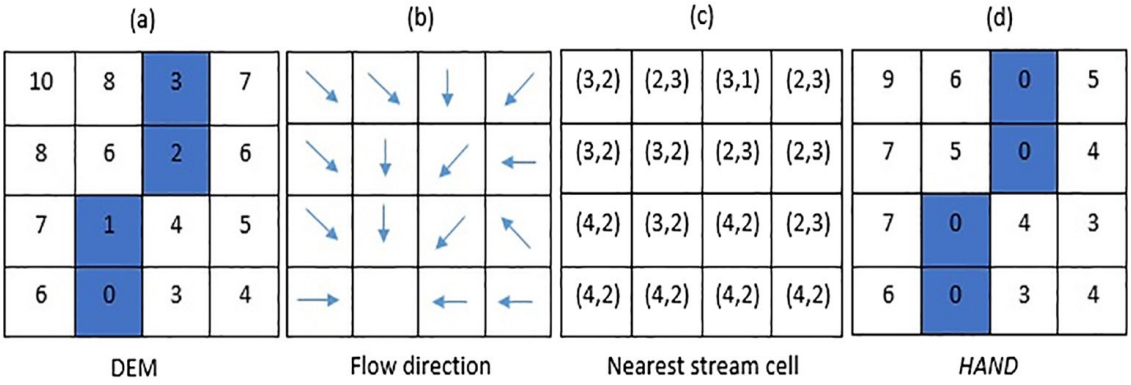

Figure 4.8: Calculation of height above the nearest drainage raster for a hypothetical DEM and stream (blue cells): (a) DEM (b) flow direction (c) Nearest stream cell computed using flow direction, the coordinates (row, column) of the nearest stream cell, drained by each cell, are determined (d). HAND raster created by deducting the elevation of nearest stream cell from DEM.

III. Determining criteria weights using AHP:

Analytic Hierarchy Process (AHP) is one of Multi Criteria Decision Making (MCDM) method that was originally developed by Saaty, (1987), it has been widely applied to solve decision-making problems related to water resources. The approach combines mathematics and psychology in dealing with complex decision and in turn converts it into a simpler system of hierarchy. This method compares two criteria at a time through a pairwise comparison matrix, each criterion is assessed by arranging every possible pairing on a ratio scale to express the comparative importance by numerical values. Numerical expression of suitability rating (Burnside et al., (2002)) and scaling of comparative importance (Saaty, (1990)) is given in Table 4.5. The judgment on dominance of one criterion over another is based on the researchers expertise and

a literature survey (Krois & Schulte, (2014); Singh et al., (2017); Jha et al., (2014); Ammar et al., (2016); Wu et al., (2018); Prasad et al. (2014); Bitterman et al., (2016); Kahinda et al., (2018)). The determination of the relative importance weight of each criterion (Slope, TWI, LULC, Curve Number, Stream Order and HAND) for potential runoff storage zones is calculated by using the pair-wise comparison matrix method. The number of comparison can be determined using Eq. (4.17):

$$\text{Number of comparison} = \frac{n(n-1)}{2} \qquad 4.17$$

where, n = number of criterion

The resulting pair-wise comparison matrix is used to obtain the Eigen value of each criterion, which represents its relative importance weight (Saaty, 1990). The relative importance weight given to the criteria one over another is acceptable if the consistency ratio (CR) is less than 10%. If it increases 10%, a new value is assigned in the pair-wise comparison matrix. CR is computed as:

$$CR = \frac{\frac{(\lambda_{max} - n)}{(n-1)}}{RI} \qquad 4.18$$

where λ_{max} is principal eigen value, n is the number of elements compared and RI is the so-called random consistency index, a value that depends on the number of criterion that are being compared (Saaty, (1987); Krois & Schulte, (2014)).

Table 4.5: Pairwise comparison scale for AHP preferences.

Numerical expression	Suitability rating	Comparative importance
1	Not suitable	Equal importance
3	Marginally suitable	Moderate importance of one over another
5	Moderately suitable	Essential or strong importance
7	Highly suitable	Very strong importance
9	Optimally suitable	Extreme importance
2, 4, 6, 8	Intermediate values between the two adjacent judgments	
Reciprocal of above numbers	If one criterion has one of the above numbers assigned to it when compared with a second criterion, then the second criterion has the reciprocal value when compared to the first.	

IV. Weighted overlay process (WOP) within GIS:

After calculating weights for each criterion, the weighted overlay process (WOP) is applied to construct suitability map, also known as Multi Criteria Decision Making (MCDM) within a Geographic Information System GIS environment. ArcGIS was used for (WOP), each criterion

raster layer is assigned calculated weight in the suitability analysis. Values in the rasters were reclassified to a common suitability scale 1 (least suitable) to 9 (highly suitable). Each raster layer is multiplied by its weight and the results are summed according to the following equation (Malczewski, (1999)):

$$A_j = \sum_{i=1}^{m} W_i X_{ij} \qquad 4.19$$

where

A_j = *final suitability score in each cell*
X_{ij} = *suitability of the i^{th} cell with respect to the j^{th} layer*
W_i = *normalised weight so that* $\sum W_i = 1$

The resulted suitability map or potential runoff storage zones map is further classified into four classes as (a) Not suitable (b) Marginally Suitable (c) Moderately Suitable (d) Optimally Suitable.

5. **Objective**: To develop an approach for operational flood extent mapping using Synthetic Aperture Radar (SAR) and preparation of flood inundation map for data scarce region using 2D flow modelling using rain on grid model.

