GUIDELINE AND STANDARDIZE OPERATIONAL PROCEDURE ON DATA BLENDING AND INTEGRATION FOR NMA MERGED DATASET

FIRST EDITION

NATIONAL METEOROLOGICAL AGENCY

ADDIS ABABA, ETHIOPIA
DECEMBER, 2018

Guideline and Standardize Operational Procedure on Data Blending and Integration for NMA Merged Dataset

Lead Author: Zekarias Abera (MDC Directorate)
Members:
Yosef Tesfaye (MRSD)
Zekarias Abera (MDCD)
Asmrom Berhana (SNMRC)
Elias Feseha (MRSD)

First Edition: December, 2018

National Meteorological Agency
P. o. box 1090, Tel +251116615779, email: nmsa@ethionet.et
Fax: 251-11-662-5292
December 2018
Addis Ababa, Ethiopia

PREFACE

National Meteorological Agency

Climate information must play a crucial and primary role in national development planning. High quality climate data allows decision makers to better manage risks and maximize opportunities from a changing climate. Decision-relevant information on the past climate, recent trends, likely future trajectories, anomalies and associated impacts is a prerequisite for decision-making at different levels. Availability of decision-relevant information about the past climate, recent trends, likely future trajectories, and associated impacts is thus a prerequisite for climate-informed decision making.

Climate data could thus support a suite of climate-smart solutions able to reinforce development gains and improve the lives of those most vulnerable to climate variability and change. However, climate information is not widely used in Ethiopia to make development decisions. This is mainly because useful information is often not available or, if it does exist, is inaccessible to those that need it most. Currently, the primary source of climate data is observation by ground-based weather stations across the continent. The main strength of these station observations is that they give the true measurements of the climate variable of interest.

The Enhancing National Climate Services (ENACTS) initiative and NMA is an ambitious effort to simultaneously improve the availability, access and use of climate information. This is accomplished by working with National Meteorological to develop high-resolution, spatially and temporally complete gridded historical meteorological datasets; produce suites of derived climate information products; and disseminate them to users. ENACTS enables the NMA to provide enhanced services by overcoming the challenges of data quality, availability and access – while at the same time fostering stakeholder engagement and use. The new data products allow for characterization of climate risks at a local scale and offers opportunities to support applications and research. Meanwhile, your comments and constructive suggestions are highly appreciated to make the objective of this the guide line a success.

Regards,

Fetene Teshome (Director General)

FOREWORD

The NMA blending data set Precipitation and Temperature data is the solution climate data is the combination of satellite raster data vs station data the weather modelling has been hindered by suitable high-resolution surface meteorological datasets. To overcome these limitations, desirable spatial attributes of gridded climate data are combined with desirable temporal attributes of n reanalysis national and daily gauge-based precipitation to derive a spatially and temporally complete, high-resolution (4-km) gridded dataset of surface meteorological variables required in Weather rainfall forecast modelling for the contiguous Ethiopia from 1961 to 2017. The data availability is improved by combining quality-controlled data from the NMA with proxies such as satellite and reanalysis products. Online tools for data analysis, visualization, and engagement make the data available to stakeholders, while ongoing engagement ensures that it will be meaningfully deployed. The IRI's efforts and experience with our agency to date has clearly demonstrated how observations from weather stations in Ethiopia can be used to develop quality-controlled datasets that provide the basis for climate science and information services. The Climate data set is part of a set of agro-climatic analysis products developed by the national meteorological agency and international research institution Columbia university is designed for climatological analysis of rainfall and temperature data was developed by using Climate data tool CDT software. The Climate data tool CDT is provides an array of accessible analysis tools for climate-smart agricultural development. These user-friendly tools can be used to:

- Blend station information with satellite data to create improved datasets
- Analyze seasonal trends and/or historical climate data,
- Analyze drought for a selected region by calculating the standardized precipitation index (SPI),
- Create visual representations of climate data, create scripts (batch files) to quickly and efficiently analyze large quantities of climate data,
- View and/or edit shape files and raster files, and extract statistics from raster datasets to create time series.

