Chapter Three: Methodology

3.1 Data selection

The DWAF's web site contains numerous river flow data sets categorized under different drainage regions across South Africa (DWAF, 2008). The drainage regions were studied to select the data sets with the longest record length and with the least amount of missing data. Five sets of long-term river flow data were selected to represent the eastern, western, eastern central-interior, and western central-interior (Table 1, Chapter one) of South Africa. Since rainfall is a major factor responsible for changes in river flow, rainfall trends were studied first, followed by the study of river flow patterns. A list of rainfall stations across southern Africa was obtained from the South African Weather Service (SAWS) and the rainfall stations upstream of the river flow gauges were selected on the basis of data length and the quantity of missing data. The rainfall stations upstream to the river flow stations were selected because the former impacts the downstream flow of rivers. The SAWS maintains a rain-gauge network, with unavoidable gaps in the records of some of the rainfall stations (Tennant & Hewitson, 2002). A guiding principle for determining trends and patterns in time series data is the length of data records. This is critical to determine the existence and rate of change of climate effects, thus Westmacott & Burn (1997) suggest a minimum of 30 years of record length. Rainfall data from 13 of the SAWS stations across South Africa and two stations from Lesotho (data obtained from Lesotho Meteorological Services) were analysed (Table 1, Chapter one). The daily rainfall data supplied by SAWS was processed to obtain monthly and annual rainfall totals by adding the daily or monthly values for any given month or year to obtain a total for the month or year respectively. The missing data for any given month were calculated by averaging all the values across the years for the specific month (yellow highlights in tables in Appendix).

All the SAWS rainfall data used in this study were recorded at rainfall stations, with the exception of the Bloemhof station, which is an Electric Temperature station (SAWS, 2008). The most common method of measuring rainfall is through the use of a rain gauge (Tennant & Hewitson, 2002). The rainfall stations are only equipped with a 127 mm rain gauge and rainfall

is measured daily at 08:00 South African time, while the Electric Temperature stations are automatic weather stations and have data for most elements every 5 minutes (SAWS, 2008). Although the Touwsrivier rainfall data extends back to 1886, it has large gaps of approximately 14 years, and thus the current study only considers the more continuous data from 1918 onwards.

3.2 Analysis

Inter-annual, seasonal and 5-year moving average rainfall trends were analysed over long-term periods for each region to determine patterns of climate change, using the Mann Kendall statistical method, and quantified using the Sen's method. The Sen's method was used instead of linear regression to limit the influence of outliers on the slope (Novotny & Stefan, 2007). The river flow data were correlated with the climate data to establish a possible long-term climate impact on river flow, using the Kendall tau statistic.

The data series were subdivided into early rainfall, mid rainfall and late rainfall seasons to account for the seasonal changes, with the early season defined as October and November, the mid season defined as December to February, and the late season defined as March and April for the summer rainfall regions (Tyson, 1986) (i.e. the Orange, Tugela, Mgeni and Vaal River regions). For the winter rainfall region the early rainfall season was defined as April and May, the mid rainfall season as June and July and the late rainfall season as August and September (Tyson, 1986). The seasonal percentage changes were calculated using the difference of the averages of the first and second half of the period of record of the shortest length in the specific catchment.

The rainfall and river flow data were analysed using the Mann Kendall (MK) statistical trend analysis, which is an effective tool for identifying trends in hydrologic and other related variables (Westmacott & Burn, 1997). MK is a rank based non-parametric test (Novotny & Stefan, 2007) and the most important reason for the use of non-parametric statistical analysis compared to parametric tests is that the former has a higher power for non-normally distributed data, which is characteristic of hydro-meteorological data (Yue *et al.*, 2002). MK is a reliable method of identifying monotonic linear and non-linear trends in non-normal data sets with outliers (Burns *et al.*, 2007). The basis of the Mann Kendall test is the null hypothesis where the

sample of data is independent and distribution free, which means that no trend exists in the data set (Westmacott & Burn, 1997). The MK method first calculates the *S* variable, which is the sum of the difference between data points:

$S=\sum_{i=1}^{n=1} \sum_{j=i+1}^{n} \text{Sgn}\,(x_j - x_i)$,

where *n* is the number of values in the data set, Sgn is sign and

$\text{Sgn}\,(x_j - x_i) = 1$ if $(x_j - x_i) > 0$,

$\text{Sgn}\,(x_j - x_i) = 0$ if $(x_j - x_i) = 0$ or

$\text{Sgn}\,(x_j - x_i) = -1$ if $(x_j - x_i) < 0$ (Novotny & Stefan, 2007).

