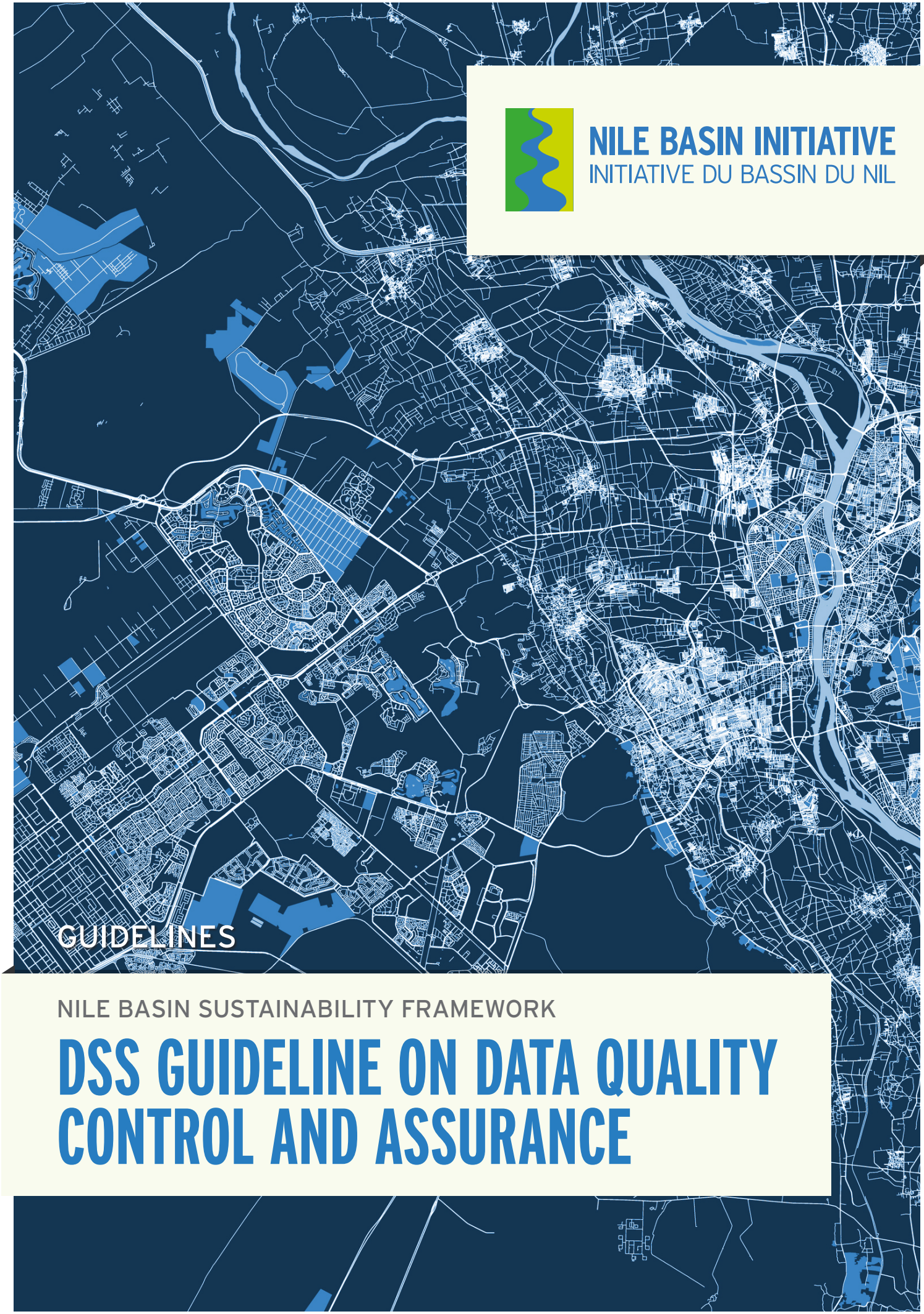


NILE BASIN INITIATIVE
INITIATIVE DU BASSIN DU NIL

GUIDELINES

NILE BASIN SUSTAINABILITY FRAMEWORK

DSS GUIDELINE ON DATA QUALITY CONTROL AND ASSURANCE

Photo: Shutterstock

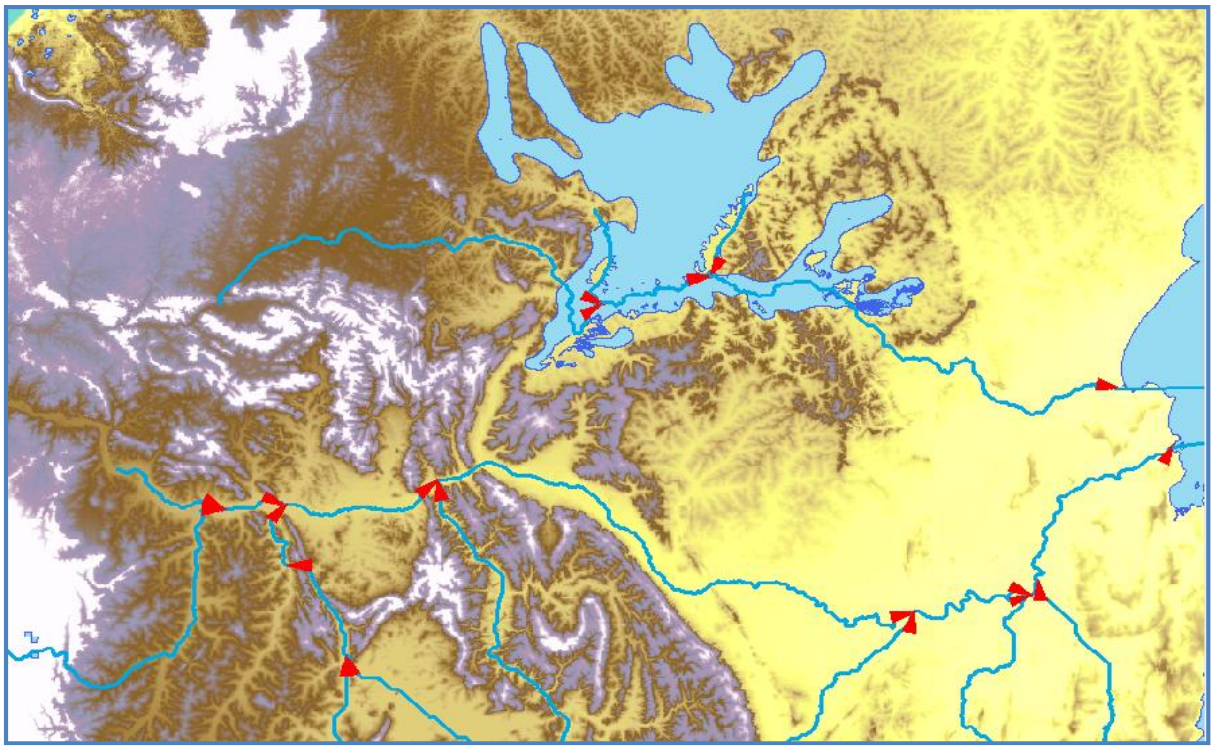
Document Control Sheet

Title	DSS Guideline on Data Quality Assurance
Document type	☐ Policy ☐ Strategy ☒ Guidelines ☐ Legal and Foundational Document
Prepared by	☐ Nile-SEC ☐ ENTRO ☐ NELSAP-CU ☒ Other: WRPM Project
Status	☒ New Policy/Strategy/Guideline/Legal and Foundational Document ☐ Revision of existing Policy/Strategy/Guideline/Legal and Foundational Document
Revision Date	
Effective Date	2012

Consideration by Nile-COM~~/EN-COM/NEL-COM~~ (cross out whichever body is not applicable)	
Date of submission for consideration	
Action by Council of Ministers	
Comments satisfactorily addressed	☐ Yes ☐ No ☐ Not Applicable

Consideration by Nile-TAC~~/ENSAPT/NEL-TAC~~ (cross out whichever body is not applicable)	
Date of submission for consideration	
Action by the Technical Advisory Committee:	
Comments satisfactorily addressed	☐ Yes ☐ No ☐ Not Applicable

Responsible Officer: Dr. Abdulkarim H. Seid



DATA QUALITY ASSURANCE GUIDELINE:

Data Processing, Quality Assurance and Metadata

Final

December 2012

Contact person:
V Jonker
Aurecon Centre
1 Century City Drive
Waterford Precinct
Century City, Cape Town, RSA
+27 21 526 9400
Verno.Jonker@aurecongroup.com

Submitted to:
Nile Basin Initiative
Water Resource Planning
and Management Project
Dessie Road
Addis Ababa
Ethiopia

In association with:

SOLARIS Engineering & Consulting

BEUSTER, CLARKE & ASSOCIATES
GIS Application Development, Software Development, Database Development

Nepid Consultants

Conningarth Economists

Climate Systems Analysis Group
(CSAG), University of Cape Town

Data Compilation and Pilot Application of the Nile Basin Decision Support System (NB-DSS): Work Package 2: Stage 2

DATA QUALITY ASSURANCE GUIDELINE:

Data Processing, Quality Assurance and Metadata

Prepared by:

aurecon

In association with:

BEUSTER, CLARKE & ASSOCIATES
GIS Application Development, Software Development, Database Development

Climate Systems Analysis Group
(CSAG), University of Cape Town

SOLARIS Engineering & Consulting

Conningarth Economists

Nepid Consultants

Prepared for:

December 2012

Final

PROJECT NAME	**:**	Data Compilation and Pilot Application of the Nile Basin Decision Support System (NB-DSS): Work Package 2: Stage 2
REPORT TITLE	**:**	**DATA QUALITY ASSURANCE GUIDELINE:** Data Processing, Quality Assurance and Metadata
AUTHORS	**:**	Hans Beuster
REPORT STATUS	**:**	Final
AURECON REPORT NO.	**:**	7332/107486
DATE	**:**	December 2012

Submitted by:

.....	.....
V JONKER Study Leader	(Date)
.....	.....
M KILLICK Project Director	(Date)

Approved for the NBI WRPMP

.....	.....
HA GHANY Regional Manager	(Date)
.....	.....
AH SEID DSS Lead Specialist	(Date)

DATA QUALITY ASSURANCE GUIDELINE: Data Processing, Quality Assurance and Metadata

Project	Data Compilation and Pilot Application of the Nile Basin Decision Support System (NB-DSS): Work Package 2: Stage 2
Main Authors	Hans Beuster
Status	Final
Date	December 2012

Reason for Circulation
Final

Circulation List
NBI DSS Core Team
NBI Country representatives

Nature of comments required	
Email comments Marked-up hard copy or electronically with track changes	
Date of Receipt of Comments	

Dr V. Jonker: Study Leader
Aurecon Centre, 1 Century City Drive,
Waterford Precinct, Century City, Cape Town
Republic of South Africa
Tel: +27 21 526-9400
Fax: +27 21 526-9500
Verno.Jonker@aurecongroup.com

TERMINOLOGY

Development Intervention	A specific infrastructure implementation for regulating the water resources of a basin (e.g. dams, canals, irrigation systems, etc).
Management Intervention	A specific plan for the allocation and/or operation of the water resources of a basin aimed at prioritizing hydropower production, minimizing environmental impacts, etc.
Ecological Water Requirement	The flow patterns needed to maintain an aquatic ecosystem in a particular condition or future desired state.
Indicator	A socio-economic, environmental or hydrological characteristic that can be quantified across different model scenarios, for the purpose of choosing between alternative development and/or management scenarios.
Model Validation	A process whereby a model's "fitness for purpose" is assessed through a set of validation tests performed with a calibrated model.
Scenario	A contemplated state of a basin induced either through targeted human intervention (e.g. combinations of development and management interventions) or through externalities (e.g. climate change, economic policies etc.).
Quality Control	Data processing and analysis procedures to ensure that data comply with specified acceptance criteria.
Quality Assurance	Processes which involve ensuring that data sets and models are properly documented and auditable.
Decision Support System	A tool which supports decision making and the integrated management of a river basin based on the integration of the results of various analyses and the evaluation of scenarios and their implications.
Multi Criteria Decision Analysis	A structured approach towards solving decisions and planning problems involving multiple criteria.
Cost Benefit Analysis	A systematic process for calculating and comparing benefits and costs of a project to determine if it is a sound investment and/or to evaluate how it compares with alternate projects.
Integrated Water Resource Management	A participatory planning and implementation process, based on sound science, that brings stakeholders together to determine how to meet long-term needs for water while maintaining essential ecological services and economic benefits.
Ecoregion	A large area encompassing one or more freshwater systems that contains a distinct assemblage of natural freshwater communities, based mainly on the distribution of fish species (Abell *et al*. 2008). The freshwater species, dynamics, and environmental conditions within a given ecoregion are more similar to each other than to those of surrounding ecoregions and together form a conservation unit.

LIST OF ACRONYMS

BCR	Benefit Cost Ratio
CBA	Cost Benefit Analysis
CORDEX	COordinated Regional climate Downscaling EXperiment
CSAG	Climate Systems Analysis Group (Univ. Cape Town, South Africa)
DEM	Digital Elevation Model
DRIFT	Downstream Response to Imposed Flow Transformation
ELOHA	Ecological Limits of Hydrologic Alteration
EWR	Environmental Water Requirement
FSL	Full Supply Level
GCM	General Circulation Model
GDP	Gross Domestic Product
GIS	Geographic Information System
IMS	Information Management System
IRR	Internal Rate of Return
IWRM	Integrated Water Resource Management
MAP	Mean Annual Precipitation
MCDA	Multi Criteria Decision Analysis
MUSLE	Modified Universal Soil Loss Equation
NPV	Net Present Value
PES	Present Ecological State
QA	Quality Assurance
QC	Quality Control
RCM	Regional Climate Model
SAM	Social Accounting Matrix
SOMD	Self-Organizing Map based Downscaling
SAP	Strategic Action Plan
WMO	World Meteorological Organisation
WP	Work Package
WRPM	Water Resource Planning and Management
XML	Extensible Markup Language

Table of Contents

List of Tables

List of Figures

Annexures

1. INTRODUCTION

1.1 BACKGROUND AND SCOPE OF WORK

A Data Quality Assurance (QA) System and Quality Control (QC) procedures were developed as part of the preceding WP 2/1 Stage 1 consultancy. The QA system was integrated within the NB DSS so that the source, lineage (processing steps) and quality of the data and models that form part of the DSS are properly documented, allowing for future improvement of data and model components with questionable quality. The QA System will also assist future users to evaluate outputs of the DSS with a full understanding of the quality of the data and models that constitute the DSS.

The QA system and QC procedures that have been developed by the preceding consultancy include procedures for stream flow gap filling, evaluation of water balance model (MIKE Basin) performance and metadata templates to document these. As part of this consultancy, the QA system and QC procedures were extended to include quality control procedures for rainfall gap filling and extension, calibration of rainfall-runoff and hydrodynamic models, processing and handling of spatial data sets, and metadata associated with these.

1.2 PURPOSE OF THIS REPORT

The purpose of this report is to build on the technical notes and QA guidelines developed in the WP2/1 Stage 1 consultancy (these guidelines are integrated into this report) by providing meta-data standards and QC procedures associated with rainfall gap-filling and extension, and processing and handling of spatial data sets. It should be read in conjunction with the "*Model Calibration and Validation Guideline*" and "*Guideline for the Evaluation of Water Management Interventions* " companion volumes.

1.3 REPORT STRUCTURE

Chapter 2: Quality Assurance and Metadata. The metadata templates developed by WP2/1 Stage 1 are expanded to meet the data requirements of WP2/1 Stage 2.

Chapter 3: Data Quality Control. Data quality control procedures for infilling and extension of stream flow and rainfall data, and for quality control of spatial data sets are introduced and discussed.

Chapter 4: References.

2. QUALITY ASSURANCE SYSTEM AND METADATA

2.1 INTRODUCTION

The objective of the QA system is to document the source, lineage (processing steps) and quality of the data and models that form part of the DSS. Documentation is done in the Metadata Manager of the DSS. Except where otherwise indicated in footnotes, the following statements and definitions have been adopted (with minor additions) from the WP2/1 Stage Technical Note on *Quality Assurance System and Metadata* (O'Donnell, 2011):

Purpose of the Quality Assurance System

The data sets and models developed in this project will be used in WP2 Stage 2 and later work, so:

> *the **principal aim** of the quality assurance system is to ensure that: (1) the data and models are properly documented in a simple but robust fashion; (2) this documentation is appropriate for, and directly useful in, later work; and (3) as far as practical, this documentation is of in a form likely to be compatible with any standards adopted in later work.*

Definition of Metadata

In the context of this work:

> ***Metadata** is stored in the NB DSS database and lists the source, properties and limitations of data or models.*

The scope of the metadata is defined by the list of metadata field headings in the universal template metadata file for a data set or model. The same structure of metadata file will be used to cover all types of data and models, ensuring transparency. The scope may be extended in the future, but will not be reduced.

Definition of "Verified Model"

In the context of this work:

> *A **verified model** is a model for which a complete relevant set of metadata exists within the quality assurance system.*

Definition of "Quality Assured Data", "Data Set" and "Data Point"

In the context of this work:

> ***quality assured data** are data contained in in the DSS for which complete relevant metadata exists within the quality assurance system.*

The label "quality assured" attached to data indicates that the data have been handled appropriately and documented to the required standard. The label does not mean that the data are of high quality or of the highest quality available or possible. Improving the quality of the data is an ongoing process that will continue throughout the working life of the DSS system, and the quality assurance system and metadata will be of assistance in this process.

For the purposes of this work:

> *A **data set** is any feature (spatial) or temporal (time series) data that is stored in the DSS database;*
>
> *A **data point** is any individual record in a data set. Quality assurance codes can be associated with data points.*

2.2 DATA SET METADATA TEMPLATES

The WP 2/1 Stage 1 consultancy proposed a universal metadata template that can be used to document the source, properties and limitations of any data set or model. The template has been extended as part of the current WP 2/1 Stage 2 consultancy to provide additional information regarding custodianship and status of spatial data sets.

Note regarding metadata for spatial datasets:

The NBI provides spatial datasets with metadata in the form of ArcGIS xml files. The data is generally well documented, and adoption of the ESRI metadata standard for use in the NB DSS would ensure that most, if not all of this information, can be transferred to the NB DSS. The decision is however not as straightforward as it may seem at first glance. The metadata standard supported by ArcGIS version 9.3 and earlier versions is the Federal Geographic Data Committee's (FGDC) Content Standard for Digital Geospatial Metadata (CSDGM), extended with non-standard fields to suit ArcGIS specific requirements. With the release of ArcGIS 10, ESRI has implemented ISO 19115, a metadata content standard for describing data, and ISO 19119, a metadata content standard for describing services as the default schemas for metadata. The implication of this change is that ArcGIS users will have to decide whether to maintain their legacy metadata in FGDC format, or to migrate to the new ISO standards. Depending on the size of users' data holdings, this may require considerable effort. To complicate matters further, the ISO standards are currently undergoing revisions. In view of these developments, it is suggested that, for the time being, only the most common and useful metadata elements that appear in both the ESRI ("extended" FDGC) and ISO standards are included in the NB DSS metadata schemas.

The universal metadata schema is shown in Figure 2-1 and element descriptions are provided in Table 2-1. The schema XSD file is provided in **Annexure A**.

Table 2-1 : Expanded Universal Metadata Template

Field Heading	Data to be recorded against the field heading
Data set title	Short descriptive name for the data set.
Data set ID	Unique identifier for the data set.
Metadata date stamp	Date this metadata file was last updated.
Data set topic category	Keywords describing the data set and its data. Any number of keywords can be recorded, but it is compulsory that one keyword is selected from the following list: observations, simulations or other sources.
Abstract describing the data	Brief narrative summary of the contents of the set.
Source data	*Data set ID* for all quality assured sets used in deriving/constructing this data set. Information on the source data used, including references to external documentation (where available).
Lineage	Information on history of derivation/construction of data set, including: (1) references to external documentation on processing and sources (where available); (2) *Data set IDs* for quality assured data sets that are superseded by this set.
Limitations of data set	Known and suspected deficiencies relating to the data set.
Metadata point of contact	Contact details (email address) for the person currently responsible for this metadata file.
Custodian	Name of organisation that is responsible for updating the data set
Custodian point of contact	Contact details (URL or email address) of the organisation responsible for maintenance and/or updating of the data set
Data set status	Current status of the data set. One keyword is selected from the following list: under development, regular update or final
Date of Ground Conditions	Date of ground conditions represented by the data (eg. year of satellite data acquisition for a land cover data set)
Reference Scale	Applicable to digitised vector data. Scale of source data.
Data set resolution	Applicable to raster data sets. Grid cell size in data set units
Geographic Coordinate System and Projection	Standard name for geographic coordinate system (Datum) and projection (if used)
Attribute list	List of attributes with description of meaning and units (if applicable). For raster data sets, the raster pixel VALUE must be described.