Members of Guideline and Standardize Operational Procedure on Data Blending and Integration for NMA Merged Dataset

PURPOSE OF GUIDELINE AND STANDARDIZE OPERATIONAL PROCEDURE

The objective of this guideline is to study a historical climate data provides an opportunity to better understand the nature of climate variability and its impact on outcomes of importance to society such as agriculture, water, health and energy. It can also be used to help understand climate trends and changes in the frequency of extreme events by placing current observations and predicted changes in historical context. Armed with high-quality climate data, decision makers can better understand both how climate has varied in a particular place across seasons, years, or decades and the magnitude and frequency of past extreme events. As a result, they will be better able to contextualize current climate and future climate forecasts. Since its inception over 30 years ago rainfall and temperature data , the Ethiopian National Meteorological Agency (NMA) has served both the government and civil societies in Ethiopia with the objective of providing weather and climate monitoring and prediction products to stakeholder communities and thereby reduce the effects of climate-related catastrophes. The NMA has a large data archive and record of observations the blended data set time temperature both minimum and maximum temperature is 1961-2017 and also the temperature data goes form 1981- 2017. With the data set size both parameter goes 53 GB.

The National Meteorology Agency station networks and distribution of existing is uneven, with most located in central parts of the country which is located cities and towns along major roads most of the time the station distribution concentrate on the central and northern mountain region of the country. As a result, coverage tends to be worse in western and eastern part of the Ethiopia exactly where livelihoods may be most vulnerable to climate variability and climate change. Where station records do exist, they are often of poor quality, with many gaps in the data. Limited technical capacity and inadequate data sharing policies within the national meteorological agencies make it difficult to access the limited data that does exist. As a result, NMA have difficulty providing even the most basic of climate information to decision makers. They note that modernizing infrastructure must coincide with better institutions and regulations, improved services, adequate budgets for operational costs, and integration over Ethiopia. Incremental changes are clearly inadequate for the development of climate information in Ethiopia. However, action on the most critical climate issues cannot wait until all necessary investment has been made. Besides, future investments may not help in filling

gaps in past observations. A new approach to the development of climate information services has been developed - one that takes advantage of the climate data currently available from National Meteorological Agency and integrates them with freely available global products to create higher quality climate information than is currently available. Importantly, users are able to access data and derived information products immediately, without waiting for the necessary improvements in observation networks. However, when these new resources and observational data do become available they can rapidly be absorbed into the integrated data system.

ABBREVIATIONS AND ACRONYMS

CDT	Climate Data Tools
DEM	Digital Elevation Model
IRI	International Research Institution
JRA55	Japan 55 Years Re-analysis
CDT	Climate Data Tools
Meteosat	Meteorological satellite (European)
MODIS	Moderate resolution imaging spectra radiometer
NMA	National Meteorological Agency of Ethiopia
rfe	Rainfall Estimation
TAMSAT	Tropical Application of meteorological using Satellite data and ground based observations
TIR	Thermal InfraRed
WMO	World Meteorological Agency
MDCD	Meteorology Data and Climatology Directorate
MRSD	Meteorology Research and Study Directorate
SNRMC	Somali and neighboring Region Meteorological Center

TABLE OF CONTENTS

1. INTRODUCTION

1.1. Background

Set of connections of ground based meteorological observations are frequently sparse in developing countries. Part of the reason is the lack of resources available in countries which have more pressing economic and social issues. However, these are also the very countries where improved estimates of climate availability are required. Within many developing countries and specifically parts of east Africa, networks of ground based rainfall observations have always been relatively sparse. The situation is not improving and the density of many networks is decreasing, the amount of missing data is increasing and the reliability of some of the data is becoming increasingly expected. Part of the reason for this trend is the lack of commitment to funding meteorological gauging networks in countries that have many more immediate economic issues.