S is normally distributed when *n* is greater than or equal to 8 (Novotny & Stefan, 2007), which allows for the calculation of the standard normal variable Z, with a mean value of zero and a variance of one (Yue *et al*., 2002), calculated by:

$Z = \frac{S-1}{\sqrt{Var\,(S)}}$ if $S>0$

$Z = 0$ if $S=0$

$Z = \frac{S+1}{\sqrt{Var\,(S)}}$ if $S<0$ (Partal & Küçük, 2006),

where *Var* is the variance.

The normalized test statistic (Z) was calculated in this study using the Microsoft Excel template, Makesens (Mann Kendall test for trend and Sen's slope estimates) (Makensens, 2002). If the probability value (*p*-value) is less than or equal to the significance level, the null hypothesis that a trend does not exist in the data set is rejected (Novotny & Stefan, 2007). A decreasing trend is indicated by a negative Z value and a probability value greater than the level of significance. An increasing trend is indicated by a positive Z value and a probability value greater than the level of significance. A probability of less than the level of significance indicates no trend (Novotny & Stefan, 2007). The p-value was calculated using the Microsoft Excel function, Normsdist (Khambhammettu, 2005). A significance level of 0.05% was chosen for this study.

The Sen's non parametric method was used to determine the magnitude of the trend slope (Salmi *et al.*, 2002). Sen's slope is the median of all possible pair wise slopes (Burns *et al.*, 2007), and assumes the trend is linear and is calculated by:

$f(t) = Qt + B$,

where Q is slope and B is a constant indicating the vertical axe crossing of the slope line and

Q is obtained by calculating the slopes of all data pairs:

$Q_i = x_j - x_k / j-k$, where $j>k$ (Salmi *et al.*, 2002).

The Kendall tau was used to determine the correlation between rainfall at the different stations and river flow in each region. The Kendal tau correlation coefficient is calculated by

$$\tau = \frac{S}{n(n-1)/2},$$

where $S=P-M$ and P is the number of times y increases as x increases and M is the number of times y decreases as x increases (Helsel & Hirsch, 1992).

Winstat was installed into Excel and the correlation coefficient was calculated using Kendall tau (Winstat, 2007). Kendall tau measures the strength of a monotonic relationship (Helsel & Hirsch, 1992). A Kendall tau coefficient (τ) of 1 reflects that the two rankings in two data sets are the same; a value of -1 indicates that the one ranking is the reverse of the other and a value of 0 reflects that rankings are completely independent. A coefficient between −1 and 1, and increasing values imply increasing agreement between the rankings. The Mann Kendall statistic, Sen's method and Kendal tau statistic were applied to the inter-annual, seasonal, and 5-year moving average of rainfall and river flow data, as presented in chapter four.

The 5-year moving average rainfall trends were also analysed using the Microsoft Excel data analysis function moving average. A moving average was used to provide trend information about the historical data which the normal average of historical data would mask (Microsoft Excel help). The annual percentage changes were calculated using the difference of trend data over the period of record. A 5-year period was chosen since the record lengths of some of the datasets are short therefore a longer duration would not be suitable. In addition, Novotny &

Stefan (2007) used a 5-year period and proved to work well in finding trends in the data. The amplitudes of wet and dry periods were calculated and the trend over time and the significance was determined using regression analysis since the quantity of amplitudes within a data set was too small to use the Mann Kendall statistic.

3.3 Methodological limitations

A major limitation of this study is the relatively short temporal record for which rainfall and river flow data are available for analysis. The majority of the records were too short for a comprehensive trend analysis, thus also limiting the determination of cyclical patterns. In addition, the various data sets in the same catchment and across catchments need to be compared with caution due to the different lengths of data records. For instance, the data set of the shortest duration had to be used in order to determine correlation coefficients. In the 5-year moving average analysis, the lengths of records affected the calculation of the long-term averages, resulting in a variation in the benchmark used to determine wet or dry periods. As a result, the frequencies of wet and dry periods within data sets are difficult to compare across stations. An additional limitation with the 5-year moving average analysis is the calculation of the amplitudes, as only the maximum peak in a given wet or dry phase was chosen. Due to time constraints and for the sake of keeping the current project focused, five year moving average peaks of slightly lower values were not taken into account for the trend analysis, neither were the peak curvatures taken into consideration, yet it is acknowledged that these may be important variables that require consideration for a more complete picture of the overall trends.