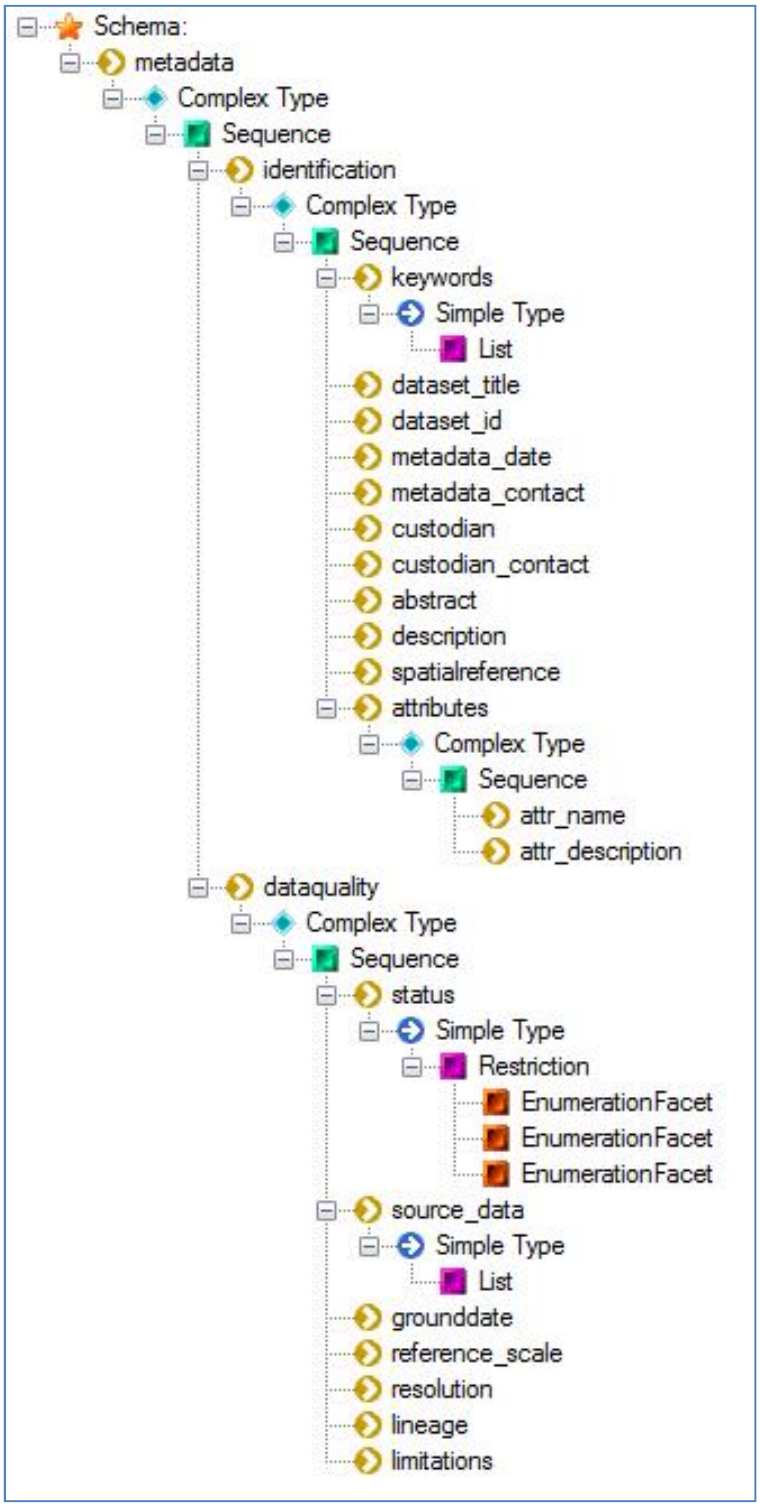


Figure 2-1: Universal Metadata Schema

2.3 DATA POINT METADATA

The observed (measured) time series in the NB DSS were obtained from many different sources, including country hydromet agencies, global data sets, feasibility studies and masterplans. Some of these sources have quality coded individual data points according to organisation specific code sets. It is an enormous task to harmonise the coding system across data types (rainfall, stream flow, water quality) and source. For compilation of the Nile Basin Encyclopedia and Ethiopian Masterplan data, the NBI has adopted an approach where the original code and description received with the data is retained in a growing quality code list, presumably with a new unique identifier to avoid duplicates. This approach will be followed with new data sets (examples include the FAO and GHCN rainfall data sets) obtained during the course of the consultancy.

For modelling purposes, it is important to know whether individual values in processed model input time series are missing, infilled, extended, or of questionable quality ("outliers"). For processed rainfall data, the following codes are proposed:

Table 2-2 : Metadata Quality Codes for Processed Rainfall Data

Description	Code for Metadata entry in NB DSS
Patched (infilled) values	10100
Extended values (i.e. outside the date range of observations	10300

Note 2-1: *For the purposes of WP 2/1 Stage 2, point metadata were included in the patched rainfall ASCII files that were used as import sources of to the DSS. In its current version, the NB DSS provides for manual flagging of point values with colour codes. It would require scripting to read the quality codes in the ASCII input files, translate these to the colour coding scheme in the DSS, and applying these to the point values. The process would however be non-transparent and will require careful maintenance of scripts and code translation tables to ensure that quality codes from new data sources can be catered for.*

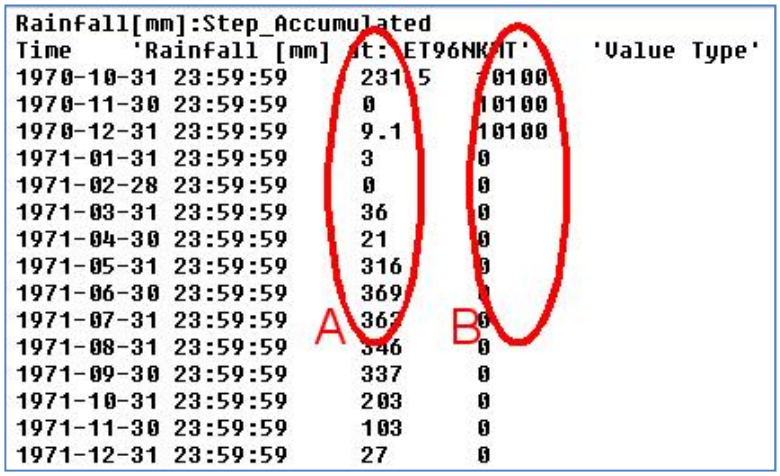


Figure 2-1: Example of ASCII import file with point values (A) and point codes (B)

3. DATA QUALITY CONTROL

3.1 INTRODUCTION

The need for Quality Control (QC) procedures is described in WP 2/1 Stage 1 Technical Note TN_0002 thus:

> *The screening of hydrological data is performed to ensure that the data are subjected to Quality Control (QC) procedures; QC is a pre-requisite for any hydrological analysis or modeling activity. Data screening can be used to identify several types of error, including accidental errors, for example a misreading by an observer; and systematic errors, e.g. caused by an inappropriately located rain gauge. An additional important aspect of QC checking is learning about the data sets. Failure of a test does not automatically imply that the datainvolved are erroneous, but rather highlights the need to perform further investigations. In particular, physical explanations for a data anomaly should be investigated before the data can be rejected.*

The WP 2/1 Stage 1 QC procedures focussed on the screening and infilling of incomplete flow records and extension of short flow records to cover the simulation period adopted for the Baseline and Pilot Case Models (1951-90; O'Connell et al., 2011). This has been expanded to include procedures for screening, infilling and extension of rainfall records, as well as procedures for quality control of vector spatial data sets, with specific reference to drainage networks. Raster data quality control procedures mainly relate to the creation of rasters from raw imagery, geo-referencing of imagery and/or creation of rasters from vector data, all of which require relatively specialised GIS tools which are not incorporated in the NB DSS. As this guideline focuses on procedures that are of direct relevance to the application of the NB DSS, not much attention is given to this aspect of spatial data quality control.

3.2 HYDROLOGICAL DATA SCREENING (FROM WP2/1 TN0002)

Four informal data screening methods (involving the use of checks or tests) are detailed below that rely primarily on the visual detection of anomalies/inconsistencies in hydrological data series (comprehensive details may be found in Basson et al., 1994; Hoaglin et al., 1986, and Gordon et al., 2004).

3.2.1 Visual inspection

Visual inspection of time series plots of hydrological data series is a simple check that should be performed as the initial step in data screening. The check is useful in identifying gross errors, for example, typing errors when translating data contained in manuscripts to digital format, identifying where the data have been assigned incorrect units and where there are long periods of missing records.

3.2.2 Unit runoff checks

The unit runoff check involves dividing the (monthly) runoff by the catchment area in order to determine the runoff as a depth. This is compared for consistency with values obtained from nearby hydrologically similar catchments. In regions with a scarcity of gauges, it may be necessary to compare upstream and downstream gauges. This check is particularly useful in identifying abrupt changes in river flows resulting from river basin management activities.

A simple example of a unit runoff check test is provided in Figure 3-1 for the Abbay/Blue Nile, using an upstream station at El Deim and a downstream station at Khartoum and Soba. Although the peak unit flows show similar inter annual variations, the unit runoff is significantly greater at the upstream site, El Deim. In this case, the differences in unit runoff depths can be explained by a sharp gradient in precipitation across the catchment, and the data are accepted at this stage of data screening. (It should be noted that, in this demonstration case, the data are not strictly from hydrologically similar catchments. Identifying hydrological similar catchments is often difficult at the larger catchment scale.)

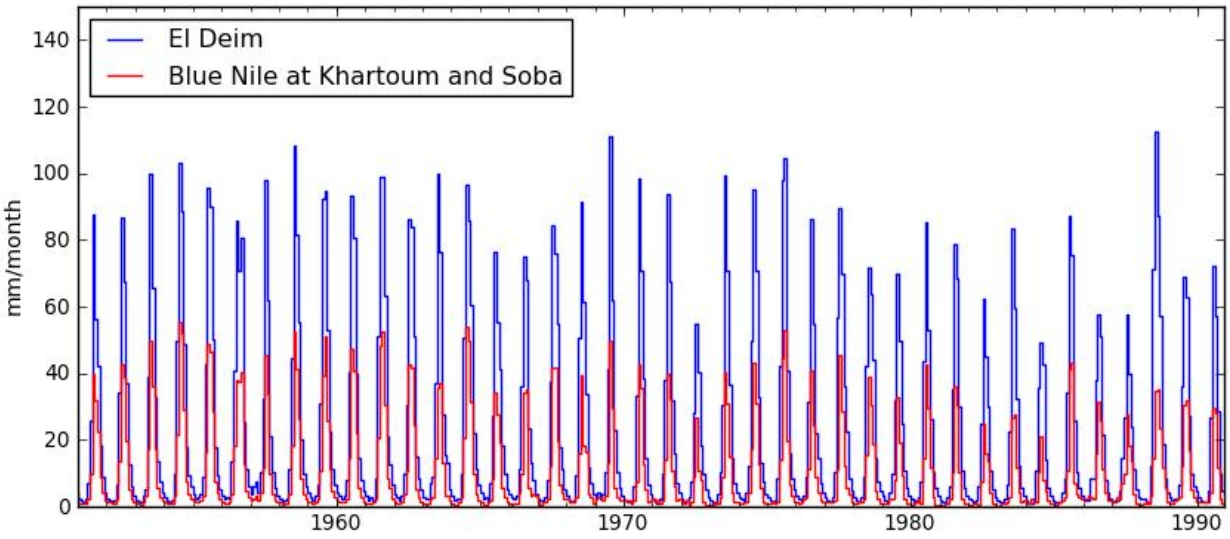


Figure 3-1 : Unit runoff check for the El Deim and Khartoum/Soba stations, 1951-90

3.2.3 Double mass checks

A double mass check involves plotting the cumulative data of one station against the cumulative data of another nearby station. If the data records are consistent, a straight line is obtained. Data from stream flow gauges can be compared with data for other flow gauges in the same general area, and, similarly, data for rainfall gauges can be compared (stream flow should generally not be compared to rainfall data as the relationship is usually nonlinear). Where an inconsistency is observed, such as a break in the slope of the line, an investigation into the cause should be performed. For data sets obtained from external agencies, a full investigation of the causes may not be possible. However, for stream flow data, basic investigation may be feasible, for example, establishing whether there has been a significant change in water resources development/management activities (e.g. dam construction).

In Figure 3-2, a double mass plot is provided for the Abbay/Blue Nile at El Deim and Khartoum. A change in the slope of the line is observed, which requires further investigation. In this case, the change in slope can be attributed to the construction of Roseires dam in 1961-66, and the data are accepted (The cumulative sum of the flows for the period 1951-66 is approximately 800 km^3 at Khartoum, which is where the break in slope occurs). Note that this change was not evident in the unit runoff check, showing the benefit of performing a suite of checks.

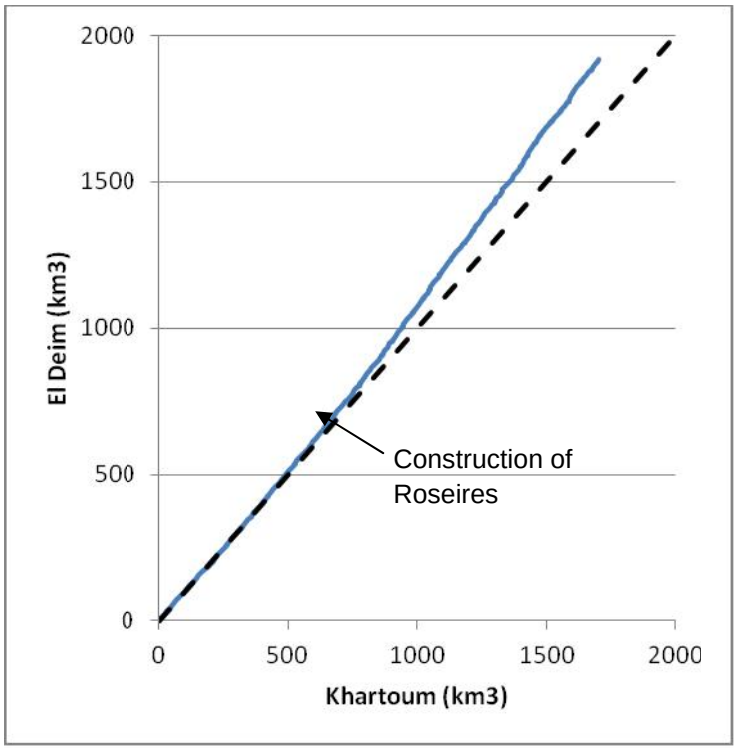


Figure 3-2 : Double mass check for the Abbay/Blue Nile at El Deim and Khartoum/Soba for the period 1951-90 (line of unit slope: black dashed line).

3.2.4 Mass balance checks

A mass balance check involves a comparison of upstream and downstream flow data series. Typically, flow would be expected to increase in a downstream direction, but the relationship may be more complex in arid regions. This test can be used to identify lack of consistency, for example caused by a change in a rating curve, and also be used to identify where significant abstractions or losses to evaporation occur.

A mass balance check is provided for El Deim and Khartoum in Figure 3-3. To aid clarity, the data are presented for the period 1956-75. Prior to the completion of the Roseires Dam in 1966, the flows at Khartoum are typically equal to or greater than those recorded upstream at El Deim. This pattern is generally reversed after the dam construction. The similarity in the peak magnitudes prior to 1966 also shows the extent of natural channel losses between El Deim and Khartoum.

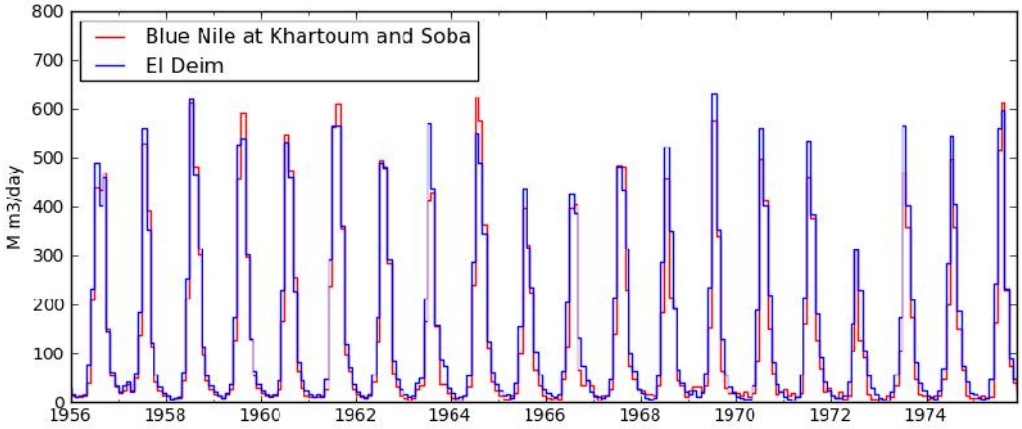


Figure 3-3 : Mass balance check for El Deim and Khartoum/Soba, 1956-75

3.3 HYDROLOGICAL DATA TESTS USING ROBUST STATISTICS (FROM WP2/1 TN0002)

This section provides details of statistical tests that can identify outliers and trends in hydrological data series. If outliers or trends are detected in the data, physical explanations should then be explored.

3.3.1 Outliers

An outlier is a measurement that is anomalously larger or smaller relative to the main body of the data. Outliers may result from an actual data processing error, for example, a transcription error or measurement problem, or may be a true extreme value.

A widely used test for the detection of outliers is the Quartile or Fourth Spread test (Basson et al, 1994). The test is firstly introduced and then an example is provided.

Firstly, this test requires calculating the inter quartile range d_F of the data series

$$d_F = F_U - F_L$$

where F_U is the 75th percentile of the data and F_L is the 25th percentile. This provides a good indicator of the spread of the centre region of the data. The upper and lower cutoff boundaries are calculated, respectively, as:

$$C_U = F_U + 1.5d_F \text{ and } C_L = F_L - 1.5d_F$$

Values lying outside the interval (C_L , C_U) are defined are outliers, and require further investigation. (This test is sensitive to highly seasonal flows, due to their asymmetric distribution.)

In Figure 3-4, the cutoff bounds are provided for the flow series for the White Nile at Mogren. Several outliers are identified during the period 1964-65, which require further investigation.

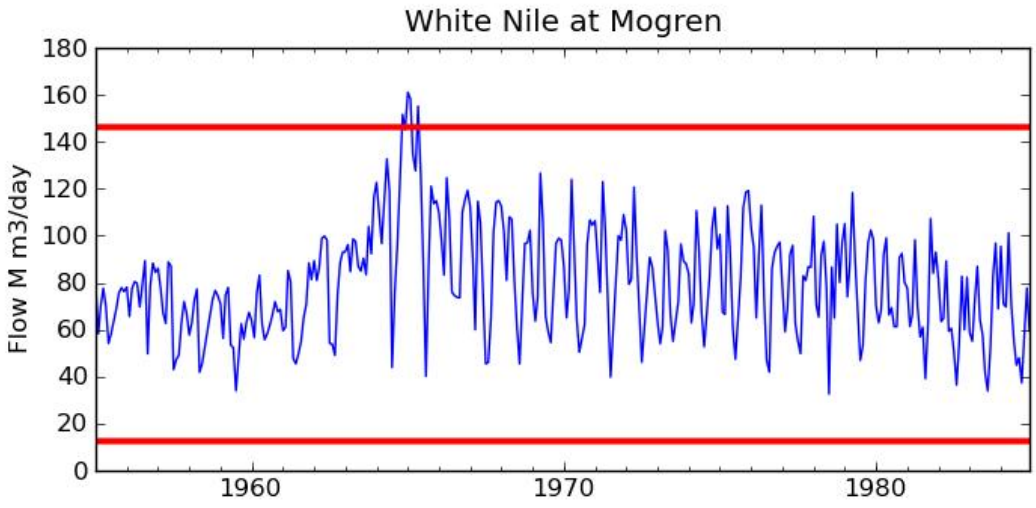


Figure 3-4 : Outliers for the White Nile at Mogren, 1955-84, cutoffs indicated by red bars

The calculation procedure for the bounds shown in Figure 3-4 is presented graphically in Figure 4-5. Firstly, the flows are ranked in ascending order and assigned a percentile. The interquartile range is calculated as the difference between the flow value for the 75th percentile (F_U; 97 M m^3/day) and the 25th

percentile (F_L; 63 M m^3/day) to give the inter quartile range, d_F (97-63 M m^3/day; 34 M m^3/day). The cutoffs are derived as:

$$C_L = 63 - 1.5 \times 34 = 12 \text{ M } m^3/\text{day}$$

$$C_U = 97 + 1.5 \times 34 = 148 \text{ M } m^3/\text{day}$$

During investigation of the outliers identified in Figure 3-4, the upstream and downstream flow records were examined. There was a physical explanation for the high values (relating to a change in the regime of Lake Victoria) and the outliers were accepted as plausible values.