Without an adequate network of weather stations, it is common practice to make use of meteorological models and many socio-economic activities. However, the successful application of such models and other applications is an accurate representation of weather element inputs. It is increasingly obvious that new approaches need to be adopted to provide these inputs as future gauge networks are unlikely to meet the requirements. Satellite for earth observation is challenging due to many dynamic environmental factors such as aerosols, sun glint, clouds, and others (Wang and Bailey, 2001). Thus, the types of satellite have the potential to fill some of the information gaps and provide input data for estimation models. However, there are several practical considerations that need to be addressed if such products are to be used successfully and with confidence: There are a number of references to satellite rainfall and temperature estimation methods and products in the meteorological literature (Grimes et al., 1999 and Thorne et al., 2001). Rain gauge observations are merged with a locally calibrated version of the TAMSAT satellite rainfall estimates to produce over 30-years (1983-todate) of rainfall estimates over Ethiopia at a spatial resolution of 4 by 4 km and a ten-daily time scale. (Tufa et al, 2014).

Climate data are used in a number of applications and availability of climate data, particularly throughout rural Africa, is very limited. Available weather stations are unevenly distributed and mainly located along main roads in cities and towns. Weather station data also suffer from

gaps in the time series. Satellite proxies, particularly satellite and reanalysis data estimate, have been used as alternatives because of their availability even over remote parts. An attempt is made here to improve these problems by combining rainfall and temperature measurements with the complete spatial coverage of satellite and reanalysis estimates. Merged station with satellite and reanalysis is centered on the part of the east of Africa over Ethiopia, and the country has the most complex topography on the continent.

2. Climate Data Library

The data library is a powerful and freely accessible data storehouse and analysis tool that allows a user to view, analyze, and download climate related data through a standard browser. It is a powerful tool that offers the following capabilities at no cost to the user:

- access any number of datasets;
- create analysis of data ranging from simple averaging to more advanced analysis using the Ingrid data analysis language;
- monitor present climate conditions with maps and analyses in the Maproom;
- create visual representations of data and product animations;
- Download data in a variety of commonly used formats.

The maproom is a collection of maps and other figures that monitor climate and societal conditions at present and in the recent past. The maps and figures can be manipulated and are linked to the original data. Even if you are primarily interested in data rather than figures, this is a good place to see which datasets are particularly useful for monitoring current conditions.

2.1. Data Blending and Data Integration

Climate data blending is the process of combining data from multiple sources into a functioning dataset. This process is gaining attention among analysts and systematic group due to the fact that it is a quick and straight forward method used to extract value from multiple data sources. Data blending is a process whereby data from sources are merged into a data warehouse or data set. It concerns not merely the merging of different file formats or disparate sources of data but also different varieties of data.

In addition, the practice of data blending involves taking data from different sources and compiling it into a single useful and standardized data set. It is a major part of strategy in the big data age, as climate work with large and diverse volumes of data to try to define climate intelligence. Data blending takes place in many different ways, but it typically starts with the process of aggregating data from different sources. Experts might segment the process of data blending into three steps: the first step being data acquisition, the second step being the compilation of data, and the third step being the refinement or cleansing of data into a more consistent and accessible end result.

2.2. The Goals of Data Blending and Integration

- Expose a deeper intelligence within your data, by utilizing data from data sources;
- Provide accurate, actionable data in the hands of climate analysts in a spatiotemporal matter;
- Drive better decision-making skills by senior leaders in an Agency;

Data blending has been described as different to data integration due to the requirements of climate data experts to merge sources very quickly for any practical intervention by data scientists.

Data integration involves combining data residing in different sources and providing users with a unified view of them. This process becomes significant in a variety of situation, include scientific domains; for example: combining research results from different bioinformatics repositories. Data integration appears with increasing frequency as the volume that is, big data and the need to share existing data explodes. It has become the focus of extensive theoretical work, and numerous open problems remain unsolved.

2.3. Benefits and Key Steps of to Data Blinding and Integration

The bullets below list some of the benefits obtained by integrating NMA’s datasets:

- Reduction of subjectivity and inconsistencies among datasets that span observing networks and platforms

- Standardized quality control based on reporting time resolution;
- Standardized products are more easily developed and consistent
- Collective experience and expertise leads to a better product
- Conformance of data to documentation, in on format for all users.

The key steps to data blending and Integration:

- Data Acquisition: Identify and obtain access to the data within your sources
- Joining Data : Combine the edited data for further use and analysis
- Data Cleansing: Redesign the data into a usable/functional format and correct/remove any bad data

The agency may have climate different parameters in different dataset. The data blending approach would begin with in taking all of these different data and integrate them with satellite and reanalysis data, merging it into something that can be kept in a single repository.