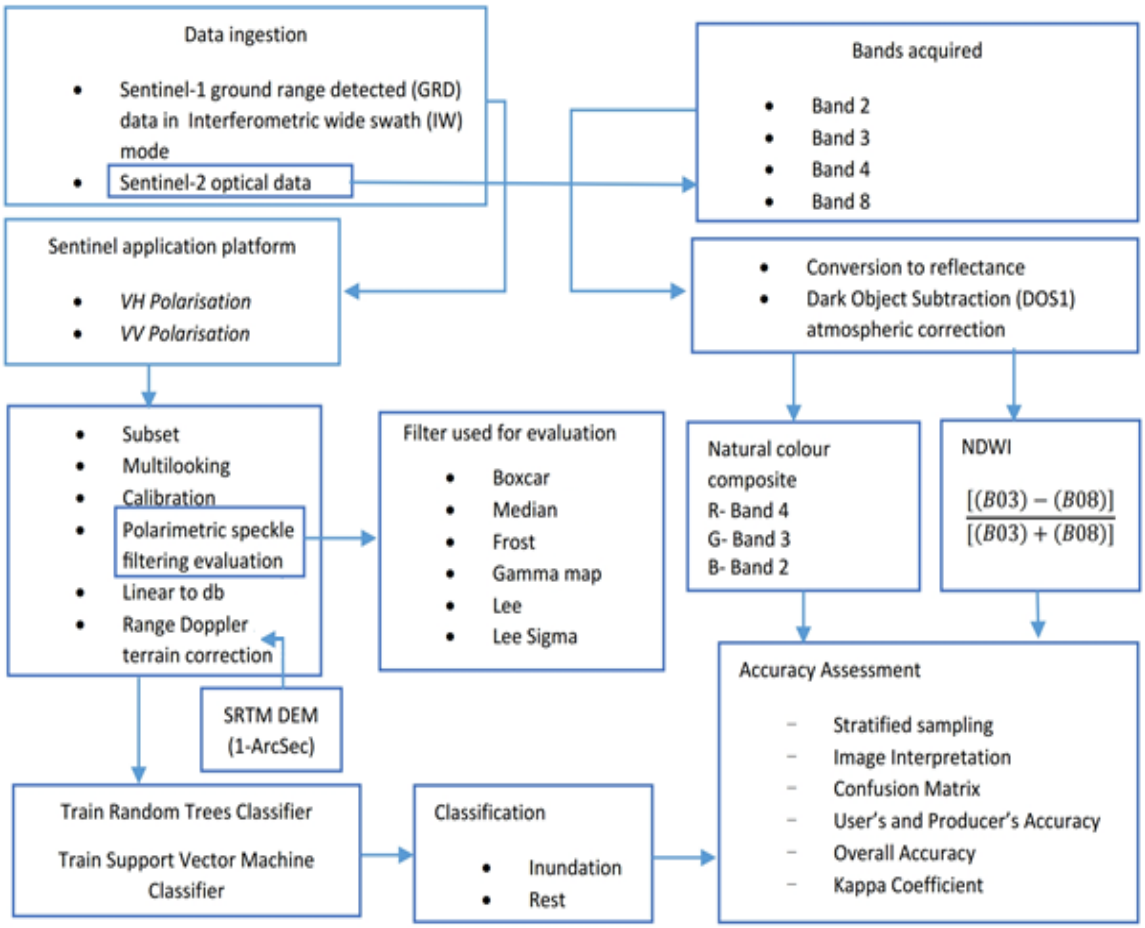

Figure 4.9: Methodology of SAR workflow.

A. Pre-processing of SAR Data

A schematic of the Sentinel-1-based processing chain is outlined in Figure 4.9. The downloaded Sentinel-1 Leve-1 GRD data acquired in IW with VV and VH polarizations were loaded onto Sentinel Application Platform (SNAP). SNAP offers a wide range of tools and features for Sentinel-1 imagery processing and analysis. Due to the large swath width (250 km) of the Sentinel-1 data, the image was first divided into a subset for the study sites to reduce the processing time. Multi-looking was then performed to reduce the standard deviation of the noise. The number of Azimuth looks and the number range of looks (2×2) with mean GR mean pixel of 20 meters were applied to a 1 m × 5 m (single look). The multi-looked data were then calibrated to transform the pixel values from the digital values recorded by the sensor into backscatter coefficient values or Sigma$_0$ (σ_0). This was achieved using the following equation:

$$\sigma 0=\frac{|DN_i|^2}{{A_i}^2} \qquad 4.20$$

DN_i = *pixel's digital number*

A_i = $absolute\ calibration\ constant$

B. Application of filters

Speckles inherently corrupt all radar images, degrading the image quality, and making it more difficult to interpret features. Thus, it is often necessary to enhance the image by filtering speckles before data can be used in different applications. All of the filters, namely, Boxcar, Median, Frost, Gamma map, Lee and Lee sigma with 3×3 and 5×5 kernel size, used in the study were available in SNAP and applied using default system parameters.

C. Evaluating the performance efficiency of filters

Several techniques are available to quantitatively assess the efficiency of a speckle filter in distinct ways, for example, edge conservation and conservation of features. The findings of the various measurements may be contradictory. Therefore, distinct techniques of evaluation should be used to discover the optimum tradeoff between the various elements of the image.

A speckle suppression filter is expected to filter the homogeneous areas with reasonable speckle reduction. A good SAR despeckling technique should have the following characteristics (i) scene feature preservation (such as texture, linear features, and point features) (ii) radiometric preservation (iii) speckle-noise reduction, smoothing, blur reduction, and edge preservation. The evaluation of the performance of the filters in de-speckling the SAR image is, therefore, necessary. Selected Homogeneous area, linear feature and Edge from Kerala SAR data is shown in Figure 4.10.

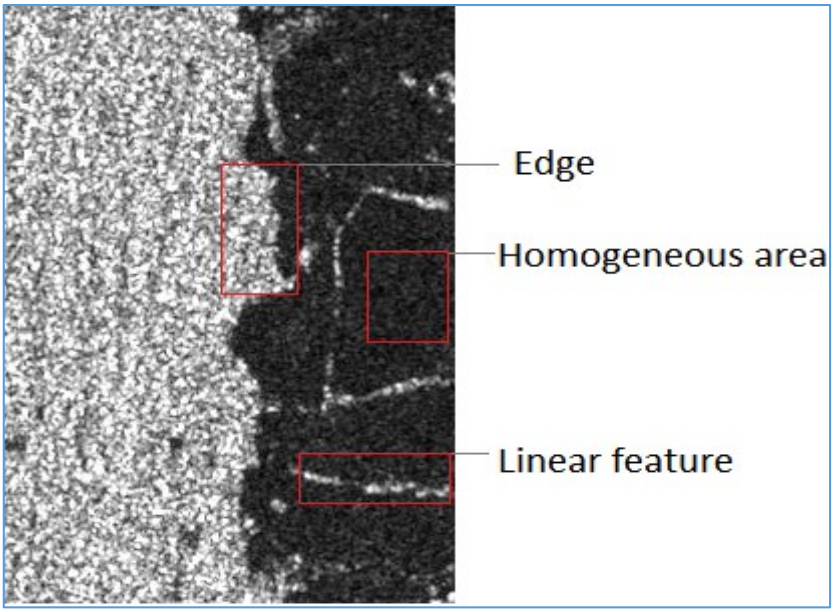

Figure 4.10: Selected Homogeneous area, linear feature and Edge from Kerala SAR data.