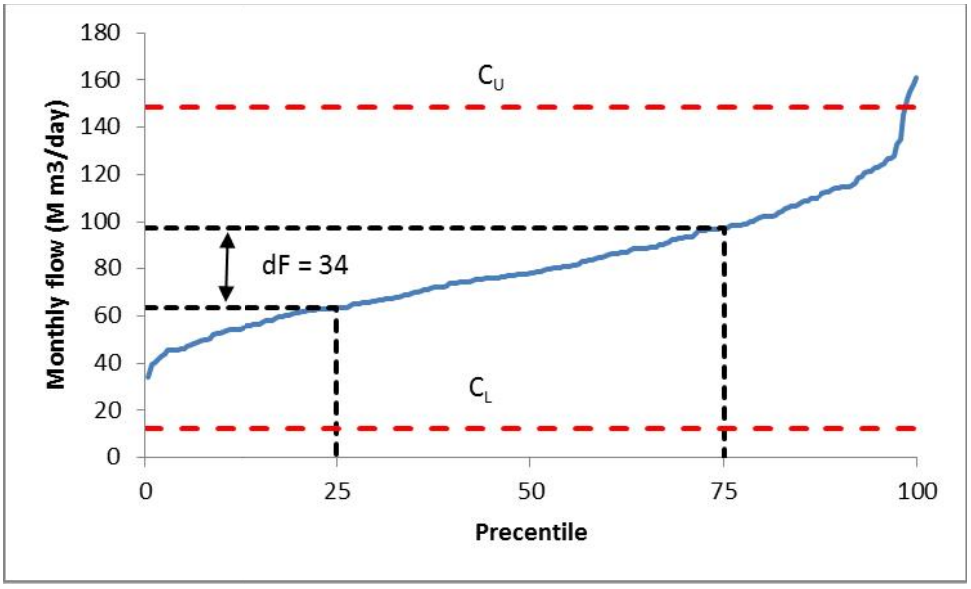


Figure 3-5 : Calculation of cutoff bounds

3.3.2 Trends

Trends in data series may be caused by inconsistencies, for example a change in the channel configuration at a gauging site, or from changes in the phenomena itself, such as the result of climatic variations.

A numerically simple test for a trend is Armsen's test. Due to the inherent seasonality in monthly data series, it is usually appropriate to perform this test on annual data series. The following description of this test is taken from Basson et al, (1994).

Suppose there are n years of annual flow data x_i=1,2,…,n. in a time series which is to be tested for an *increasing* trend. For each j=1,2,…n, count the number L_j of values of x_i to the left of x_j (i<j) which are greater than x_j. Define

1.5

$$L = \sum L_j$$

For the testing of a *decreasing* trend, L is replaced by L* where

$$L^* = n(n-1)/2 - L$$

Table 3-1 contains significance levels L_0 for small samples (n≤30). If L (or L*) < L_0, then the series can be said to have a trend significant at the chosen level for a one tailed test.

For larger samples:

$$z = (|\mu - L| - 1/2)/\sigma$$

which can be tested as a variate with a standard normal distribution, where

$$\mu = n(n-1)/4$$

$$\sigma^2 = n(n-1)(2n+5)/72$$

Table 3-1 : Significant points in Armsen's trend test (Basson et al., 1994, p90)

n	0.5 %	1.0 %	2.5 %	5 %
10	8	9	11	12
11	11	12	14	16
12	14	15	18	20
13	17	19	22	25
14	22	24	27	29
15	26	28	32	35
16	31	34	37	41
17	36	39	43	47
18	42	45	50	54
19	48	51	57	61
20	55	58	64	69
21	62	66	72	77
22	69	73	80	85
23	77	82	89	94
24	85	90	98	104
25	94	99	107	114
26	103	109	117	124
27	113	119	128	135
28	123	129	139	146
29	133	140	150	158
30	144	151	162	170

For the White Nile at Mogren (n =30), L is 220 and L* 215. (The annual flow series used in this test is shown in Figure 3-6.) From Table 1, for n=30, the 5% significance level L_0 is 170. Given that both L and L* > 170, there is no evidence at the 5% significance level to exclude the stationary of the data.

Visual inspection of annual series is also useful for detecting short term trends, for example the rise in the flows in the mid-1960s (Figure 3-6).

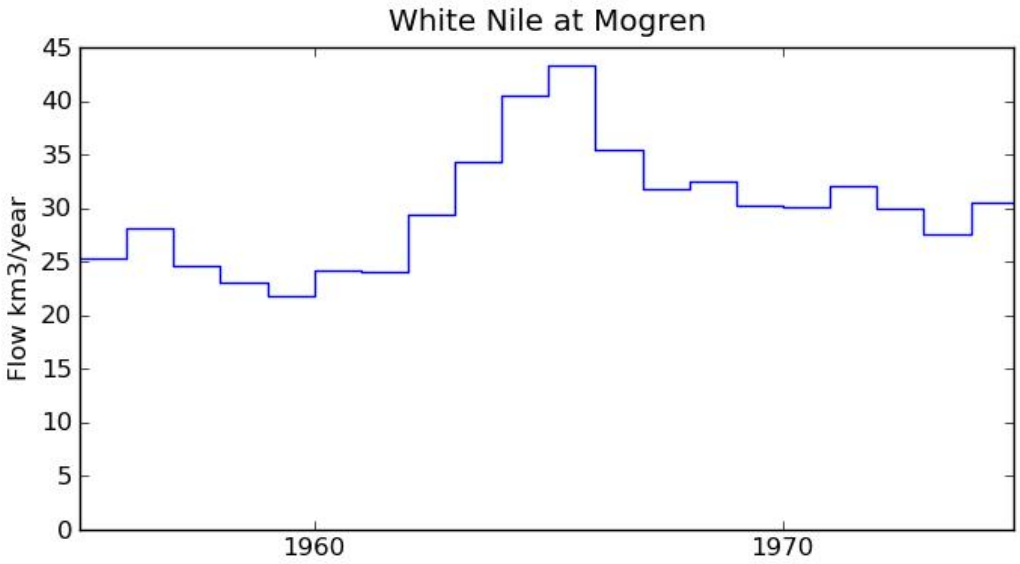


Figure 3-6 : Annual flows for the White Nile at Mogren

3.4 STREAM FLOW DATA INFILLING AND EXTENSION

3.4.1 Overview

Flow data sets typically have deficiencies in the form of missing data and short record lengths. Infilling of missing sections and extending data records is necessary prior to the use of hydrological time series in water resources modeling. A summary of the more commonly used methods for data infilling and extension is provided below. (The list of methods is not comprehensive.) References where further details can be found are provided.

In the terminology used below, a 'target' site is the site to be infilled, using information from a 'donor' site.

Scaling factors: In this very simple method, the flows at the donor site are multiplied by a scaling factor, for example, the ratio of donor and target catchment areas. The donor catchment should be in close proximity and have a similar hydrological regime to the target catchment (Kottegoda and Elgy, 1977).

Hydrological Modelling: This method, which is typically used when there is a short flow record and longer meteorological records for the catchment, involves developing and calibrating a catchment rainfall-runoff model against an observed flow series and then using it to generate a longer flow record using the meteorological inputs. Model complexity can range from simple black-box models to process-based models (Gyau-Boakye and Schultz, 1994).

Linear Regression methods: The most widely used methods for infilling are based on regression. In its simplest form, a linear regression equation is derived relating the target station flows (the dependent variable) to the donor station flows (the independent variable). This relationship is then used to infill the missing target flows, and to extend the target station flows if the donor station has a longer record. An extension of this approach is stepwise multiple linear regression, in which flow records from nearby catchments are included or excluded, based on the total variance they explain (Basson et al., 1994). These methods can also be used to infill/extend rainfall series. However, a difficulty arises with this method when the donor stations themselves have missing values; this is overcome using the Pseudo-EM algorithm (Pegram,1997).

Pseudo-EM algorithm: This algorithm accomplishes infilling using multiple linear regression in combination with the pseudo-EM (PEM) algorithm. Starting with a reasonable initial guess at the missing

data (e.g. the mean) the M-step of the algorithm performs the Maximum likelihood estimation of the regression parameters. The algorithm then switches to the E-step and estimates the missing data at each site in turn. These steps are repeated until reasonable convergence is achieved. This method has been widely used to infill rainfall records (Pegram, 1997), and is the basis of the PATCHS streamflow infilling software package utilized extensively in the Vaal River Study (Basson et al., 1994).

Loss of variance for regression methods: Reservoir capacity needed to maintain a target yield (outflow) is known to be a function of the coefficient of variation of the inflows $C_v = \left(\frac{\sigma}{\mu}\right)$, where μ is the mean and σ is the standard deviation (McMahon and Mein, 1978). Loss of variance in the inflows would therefore lead to an over-estimation of yield, and should be avoided, if possible. In view of this example, it is appropriate to highlight a shortcoming of regression methods, given their widespread use. The intent of infilling is to produce a time series that has statistical characteristics similar to those of the actual record at that station (Hirsh, 1982). The aim of linear regression is to produce a best estimate, in terms of the minimum mean squared error, of each missing flow value. However, the variance of the infilled values is biased downward because the regression estimates lie on the regression line and the actual data are scattered about the regression line and hence are more variable (Hirsh et al., 1993). The magnitude of the bias (loss of variability) depends on the explained variance of the regression R^2; if this is high (~0.90), then the loss of variance in the infilled values will be relatively small (~10%), as the explained variance decreases, then the loss of variance increases. The overall loss of variance in the target flow record will not only depend on R^2, but on the proportion of the record that is infilled/extended.

One approach to replacing the lost variance is to add a random error to the regression estimates, but the random term modifies the serial correlation structure of the record (Hirsh, 1982). Moreover, many different realizations of the error term are possible, and so any single realization is not unique. The sensitivity of water resources modeling results to the uncertainty resulting from the loss of variance could be explored by sampling multiple realizations of the error term for each infilled value.

Hirsh (1982) pioneered a class of record extension called “maintenance of variance extension” (MOVE). Rather than minimizing the mean squared error, the MOVE method aims to reproduce the mean and variance of the observed series. Vogel and Stedinger (1985) derived unbiased minimum variance estimates of the mean and variance of the infilled record, and then used these results to formulate a regression that delivers unique infilled values that have the required mean and variance (the Maintenance of Variance procedure MOVE.4). However, these MOVE results relate only to simple linear regression which limits their practical usefulness.

Generalized Linear Models (GLM). GLM was formulated by Nelder and Wedderburn (1972), and is a flexible generalization of ordinary linear regression that allows for response variables that have other than a normal distribution. The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a **link function** and by allowing the magnitude of the variance of each measurement to be a function of its predicted value.

3.4.2 Methods used to infill Nile records

Based on the above review and software availability, two infilling methods have been selected for testing here (i) the NB-DSS Gap Fill tool and (ii) the PATCHS method. Due to its ability to infill from multiple rainfall and streamflow records, PATCHS has been used to infill flow records in the WP2/2 Stage 1 Study.

NB DSS GAP Fill Tool

Within the NBI-DSS, there is a Gap Fill tool for infilling and extending flow records. This uses a simple linear regression approach. A regression equation is derived between the flow data for the concurrent period of record at a target station and those at a (single) donor station. This regression relationship is

then used to calculate the missing target flows. Several difficulties arise when using simple linear regression:

- If data are missing at the donor station, infilling cannot be performed.
- There is usually a strong temporal dependence structure (serial correlation) due to catchment storage, which is not accounted for.
- Flows often exhibit seasonality, which cannot be captured by a single regression relationship. (The tool does not provide for seasonal or monthly regressions.)

PATCHS

Stream flow records invariably exhibit cross correlation with records for other stations in a catchment, and also possibly with records for stations in neighboring catchments, due to regional climatic factors. They also exhibit serial correlation, due mainly to the effects of storage or travel times from upstream to downstream stations. The PATCHS program (developed by Professor G.G.S. Pegram) exploits available cross and serial correlations between and within stream flow and rainfall records at low order lags. Desirable properties of this method include:

- The ability to infill data even when there are simultaneous gaps in the target and donor stations.
- The ability to utilise information from several flow gauges.
- The ability to utilise information from rainfall records; which is useful when the donor catchments are not highly correlated or a suitable donor flow site is not available.
- The use of serial correlation to preserve the structure and seasonality of the data.

Model parameters are estimated using the EM algorithm, which involves recursively substituting the data that is missing and then re-estimating the parameters to maximize the model likelihood. (A detailed description of PATCHS is provided in a user manual: Pegram, 1993.) The selection of PATCHS for infilling was based on the performance of trial runs of infilling flow records and a limited comparison against the Gap Fill tool, examples of which are provided below.

3.4.3 Comparison of NB-DSS Gap Fill Tool and PATCHS

A simple test of the relative performance of the Gap Fill tool and the PATCHS infilling techniques is demonstrated. An observed period of record at a target station was flagged as missing and then infilled using a donor station. The chosen target station was Dongola and the donor station Tamaniat. These stations on the lower Nile have a concurrent record over the period 1962-97 and a correlation of 0.9. The period 1970-75, which includes a very wet and a very dry year, was flagged as missing.

The infilled and observed series are provided for the two techniques in Figure 3-7, where the upper panel shows that the performance of the Gap Fill tool is inferior to PATCHS. Firstly, residuals are negative on the rising limb of the hydrograph, indicating over-prediction, and then positive on the falling limb, indicating under-prediction. This error is in part due to a lag between the two stations. Tamaniat is located several hundred kilometres upstream of Dongola and hence there is a routing delay which is not accounted for by the lag zero regression relationship used in the Gap Fill tool. The performance of PATCHS, which can account for low order lags, is superior although there is some under-prediction of peak flows, Figure 3-7, lower panel.

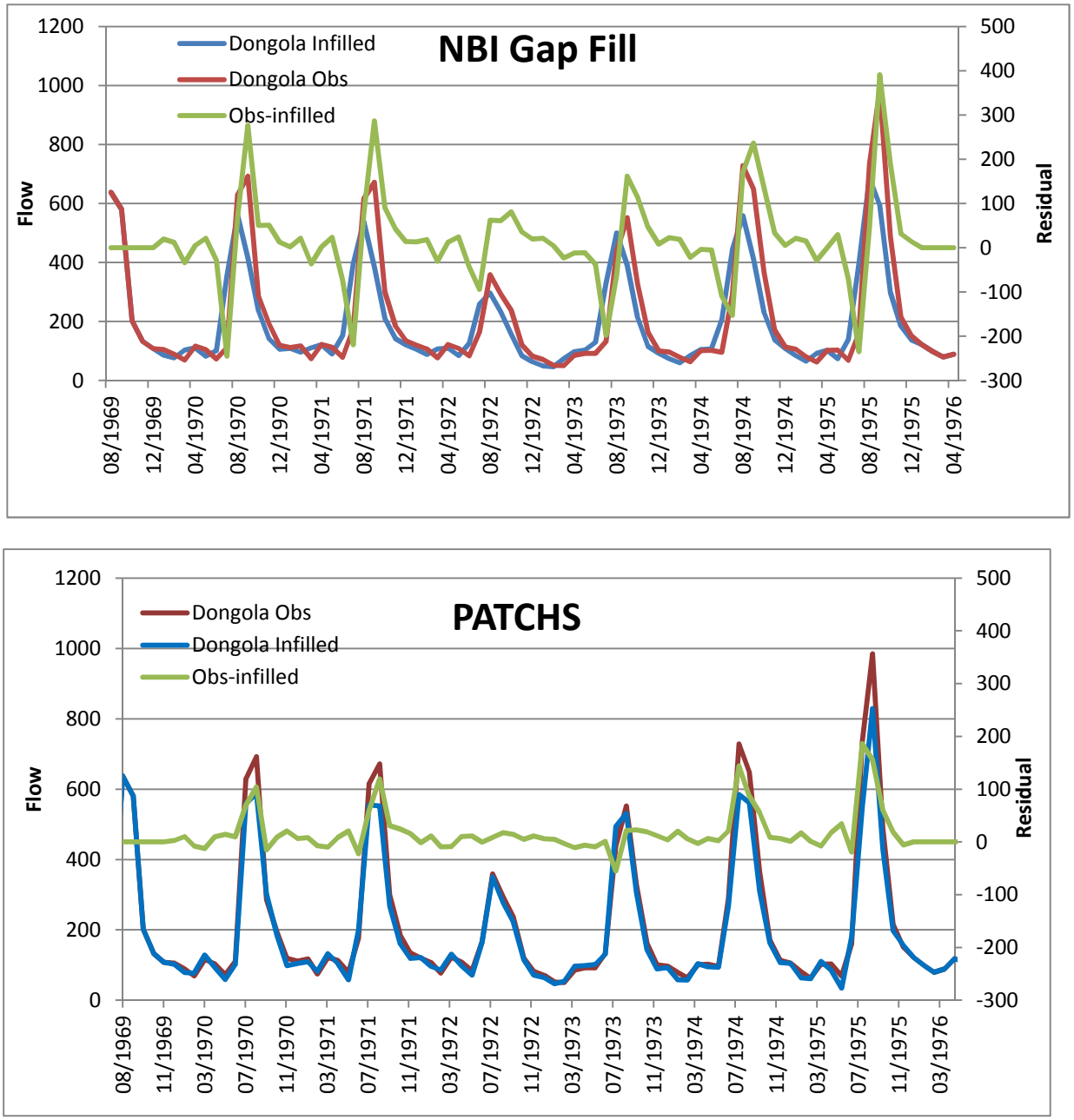


Figure 3-7 : Comparison of NBI-DSS Gap Fill tool (top panel) and PATCHS (bottom panel) for Dongola (units: M m3/day)

Utilization of Rainfall Data in PATCHS

This section provides a case in which PATCHS is used to extend a flow series using both a donor flow record and several rainfall records. Monthly flows for the Bahr el Jebel at Mongalla are required for the Baseline and Sudd Pilot Case models for the period 1951-90. However, no flow data are available at this location after 1983. The only upstream data available for extending the record over this period is the flow record for the Victoria Nile. Although this has a reasonable relationship with Mongalla (correlation coefficient 0.7), there are significant differences in the timing and seasonality of flows due to the presence of Lake Kyogo and Lake Albert and the seasonal inflows from the Torrents between locations (see Figure 3-8, compare black and red lines). Due to the presence of the Sudd, flows recorded at the

station immediately downstream, Malakal, cannot be reliably used in the infilling process (correlation coefficient: 0.36).

The Mongalla record was extended using both the upstream flows from the Victoria Nile and rainfall from several rain gauges, to represent the inputs from the ungauged Torrents. To test the quality of the record extension procedure, a 5 year period of observed data at Mongalla was flagged as missing (1965-69) and infilled. The magnitude and seasonality of the Mongalla record are well reproduced by PATCHS (Figure 3-8: compare black and blue lines).