The lack of high quality long-term rainfall and river flow data was a major concern for this study. For instance, the longest river flow record for the Mgeni River catchment was only 57 years. The analysis of numerous rainfall stations within a region is vital for a complete study of regional rainfall change (Tyson *et al.*, 1975) because two stations might show opposite trends as observed in the New Hanover and Mistley stations in the Mgeni River catchment. Additional station data would have provided a more complete picture of the rainfall trends in the area. Easterling *et al.* (2000) identified the lack of high quality long-term data as a major limitation in determining historical climate trends (Easterling *et al.*, 2000).

The missing data for any given month were substituted with averages of all the values of the specific month across the years of the record (see Table 2 and the yellow highlights in the Appendix tables). This may have affected trends observed in the 5-year moving average analysis, by either increasing or decreasing averages, and might not be an absolutely accurate reflection of the real trend for the area. The high variability of South African river flow, together with heavy sediment and debri loads affect the measurement of flow. Consequently, the need has arisen to modify gauging stations (Wessels & Rooseboom, 2009), which raises the question as to the consistency of data capturing and quality assurance, and in fact raises doubt as to the quality of data prior to the substitution of new gauging stations. The accuracy of the data, quality of monitoring equipment, and consistency with data recording, is exceptionally difficult to verify as different authorities and individuals have been responsible for the process over time, and written records on such matters are scarce or unavailable. The calibration of rain gauges is required to produce consistent data between stations, since the comparison of data from different rain gauges does not produce accurate results. In addition, windy conditions result in errors in measurement because the amount of rain caught for measurement is reduced by wind driving the rain away from the gauge (Nel & Sumner, 2005). Sene *et al*. (1998) demonstrated that rainfall varies with elevation and location across Lesotho. Since rainfall stations were chosen on the basis of record length for this research report, the chosen rainfall stations are at variable altitudes, and this too would have affected the observed differences between sites.

Table 2: Rainfall and river flow data sets in southern Africa showing data gaps.

Stations	Historical record	% data gaps	Data gaps
Tugela Ferry River flow	1927 to 2008	73 months missing 7.5%	April to Dec 1930, March to Dec 1931, Nov 1932, June, July and Dec 1943, Jan to Oct 1944, March to Dec 1968, Jan to June 1969, Jan 1970, July 1976, Nov to Dec 1977, Jan to Oct 1978, Oct to Nov 1987, Jan to May 1988, Feb 1996, Dec 2007
Swartwater rainfall station	1932 to 2008	26 months missing 3%	April 1935, Jan to March 1953, Nov 1956, Dec 1960, Feb and July 1984, March to July 1985, May and June 1987, April to June 1988, July 1990, Feb 1993, Aug to Dec 2008
Moorside rainfall station	1914 to 2008	28 months missing 2.5 %	Jan to Feb 1914, April 1920, Dec 1924, Jan to Feb 1925, March 1928, March 1929, Dec 1931, Jan to Sept 1932, May and Sept 1938, Feb 1941, March 1950, Nov 1964, Feb 1970, Aug 1991, Oct to Dec 2008