Following are the parameters to evaluate the performances of a despeckling filter:

I) Mean square error (MSE)

Mean square error (MSE) is the measurement of the difference between the output image and the input image. Higher the value of MSE, higher is the dissimilarity between the unfiltered image and the filtered image. A lower MSE value represents better image quality of the filtered image (Senthilnath et al., (2013)). MSE based measurements, however, yield little information about the preservation of specific features as it assesses the whole image.

$$MSE=\frac{1}{K}\left[\sum_{i=1}^{n}\left(I_u - I_f\right)^2\right] \quad 4.21$$

I_u = *unfiltered image*

I_f = *filtered image*

K = *total number of pixels*

II) Speckle suppression index (SSI)

The ability of a filter to suppress speckles is measured in terms of the standard deviation of the image to its mean intensity. For homogeneous areas, the ratio $\left(\sqrt{Var(I_u)}/_{mean(I_u)}\right)$ is regarded as the measurement of speckle strength. The speckle suppression index (SSI) is the coefficient of variance of the filtered image standardized by that of the unfiltered image, which is defined as:

$$SSI=\left(\sqrt{Var(I_f)}/_{mean(I_f)}\right) \times \left(\sqrt{Var(I_u)}/_{mean(I_u)}\right) \quad 4.22$$

$Var(I_f)$ = *variance of filtered image*

$Var(I_u)$ = *variance of the unfiltered image*

SSI has an inverse relationship with the suppression ability of the filter. The filtered image has lower variance because of speckle suppression. SSI smaller than 1.0 indicates efficient speckle suppression.

III) Speckle Mean Preservation Index (SMPI)

SSI is not accurate when the mean value is overestimated due to the existence of extreme values in a relatively lower region of the image. In addition to SSI, therefore, SMPI (Speckle Suppression and Mean Preservation Index) is used to evaluate the filter efficiency (Wang et al., (2012)). In terms of mean conservation and noise removal, lower SMPI values show better filter efficiency (Shamsoddini & Trinder, (2010)).

$$SMPI = Q \times \left(\sqrt{Var(I_f)} \Big/ \sqrt{Var(I_u)} \right) \quad 4.23$$

Where, Q is calculated as under:

$$Q = 1 + \left| mean(I_u) - mean(I_f) \right| \quad 4.24$$

IV) Equivalent Number of Looks (ENL)

Another commonly used evaluation criterion is the equivalent number of looks (ENL), also known as measure of the signal-to-noise ratio. This index is calculated using the following equation:

$$ENL = \left(\frac{mean(I_f)}{standard\ deviation(I_f)} \right)^2 \quad 4.25$$

Higher ENL value for a filter represents higher efficiency in smoothing speckle-noise over homogeneous areas (Bruniquel & Lopes, (1997)).

The performance of a filter method is evaluated by considering changes in mean and standard deviation. Ideally, the implementation of filters should not result in any change in the mean of the target image, while it should reduce the standard deviation.

The $sigma_0$ values of VV and VH polarised data with the applied filter was terrain corrected using SNAP's 'Range Doppler Terrain Correction' algorithm with an SRTM 1 ArcSecond digital elevation model. The bilinear interpolation method was used for DEM and Image resampling with a pixel spacing of 20 × 20 meters. Terrain correction helps in improving the geometric representation of the real-world surface. This is needed because, during image capture, topographical variations and off-nadir distortion unsettle the image.

D. Machine learning algorithms for classification

The terrain corrected images were classified using the random forest and support vector machine algorithms as a next step. For both the classifiers, the same number of training samples

was used. The training inundated pixels covered 5.2 Km^2 and the rest of the training pixels covered 3.1 Km^2 of the study area, Kerala. Similarly, for the study area Assam, the training inundated pixels covered 11 Km^2 and the rest of the training pixels covered 54 Km^2.

During the southwest monsoon season, it is nearly impossible to obtain 100% cloud-free data, however, a small extent of the cloud-free data can be used for validation. The normalized difference water index (NDWI) is defined for Sentinel–2 data as ((B03) − (B08)/(B03) + (B08)), where B03 is a green band and B08 is the near-infrared band. When NDWI is applied over a multispectral image, the water feature has positive values, while soil and terrestrial vegetation features have zero or negative values. This is because NIR is absorbed strongly by water but reflected strongly by terrestrial vegetation and dry soil, while in a green light, water has high reflectance than terrestrial vegetation and soil. Therefore, the NDWI was applied to extract water from the optical data. A cloud-free part of satellite optical image was collected by Sentinel-2 at 05:06:49 on 22 August 2018, 28 h after the Sentinel-1 pass over the study area Kerala. Similarly, a cloud-free part of satellite optical image was collected by Sentinel-2 at 04:27:09 on 15 July 2019, 40 h after the Sentinel-1 pass over the study area Assam. Sentinel-2 data were converted to reflectance and dark object subtraction atmospheric correction (DOS1) was applied. The corrected Sentinel-2 image was used to validate the extent of the flood. The normalized difference water index (NDWI), established earlier to extract the water from the optical data is calculated as:

$$NDWI = \frac{\rho_{Green} - \rho_{NIR}}{\rho_{Green} + \rho_{NIR}} \quad 4.26$$

Moreover, stratified random sampling was used with 500 points for the accuracy assessment. The classification maps were evaluated in terms of their overall accuracy (OA), producer's accuracy (PA), user's accuracy (UA), and the kappa index of agreement (*k*) or kappa coefficient. Confusion matrix was created to compare the kappa coefficient, producer's accuracy, user's accuracy, and the overall accuracy of the classifiers. The overall accuracy gives the correctly classified regions for the image and is calculated by the proportion of the correctly classified pixels to the total number of pixels in the confusion matrix.