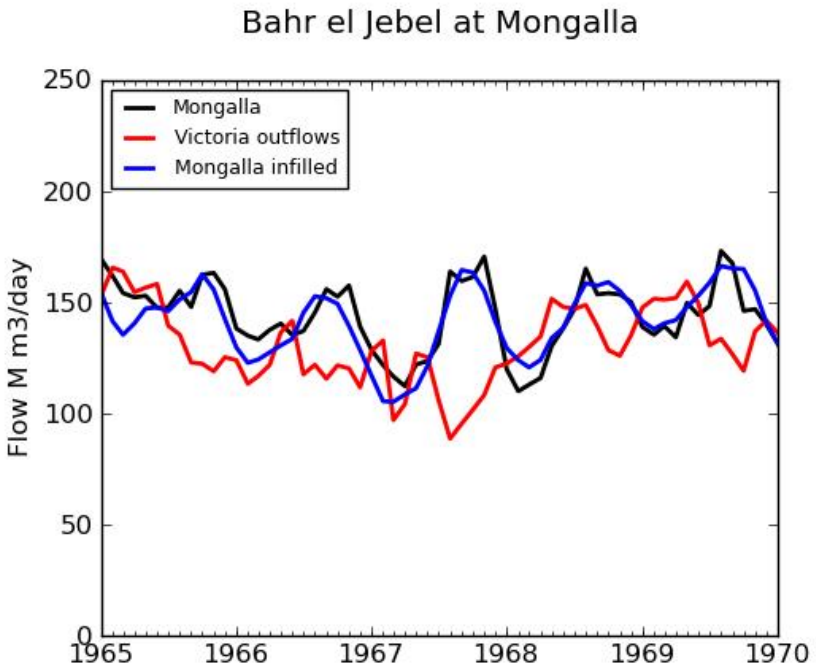


Figure 3-8 : Test of infilling of flows for Bahr el Jebel at Mongalla

3.4.4 Assessment of uncertainty due to infilling

In order to make an assessment of the uncertainty resulting from infilling, a Model Conditional Processor (MCP) has been employed to demonstrate how the uncertainty can be quantified. Details of the MCP approach, which is usually employed to quantify the uncertainty in model predictions, can be found in Coccia and Todini (2010). In the MCP approach, a model is trained to reproduce an historical sequence of flows. During infilling, predictions and information about the uncertainty of the predictions are made, which are based on the behavior of the model during the historical period. The uncertainty of predictions takes the form of a probability distribution, the width of which relates to the expected range in which the true value may lie. Since the MCP approach generates its own predictions, in this case the infilled values, the opportunity has been taken to make a comparison with the infilled estimates provided by PATCHS.

The chosen test site for the uncertainty assessment and comparison is the Kagera at Kyaka Ferry, Figure 3-9. For the MCP model, the red line is the best estimate of the missing values (expected value) and the grey lines provide the 5% and 95% predictive uncertainty bounds (i.e. it would be expected that 10% of values may be outside this range). In general, the PATCHS inflilled flows are within the uncertainty bands of MPU except the mentioned discrepancy.

In principle, PATCHS could generate its own uncertainty bounds, as residuals could be calculated for the linear regression model/EM estimation algorithm on which it is based.

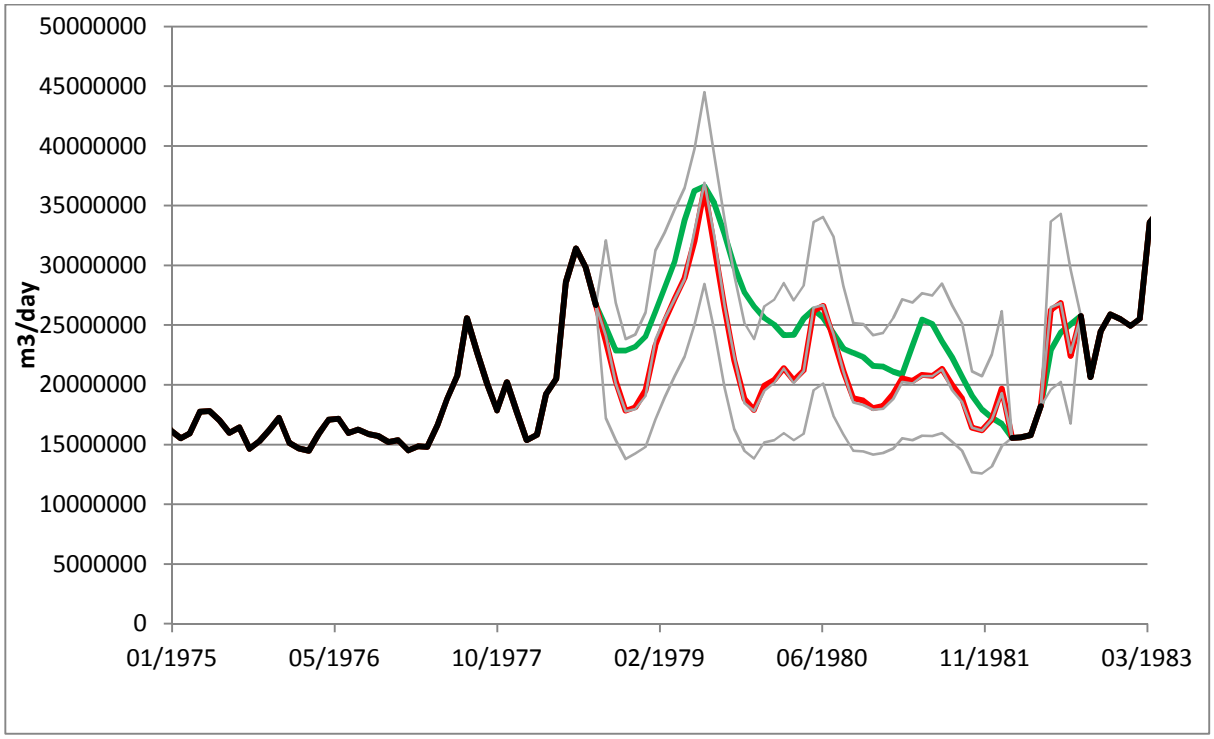


Figure 3-9 : Comparison of infilled series using the MPU approach and PATCHS

Black - observed flows; green – PATCHS; red MPU. The MPU predictive bounds are provided in grey. The upper and lower grey bounds represent the (95% and 5%) confidence interval

3.4.5 Recommendations on best practice/further work

Based on the above assessment of flow data QC and infilling methods, the following recommendations can be made:

(a) the use of a set of data QC checks and tests to identify data inconsistencies and anomalies is recommended, rather than the use of individual checks/tests in isolation;

(b) the set of tests demonstrated in this technical note through applications to Nile data sets represents a robust set, most of which have previously been proven in the Vaal Rover study (Basson et al, 1994). They can be utilized for future QC assessments of Nile data, provided the data analyzed comply with the underlying assumptions;

(c) the comparison of the NBI DSS Gap Fill Tool with a more sophisticated data infilling method, PATCHS, suggest that there is potential to improve on the results from the Gap Fill Tool. As the PATCHS program can be obtained free of charge from the South Africa Department of Water Affairs (the PATCHS program used in the work reported here was provided courtesy of its developer, Prof. G. G. S Pegram), it is recommended that PATCHS should be acquired and used as a DSS utility program;

(d) If infilling exceeds a threshold of 15% of the total data set, then an uncertainty assessment should be performed. Although uncertainty assessment lies outside the ToR for this study, a demonstration of the Model Conditional Processor (MCP) method has been provided here courtesy of its developer, Prof E Todini, and his colleague, Dr G Coccia. It is recommended that sensitivity analyses of model outputs to this uncertainty should be carried out in the future. The PATCHS program has the potential to provide the statistics of residuals from which an uncertainty assessment could be made, making this a self-contained infilling and uncertainty assessment package.

3.4.6 Guidelines for Stream Flow Infilling Using the PATCHS Software

A detailed description of PATCHS is provided in a user manual for the software (Pegram, 1993). Extracts from this manual are provided in Annexure B. The purpose of this section is to provide a brief guide to selection of PATCHS parameters that will assist in achieving credible infilled stream flow sequences.

Number of stream flow (s) and rainfall stations (r)

A maximum of 3 stream flow and 8 rainfall records can be used to cross-infill the stream flow records.

As many stream flow and rainfall sequences that are thought to be associated, should be included. Rainfall data must have been screened and patched using techniques such as CLASSR and PATCHR described in Section 3.2.3.

Number of stream flow (p) and rainfall lags (q)

The infilling process is achieved by conducting a series of exploratory 'trials' and inspection of the outcomes of these. Parameter **p** can take values of 1 or 2, and **q** 0 or 1. It is recommended to build up from small to large lags, i.e. for the first trial, p and q should be set to 1 and 0, respectively.

Smoothed or recorded data (ir)

Parameter **ir** indicates a choice between the calculation of smoothed data, deleted residuals and MCV (mean cross-validation criterion) (**ir** = 1), and stream flow data as recorded (**ir** = 0). The parameter should initially be set to 1 to assist with selection of the best model from the set of trials. Deleted residuals assist with the identification of outliers. Once a model has been selected, **ir** can be set to 0 to produce final results.

Maximum number of iterations (maxit)

With **ir** equal to 1, reasonably coherent data (i.e. stations are well-associated) should allow the algorithm to converge within 20 to 30 iterations, therefore **maxit** should initially be set to about 40.

Lognormal transform (ilog)

A lognormal transformation is usually not helpful in stream flow patching, but this option (**ilog**=1) is provided for flexibility in approach.

Shifting and scaling (itr)

Three options are provided:

- **itr** = 1: no shifting or scaling
- **itr** = 2: no shifting, but scaling month-by-month by standard deviations
- **itr** = 3: a month-by-month standardisation using means and standard deviations.

A setting of 1 (no shifting or scaling) is recommended as an initial choice. This is based on the assumptions that process parameters are constant and the process is linear.

3.5 RAINFALL DATA INFILLING AND EXTENSION

3.5.1 Preliminary Screening

Preliminary screening of rainfall station records is a subjective exercise and decisions whether to retain or discard a particular record are based on:

- A visual inspection of a cumulative mass plot (i.e. a cumulative record of annual totals) to gauge whether a change in slope in the cumulative mass plot is due to external influences on the rainfall record (such as relocation of the rain gauge or screening by a tree growing nearby), or due to missing data (which is acceptable at this stage). Figures 3-10 and 3-11 show examples of rainfall records with acceptable and unacceptable non-stationarities, respectively.
- Period of record. Depending on the target (modelling) period of record, amount of missing data, and availability of statistically similar records that can be used to infill a given rainfall record, a decision can be made to retain or discard the record. This decision is usually only taken during the classification process (see following sections).

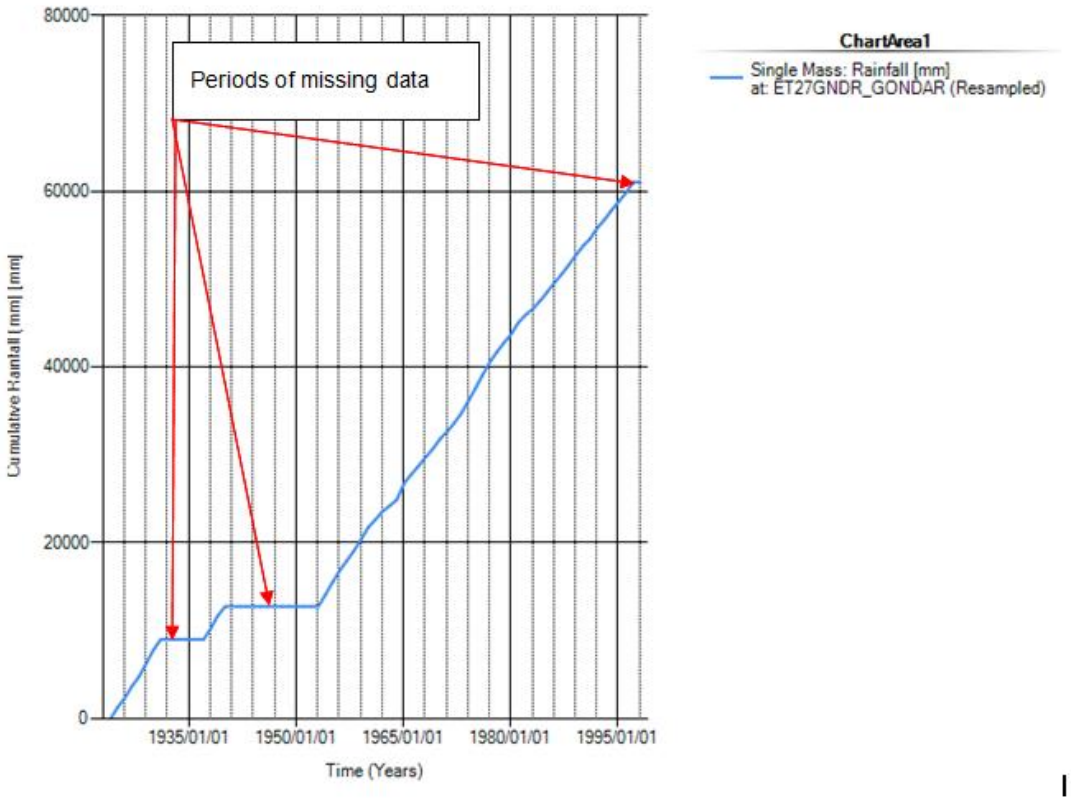


Figure 3-10 : Example of a stationary record

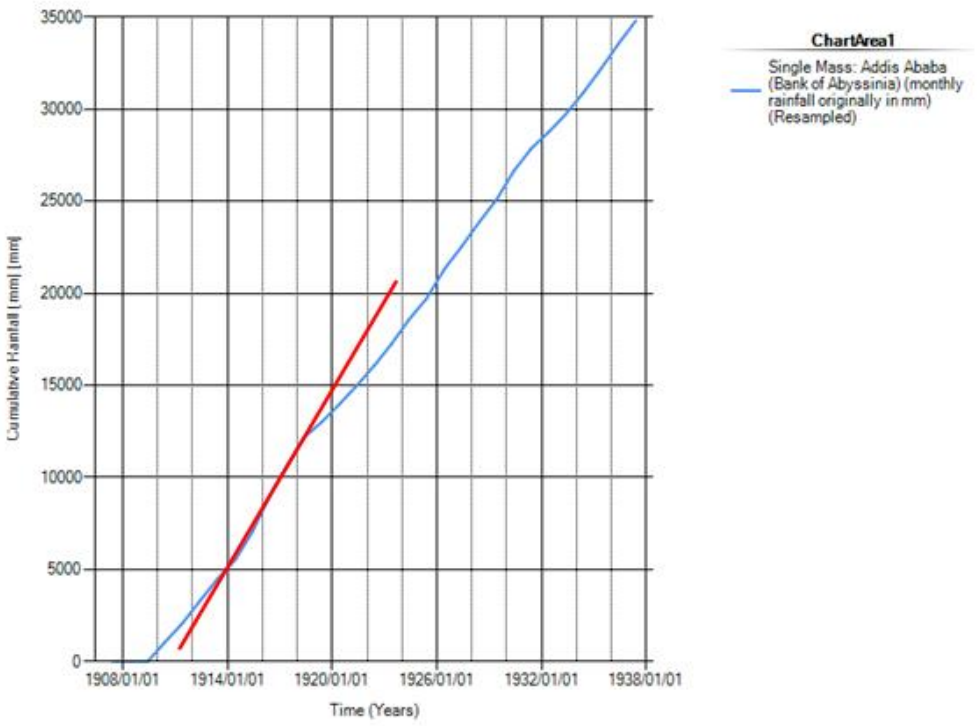


Figure 3-11 : Example of a non-stationary record

> ***Guideline 3-1:*** *Horizontal mass plot segments usually indicate periods of missing data, while breakpoints between sloping segments indicate systematic measurement errors that have been introduced. This should be verified by visual inspection of data values.*

3.5.2 Methods Available to Infill Nile Records

It is frequently the case in the field of engineering hydrology that rainfall records of interest have missing data and may also contain “outliers”. The NB DSS gap-filling tool (described in Section 3.4.2) can be used to infill stream flow or rainfall records, and uses multiple linear regression for gap-filling among a group of stations. The CLASSR and PATCHR (Pegram 1997) software packages provide an alternative which can conjunctively be used for infilling and extension of rainfall records, and are "sister" applications of the PATCHS application (Section 3.4.2) used by WP 2/1 Stage 1 for infilling of stream flow records. Section 3.5.3 provides guidelines for application of CLASSR/PATCHR, followed by a comparison of CLASSR/PATCHR and the DSS gap-filling tool.

3.5.3 Classification, Infilling and Extension with CLASSR and PATCHR

CLASSR classifies rainfall records that are statistically similar according to covariance bi-plots, assists with the grouping of records for mutual infilling. PATCHR uses multiple linear regressions in combination with an EM (Expectation-Maximization) algorithm to fill gaps in a group of stations in one operation using an iterative calculation.

Broadly speaking, the steps that are required to patch rainfall records are to:

- assemble related gauges in a group for maximum information transfer
- group similar months into seasons

- detect outliers
- assess patching success

The following sections provide an outline of the process, and recommendations are made that will guide decisions regarding (a) classification or grouping of records for mutual infilling, and (b) acceptability of patching results, i.e. measures of patching success.

Rainfall Station Grouping

CLASSR is used to group statistically similar records for mutual gap filling. It attempts to answer the following questions (Pegram, 1997):

- Which gauges are hydrologically similar in the sense that there is a strong correlation between them?
- By the same token, which gauges do not belong to the subset?
- If the data sets are short (in the sense that there will not be enough concurrent data to make a good month-by-month patch) which months can be grouped into seasons and the information pooled?
- Are there any gross outliers which are clearly in error and will have a polluting affect on the patching? Should they be flagged?
- Are the records very patchy, i.e. do the data need a rough preliminary patch to help with the classification?

A maximum of 8 rain-gauge records can be classified together. Rainfall stations that are geographically similar (they are located in the same area and have similar rainfall characteristics, such as mean annual precipitation - MAP and station elevation) are initially chosen and placed together in groups. Preference should be given to stations that have a reasonably long period of record common to all the gauges in the group. Within a group, the number of intact (no missing data) years should be at least 2.5 times the number of stations in the group, i.e. if there are four stations in a group, the group members should all have at least 10 corresponding years of complete data.

The software convention dictates that missing or doubtful monthly values should be flagged by putting a '+' sign immediately after the value.

CLASSR then uses a measure of "distance" to identify gauges of a similar "nature". Output from CLASSR includes two bi-plot diagrams. Bi-plot axes show G-vectors (a measure of variation in values of candidate stations) and H-vectors (a measure of variation in values for months). For a full description of the biplot and its properties, see Gordon, 1981.) The first bi-plot (Figure 3-12) is used to group stations for patching. If a station lies far away from all other points in the bi-plot, it may well end up in a group by itself. For the first pass of CLASSR, it is advisable to include more gauges in the analysis than needed. It may be necessary to conduct a second pass to confirm the grouping and reduce the number in the group to a reasonable size - between three and five is usual. An indication as to the number of gauges that can be treated in a group for submission to PATCHR can be obtained by inspection of the number of intact years reported by CLASSR.

Guideline 3-2 : *The minimum number of intact years that should be accepted is 2.5 times the number of stations in a group. If there are 4 stations in a group, then there should be a minimum of 10 intact years amongst the records in the group. A factor of 4 times the number of stations is considered to be "good"*

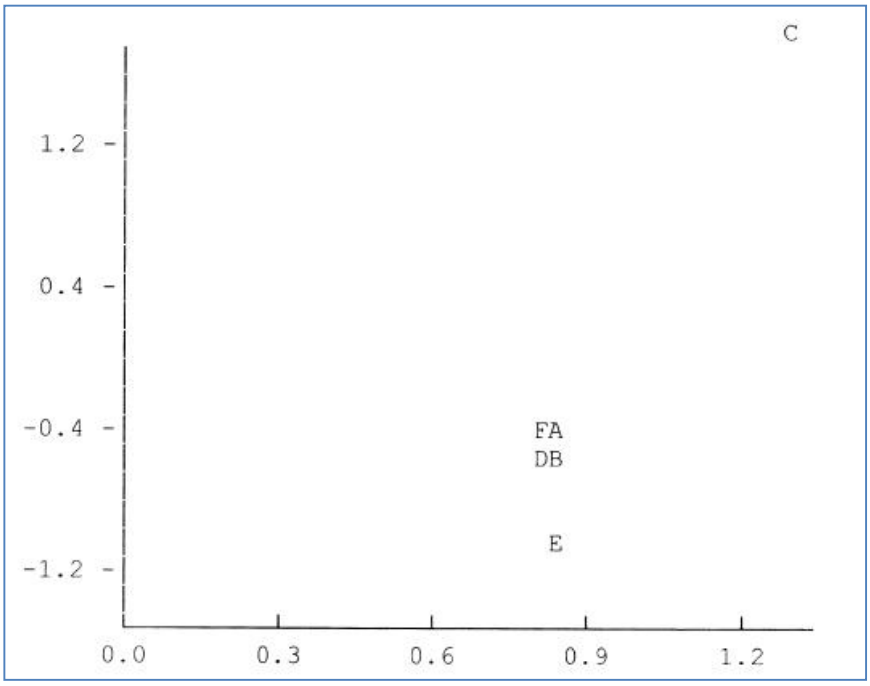


Figure 3-12 : Stations versus months biplot for stations A to F (Pegram 1997)

Seasonal Grouping of Months

Once the hydrologically similar stations are grouped together and the number of intact years is equal or greater than the recommended 2.5 times the number of stations, the second bi-plot (Figure 3-13) is used to group months with similar rainfall characteristics together. If certain months are grouped in the bi-plot they are included in a season. Seasonal grouping of months are used in the PATCHR process.