3. PROCEDURE FOR DATA BLENDING AND INTEGRATION

3.1. Climate Data availability

The availability of detailed well distributed in space information on rainfall and temperature are of essential importance for applied meteorology. Conventional measurements of climate parameters are in a limited number of points represented by synoptic, principal, ordinary and rain gauge stations. This information is not sufficient for correct spatial distribution of climate. The distribution of the ground stations are quite irregular and thus distances between stations could be quite big, sometimes more than 50km. On the other hand, rainfall has high variability in area. The result of spatial distribution depends on the density of the ground measurements.

Rainfall estimated from satellite information includes spatial information that could be used to remodel rainfall field based only on ground stations. This manual presents the results of an application that merges satellite information with conventional ground measurements of the climate for meteorological applications. Meteorological reproduction will be performed with three types of climate field. Reproduction using satellite information, reproduction with real measured and with merged information for rainfall and temperature. Different approaches

have been tested to merge satellite based rainfall and temperature estimates and station observation. These include simple bias adjustment and regression kriging (Hengl *et al.*, 2007). Here, we test both bias adjustment and regression kriging.

1. **Procedure of Regression Kriging/Gridding with RFE for Rainfall Merging**
    1. Prepare measured data from station;
    2. Extract satellite rainfall estimates (rfe) at rain and gauge locations;
    3. Compute the difference between the satellite estimate and rain gauge values at each station location;
    4. Create Regression equation from rain gauge data and RFE
    5. Y=ax + b, where x is REF data and Estimate Y for each 4x4km pixel
    6. Compute Residuals: GG-Y at gauge locations
    7. Interpolate these differences to each grid point (same as satellite pixel centers) using inverse distance weighting; and
    8. Interpolate residuals and add the interpolated differences back to the satellite estimate (Add interpolated residual to Y)
2. **Procedure of Regression Kriging with JRA55/MODIS and DEM for Temperature**
    1. Prepare measured data from station;
    2. Extract MODIS and reanalysis data at temperature gauge locations and Down scale by elevation to 4X4 km;
    3. Error Climatology(81-2010 Bias) by Ration Correction factors(Station data and Down scale JRA55)
    4. Create Regression equation from Temperature data and MODIS/Reanalysis;
    5. Y=ax1+bx2+c, where x1 is MODIS data and x2 is DEM estimate and Y for each 4x4km pixel;
    6. Compute Residuals: TT-Y at Thermometer locations and Interpolate residuals;
    7. Add interpolated residual to Y; Interpolate at each Station and we get 36 image(Factors);
    8. Reanalysis 1961-2018 ;
    9. Station 1961 Jan 1 dekade *1 image out of 36 image and get 4km by 4km resolution readjusted temperature.

Therefore, Merging has been done for three data sets separately. The first data set uses all available stations, while the second uses those with at least 60% of time series complete. The first should produce best product at a given time, while the second for generating a more temporally consistent time series for climate analyses. The third data set is limited to operational stations. There are about more than 90 operational stations, which report on daily basis using communication. The NMA uses the operational stations for monitoring activities, daily and dekadal weather reports. Thus, merging these stations with satellite estimates and reanalysis are for improving NMA's monitoring capabilities. Three daily and dekadal base climate products are generated these are locally calibrated satellite based rainfall estimates, gridded gauge, only product and the combined satellite-gauge rainfall product. Both the satellite and the combined products are available from 1981 to current (Tufa, et al., 2014).

Satellites usually have near global coverage for remote rainfall monitoring and they are especially valuable for regions that lack adequate surface-based measuring techniques. At the same time, satellite rainfall datasets are usually free of charge and their availability is not limited by administration factors. Due to these advantages, in recent years significant developments have been achieved in the field of satellite rainfall estimation. However, satellite rainfall estimates have been produced at rather coarse spatial resolutions. Moreover, satellite rainfall is usually very inaccurate compared with gauge measurements. Thus, the full utilization of satellite rainfall in climate resources management applications has been slowed down.