To calculate inundation for entire scene, thresholding was done in SNAP by using the conditional function given below after carefully analysing the histogram:

If $Sigma_0$ VV/VH db < X then 1 else 0

X = threshold value
1 = inundated pixels
0 = Rest

The resulted output was later used to calculate the difference in inundated areas calculated by thresholding technique and data classified using the random forest and support vector machine algorithms having highest accuracy for both the study areas.

2D Hydraulic modelling for flood hazard assessment:

The intent of this work is to examine the findings of situations for which no observed data or very limited data, related to flooded areas and discharge, are available. This is a common occurrence in small watersheds, which are frequently ungauged catchments for which data for model calibration and validation is unavailable (Costabile et al., (2020)). In circumstances like these, the reliability of the commercial applications should be measured using a state-of-the-art research model that is developed for benchmarking purposes. For these reasons, an observed storm event (30-07-2019 to 03-08-2019) for modelling has been taken under study. This period of storm event witnessed the stronger than normal cross equatorial flow and active monsoon conditions over major parts of the watershed during last week of July to first phase of August in the year 2019. For rainfall-runoff simulations at the watershed scale, the runoff was evaluated with the well-known SCS-CN method, the potential maximum soil retention is calculated using following formula:

$$S = \frac{25400}{CN} - 254 \qquad 4.27$$

Where, S is in mm, and CN is the curve number (dimensionless).

The assumption of SCS-CN is that, for a single storm event, potential maximum soil retention is equal to the ratio of direct run-off to available rainfall. This relationship, after algebraic manipulation and inclusion of simplifying assumptions, results to the following expression:

$$\text{Daily Runoff (mm)}, \mathrm{Q} = \frac{(P - I_a)^2}{(P + S - I_a)} = \frac{(P - \lambda S)^2}{P + (1 - \lambda)S} \; for \; P > \lambda S \qquad 4.28$$

Q = direct run-off depth
P = total rainfall
I_a = initial abstraction

I_a and S can be related using the following equation:

I_a = λ S

λ = 0.2 was assumed in original SCS-CN model

The Hydrologic Engineering Center's Geospatial Hydrologic Modeling Extension (HEC-GeoHMS) is extension to ESRI's ArcGIS software that compute the curve number and other loss rate parameters based on various soil and land use/land cover databases. HEC-GeoHMS is used to create the curve number with the help of the Support Vector Machine classified land use and land cover map using Principal Component Analysis (PCA) based approach and soil map

containing hydrological soil groups. One of the most popular and most used model in both the scientific literature and in practice amongst the software packages using physically oriented equations. The Hydrologic Engineering Centre-River Analysis System (HEC-RAS) developed by the U.S. Army Corps of Engineers. In the latest release version (5.0.7), the HEC-RAS model is complimented by new modules, which include complete 2-D calculations based on 2-D fully dynamic equations and 2-D diffusion wave equations that ignore inertial conditions. It also provides the possibility of 1-D/2-D combined simulations, which aim to combine both a full 2-D and a full 1-D. The numerical simulation of the flood event was undertaken using HEC-RAS-v-5.07 using 2D shallow water equations:

$$\frac{\partial H}{\partial t}+\frac{\partial p}{\partial x}+\frac{\partial q}{\partial y}=r \qquad 4.29$$

$$\frac{\partial p}{\partial t}+\frac{\partial}{\partial x}\left(\frac{p^2}{h}\right)+\frac{\partial}{\partial y}\left(\frac{pq}{h}\right)=-\frac{n^2 pg\sqrt{p^2+q^2}}{h^2}-gh\frac{\partial H}{\partial x}+pf+\frac{\partial}{\rho\partial x}(h\tau_{xx})+\frac{\partial}{\rho\partial y}(h\tau_{xy}) \qquad 4.30$$

$$\frac{\partial q}{\partial t}+\frac{\partial}{\partial x}\left(\frac{pq}{h}\right)+\frac{\partial}{\partial y}\left(\frac{q^2}{h}\right)=-\frac{n^2 qg\sqrt{p^2+q^2}}{h^2}-gh\frac{\partial H}{\partial y}+qf+\frac{\partial}{\rho\partial x}(h\tau_{xy})+\frac{\partial}{\rho\partial y}(h\tau_{yy}) \qquad 4.31$$

where $H(x,y,t)=z(x,y)\ +\ h(x,y,t)$ is the surface elevation (m), z is the cell elevation in the cartesian coordinates x, y, h is the water depth (m), $p=hu$ and $q=hv$ are the specific flow in the x and y directions ($m^2\ s^{-1}$), u and v are the velocities in x and y respectively, r is the net rain (m), g is the gravity acceleration (ms^{-2}), n is the Manning's roughness coefficient ($s\ m^{-1/3}$), ρ is the water density ($kg\ m^{-3}$), τ_{xx}, $\tau_{yy}$and τ_{xy} are the components of the stress tensor and f is the Coriolis parameter (s^{-1}). When the diffusive wave is selected, the inertial terms in Equations (2) and (3) are neglected. The 2D diffusion wave equations were preferred in the present study due to their faster computing time and higher stability properties (Brunner, (2016)). The above equations are solved with an implicit finite-volume scheme. The area of the model is divided into grid cells, where each cell uses the underlying terrain data with less loss in resolution (sub grid model). For each cell and cell face HEC-RAS generates a detailed hydraulic property table (such as elevation-volume relationship, elevation-area, etc.). As regards the boundary conditions, the upstream boundary condition not specified due to the nature of the simulation. The boundary of the model is generally characterized by a closed boundary (watershed ridge line) except where an open line boundary condition with normal depth is drawn at the downstream section of the watershed to allow outflows from the watershed, which means uniform flow condition. Based on the modelled results for the storm event, inundation map is prepared for Vadodara city, which is further assisted by field sites visit.