Guideline 3-3: *Experience indicates that 2 to 4 seasons deliver best results. As far as possible, groupings of months should make "hydrological sense". If it is known that a region experiences two rainy seasons with different characteristics such as season length and monthly amounts, then those should be grouped into separate classification seasons. A sensible grouping often serves to improve overall patching success.*

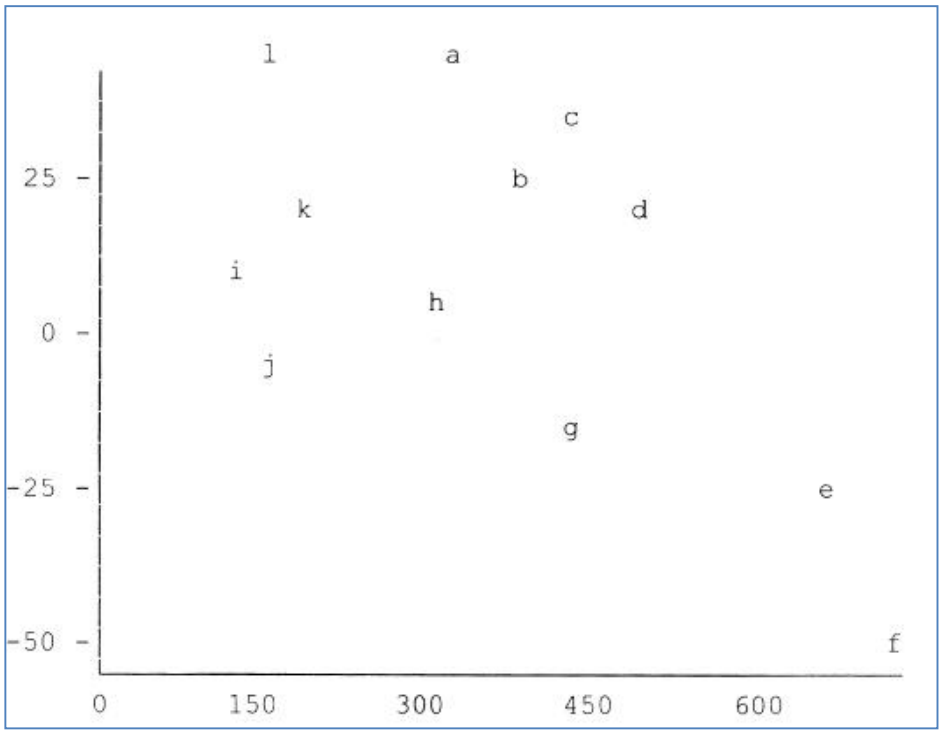


Figure 3-13 : Stations versus months biplot for the months (Pegram 1997)

Preliminary Outlier Screening and Rough Patching

CLASSR provide preliminary identification of outliers, and can be used to flag obvious outliers for later patching with PATCHR.

> ***Guideline 3-4:*** *Look for obvious, rather than ambiguous outliers in CLASSR. These could be order of magnitude differences between stations, such as those introduced by capturing erroneous extra zeroes*

CLASSR can also be used to perform "rough patching" of missing values. The purpose is solely to strengthen regression equations for station grouping purposes. Rough patched values are discarded before the more rigorous patching process employed by PATCHR starts.

Outlier Identification, Infilling and Extension

Once the grouping of stations has been finalised, it is submitted to PATCHR to identify outliers. The PATCHR output lists all potential outliers and the user must decide which outliers are indeed valid (usually only a few from those listed). The doubtful values are then flagged (a + sign), and PATCHR is rerun.

Guideline 3-5: *It is unusual to flag more than 3 to 5 "genuine" outliers out of the long list of potential outliers identified by PATCHR. There are several considerations that should be taken into account when flagging outliers:*

- *Is the station located in an area where frontal systems predominate? This indicates that stations that are located close to each other are likely to receive rainfall of similar order of magnitude in a given month. Conversely, it is quite possible that a station in a region that receives most of its rain in the form of thunder showers may "miss" a few localised events and show zero or little monthly rainfall while adjacent stations record large falls.*
- *Will the rainfall of this particular station have a large influence on modelled runoff used for scheme yield estimates? If so, a conservative approach should be adopted, with special attention given to unusually high values.*

The program produces patched rainfall files as well as an output file which presents the user with a measure of patching success.

Assessing patching success

CLASSR and PATCHR provide the user with measures of the goodness-of-fit of the regression equations and diagnostics to measure overall success of the patching process. Information such the R^2 value (in the CLASSR output) measures the accuracy of the multiple linear regression equations, while the iteration number and convergence criteria (listed in the Beta matrix of the PATCHR output) measure the overall success of the patch.

Guideline 3-6:

- *Beta Matrix values of larger than 1.0 indicate that the relevant station-pair(s) are poorly correlated, and that one should consider discarding one of the stations in the pair.*
- An acceptable number of iterations to converge is 25, and 15 or less is considered to be good.

3.5.4 Comparison of NB-DSS Gap Fill Tool and CLASSR/PATCHR

While the general approach and outputs of the DSS gap-filling tool and the CLASSR / PATCHR suite are similar (both tool sets provide for mutual infilling of missing values among a group of stations based on the strength of cross-correlations between pairs of stations), the functionality and degree of user control over the process are quite different. A comparison of functionality is provided in Table 3-2:

Table 3-2: Functional Comparison of DSS Gap-filler and CLASSR/PATCHR

Function	DSS Gap-filler	CLASSR/PATCHR	Comment
Identification and classification of stations for mutual infilling	This must be done manually on the basis of user judgment	CLASSR provides a structured approach to identify related stations on the basis of coviariance bi-plots	A structured, automated process is essential when working with a large number of stations
Identification and exclusion of statistical outliers	User intervention by setting a minimum cross-correlation for cross-referencing. (An outlier flagging tool is available in the DSS, but this is more suited to quality control of water level and/or discharge measurements)	Statistical assessment and identification of outlier values based on distributions of monthly values and cross-correlations between stations	CLASSR / PATCHR identifies individual values as potential outliers, which can then be manually included or excluded from the process. Use of the cross-correlation threshold in the DSS tool will reduce the risk of including outliers, but do not allow for statistical assessment of individual values.
Cross-referencing (mutual infilling)	Based on assigning inter-station priorities for cross-referencing. Priorities remain the same for the entire period of record	Based on seasonal or monthly cross-correlations between stations	CLASSR/PATCHR provides for maximum information transfer between stations by utilising sets of seasonal cross-correlations rather than a single matrix for all months of the year
Time step	The gapfill tool is generic for all time series types and it's time step is set by the user.	CLASSR and PATCHR are specific to monthly rainfall (PATCHS is used for monthly stream flows)	Infilling of rainfall data on a sub-monthly / daily time step must be approached with extreme caution. It is very rare to find statistically significant cross-correlations between stations on a daily basis across a period of record that is suitable for long term water resource assessments
User experience	The tool is easy to use, with a minimal set of user decisions / inputs. Uses DSS native time series format	Requires manual formatting of input files, which requires meticulous checks to ensure correct formatting. Software runs on outdated (MS-DOS) operating system	CLASSR/PATCHR provides a powerful and scientifically rigorous approach to rainfall infilling and extension, which is a vital building block in most water resource assessments. The outdated software environment that it operates on is a considerable drawback, and the software will have to be ported to a DotNet platform if it is to be used in the long term as part of the DSS tool set.

The NBI has discussed the possible integration of CLASSR/PATCHR (and PATCHS) with the author, Dr Geoff Pegram, and it is envisaged that the three tools will in future be integrated with the DSS.

It is difficult to make a direct comparison of the performance of the two alternatives, due to the very different input parameters that are encapsulated in the two tools. These cannot be directly replicated in both tools. (One example is the inclusion and exclusion of outliers in PATCHR, which cannot be replicated in the DSS gap-filler, and another is the pre-screening of stations for mutual infilling with CLASSR.) Subject to this understanding, a comparison was done by infilling four stations in the Blue Nile catchment that were pre-identified as a patching group with CLASSR. The stations are Gondar, Bahar Dar, Gorgora and Maksegnit. The DSS tool was run with a relatively low cross-correlation threshold of 0.2, to ensure successful infilling of all four stations over the full common period of record (1965-1993 calendar years). The bounds for infilling wëre set as "Curb to value" (where infilled values do not exceed values already present in the observed sequence), rather than "Leave gap" (where no infilling is done if a regression value exceeds observed values). The entire process of outlier identification and seasonal grouping of months was followed to perform a parallel infilling of the records with PATCHR. Differences in the outputs of the two processes are illustrated in Figure 3-14.

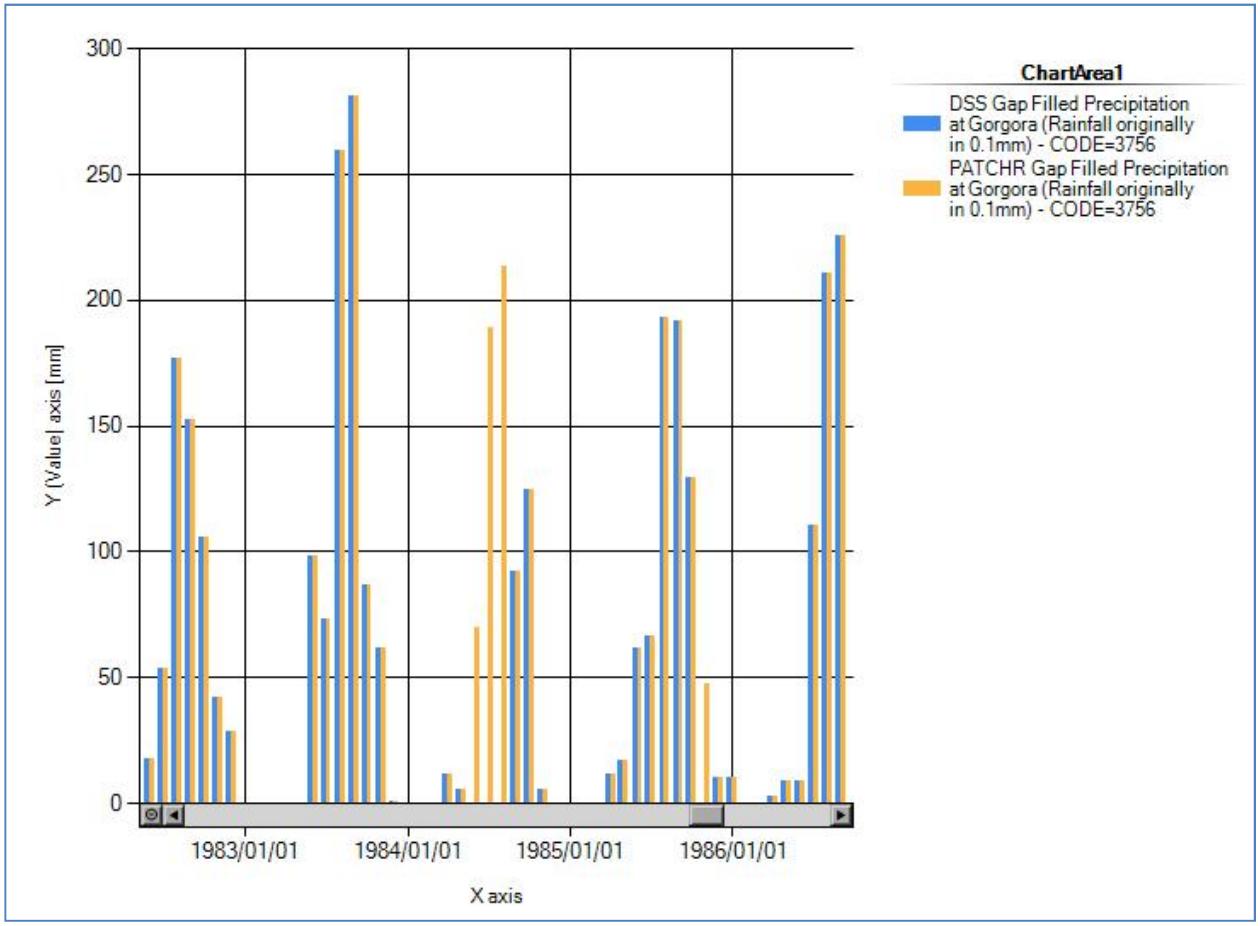


Figure 3-14: Gap-filling Comparison - DSS Tool and PATCHR

It can be seen that the DSS tool either assigned zero values to the missing months in 1984 and 1985, or cross-correlations for even the strongest correlated station-pair in these months were lower than 0.2 (more likely). PATCHR was able to infill these months, due to its ability to utilise sets of seasonal cross-correlations.

3.6 SPATIAL DATA QC PROCEDURES

3.6.1 Projections and Datums

Datums

Spatial data represent the location of features on the three dimensional earth surface. 3-D representations of the earth's surface are encapsulated in geographic coordinate systems (ellipsoids and/or datums). It is common practice to store and distribute spatial data sets in un-projected geographic coordinate systems. Units of these data sets are usually decimal degrees. With a few exceptions, the spatial data sets that have been collected or derived for use in the NB DSS applications are based on the **WGS84** datum. Working with any other datum will require datum transformations, which will introduce inaccuracies, and is not recommended.

Projections

All spatial data are associated with a specific reference scale. Global data sets usually have small scales (i.e. depicting large areas on a small paper space), while local data sets (such as detailed topographical surveys for scheme design) are done at large scales. Small scale two-dimensional representations (such as the Nile Basin shown on a paper map) are affected by the curvature of the earth, and introduce variations of scale across the 2-D representation. The magnitude and type of distortions that are introduced depend on the **projection** that is used to represent a 3-dimensional surface in 2-D space. Selection of projections for data processing are therefore influenced by the location and size of the area that one is working with, and requires an appreciation of the accuracy that is required for outputs.

In order to process spatial data sets to derive catchment parameters for modelling and scenario evaluation (examples include catchment areas, river slopes, summaries of land cover, population estimates and more), a meter based projection must be used. The projection should produce reasonably consistent area and length calculations across the study area. For this purpose, it is recommended that the UTM (Universal Transverse Mercator) Zone 36N projection be used for spatial data processing on a basin-wide scale. The Transverse Mercator projection delivers high accuracy in zones less than a few degrees in east-west extent. Maximum linear error for this projection is about 1 : 2 500, implying that a distance error of +- 4m can be incurred in every 10km measurement.

Parameters of the UTM Zone 36N projection are as follows:

Projection: Transverse_Mercator

False_Easting: 500 000

False_Northing: 0

Central_Meridian: 33.0

Scale_Factor: 0.999600

Latitude_Of_Origin: 0.000000

Linear Unit: Meter (1.000000)

It is often useful to overlay spatial data sets on Google Earth™ (GE) imagery for visual inspection of areas that may be affected by a scheme development. GE uses a proprietary projection - the so-called "Auxillary Web Mercator" projection. Before overlaying data on GE imagery, the data should therefore be re-projected to the GE projection which has the following parameters:

Datum: WGS_1984_Web_Mercator_Auxiliary_Sphere

Projection: Mercator_Auxiliary_Sphere

False_Easting 0.0

False_Northing 0.0

Central_Meridian 0.0

Linear Unit: Meter (1.000)

3.6.2 Spatial Data Quality Control Checks

Introduction

Quality control of spatial data sets is a specialised field covering many different applications such as digitising of vector data from paper maps, field surveys of ground data, remote sensing applications, development and processing of digital elevation models and so forth. For the purposes of this project, a generic set of spatial data quality control checks that can be used to ensure the integrity of data is provided, and is then practically illustrated by application to quality control of a drainage network :- a primary data set for any hydrological study.

Generic Quality Control Checks

QC of spatial data should typically address the following aspects (NLIS, 1997):

Data must be feature related. A feature is defined as an object that is related to a position on the earth's surface, and is comprised of two components - the spatial component and the descriptive component (also called the non-spatial component). Features can share all or part of their spatial geometry with other features thus eliminating redundancy of data.

Unique Identifiers. .A unique identifier of a feature instance must be allocated in accordance with NB-DSS requirements. The value of this lies therein that it uniquely identifies each feature instance within a feature class. It is the attribute that will be used for linking data of associated data sets for the same feature and as such must be standardised.

Spatial Referencing. All data sets will be projected to a common reference system (combination of geographic projection and datum). Preference will be given to a meter based system which will preserve accuracy.

Measurements and Quantities, and Time formats. All measurements and quantities must be in the International System of Units (SI units). Time formats must conform to NB-DSS conventions.

Data Ambiguity. Data must be unambiguous**.** Ambiguity of data leads to misinterpretation and loss of integrity of the data. A feature can occupy only one position in the real world and its digital representation should reflect this. Also, the feature's topological relationship (i.e. is it to the right or left, inside or outside, adjacent or co-incident) with respect to other features must be maintained.

Data Accuracy. The accuracy of the registration, or geo-referencing information is particularly important. Source documents (paper maps/photographs) should have registration marks accurately placed. A measure of how well the source document is registered is the Root Mean Square (RMS) error that is computed by the software. There is no fixed criteria for acceptability of this error, as this would depend on the quality and scale of the map being georeferenced and the purpose (intended use) of the georeferencing. RMS criteria for catchment scale features would therefore be much less restrictive than for, say, scheme layouts.

Planimetric accuracy. Accuracy of digitising will be ensured by adopting tolerances that are suitable for map scales ranging from 1:10 000 to 1: 1000 000. Scanning resolutions and raster tolerances will be chosen such that the data complies with this standard.

Digitising conventions. Vector data must be tested for logical consistency to ensure that there are no "undershoots", "overshoots", repeated digitising of points or lines, that intersections are correct, that there are no unwanted line crossings, that regions (polygons) are closed, and that digitising direction corresponds to flow direction in hydrological applications.

Topological Structuring. The spatial component of the vector data must have topological structuring. Topologically structured data have spatial relationships inherent in the data explicitly encoded. A topological rule could for example be used to detect polygons that are not closed (the line start and end points are not coincident). This permits better definition of the data, the removal of data redundancy, reduction in data volumes and a higher degree of integrity in the data.

Spatial Meta Data. A statement of data quality for all spatial data must be given**.** Such a statement will contain sufficient information so as to provide truth in labelling the data. This statement is essential for the user so that users can determine whether such data is fit for use in his application. A statement of data quality can be either generalised for the whole data set, specific for different parts of the data set or a combination of being generalised for some aspects and specific for others, depending on the type of data in the data set and the amount of detailed data quality information available. The data quality statement will comply with established international standards. Other metadata items will be added according to the metadata fields defined in the NB-DSS. All data processing steps (projecting, cleaning, vectorising or rasterising) will be flagged and documented.

Quality Control With Reference to a Drainage Network

A topologically correct drainage network is an essential data layer for many hydrological applications. The network can be used for correcting a digital elevation model with "stream burning" techniques, it can serve as a longitudinal centre line for taking off of river cross-sections, and used as a "snapping" layer for catchment outlets when generating modelling sub-catchments. The following quality control checks should be performed on the network (adapted from Hornby, 2010):

Null length polylines

Incorrect digitising, and automated cleaning methods can sometimes introduce errors such as collapsing the geometry of a polyline into "nothing". Such features still remain in the network and possibly retain their attributes. These features should be removed if the connectivity of the network will not be broken.

Multi-part polylines

Multi-part polygons are often created when the boundary of a lake along with the inflow and outflow have been captured. The resulting feature is a multi-part polyline. Figure 3-15 shows a multi-part polyline "exploded" into its individual parts, and it can be seen that the middle part (2) forms a loop. The error can be removed by deleting one of the two sides forming the loop.