The satellite product depicts the overall spatial structure of the rainfall field reasonably well. However, the satellite product underestimates rainfall amounts over most parts of the country. Underestimation is more severe over areas of high rainfall amounts. Efforts are underway to improve the availability of climate data over Africa by combining station observations with satellite and other proxies. An example has been provided where historical time series of rainfall data have been generated over Ethiopia by combining satellite rainfall estimates and available rain gauge data. Meteosat TIR data were obtained by the TAMSAT group and calibrated over Ethiopia to generate dekadal satellite based rainfall time series. Although this satellite rainfall product was found to underestimate high rainfall values, it has been shown that this product performs better than or as good as other satellite products that use algorithms

that are more sophisticated. This better performance has been attributed to the use of national rain and temperature gauge data for calibrating the retrieval algorithm. The satellite rainfall estimates were then combined with station measurements. The combined product using regression kriging has been compared to simple bias removal. There was no substantial difference between the two approaches. As a result, simple bias adjustment was used. The merged product was shown to be significantly better than satellite estimate. There is no significant difference between gridded gauge and the merged product over regions with relatively dense station network.

However, the merged product performs much better over data sparse parts of the country. This is a desired result as the main objective of this work has been to improve data availability over regions with few or no meteorological observations. The advantage of the gridded data is that it is not limited by availability of satellite data; thus, it could be used to generate a time series over a much longer period. Regression kriging models the value of a variable at a desired location as the sum of the deterministic and stochastic components (Hengl *et al.*, 2007). The deterministic component is obtained through linear regression on an auxiliary variable and the stochastic components are interpolated residuals. In this, the satellite climate estimates are used as the auxiliary variable. The open source CDT (https://github.com/rijaf-iri)with the R language (http://www.rproject) package was used to implement regression kriging.

3.2. Climate Data Tools (CDT)

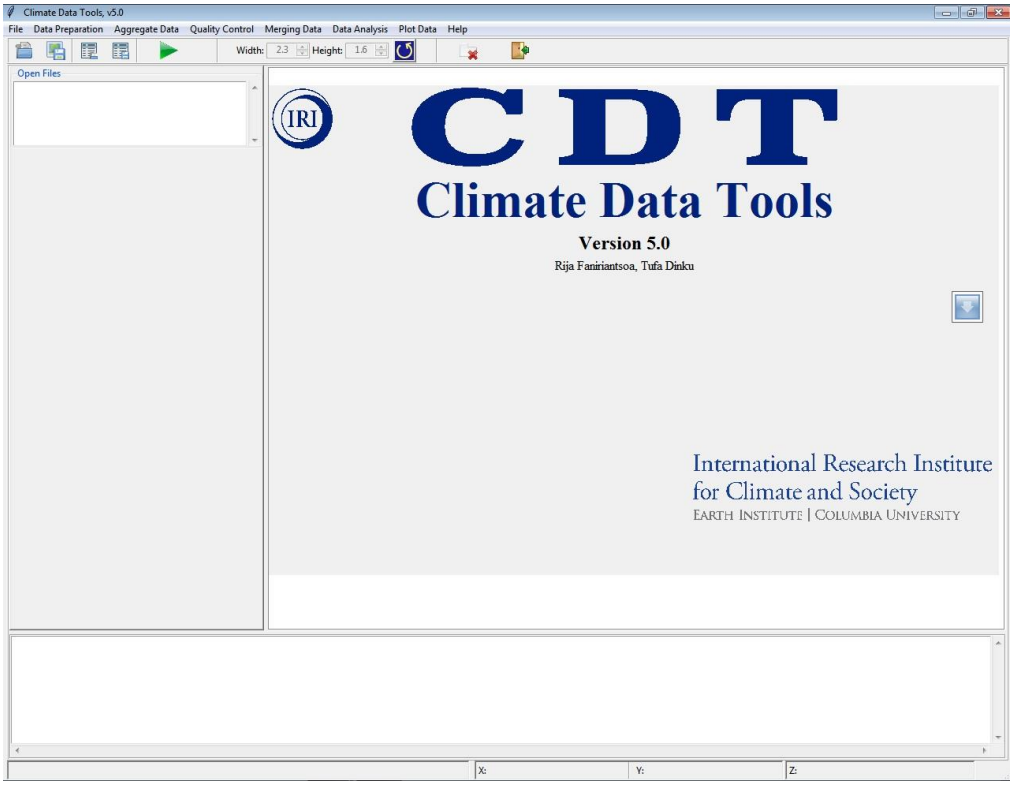
The Climate Data Tools (CDT) is a powerful tool for quality control of station data, merging station data with satellite and others proxies, and processing station and gridded data.CDT was used as key resource for Climate data for gridded data outputs, down manipulation, data extraction, spatial and temporal analysis, rainy dekadal, monthly, season, annual characteristics and climate extremes indices like drought. National Meteorological Agency is the first from African Countries by applying the application and interpretation high resolution map products and reach merged data for different socio-economic activities of the community on Map-Room and using CDT products.