Tugela Ferry rainfall station	1926 to 2008	20 months missing 2%	Nov 1926, Nov 1933, Feb 1944, May 1951, Feb 1953, Nov 1966, Feb to Oct 1972, Nov 1979, Aug 1991, Oct to Dec 2008
Mistley Estate rainfall station	1940 to 2008	41 months missing 5%,	June to July 1946, Aug 1949, June 1950, July 1953, March 1954, May, July, and Sept to Nov 1955, April to Aug 1958, Aug 1964, April 1965, July 1986, Jan to Feb, and April to Dec 1989, Jan to July 1990, Feb, and Oct to Dec 2008
New Hanover rainfall station	1940 to 2008	12 months missing 1,5%	Jan to July 1941, Oct 1978, Jan 2005, Oct to Dec 2008
Mgeni River at Table Mountain	1951 to 2008	4 months missing 0.6%	Jan 1957, Oct 1960, Jan to Feb 1961, Jan to Oct 1962, Oct to Dec 1994, Jan to Sept 1995, Jan 2005, June to Dec 2008.
Orange at Aliwal River flow	1914 to 2008	7 months missing 0.5%	Feb 1952, April to Sept 2007
Lille rainfall station	1932 to 2008	42 months missing 5%	Feb 1933, Dec 1934, March 1935, Oct 1937, March and Oct 1938, Aug 1939, Oct 1940, Sept 1941, Jan 1943, Oct 1949, Feb, April, May, and Oct to Dec 1953, Feb, May, July and Dec 1954, Jan, May and July 1955, Jan, Feb, April, and Sept to Nov 1956, Feb, May, Aug, Sept, Dec 1957, Jan, April, Sept, Nov, Dec 1958, Oct 2003
Middelplaats rainfall station	1911 to 2008	11 months missing 1%	Feb 1912, Jan 1930, March, April, and Oct 1931, Feb 1937, Nov 1939, Oct to Dec 2008
Zastron rainfall station	1906 to 2008	58 month missing 5%	June to Aug 1906, April, Oct, Nov 1910, Jan to May 1912, July 1913, Oct to Dec 1914, Oct and Dec 1915, June, Sept 1921, Nov 1922, March 1923, Sept 1926, Aug, Nov 1929, Oct 1939, Jan, March, April, Nov 1955, Jan, Feb, April, June, July, Nov, Dec 1956, Jan to March, Sept, Oct, Dec 1957, Jan to April, Dec 1958, Dec 2008
Thaba Tseka rainfall station	1965 to 2006	60 months missing 12%	1971 to 1975
Semonkong rainfall station	1967 to 2006	5 months missing	Jan to May 1967

		1%	
Vaal River at Schoolplaats	1940 to 2008	8 months missing 1%	Jan to July 1940, Dec 2008
Villiers rainfall station	1905 to 2006	84 months missing 7%	Jan, July, Oct and Nov 1905, July 1906, Feb, April to Dec 1909, Jan to Aug and Nov 1910, March 1911, Jan to Sept 1912, May, June and Sept 1913, 1914, April 1918, June 1923, Feb 1925, May to Sept 1927, March 1932, Feb, May to July 1933, Aug 1942, March 1943, Jan 1953, Aug 1957, Sept 1962, Sept 1968, Feb 1978, Aug 1991, May, June 1991, May to Dec 2002, Jan, Feb 2003
Bloemhof rainfall station	1931 to 2008	21 months missing 2%	March 1935, July 1936, Sept 1937, June, Sept, 1938, May 1941, Sept 1944, Jan 1947, June 1948, June 1954, Jan and Aug 1956, March and July 1957, Feb 1959, Nov 1968, Sept 1975, July 1991, March and April 1999, Dec 2008.
Klerksdorp Hartbeesfontein	1917 to 2008	42 months missing 4%	Oct 1918, Feb and Oct 1925, Nov 1927, Sept 1929, Jan and May to Sept 1940, Jan to July 1948, Jan 1950, Jan 1953, Jan 1969, Sept and Dec 1978, April and Dec 1997, March and Dec 2000, Sept to Dec, Jan 2004, July to Dec 2008.
Touwsrivier rainfall station	1918 to 2008	50 months missing 5%	Jan, April 1918, Sept to Dec 1923, Jan to May 1924, Dec 1925, Sept and Dec 1930, Sept 1931, April, June, Aug, Nov, Dec 1932, Aug, Sept, Nov 1935, April, May, July to Sept 1936, Dec 1942, Oct to Dec 1943, June 1944, Dec 1947, Oct to Nov 1961, Jan to June 1962, Jan 1963, March 1996, Nov 2007, Nov to Dec 2008
Malabar rainfall station	1944 to 2008	38 months missing 5%	Jan to May 1944, July and Dec1948, May 1949, June 1950, July, Sept to Dec 1959, Jan to May 1960, July to Dec 1977, Feb to Aug 1978, Dec 1996, Jan to Feb 1997, Oct to Dec 2008
Breede River	1923 to 2008	31 months missing 3%	Jan 1923, Oct to Dec 1933, Jan 1934, March to July 1944, Oct to Dec1952, Jan 1953, Jan to April 1956, July, Nov, Dec 1957, Jan 1958, June to Dec 1967, Oct to Dec 2008