6. **Objective:** To quantify the effects of urban land forms on land surface temperature and modeling the spatial variation using machine learning. The models can help to predict land surface temperature under temporary cloud cover spots, which are present in the data at the time of the acquisition, using neighboring biophysical (cloud-free) independent variables relationship with land surface temperature.

   The methodology used in the study is presented in Figure 4.11. The workflow was divided into six steps. First, the satellite data were subjected to image pre-processing and atmospheric correction to remove the atmospheric effect and sensor defects for land surface temperature retrieval. Second, the classification of the heat zones. Third, derivation of land use/land cover and accuracy assessment. Fourth, derivation of NDVI, NDWI and DBSI. Fifth, calculate Land Contribution Index (CI) and Landscape index (LI). Sixth, model fitting and evaluation. Each step has been discussed in detail in the following sections.

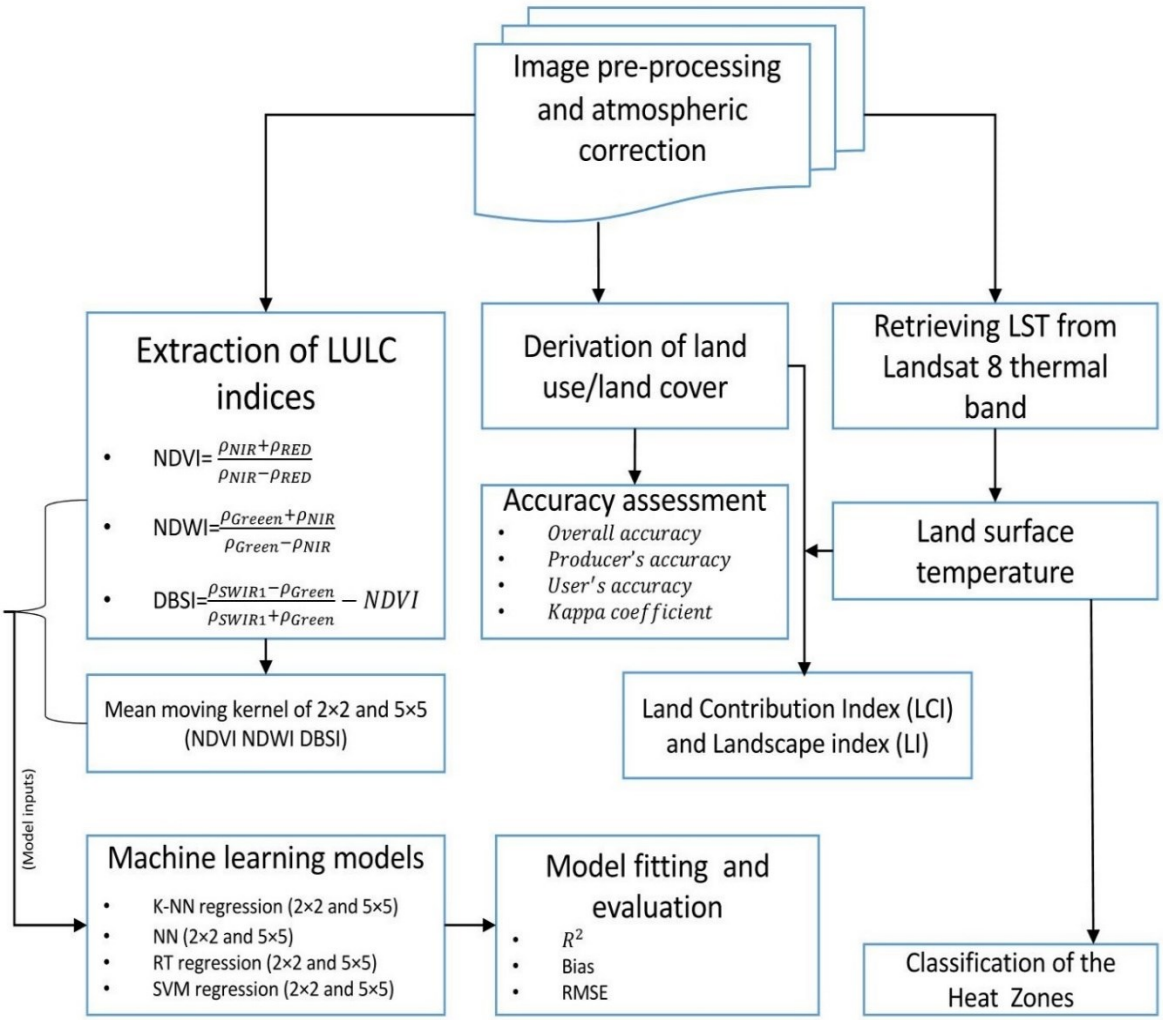

Figure 4.11: Methodology adopted to quantify the effects of urban land forms on land surface temperature and modeling the spatial variation using machine learning.

A. Retrieval of Land Surface Temperature

The two cloud-free Landsat 8 level 1T data products (ID: LC08_L1TP_148045_20180423_20180502_01 and LC08_L1TP_148045_20181101_20181115_01) were acquired from the United States Geological Survey (USGS) Earth Resources Observation and Science (EROS) Center. Landsat 8 level 1T data products are orthorectified images of the thermal infrared radiance-at-the-sensor. The land surface temperature data in summer and winter were derived from the thermal infrared sensor (TIRS) Band 10 (10.30–11.30 µm) at a spatial resolution of 100 m, resampled to 30 m using a cubic convolution resampling method (Parastatidis et al., (2017)), which were respectively acquired at 11:02:51.19 AM local time for both summer (23 April) and winter (1 November) in 2018. TIRS data were converted to top of atmospheric spectral radiance using the radiance rescaling factors provided in the metadata file using Eq. (4.32).

$$L_\lambda = \mathrm{M}_L \times Q_{cal} + A_L - O_i \qquad 4.32$$

L_λ = *at-sensor spectral radiance (W/(m².sr. µm))*
M_L = *multiplicative rescaling factor*
Q_{cal} = *quantized and calibrated standard product Digital Numbers (DNs)*
A_L = *additive rescaling factor*
O_i = *correction for Band 10*

At-sensor spectral radiance of Band 10 was converted into at-sensor brightness temperature (T_B) using the thermal constants provided in the metadata file (Table 4.6) using Eq. (4.33). To obtain the results in Celsius from Kelvin, the radiant temperature is adjusted by adding the absolute zero (-273.15°C).

$$T_B = \frac{K_2}{ln\left[\left(\frac{K_1}{L_\lambda}\right) + 1\right]} - 273.15 \qquad 4.33$$

Table 4.6: Metadata of the satellite data.