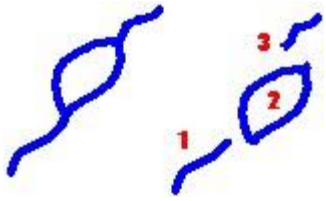


Figure 3-15 : Multi-part polylines

Self-intersecting polylines

These are created by careless digitising, and in many instances consist of one incorrectly placed vertex (circled in red in Figure 3-16). The error can be removed by deleting or moving the relevant vertex.

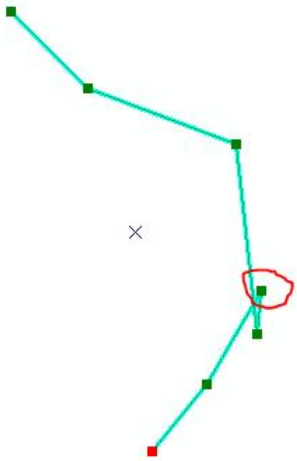


Figure 3-16 : Self-intersecting polyline

Closed polylines

Closed polylines often appear to be short tributaries. They are created when the endpoint is snapped back to the original starting point (i.e "folded over"), and can be identified when overlaid with the network node layer and looking for a polyline which does not have a node at one end.

To correct this, the downstream node should be deleted so that the second vertex now becomes the downstream node of the polyline. (Figure 3-17). The main stream polyline needs to be split at the red downstream vertex and polylines snapped to a new node.

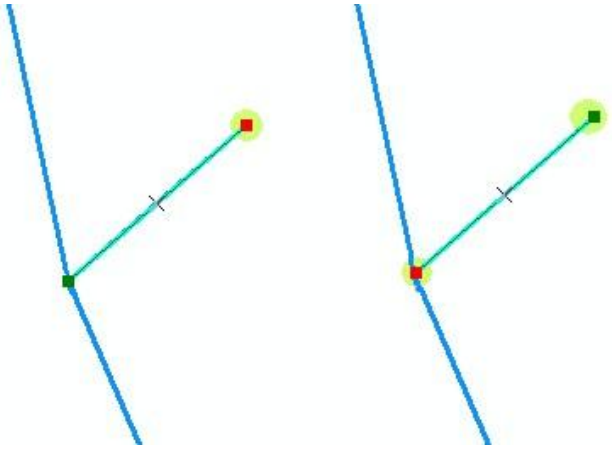


Figure 3-17 : Closed polylines - Incorrect (left), and fixed (right)

Disconnected polylines

River network polylines must join each other at their ends. A polyline intersecting another somewhere along its length breaks this rule and consequently the topology of the network.

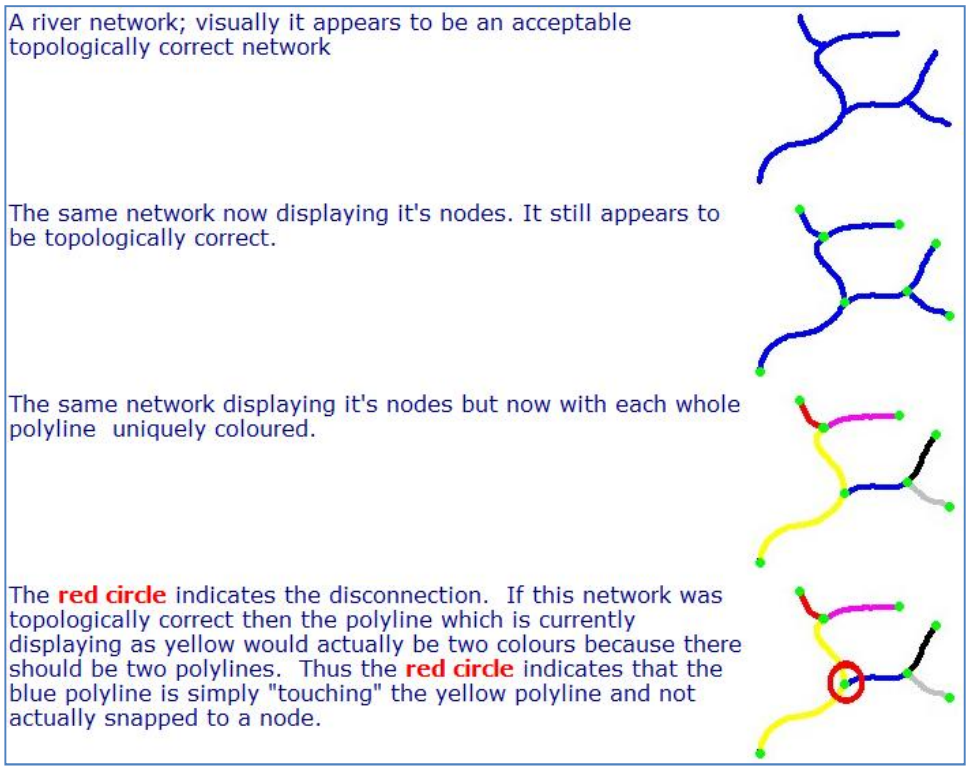


Figure 3-18 : Disconnected polylines (Hornby, 2010)

To fix the disconnection, the polyline must be split, and the end of the intersecting polyline must be snapped to the new node.

Double-digitised polylines

Double digitised polylines are polylines that share the same From and To nodes. These could be genuine braids that are not errors. However, if they intersect each other, other than at their nodes, they must be double digitised (Figure 3-19). The error can be corrected by selecting one of the polylines, and deleting it.

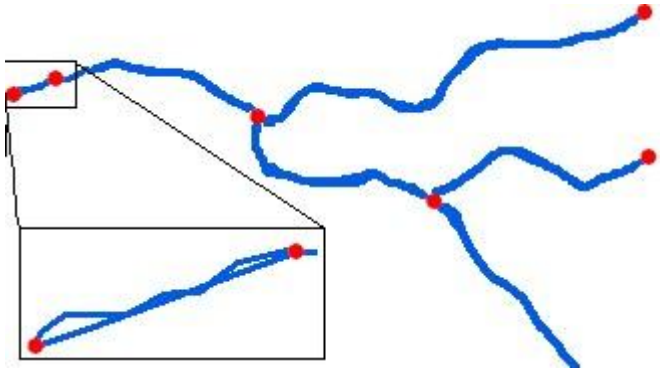


Figure 3-19 : Double-digitised polylines

Intersecting polylines

To be a valid river network, polylines must connect to each other only at their nodes. A network can fulfil this requirement (i.e. still have valid topology) yet still have polylines that intersect each other along their length. This is illustrated in Figure 3-20.

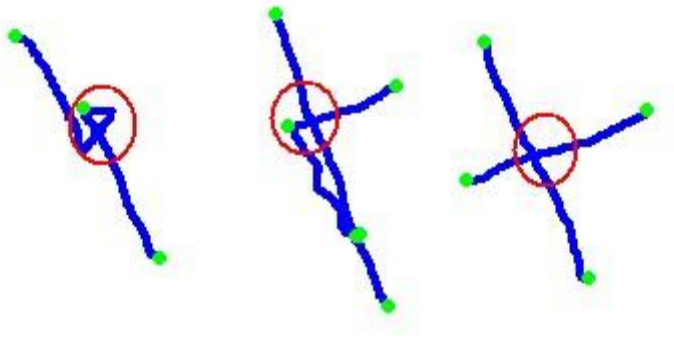


Figure 3-20 : Intersecting polylines

Fixing the intersection depends upon the type of intersection that is occurring. A self-intersection can often be fixed by deleting a single vertex. Alternatively, the polyline(s) may have to be split, and the intersection snapped to the new node.

Sources within the network

Tributary sources (starting points) should only be found on the outer ends of a river network. Poor digitising (or automated vectorisation) can create polylines that flow towards each other. Figure 3-21 shows an example of erroneous sources (red) flowing towards mouths (magenta) within a single river reach.

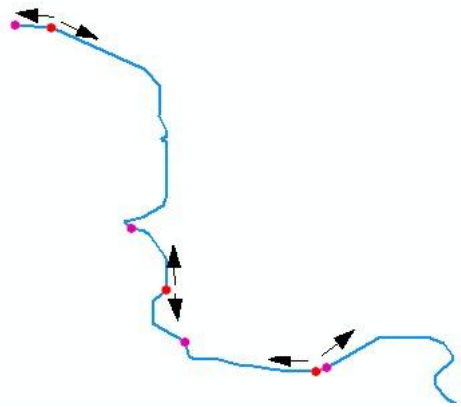


Figure 3-21 : Sources within a network

Sources in a network can be removed by flipping the direction of relevant polylines.

3.6.3 The Nile Basin Rivers Network

In discussion with the DSS core team and national specialists, it was agreed that the DSS will contain the NBI's fine scale quality controlled river network as well as a "model-scale" river network. The latter was derived by correcting flow direction errors in the NBI dataset followed by weeding of vertices (i.e. removing vertices that, within a specified tolerance, do not significantly contribute to defining the shape of a line) to approximately match the scale of the 90m SRTM DEM (a weeding distance of 50m was eventually adopted). Further modifications that were made to the "model-scale" network include:

- Flow directions were corrected by tracing the network from a point directly upstream of the Nile Delta in Egypt to the outer boundaries of the basin and ensuring that no coinciding "FROM""or "TO"nodes were found.
- Self-intersecting, intersecting, closed and disconnected polyline errors were corrected
- River segments upstream and downstream of lakes were connected through the lakes with artificial segments
- The final network was Strahler-ordered (orders 1-6) to allow for selection of subsets and to aid visual representation on maps.

Metadata were compiled for both data sets to clearly describe the differing origins and intended use of the data sets.

4. REFERENCES

Basson, M. S., Allen, R. B., Pegram, G. G. S. and van Rooyen, J. A. (1994). *Probabilistic Management of Water Resource and Hydropower Systems*, Water Resources Publications, Highland Ranch Colorado, pp424.

Gordon N.D, McMahon, G. T. Finlayson, B. L, Gippel, C. J (2004). *Stream hydrology: an introduction for ecologists.* Wiley-Blackwell.

Coccia G. and Todini E. (2010).*Recent developments in predictive uncertainty assessment based on the model conditional processor approach*. Hydrol. Ear th Syst. Sci. Discuss., 7, 9219–9270, 2010

Djokic, D., 2008. *Comprehensive Terrain Preprocessing Using Arc Hydro Tools*. ReCALL, (5), p.61. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20627044.

Gordon, A.D., 1981. *Classification*. Chapman & Hall.

Gyau-Boakye P. and Schultz G. A. (1994). *Filling gaps in runoff time series in West Africa,* HydrologicalSciences –Journal 39, 621-636.

Hirsch, R.M. 1982. A *comparison of four streamflow record extension techniques.* Water Resources Research, 15, 1781–1790.

Hirsch, R. M., Helsel, D. R., Cohn T.A., and Gilroy E.J. (1993) Chapter 17: *Statistical Treatment of Hydrologic Data*. In Handbook of Hydrology, Ed., Maidment, D. R., McGraw-Hill.

Hoaglin, D. C., Iglewicz B. and Tukey, J.W. (1986). *Performance of Some Resistant Rules for Outlier Labeling*. Journal of the American Statistical Association. Vol. 81, No. 396, pp. 991-999.

Hornby, D. D. (2010). RivEX (Version 6.9) [Software]. Available from http://www.rivex.co.uk

Kottegoda, N.T. and Elgy, J. (1977). *Infilling missing flow data*. In: Morel-Seytoux, H.J. (ed). Modelling Hydrologic Processes. Water Resources Publications.

National Land Information System (NLIS), 1997. *The Standard for Data for the National Land Information System*. The Standards Committee of the Co-ordinating Committee for the National Land Information System, South African Department of Water Affairs. Internal Document.

Nelder, J and Wedderburn, R (1972). *Generalized Linear Models.* Journal of the Royal Statistical Society. Series A (General) (Blackwell Publishing) **135** (3): 370–384.

O'Connell, P.E., Carron, J., Parkin, G., and O'Donnell, G.M. (2011) *Inception Report, Nile Basin Decision Support System (DSS), Data Processing and Quality Assurance, Pilot Application of the Nile Basin Decision Support System*: Stage 1, NBI Water Resources Planning and Management Project, 204pp.

O'Donnell, G.M. (2011) *Technical Note on Quality Assurance System and Meta Data TN0001, Nile Basin Decision Support System (DSS), Data Processing and Quality Assurance, Pilot Application of the Nile Basin Decision Support System*: Stage 1, NBI Water Resources Planning and Management Project.

O'Donnell, G.M. (2011) *Technical Note on Data Quality Control, Infilling and Record Extension TN0002, Nile Basin Decision Support System (DSS), Data Processing and Quality Assurance, Pilot Application of the Nile Basin Decision Support System*: Stage 1, NBI Water Resources Planning and Management Project.

Pegram, G & Zucchini. WS 1991. *Patching monthly rainfall data using CLASSR and PATCHR*. Proceedings of the Fifth SA National Hydrological Symposium, Stellenbosch, pp 7-2-1 – 7-2-10.

Pegram G. 1993. *Patching Streamflow Data Using PATCHS - A Guide*. Report prepared for the South African Department of Water Affairs and Forestry through BKS Inc. January 1993.

Pegram, G. 1997. *Patching rainfall data using regression methods*. 3. Grouping, patching and outlier detection, Journal of Hydrology, Volume 198, Issues 1–4, 1 November 1997, Pages 319-334, ISSN 0022-1694, 10.1016/S0022-1694(96)03284-2.

Vogel, R.M. and Stedinger, J. R. (1985). *Minimum Variance Streamflow Record Augmentation Procedures*. Water Resources Research, 21, 715-723.

ANNEXURE A

UNIVERSAL METADATA TEMPLATE XSD FILE

```
<?xml version="1.0" encoding="utf-8"?>
<xs:schema attributeFormDefault="unqualified" elementFormDefault="qualified"
xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:element name="metadata">
    <xs:complexType>
      <xs:sequence>
        <xs:element minOccurs="0" name="identification">
          <xs:complexType>
            <xs:sequence>
              <xs:element default="Keywords describing the data set and its data" name="keywords">
                <xs:simpleType>
                  <xs:list itemType="xs:string" />
                </xs:simpleType>
              </xs:element>
              <xs:element default="Short descriptive name for the data set." name="dataset_title"
type="xs:string" />
              <xs:element default="Unique identifier for the data set." name="dataset_id" nillable="true"
type="xs:string" />
              <xs:element minOccurs="0" name="metadata_date" type="xs:dateTime" />
              <xs:element default="Contact (email address) for person who compiled metadata"
name="metadata_contact" />
              <xs:element default="Name of organisation that is responsible for updating the data set"
name="custodian" type="xs:string" />
              <xs:element default="Contact details (URL or email address) of the organisation responsible for
maintenance and/or updating of the data set" name="custodian_contact" type="xs:string" />
              <xs:element default="Brief narrative summary of the contents of the set." name="abstract"
type="xs:string" />
              <xs:element minOccurs="0" name="description" type="xs:string" />
              <xs:element minOccurs="0" default="Geographic coordinate system (Datum) and projection (if
used).  Copy .prj file in here." name="spatialreference" type="xs:string" />
              <xs:element name="attributes">
                <xs:complexType>
                  <xs:sequence>
                    <xs:element minOccurs="0" maxOccurs="unbounded" default="Attribute name" name="attr_name"
type="xs:string" />
                    <xs:element default="Attribute description.  Provide units where applicable"
name="attr_description" type="xs:string" />
                  </xs:sequence>
                </xs:complexType>
              </xs:element>
            </xs:sequence>
          </xs:complexType>
        </xs:element>
        <xs:element minOccurs="0" name="dataquality">
          <xs:complexType>
            <xs:sequence>
              <xs:element name="status">
                <xs:simpleType>
                  <xs:restriction base="xs:string">
                    <xs:enumeration value="under development" />
                    <xs:enumeration value="regular updates" />
                    <xs:enumeration value="final" />
                  </xs:restriction>
                </xs:simpleType>
              </xs:element>
              <xs:element name="source_data">
                <xs:simpleType>
                  <xs:list itemType="xs:string" />
                </xs:simpleType>
              </xs:element>
              <xs:element name="grounddate" type="xs:gYear" />
              <xs:element minOccurs="0" default="Applicable to digitised vector data.  Scale of source data."
name="reference_scale" type="xs:string" />
              <xs:element default="Applicable to raster data sets.  Grid cell size in data set units"
name="resolution" type="xs:string" />
              <xs:element minOccurs="0" default="Information on history of derivation/construction of data
set, including:references to external documentation where available" name="lineage" type="xs:string" />
              <xs:element default="Known and suspected deficiencies relating to the data set."
name="limitations" type="xs:string" />
            </xs:sequence>
          </xs:complexType>
        </xs:element>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
</xs:schema>
```

ANNEXURE B

WORKED EXAMPLE FROM PATCHING STREAMFLOW DATA USING PATCHS

PATCHING STREAMFLOW DATA USING PATCHS

- A GUIDE

by

GGS Pegram

for

The Department of Water Affairs and Forestry

through

BKS Inc

January, 1993

How to run PATCHS

The modeller will need an executable version of PATCHS a command file and some data (streamflow and optionally some rainfall) in HRU format. Also required is a reasonably fast PC with an 8087 coprocessor running DOS 2.1 or later. Once set, enter PATCHS and answer three questions from the keyboard

- the name of the command file,
- an optional date and time and information list (a blank will do)
- the name of the desired output file.

Output to the screen (not repeated in the output file) gives the iteration number and the AICc as a measure of the convergence rate, to show that the machine is busy.

Accompanying this report is a diskette containing various data files, an executable version of PATCHS and a README.DOC file which contains the following text:

This diskette contains the following files

This file in ASCII:
README.DOC

The command file in ASCII.
VAALB.FIL

The executable PATCHS program compiled under DOS for an 8087 coprocessor:
PATCHS.EXE

The streamflow data in HRU format:
GROOTB.INC
VAALB.INC

The rainfall data in HRU format:
298512.DAT
367484.DAT
368634.DAT
406221.DAT
439389.DAT
440767.DAT

Three typical output files in ASCII:
VAAL2421.110
VAAL2621.011
VAAL2621.000

Description of the contents of the files:

The COMMAND FILE - **VAALB.FIL** - contains the following lines:

```
2 6 2 1 40 0 0 0 0
1 GROOTB.INC
1 VAALB.INC
0 298512.DAT
0 367484.DAT
0 368634.DAT
0 406221.DAT
0 439389.DAT
0 440767.DAT
```

The first line in the command file sets the values of the controlling parameters which are, in order:

1	*s*	the number of streamflow stations, an integer between 1 and 3 inclusive
2	*r*	the number of rainfall stations, an integer between 1 and 8 inclusive
3	*p*	the number of streamflow lags, an integer 1 or 2
4	*q*	the number of rainfall lags, an integer 0 or 1
5	*maxit*	the maximum number of iterations to limit the computation, 40 is OK
6	*ipatch*	an integer flag to create patch files (=1) or not (=0)
7	*ilog*	an integer flag to fit a 3-parameter lognormal distribution to the $\{y_t\}$ sequence and transform to normal (=1) or to leave alone (=0)
8	*ir*	an integer flag to return smoothed data (=1) and compute deleted residuals and the MCV (mean cross-validation criterion) or (=0) to return the streamflow data as recorded
9	*itr*	an integer flag to standardize month-by-month (=2) to scale by standard deviation month-by-month (=1) or to leave alone (=0)

Initial choices

s & r

A sensible first choice of control parameters might be to include as many of the streamflow and rain gauges that are thought to be properly associated (the rain gauge data must have been screened and patched if necessary and selected using the routines CLASSR & PATCHR designed for the purpose) thus in our example, *s* & *r* are set to 2 and 6 respectively, initially.

p & q

Because the routine speed is dependent on the size of the transition matrix, it seems wise for exploratory calculation to build up from small to larger lags, hence set *p* & *q* to 1 & 0 initially.

maxit
With *ir* = 1 (the suggested initial choice) reasonably coherent data should allow the algorithm to converge within 20 to 30 iterations, hence set *maxit* = 40.

ipatch
Only on the last (!) run would one wish patched data files (the patched data appear in the output file as a matter if course, so one could always edit that to get the patched data files if desired) so initially set *ipatch* =0.

ilog
Experience indicates that the lognormal transform is not helpful in streamflow patching, (although the feature has been retained for flexibility) so this is usually set as *ilog* = 0.

ir
To enable the modeller to choose the best model from among the various offerings, deleted residuals and a mean cross-validation criterion (MCV) are computed if *ir* =1, but not otherwise. The deleted residuals are helpful in identifying outliers which are possible errors, so set *ir* = 1 initially.

itr
There are three possible combinations of shifting and scaling accepted by the program; none at all *itr* = 0, no shifting but scaling month-by-month by standard deviations if *itr* = 1, and a month-by-month standardization using means and standard deviations if *itr* = 2. These transformations are offered to help with producing better models and hence better patches. Experience shows that no transformation (assuming the process parameters are constant and the process is linear) is often a good choice, so an initial choice of *itr* = 0 is suggested.