For more than three years, International Research Institute for Climate and Society (IRI) has collaborated with the National meteorological agency of Ethiopia to improve of quality, accessibility and availability of climate data in Ethiopia. Since 2015, several improvements have been made on CDT and new features were added to facilitate the manipulation and visualization of data and generation of merged data. The main purpose of this guideline was to expose to the new features of CDT, in order to strengthen create awareness to conduct a quality control procedure for climate data at the level of their database, and combine ground stations data with global proxies (satellite rainfall estimates data and reanalysis product) on their specific region .

The CDT is a spatial analysis tool designed for climatological analysis of historical rainfall and temperature data. These user friendly tools can be used to obtain and analyze climate data, blend station data with satellite data to create more accurate datasets, analyze seasonal trends and/or historical climate data, analyze drought evapotranspitation water balance for a selected region by calculating Standardized Precipitation Index (SPI), create visual representations of climate data, create R scripts to quickly and efficiently analyze climate data, view and/or edit shapefiles and raster files, and extract statistics from raster datasets to create time series.

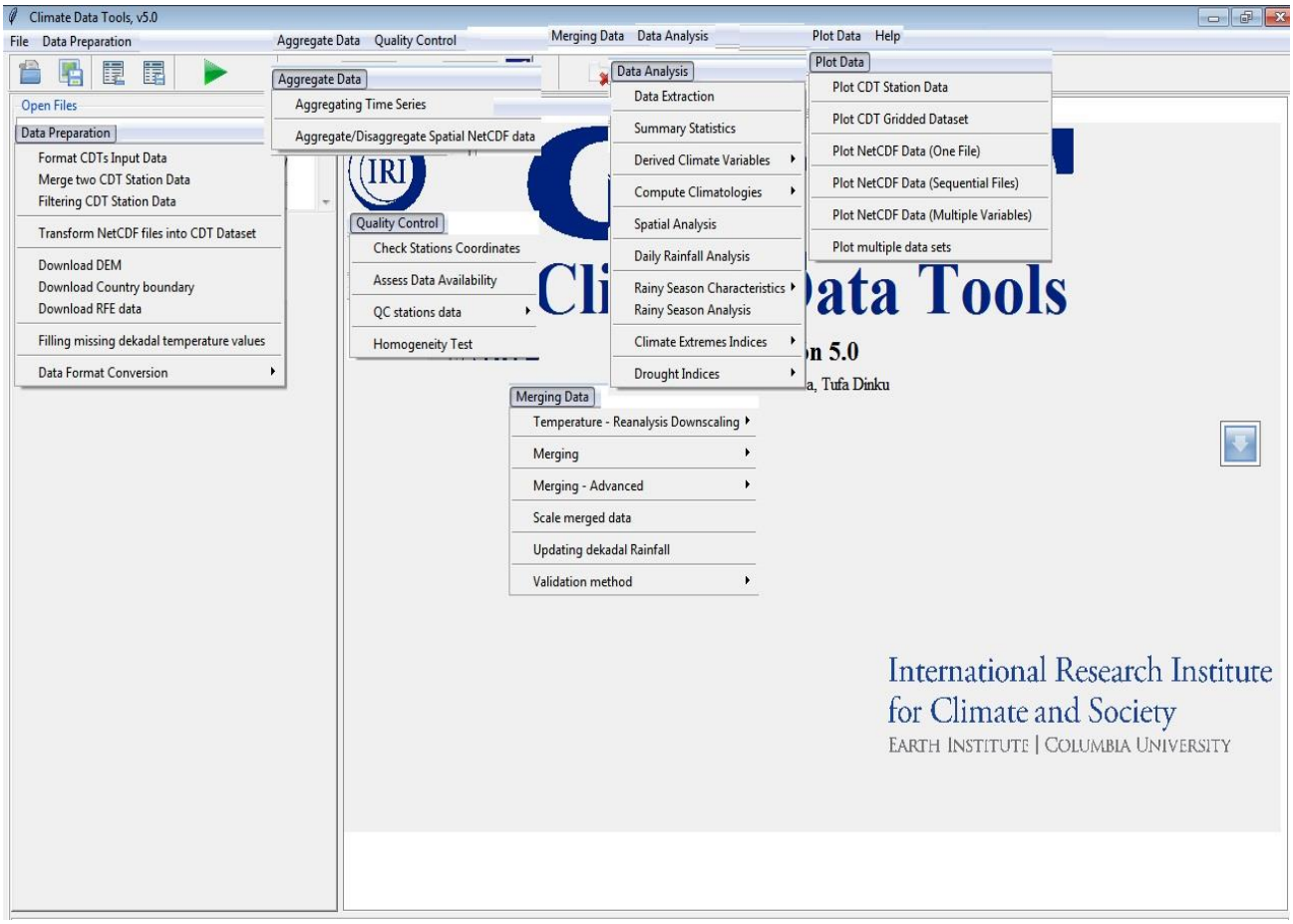
Climate Data Tools is a tool for download rfe and DEM data transforming, quality control, homogeneity test, filling missing observations, interpolation of station data, merging station data with satellite and other proxies, plotting and derived climatological PET and water balance. In the current version CDT v5.0, the major parameters are treated by CDT are precipitation, temperatures and derived past climatological elements. Next versions are expected to include some more climate variables. CDT consists of by functions for processing in-situ data are needed before embarking on the merging itself. To ensure the qualities of the final output, quality control of station data is a must. This includes checking for outliers as well as ensuring the homogeneity of the time series and handling missing values. CDT is a set of scripts running under the R environment. It has a graphical user interface to make it easy and user friendly. To launch the CDT v5.0 application, you must first start R and load the script using the source function of R.