Thermal constant, Band 10	K1	774.8853
	K2	1321.0789
Rescaling factor, Band 10	M_L	3.3420E-04
	A_L	0.1
Correction, Band 10 Barsi et al., (2014)	O_i	0.29

Emissivity-corrected LST was based on Fractional Vegetation Cover (P_V) (Eq. (4.34)) and was calculated using Eq. (4.35):

$$P_V = \left[\frac{NDVI - NDVI_{min}}{NDVI_{max} - NDVI_{min}}\right]^2 \quad 4.34$$

$$LST(°C) = \frac{T_B}{1 + \left(\lambda \times \frac{T_B}{\rho}\right) \ln \varepsilon} \quad 4.35$$

Where, NDVI is normalized difference water index, T_B is the Landsat-8 Band 10 at-sensor brightness temperature; λ is the wavelength of emitted radiance; $\rho = \left(\frac{hc}{\sigma}\right)$ = 1.438 × 10^{-2} m K (where, σ is the Boltzmann constant (1.38 × 10^{-23} J/K); h is Planck's constant (6.626 × 10^{-34} Js); c is the velocity of light (2.998 × 10^8 m/s); Emissivity (ε) is calculated using (Eq. (4.36)) :

$$\varepsilon = m\,P_V + n \quad 4.36$$

$$m = \varepsilon_v - \varepsilon_s - (1 - \varepsilon_s)F\varepsilon_v \quad 4.36a$$

$$n = \varepsilon_s + (1 - \varepsilon_s)F\varepsilon_v \quad 4.36b$$

Where, F is a shape factor whose mean value, assuming different geometrical distributions, is 0.55 (Käfer et al., (2019)). ε_s and ε_v are soil and vegetation emissivity, respectively. The ε_s and ε_v values obtained for band 10 from the ASTER spectral library are 0.97 and 0.99, respectively (Sobrino et al., (2004); Haashemi et al., (2016)). The final expression for land surface emissivity is given by Eq. (4.37).

$$\varepsilon = 0.004\,P_V + 0.986 \quad 4.37$$

B. Classification of the Heat Zones

To analyse the land surface temperature variations caused by different land use/ land cover, the LST has been divided into six zones using the average (avg) and standard deviation (σ) of the surface temperature as the cut-off criteria. Study uses the equally spaced grading method. The land surface temperature zones as shown in Table 4.7. Urban heat island (UHI) were identified by the secondary high temperature zone to extremely high temperature zone range as suggested by Ma et al., (2010).

Table 4.7: Determination of the ranges of different surface temperature intervals.

Temperature Zone		Description Breakpoints
Low temperature zone	Non-UHI	$LST < LST_{avg} - \sigma$
Secondary low temperature zone		$LST_{avg} - \sigma \leq LST < LST_{avg} - 0.5\sigma$
Medium temperature zone		$LST_{avg} - 0.5\sigma \leq LST < LST_{avg}$
Secondary high temperature zone	UHI	$LST_{avg} \leq LST < LST_{avg} + 0.5\sigma$
High temperature zone		$LST_{avg} + 0.5\sigma \leq LST < LST_{avg} \sigma + \sigma$
Extremely high temperature zone		$LST \geq LST_{avg} + \sigma$

LST: land surface temperature; LST_{avg}: average land surface temperature; σ: standard deviation of land surface temperature

C. Derivation of NDVI, NDWI and DBSI from the Landsat 8

To investigate the connection of LST with biophysical variables, indices such as NDVI, NDWI and DBSI were derived from the Landsat 8.

The Normalized Difference Vegetation Index (NDVI)

Normalized Difference Vegetation Index (NDVI) is a standardized index that uses the NIR and Red bands in its formula. The index takes advantage of the contrast of the characteristics of the NIR and Red bands—high absorption by chlorophyll of red radiant energy and the high reflectivity of plant materials in the near-infrared (NIR) band. NDVI (Eq. (4.38)) produces values in the range from −1 to 1, where positive values indicate vegetated areas while negative and near-zero values signify non-vegetated surface features such as water, barren, clouds, and snow.

$$NDVI = \frac{\rho_{NIR} + \rho_{RED}}{\rho_{NIR} - \rho_{RED}} \quad 4.38$$

Where ρ_{NIR}= surface reflectance of band 5 of Landsat 8 and ρ_{Red}= surface reflectance of band 4 of Landsat 8.

Normalized Difference Water Index (NDWI)

When NDWI is applied over the optical data, NDWI (Eq. (4.39)) produces values in the range from −1 to 1. The water pixels have positive values, while soil and terrestrial vegetation pixels have zero or negative values. This is because the NDWI equation maximizes the reflectance of the water body by using the green band and minimizes the reflectance of the water body using the NIR band. The NDWI is calculated as follows (McFeeters, (1996)):

$$NDWI = \frac{\rho_{Greeen} + \rho_{NIR}}{\rho_{Green} - \rho_{NIR}} \quad 4.39$$

Where ρ_{NIR}= surface reflectance of band 5 of Landsat 8 and ρ_{Green}= surface reflectance of band 3 of Landsat 8.