The second and third lines in the command file contain the names of the streamflow files, preceded by a flag (1) to indicate that they are streamflow files. The number of streamflow files must equal *s* in the header line, in the example, 2. If these numbers do not match, the program will stop with an I/O error, because of the differing format between rainfall and streamflow files.

The fourth to the last lines contain the names of the rainfall files, preceded by a flag set to 0 for rainfall. Again, the number of rainfall files must equal *r* in the header line, in the example, 6. If the rain gauge list exceeds *r*, the files will be ignored

THE DATA FILES

The data span the years [illegible] as described in the accompanying report and are therefore a subset of the available data. The rainfall data have been patched where necessary using the program PATCHR and were selected from among a large number of possible stations using CLASSR, a companion program. They are in [illegible] and those data which are suspect or missing should be flagged with an appropriate character of the modeller's choice. In the example, three years' data have been flagged with A after the numbers in question. Any ASCII symbol other than a blank will be treated as a flag, and these flags will appear in the output file against the patched values.

THE OUTPUT FILES

There are three output files given in the example on the diskette. We shall discuss **VAAL2421.110** This file is the output of an indifferent patch, the algorithm not converging before 40 iterations. The AICc came out as 656.91 and the MCV as 0.6414, both larger than the comparable results for trial 3 in Table 4 of the paper in the appendix of the Guide accompanying this diskette.

On examining the parameter estimates, it will be seen that there is a relatively large negative coefficient at the end of the first row (Grootdraai on Vaal-lag-2). In addition, the measurement noise matrix R has a large element for Vaal.

Considering the deleted residuals, it will be seen that there are relatively few, if any, large positive residuals. The ones that have been flagged are negative, corresponding to unexpectedly low flows. The log transform has eliminated the deletion residual in February '74 referred to in the paper in the report. On examining the patched values, it will be seen that the large values have been seriously underestimated when compared with the successful patch shown in VAAL2621.010 and discussed in detail in the paper in the report. This patching should be discarded. A better solution is the one suggested in the guide, namely Trial 9 of Table 4 in the paper in the appendix.

CONCLUSION

To conclude this brief guide, the modeller will find that it is necessary to explore the various options presented in the program and search for a 'best' solution.

2.2 MODEL SELECTION AND OUTLIER DETECTION.

In order to implement the above state-space model, a suitable choice of value for r, the number of rainfall gauges, and p and q, and the number of lags to include in the model, needs to be made. Traditionally, one would base one's choice of a particular model structure on the AIC (Akaike, 1971), BIC (Hannan and Quinn, 1979) or AICc (Hurvich and Tsai, 1989) values that one obtains. As an alternative, however, we will consider using a cross-validatory technique for model selection, because the deleted residuals that arise from employing such an approach can then be used to identify potential outliers that may exist in the data set.

The derivation of the deleted residual vectors that arise from using our chosen state space model is given in the appendix. Briefly, after having obtained a set of parameter estimates for our state space model, the deleted residual vector $\mathbf{r}_t^*$, is computed using a forward sweep to calculate $\mathbf{C}_t^{-1}, \mathbf{K}_t$ and $\mathbf{e}_t$, and then a backward sweep to calculate:

$$\mathbf{r}_t^* = [\mathbf{C}_t^{-1} + \mathbf{K}_t^T \mathbf{H}_t \mathbf{K}_t]^{-1}[\mathbf{C}_t^{-1}\mathbf{e}_t - \mathbf{K}_t^T \mathbf{m}_t]$$

$$\mathbf{L}_t = [\mathbf{I} - \mathbf{K}_t \mathbf{M}_t]\mathbf{A}$$

$$\mathbf{m}_{t-1} = \mathbf{A}^T \mathbf{M}_t^T \mathbf{C}_t^{-1} \mathbf{e}_t + \mathbf{L}_t^T \mathbf{m}_t$$

$$\mathbf{H}_{t-1} = \mathbf{A}^T \mathbf{M}_t^T \mathbf{C}_t^{-1} \mathbf{M}_t \mathbf{A} + \mathbf{L}_t^T \mathbf{H}_t \mathbf{L}_t$$

where $\mathbf{m}_n$ and $\mathbf{H}_n$ are set equal to zero.
A mean cross-validation statistic that can be used for model selection can then be given by

$$MCV_i = \mathbf{r}_t^{*T} \mathbf{r}_t^* / ns$$

3. AN APPLICATION TO TWO STREAMFLOW RECORDS

3.1 DATA REQUIREMENTS

In order to demonstrate the performance of our patching algorithm on a set of practical data, the streamflow records for two dams were examined.

The monthly rainfall figures for six rainguages in the catchment areas of both dams were also selected from among a large number of candidates using an algorithm based on the covariance biplot (Pegram and Zucchini,1991). A portion the existing streamflow record for the first dam was flagged as missing and our state space model was then used to provide an estimate for the 'missing' observations. In particular the two streamflow records that were examined were those of the inflow to Vaal Dam and the inflow to Grootdraai Dam, which is situated upstream from the former dam. Twenty-nine years of concurrent data ranging from 1955/6 to 1983/4 were used, with the three years, from 1964/5 to 1966/7 being flagged as missing for the Vaal Dam records. These years were chosen because they include a wet, dry and normal year of flows. For completeness, the streamflow records for the Vaal and Grootdraai dams are given in Tables 1 and 2, with the concurrent rainfall data for each of the raingauge stations being listed in Table 3.

3.2 MODELLING CHOICES

The model parameters in the state space model were estimated using the following set of starting values for the EM algorithm, viz.

- $\Sigma(0)$ was set equal to the identity matrix,
- $\mu(0)$ was set equal to the initial observation y_0
- $A(0)$ $B(0)$ and $Q(0)$ were set equal to $\frac{1}{2}\mathbf{I}$, $\mathbf{0}$ and $\mathbf{I}$ respectively.

Because the streamflow data may typically be skewed (due to the frequent occurrence of a zero streamflow measurement during periods of drought) the option of a lognormal transformation, coupled with a suitable scaling of the transformed variable, was also incorporated into the algorithm. If this option is used without shifting the data, a small constant is added to cope with the zero flows which do occur.

Finally, because it is unusal to encounter more than two month's lag in a streamflow record, and more than one month's lag in the effect that a rainfall record will have on the streamflow, the choice of parameter values for p and q were confined to the range of choices that are listed below, together with other controlling parameters and their ranges

8

Controlling Parameter	s	r	p	q	log	R	T
range	1/2/3	0,1,...,8	1/2	0/1	0/1	0/1	0/1/2

where

log = 0 ⇒ no lognormal transformation
log = 1 ⇒ lognormal distribution fitted
R = 0 ⇒ measurement noise not estimated
R = 1 ⇒ measurement noise estimated
T = 0 ⇒ no transformation employed on the streamflow data
T = 1 ⇒ monthly scaling by dividing by the standard deviation
T = 2 ⇒ monthly scaling by first centering the data, and then dividing by the standard deviation

These options then produced the results that are recorded in Table 4.

4 RESULTS AND CONCLUSION

Blocks A, B and C in Table 4 contain the results of 9 different trials. As distinguishing features, all the trials in Block A use a lognormal transformation on the streamflow data, while those in Block B and C leave the streamflow value unchanged. In Block C, the number of raingauges is increased from 4 to 6 with the variation within each of the blocks A,B and C being that the value of T changes.

In Block D, the variants are r and p, with the transformation being restricted to a scaling ($T = 1$) of the streamflow data by it's standard deviation only. In Block E, no transformation ($T = 0$) is employed on the data while, in Block F, the trials of Block C are re-estimated with the variance-covariance noise matrix R being set equal to zero. This then ensures that the original streamflow data are returned by the algorithm as the patched values for y.

From the results that are presented in Table 4, it would appear that the model structure denoted by the trial number 9 performs the best from a MCV point of view when combined with the AICc value and a visual scan of

the deleted residuals against the recorded data for Vaal Dam which is given in Figure 4. The largest deleted residual in that plot corresponds to the recorded monthly streamflow measurement of 2200 which occurred in February 1975. An examination of the Tables 3a-f, however, will show that no inordinately high rainfall was recorded in that month and thus our cross-validation technique would suggest that we treat this observation as representing a possible outlier.

Turning our attention to an examination of the model structures that were given in block F, because the MCV criterion is irrelevant, and the AICc-values are not directly comparable because the y-series are different, a plot of the patched and hidden data for these model structure is given in Figures 1 to 3. An examination of these three figures would indicate that the model trial number 17 is to be preferred over the other two. This model incidentally corresponds with that of trial 9.

Table 1.

Monthly total flow (millions m^3/s) into Vaal Dam
with "missing data" high-lighted in italic type

Year	oct	nov	dec	jan	feb	mar	apr	may	jun	jul	aug	sep
1955	25	33	121	192	90	216	113	57	21	16	13	6
1956	15	322	1385	639	120	134	36	5	2	218	46	1471
1957	1363	334	276	513	126	62	108	19	15	22	26	51
1958	29	146	264	221	94	23	0	126	21	25	18	23
1959	71	323	161	96	155	180	41	67	15	17	25	13
1960	53	69	521	268	66	46	166	33	51	8	28	12
1961	33	112	167	95	106	57	2	9	5	8	9	31
1962	20	102	126	317	114	24	34	13	4	137	21	21
1963	13	95	113	290	97	85	46	11	2	18	16	23
1964	*334*	*800*	*380*	*304*	*188*	*21*	*18*	*12*	*7*	*12*	*2*	*35*
1965	*5*	*34*	*35*	*129*	*236*	*9*	*2*	*7*	*6*	*3*	*7*	*5*
1966	*6*	*33*	*257*	*443*	*1376*	*177*	*234*	*33*	*33*	*12*	*15*	*15*
1967	6	69	102	36	14	20	12	12	4	3	2	5
1968	4	12	42	61	23	80	74	37	22	6	6	5
1969	116	82	147	82	134	30	4	0	0	4	4	15
1970	52	124	12	88	179	32	250	29	10	8	10	15
1971	1	162	310	352	70	303	61	20	10	10	14	12
1972	22	42	40	5	139	24	34	0	6	2	45	12
1973	37	83	298	662	499	71	76	26	19	13	12	16
1974	6	268	511	658	2200	311	132	47	29	21	21	51
1975	115	289	764	659	790	513	118	240	50	35	32	25
1976	290	201	152	104	695	107	77	12	16	16	19	22
1977	40	56	200	883	187	197	185	28	20	19	15	32
1978	76	30	105	54	33	45	7	3	3	8	47	49
1979	53	96	99	124	362	131	5	4	4	11	10	8
1980	21	19	125	147	229	338	12	13	9	12	10	56
1981	13	43	61	62	2	9	8	13	4	9	16	20
1982	22	77	16	28	17	12	3	2	1	3	18	8
1983	61	189	322	198	54	58	71	14	8	12	12	40

1

Table 2.

Monthly total flow (millions m^3/s) in ~~the Vaal river at Standerton~~ Grootdraai Dam

Year	oct	nov	dec	jan	feb	mar	apr	may	jun	jul	aug	sep
1955	13	11	205	155	24	91	29	12	8	5	3	4
1956	12	97	269	56	12	19	13	5	3	35	8	213
1957	171	32	43	193	19	13	75	5	3	2	2	25
1958	4	48	173	44	23	10	0	2	2	0	0	1
1959	2	35	62	16	43	23	22	10	2	2	3	1
1960	5	38	212	38	31	68	197	14	11	6	3	2
1961	12	56	95	75	34	9	3	3	1	1	1	0
1962	6	119	64	45	7	2	4	1	3	70	7	3
1963	3	78	13	147	24	8	3	5	1	1	1	0
1964	164	277	87	107	23	3	2	1	1	1	1	1
1965	0	0	11	13	17	1	0	0	0	0	0	0
1966	7	6	44	125	373	35	16	6	4	4	2	5
1967	4	64	91	19	8	32	5	2	1	1	1	1
1968	0	16	60	42	9	69	45	34	8	5	3	3
1969	136	60	165	50	82	6	2	2	2	2	1	1
1970	8	27	4	46	38	2	40	5	2	1	1	2
1971	1	122	243	148	20	26	7	7	4	3	2	2
1972	1	5	8	8	15	6	19	4	1	0	3	0
1973	6	39	65	70	72	11	15	8	6	5	2	1
1974	6	100	307	165	507	120	23	10	4	4	4	2
1975	4	133	312	201	145	25	79	85	8	6	4	2
1976	13	59	85	114	288	19	16	3	3	3	1	1
1977	1	11	28	201	94	75	6	4	3	3	1	2
1978	66	15	2	21	4	2	2	0	0	0	2	4
1979	13	30	27	38	155	16	3	1	2	1	2	1
1980	3	14	27	21	32	68	2	1	1	2	1	3
1981	6	3	36	45	4	1	0	1	0	1	1	2
1982	3	1	2	7	2	0	3	0	0	0	0	1
1983	1	142	83	67	102	10	8	2	3	5	2	5

2

Table 3a

Monthly rainfall totals (in 1/10 mm) at gauge 298/512

Year	oct	nov	dec	jan	feb	mar	apr	may	jun	jul	aug	sep
1955	655	1755	1395	755	2165	1500	305	500	0	0	25	290
1956	1160	1915	2940	1090	800	1090	860	0	340	565	540	2810
1957	1650	1050	1310	1385	400	980	775	215	0	0	0	730
1958	630	1325	1685	900	765	495	1065	1475	0	405	20	45
1959	1470	1115	1760	1308	1153	1095	600	120	0	130	355	330
1960	1095	805	1707	1425	260	306	1032	675	196	85	0	330
1961	90	2065	855	1360	1360	340	705	75	0	0	30	360
1962	560	1270	300	2640	560	965	455	185	400	200	0	0
1963	865	1140	1102	2310	805	1585	345	0	305	0	320	310
1964	1595	295	1130	880	520	415	690	20	520	155	605	205
1965	375	785	450	1559	825	0	110	95	50	0	75	310
1966	207	655	1220	1610	2130	1415	510	360	10	0	130	39
1967	475	929	585	505	712	856	480	240	0	20	265	0
1968	165	898	1050	440	795	1073	682	423	35	0	0	80
1969	935	580	1273	992	855	545	150	100	330	125	485	605
1970	212	335	1355	1460	633	651	582	553	13	142	115	95
1971	422	489	352	875	680	305	150	0	0	0	55	0
1972	250	430	0	400	1990	485	115	0	0	0	375	180
1973	285	604	1032	1178	1445	613	734	30	358	0	139	63
1974	249	2080	1665	2215	2015	555	225	0	0	85	200	1260
1975	800	1763	1650	1728	1500	2461	455	525	0	0	0	295
1976	1570	2115	1078	2220	740	907	585	0	0	0	0	615
1977	1215	415	743	1945	505	760	730	0	0	15	195	260
1978	665	90	1385	120	1748	740	140	355	0	510	1380	170
1979	345	1045	870	1500	1005	320	175	190	95	10	0	795
1980	155	765	690	1995	1960	325	125	0	64	0	690	210
1981	1060	760	1275	885	200	615	675	75	0	70	180	180
1982	810	1115	440	530	720	560	160	280	50	218	6	220
1983	1425	1300	1415	780	175	1015	290	0	30	120	620	260

3

Table 3b

Monthly rainfall totals (in 1/10 mm) at gauge 367/484

Year	oct	nov	dec	jan	feb	mar	apr	may	jun	jul	aug	sep
1955	1175	1060	1050	510	1225	535	55	830	0	10	0	75
1956	365	1055	1565	1353	715	485	230	130	270	920	285	1955
1957	905	795	1100	1230	395	550	725	0	0	0	0	805
1958	245	700	1590	1540	360	390	850	530	45	205	0	25
1959	1140	820	1225	455	1015	1020	490	25	0	15	330	140
1960	835	1095	2130	655	400	625	1325	320	195	0	30	535
1961	190	1435	270	310	740	860	645	0	0	0	120	655
1962	710	1295	700	1180	340	590	725	625	255	320	0	0
1963	230	1460	565	1465	385	1375	410	85	365	0	270	390
1964	2170	535	1330	1240	435	235	540	150	160	45	0	105
1965	160	1000	340	920	950	0	0	35	60	0	20	490
1966	470	835	1390	2180	1405	1030	635	260	0	0	125	110
1967	460	590	1060	290	320	1080	260	430	30	0	70	125
1968	155	385	900	520	250	1370	355	655	170	0	0	35
1969	1024	920	605	780	760	710	265	240	145	375	70	550
1970	635	1025	890	1075	1400	365	980	60	75	0	0	410
1971	405	1125	1110	1755	855	1050	130	85	0	0	75	0
1972	272	975	630	845	1120	655	575	0	0	0	955	420
1973	255	1300	1505	1370	1130	100	370	0	170	0	0	80
1974	720	1810	1410	1245	1410	355	570	145	0	0	0	32
1975	665	1815	1035	1245	1175	1140	210	515	70	0	0	195
1976	1200	1185	780	1965	495	770	0	0	0	0	0	1375
1977	845	625	1195	1763	520	2605	607	0	30	0	369	254
1978	789	544	1777	864	721	531	202	252	24	264	849	519
1979	1109	629	688	746	1066	200	305	31	32	0	0	534
1980	118	1413	508	2162	1235	698	267	76	105	0	812	500
1981	279	449	1310	1027	302	632	380	0	0	142	0	105
1982	1971	329	504	668	43	293	283	381	168	110	0	97
1983	1285	1753	1232	806	157	1100	90	0	85	55	754	25

4

Table 3c

Monthly rainfall totals (in 1/10 mm) at gauge 368/634

Year	oct	nov	dec	jan	feb	mar	apr	may	jun	jul	aug	sep
1955	742	897	1166	713	972	613	146	832	0	74	0	290
1956	849	1161	2540	1110	393	636	159	66	190	897	327	1861
1957	989	1672	1330	1434	230	1680	1123	96	0	0	0	945
1958	634	1272	1310	1428	493	329	695	505	35	230	1	15
1959	1324	1315	882	567	839	417	612	40	4	0	225	96
1960	1098	981	2124	945	845	376	1270	428	145	0	6	410
1961	259	1260	375	624	647	582	133	2	0	0	122	270
1962	427	1285	788	1214	207	230	547	310	294	268	0	0
1963	370	1072	706	1870	823	1026	495	70	95	0	190	352
1964	2642	768	1913	1521	530	176	461	13	121	135	2	56
1965	250	1211	457	1122	1393	0	80	55	103	0	26	520
1966	458	847	1698	2002	1391	622	955	227	0	60	9	215
1967	568	775	1577	769	622	1630	503	268	17	1	212	30
1968	501	393	715	988	464	1323	333	490	40	8	17	120
1969	660	547	1620	1572	862	455	182	220	135	250	100	221
1970	881	751	1148	871	843	461	1374	196	120	26	2	455
1971	709	1558	739	1382	540	922	65	100	125	2	45	26
1972	639	938	438	658	1020	330	545	14	0	0	657	480
1973	363	798	1015	1269	1255	496	710	85	115	12	3	310
1974	525	1815	2097	1785	1512	320	680	93	0	0	14	836
1975	189	857	1503	1149	767	941	216	441	30	0	0	170
1976	1285	822	1012	1296	190	610	150	60	0	0	10	705
1977	685	468	1880	1695	289	880	405	2	12	0	485	403
1978	968	435	1552	349	533	692	75	257	22	282	895	742
1979	707	352	1013	945	1180	506	132	70	1	0	60	358
1980	376	1196	999	867	1011	372	242	17	88	2	335	496
1981	470	455	714	1046	128	378	522	10	31	168	40	232
1982	1734	184	653	894	739	925	322	217	161	55	18	95
1983	1356	1611	1139	397	252	1046	45	26	56	130	604	160