Dry Bare-Soil Index (DBSI)

DBSI index was developed by Rasul et al., (2018). The index helps to distinguish between built-up areas and bare land in arid and semi-arid climate. The DBSI values can range from −2 to +2, and higher numbers represent more baresoil. The DBSI is calculated as follows:

$$DBSI = \frac{\rho_{SWIR1} - \rho_{Green}}{\rho_{SWIR1} + \rho_{Green}} - NDVI \quad 4.40$$

Where ρ_{SWIR1}= surface reflectance of band 5 of Landsat 8 and ρ_{Green}= surface reflectance of band

D. Land Contribution Index (CI) and Landscape index (LI)

The effect of land use/land cover in the warming or cooling of an area depends on the land use/land cover class and the proportion of the total area occupied by each class. In order to determine the contribution of different land use/land cover class in affecting the land surface temperature, the mean temperature of all land use/land cover class in relation to the entire study area for summer and winter seasons were calculated separately. Vegetation and water/wetlands for instance have a cooling impact on the surface because of latent heat transfer. Although they have a cooling effect, the overall value depends on the proportion of the total area they occupy. The warming or cooling extent of a land use/land cover class is quantified by the contribution index (CI) taking account of the proportion of the total area it occupies. The CI for each land use/land cover class is computed for the summer and winter seasons using Eq. (4.41).

$$CI = (T_i - M) \times P_i \quad 4.41$$

Where, T_i is the average temperature of the i^{th} land use type, M is the average temperature of the entire study area, and i represents four land use types; P_i refers to the proportion of the i^{th} land use type to the entire area(Huang et al., (2019)). A CI value greater than 0 shows corresponding land use type has a positive effect on increased heat in the city and a CI value less than 0 shows corresponding land use type helps in heat mitigation.

All land use/land cover classes can be further classified into two landscape forms, source landscape and sink landscape, to measure their contribution quantitatively. The source landscape serves as stimuli, and the sink landscape acts as a terminator to land surface temperature. Here, baresoil and builtup class are considered as source landscapes because of the positive influence on land surface temperature. Vegetation and water bodies are on the other hand considered as sink landscapes because of detrimental influence to land surface temperature. For analyzing the intensity of land surface temperature at the local scale landscape index (LI) (Eq. (4.42)) was used (Pramanik & Punia, (2019); Chen & Zhao, (2008); XU, (2009)). Different LI values indicate varying degrees of promoting and weakening the intensity of land surface temperature. A value greater than one indicates that the contribution of the sink and source landscapes weakens the intensity of the land surface temperature, a value less than one indicates the contribution of the sink and source landscapes promotes the intensity of the land surface temperature and a value equal to zero indicates the intensity of the land surface temperature is unchanged.

$$LI = \frac{\text{absolute value of the contribution index of the sink landscpe}}{\text{absolute value of the contribution index of the source landscpe}} \quad 4.42$$

E. Model fitting and evaluation

The land surface temperature is estimated and explored by four machine learning and statistical models, including K-NN regression, NN, RT regression and SVM regression. It is hypothesized in the study that the explanatory variables (NDVI, NDWI and DBSI) influence the spatial changes of land surface temperature significantly in the study area. Meanwhile, all these three explanatory variables were also calculated at the 2 levels of the observation grids unit. Since, apart from sunlight, the land surface temperature is also affected by the surrounding land cover. A mean moving kernel of 2×2 and 5×5 were used as the observation grids unit for each explanatory variable. A mean moving kernel calculates for each input pixel location, a mean of the values within a specified neighborhood around it.

To develop the models, the original dataset was divided into three parts, 70% of the whole dataset (124,578 pixels) were used as the training dataset, 20% (35,568 pixels) data were used as the testing dataset, and 10% (18,056 pixels) data were used as validation dataset.

K-Nearest Neighbor (K-NN)

Nearest Neighbor Analysis is a method for predicting cases based on their similarity to other cases. The K-NN methodology relies on a simple distance learning approach. It is based on the assumption that data points similar to each other are of the same class. It collects data from a training data set and later uses this data to make predictions for new records. In K-NN regression prediction for an arbitrary instance, the average value of the target function values of the nearest neighbors is returned as the predicted value. The K in the K-NN algorithm indicates the number of close training records that need to be taken into consideration when predicting an unlabeled test record. It has been developed in machine learning to recognize data patterns without the need to exactly match any stored patterns or cases. The number of nearest neighbors, called K, was selected automatically in SPSS modeller by cross validation. Cross-validation was used for automatic selection of the number of nearest neighbors, between a minimum K_{min} and maximum K_{max} using average error rate or sum-of square error of K. For the prediction of range target, means of nearest values were used.

A fundamental aspect of the K-NN algorithms is the metric with which the distance of data points is calculated (Wendler & Gröttrup, (2016)). The K-NN model performance depends on the choice of a distance that is used (Cunningham & Delany, (2020)). Euclidean distance was used as the distance function for K-NN. The euclidian metric describes the usual distance

between data points and is the black solid line between the two points in Figure 4.12. Object x and object y are described by (variable$_1$, variable$_2$, …, variable$_n$) = (x_1, x_2, …, x_n) and (y_1, y_2, …, y_n) (Ramli et al., (2019); Cigdem & Ozden., (2018)). Using the vector components x_i and y_i, the metric is defined as given in Eq. (4.43):

$$\text{Euclidean distance} \quad d(x,y) = \sqrt{\sum_{i=1}^{n}\left(x_i - y_i\right)^2} \qquad 4.43$$

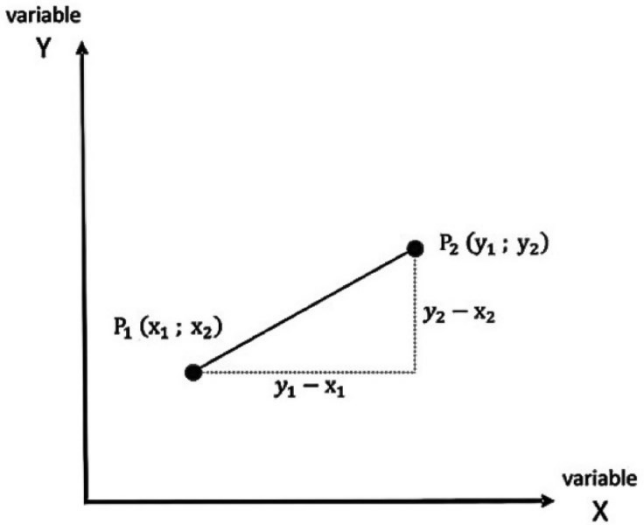

Figure 4.12: Euclidian distance between two data points in a 2-dimensional space.