5

Table 3d

Monthly rainfall totals (in 1/10 mm) at gauge 406/221

Year	oct	nov	dec	jan	feb	mar	apr	may	jun	jul	aug	sep
1955	730	1195	1995	385	650	880	0	825	0	28	0	430
1956	965	1645	1565	885	785	490	260	165	155	880	265	2115
1957	1270	615	580	1940	625	895	870	90	0	0	0	515
1958	575	1275	1475	535	570	875	415	460	0	80	0	300
1959	1045	1920	1046	680	1225	1135	675	0	0	100	185	190
1960	845	1260	1544	1000	565	735	1110	295	115	0	0	715
1961	560	1170	1400	975	1040	790	565	75	0	0	25	109
1962	340	1630	1465	1775	710	595	510	230	330	690	0	0
1963	555	1540	585	2138	576	445	400	0	5	0	0	155
1964	3052	661	883	1195	610	925	325	0	0	35	35	260
1965	295	1180	630	432	655	105	25	50	51	0	50	458
1966	857	1105	1270	1741	2040	295	710	95	35	200	135	235
1967	925	900	1383	480	335	1008	320	300	0	0	315	62
1968	300	1191	1237	2069	583	1470	680	451	0	20	35	555
1969	1043	1075	1810	1042	700	200	310	0	37	100	570	1050
1970	805	420	545	1008	150	715	800	355	0	0	0	330
1971	755	1476	1490	645	475	845	0	108	0	0	75	0
1972	90	595	192	795	760	575	403	40	0	0	635	435
1973	345	1120	717	2182	345	530	600	115	165	115	105	0
1974	595	839	1909	1527	1998	255	550	0	0	0	15	264
1975	510	1315	2155	900	330	195	300	485	0	0	0	0
1976	615	885	1925	1860	440	1225	240	0	0	0	0	140
1977	532	870	942	1326	780	315	340	0	0	0	95	510
1978	1280	1082	610	140	520	810	0	0	0	310	985	480
1979	980	530	664	1687	1537	130	175	0	0	0	160	230
1980	190	1705	460	1377	1020	780	245	0	138	0	165	120
1981	620	698	1140	1186	65	267	90	90	0	75	0	55
1982	728	390	632	502	120	662	470	90	255	75	80	15
1983	1452	1966	756	1008	325	523	155	0	92	145	720	103

Table 3e

Monthly rainfall totals (in 1/10 mm) at gauge 439/398

Year	oct	nov	dec	jan	feb	mar	apr	may	jun	jul	aug	sep
1955	1175	870	725	235	1038	603	0	680	0	0	0	159
1956	774	513	1420	1049	353	866	537	85	258	515	284	1744
1957	933	899	456	1101	358	691	505	101	0	0	0	0
1958	640	983	557	1378	736	300	504	540	12	20	0	54
1959	786	841	1212	415	725	0	525	0	1	0	290	157
1960	0	725	1532	0	400	470	1270	300	110	0	30	116
1961	235	1166	483	638	1049	0	0	0	0	0	71	655
1962	444	1627	825	1385	520	683	641	367	459	270	0	20
1963	115	0	370	1491	533	715	525	155	200	0	152	290
1964	1511	607	1585	1500	584	228	675	0	69	170	5	35
1965	137	692	337	950	733	286	23	15	35	0	45	89
1966	0	483	550	2202	1545	1574	1243	195	0	0	21	156
1967	472	845	1455	455	665	955	356	404	0	0	135	125
1968	363	613	970	325	155	816	420	775	30	0	30	0
1969	0	819	1005	1220	550	702	285	147	135	310	25	0
1970	345	1115	1500	1135	570	310	1025	75	210	0	25	226
1971	560	1079	1505	1215	330	745	37	150	145	0	230	14
1972	620	599	550	740	1225	742	235	0	0	0	410	480
1973	495	1250	1370	1125	775	570	85	160	80	0	65	75
1974	883	1270	1530	2070	1435	435	805	60	0	0	0	283
1975	285	645	1130	875	911	628	420	445	5	0	0	20
1976	953	815	1540	1300	290	975	0	0	0	0	0	525
1977	525	705	1860	1555	323	330	465	0	200	0	90	190
1978	1222	214	941	943	535	723	196	155	18	198	1073	491
1979	934	1711	832	1498	1110	346	248	44	0	0	69	253
1980	497	1093	1021	1632	363	316	290	10	0	0	429	517
1981	423	1463	1446	1225	288	396	0	0	0	278	10	64
1982	1279	208	676	952	233	198	130	225	246	95	115	62
1983	1060	1332	1045	1195	177	725	255	50	80	45	411	126

7

Table 3f

Monthly rainfall totals (in 1/10 mm) at gauge 440/767

Year	oct	nov	dec	jan	feb	mar	apr	may	jun	jul	aug	sep
1955	540	0	1280	465	875	813	0	590	0	80	0	290
1956	1325	598	1050	1160	540	750	430	30	140	710	305	1270
1957	1198	425	1185	1349	285	710	900	30	0	0	0	845
1958	255	1082	940	880	560	195	432	440	0	0	0	0
1959	625	797	1557	513	1395	740	435	0	20	55	200	100
1960	545	875	1790	1350	324	685	925	205	180	0	0	535
1961	550	921	945	910	758	435	595	0	0	0	0	460
1962	305	1475	1270	1820	80	745	445	250	545	575	0	0
1963	24	1495	335	1925	405	1045	1237	0	0	0	280	220
1964	1500	495	1980	1305	690	220	449	0	0	180	50	0
1965	765	675	473	835	1125	140	30	73	75	0	26	220
1966	660	715	965	2367	2196	608	850	200	0	0	94	285
1967	1065	852	2220	520	700	1120	305	90	0	0	40	64
1968	340	982	1227	993	504	1543	0	747	0	0	0	365
1969	1105	797	830	955	389	382	289	265	148	140	41	150
1970	1291	926	1070	1160	856	60	89	35	50	12	8	442
1971	910	1625	724	1492	855	1185	190	130	60	0	245	230
1972	460	960	285	905	1535	620	435	0	0	0	315	670
1973	60	1445	1615	1623	1110	535	715	210	90	0	0	130
1974	310	1882	1630	2435	1730	555	815	0	50	0	0	260
1975	630	1955	855	750	810	730	240	895	0	0	0	190
1976	835	1090	2025	1910	150	830	190	0	0	0	0	575
1977	635	912	1302	1391	1042	458	559	41	0	0	200	275
1978	1171	897	794	1016	596	826	530	46	83	297	609	390
1979	950	905	576	2073	1370	378	371	0	0	0	120	254
1980	576	1283	1573	1530	2169	688	405	0	51	0	192	448
1981	433	1173	689	744	743	466	351	0	0	81	0	228
1982	1421	886	587	377	481	323	170	125	170	93	105	160
1983	1247	1799	1065	945	640	922	71	0	80	145	250	115

8

Table 4

Selection of controlling parameters and resulting criteria for optimizing model choice

Block	Trial	s	r	p	q	log	R	T	m	AICc	MCV
A	2	2	4	1	1	1	1	2	82	730.9	.668
	2	2	4	1	1	1	1	1	58	984.0	.845
	3	2	4	1	1	1	1	0	34	679.6	.653
B	4	2	4	1	1	0	1	2	76	560.8	.579
	5	2	4	1	1	0	1	1	52	513.0	.610
	6	2	4	1	1	0	1	0	28	354.3	.557
C	7	2	6	2	1	0	1	2	88	553.4	.496
	8	2	6	2	1	0	1	1	64	514.1	.542
	9	2	6	2	1	0	1	0	40	323.5	.552
D	8*	2	6	2	1	0	1	1	64	514.1	.542
	10	2	6	1	1	0	1	1	60	526.0	.603
	5*	2	4	1	1	0	1	1	52	513.0	.610
E	9*	2	6	2	1	0	1	0	40	323.5	.557
	11	2	6	1	1	0	1	0	36	371.9	.550
	12	2	6	2	0	0	1	0	28	294.1	.568
	13	2	6	1	0	0	1	0	24	346.8	.556
	14	1	6	2	1	0	1	0	17	437.8	.563
	6*	2	4	1	1	0	1	0	28	354.3	.557
F	15	2	6	2	1	0	0	2	88	611.3	-
	16	2	6	2	1	0	0	1	64	558.6	-
	17	2	6	2	1	0	0	0	40	386.0	-

* Repetition of test results for easier comparison.

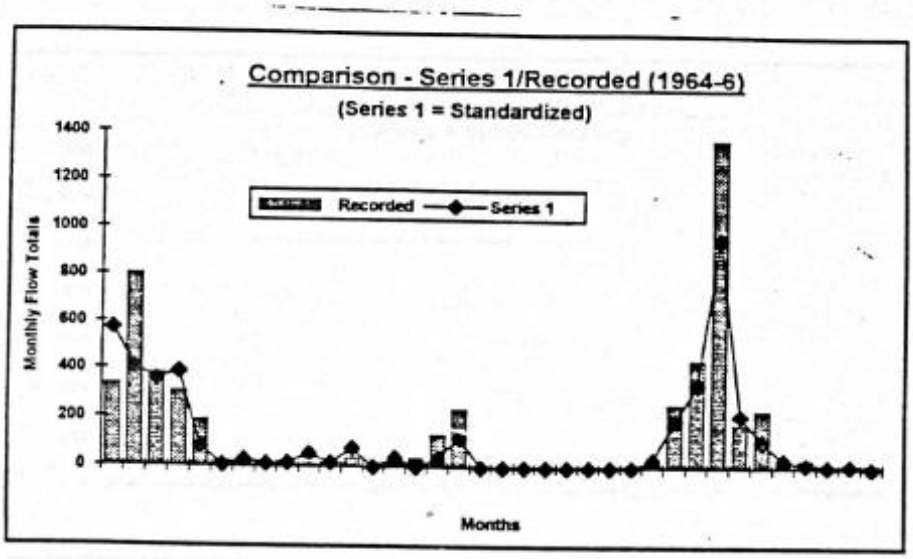


Figure 1

Comparison of Patched with the hidden recorded flows for Vaal Dam during 1964-6. The flows are patched using the model of Trial 15 as specified in Table 4.

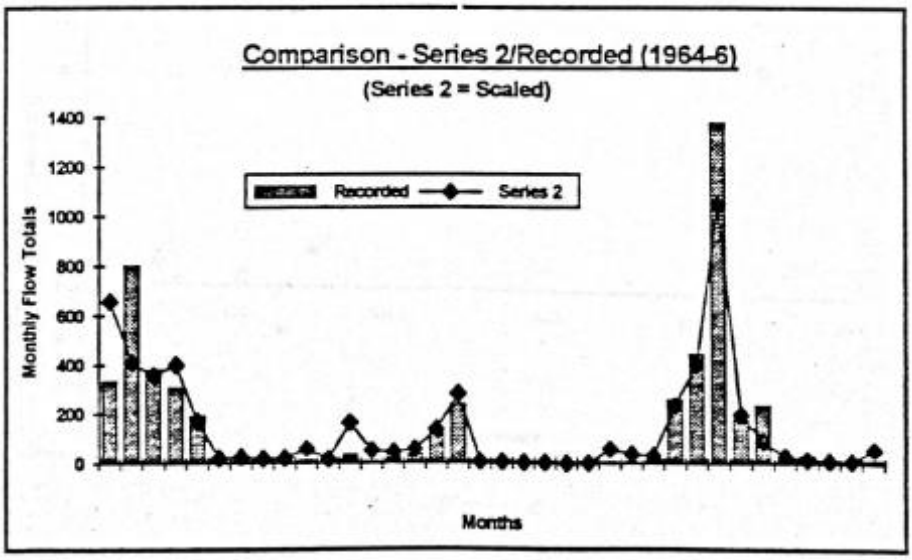


Figure 2

Comparison of Patched with the hidden recorded flows for Vaal Dam during 1964-6. The flows are patched using the model of Trial 16 as specified in Table 4.

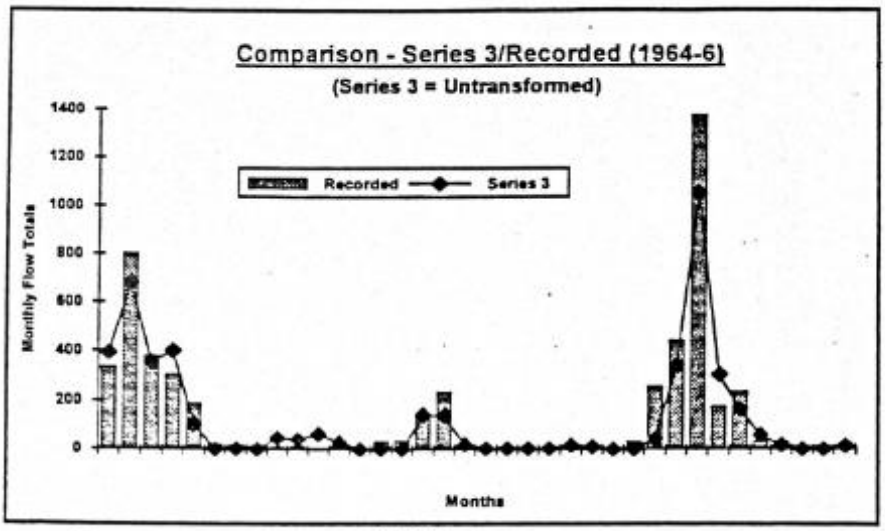


Figure 3

Comparison of Patched with the hidden recorded flows for Vaal Dam during 1964-6. The flows are patched using the model of Trial 17 as specified in Table 4.

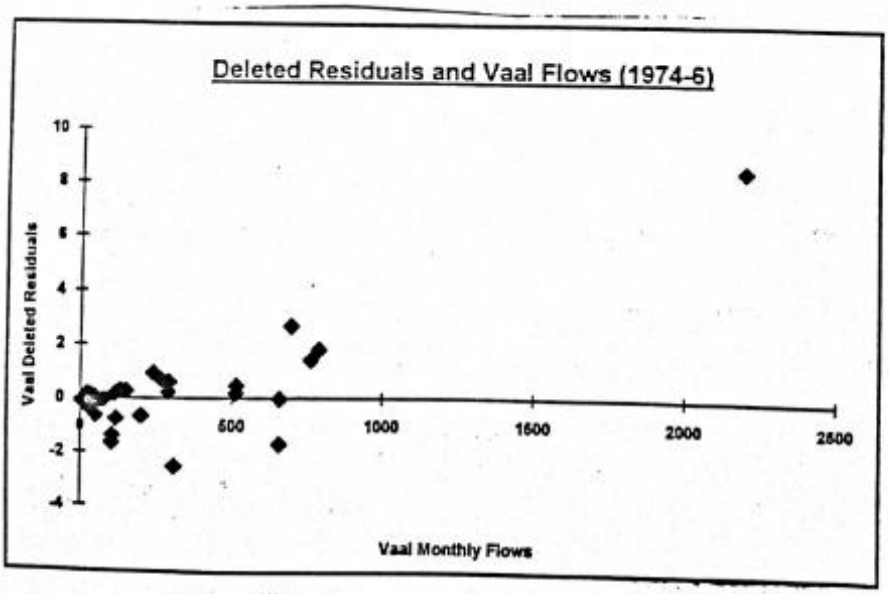


Figure 4

Scatterplot of the deleted residuals and recorded flows for Vaal Dam for the years 1974-6 which include the largest deleted residual. The model was trial 9 as specified in Table 4.

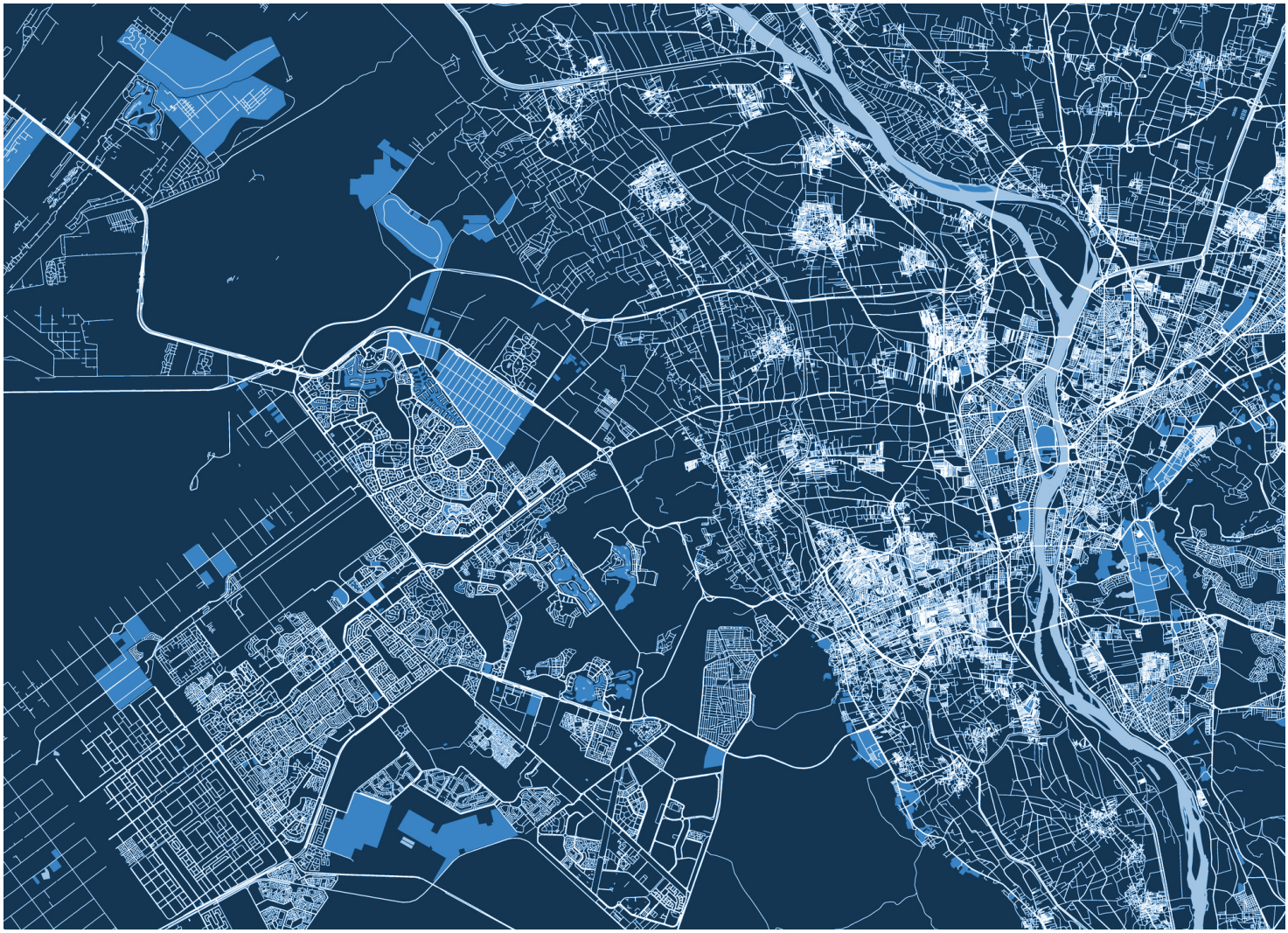


Photo: Shutterstock

ONE RIVER ONE PEOPLE ONE VISION

NILE BASIN INITIATIVE
INITIATIVE DU BASSIN DU NIL

Nile Basin Initiative Secretariat
P.O. Box 192
Entebbe - Uganda
Tel: +256 417 705 000
+256 417 705 117
Email: nbisec@nilebasin.org
Website: http://www.nilebasin.org
Facebook: /Nile Basin Initiative
Twitter: @nbiweb

Eastern Nile Technical Regional Office
Dessie Road
P.O. Box 27173-1000
Addis Ababa - Ethiopia
Tel: +251 116 461 130/32
Fax: +251 116 459 407
Email: entro@nilebasin.org
Website: http://ensap.nilebasin.org

Nile Equatorial Lakes Subsidiary Action Programme Coordination Unit
Kigali City Tower
KCT, KN 2 St, Kigali
P.O. Box 6759, Kigali Rwanda
Tel: +250 788 307 334
Fax: +250 252 580 100
Email: nelsapcu@nilebasin.org
Website: http://nelsap.nilebasin.org

NBI MEMBER STATES

Burundi · DR Congo · Egypt · Ethiopia · Kenya · Rwanda · South Sudan · The Sudan · Tanzania · Uganda

/Nile Basin Initiative @nbiweb

#NileCooperation; #NileBasin; #OneNile