Neural networks

Neural network modeling was completed in IBM SPSS Modeler, the neural networks used are feed-forward neural networks, also known as multilayer perceptrons (MLPs). MLPs are based on plotting performance error as a function of weights. Each iteration in the algorithm uses forward activation to produce a solution and a backward propagation of the computed error to modify the weights. The general architecture for MLP networks is: an Input layer, one or multiple Hidden layer(s), and an Output layer Figure 4.13. A single hidden layer was used in our study within the SPSS modeller environment. The input layer consists of the initial neurons receiving raw data without processing. Every neuron in the input layer handles an input variable that transforms through the activation function and values are propagated from each neuron to every neuron in the next layer. The activation function used by the IBM SPSS modeller is the hyperbolic tangent function$(\tanh(c) = \frac{e^c - e^c}{e^c + e^c})$ (IBM, (2015)). In contrast, neurons in the output level receive the data processed by several neurons in the network and calculate a final score for each target class, e.g. probability and prediction. Each neuron in the output layer is a target

category and gives the result for that category. The network learns by examining individual records, generating a prediction for each record and adjusting weight whenever an incorrect prediction is made.

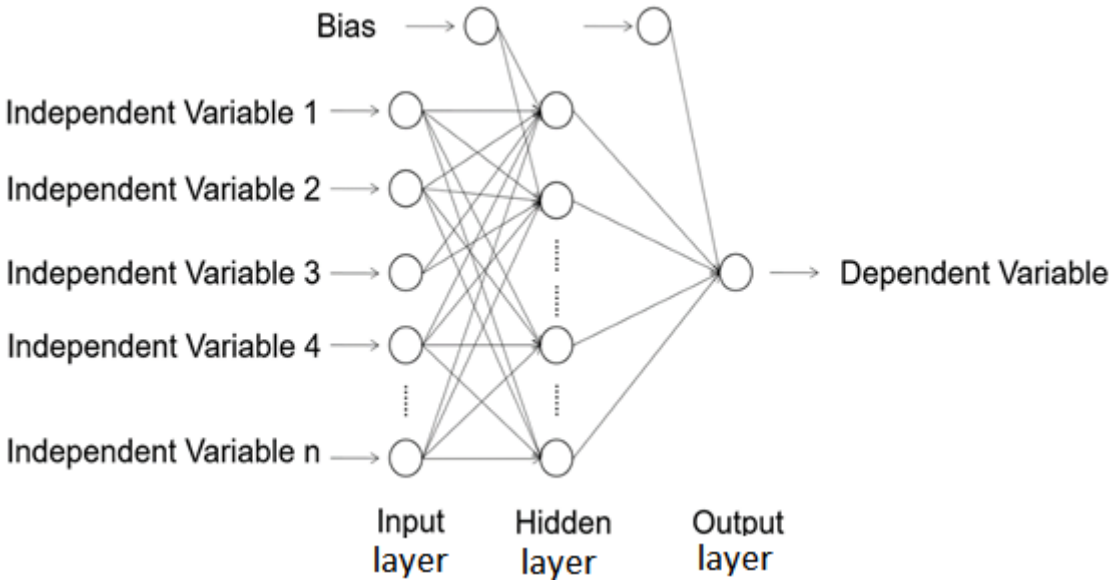

Figure 4.13: Sketch of a typical neural network.

Random Trees

Random Trees is a tree-based prediction method based on the methodology for the regression tree. To generate sample data, it uses replacement bootstrap sampling. This action creates bootstrap samples that are equal to the original dataset, following which each replica is constructed on a component model. These component models together constitute an ensemble model. The data from the sample is used to generate a tree model. During tree growth, it will not sample the data again. It selects part of the predictors randomly and uses the best part of the selection to split the node. Each tree grows without pruning, and on scoring, it combines individual tree scores by average. Random Trees uses classification and regression tree-like trees. Hence, most of the Random Trees characteristics are inherited from the classification and regression tree.

SVM regression

The SVM used to solve the regression problem is called support vector regression. It is a supervised machine learning method, which was first introduced by Drucker et al., (1997). The basic idea is to map the input data into a high dimensional feature space via a non-linear function, i.e., the kernel function, and then a linear regression problem is obtained in the feature space. Consider a training data set of the input vector. Consider a training data set of input vector:

$$G = \{x_i, d_i\}_{i=1}^{N} \quad (4.44)$$

Where, x_i= input vector and , d_i = actual values and N = number of data points

The SVR function is given by

$$f(x) = \langle w, \varphi(x) \rangle + b \quad (4.45)$$

Where, ⟨⟩ is the dot product and $\varphi(x)_{i=1}^{N}$ is the feature which is obtained by nonlinear mapping of input space x. The coefficients w and b are calculated by minimizing the regularized risk function.

Three measures, namely, coefficient of determination (R^2) (Eq. (4.46)), bias (Eq. (4.47)), and root-mean-square error (RMSE) (Eq. (4.48)) were used to evaluate the performance of the models for training, testing and validation.

In the equation below, R^2 is the coefficient of determination between the original and predicted land surface temperatures. A high R^2 indicates a satisfactory prediction.

$$R^2 = 1 - \frac{\sum(LST_p - LST_a)^2}{\sum(LST_p - \overline{LST_a})^2} \quad (4.46)$$

Where LST_p is the predicted land surface temperature, $LST_a$is the actual land surface temperature and $\overline{LST}_a$is the average of actual land surface temperature.

Bias and RMSE were used to test the errors between the predicted land surface temperature and the actual land surface temperature. The calculation formulas for bias and RMSE are as follows:

$$\text{Bias} = \frac{\sum_{i=1}^{n}(LST_p - LST_a)}{n} \quad (4.47)$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(LST_p - LST_a)^2} \quad (4.48)$$

Where n represents the number of pixels of the data.