Open R and install “devtools” package with: install. Packages ("devtools"). Now, you can install the development CDT v5.0 from install_github ("rijaf-iri/CDT"). Also, to start the application use # Load CDT library (library (CDT)) and # Starting CDT as startCDT(). This command will launch the main window shown in the figure above.



As shown in the following figures there are seven menus. These are:

1. File,
2. Data preparation,
3. Aggregate data,
4. Quality Control,
5. Merging Data,
6. Data analysis,
7. Plot data and Help menu.



3.3. Climate Data Preparation

Data preparation is the act of preparing or pre-processing raw data sources into refined information assets that can be used effectively for blended and integration purposes, such as past climate analysis. Furthermore, data preparation is necessary to manipulate and transform raw data so that the information content enfolded in the data set can be exposed, or made more easily accessible. This is the first step in data analytics for data wrangling and can include many discrete tasks such as loading data or data ingestion, data fusion, data cleansing, data augmentation, and data delivery. Data cleansing is one of the most common tasks in data preparation. Common data cleansing activities involve ensuring the data is:

- Valid: falls within required constraints (e.g. data has correct data type), matches required patterns, no cross-field issues (e.g. the state/province field only has valid values for the specific country in a country field);

- Complete: ensuring all necessary climate data is available and where possibly, looking up needed data from external sources;
- Consistent: eliminating contradictions in the data;
- Uniform: ensuring common data elements follow common standards in the data, for instance, uniform data/time formats across fields, uniform units of measure;
- Accurate: where possible ensuring data is verifiable with an authoritative source;

Given the variety of data sources (databases) that provide data and formats that data can arrive in, data preparation can be quite involved and complex. There are many tools and technologies that are used for data preparation.

Data preprocessing is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining. The representation and quality of data is first and foremost before running an analysis. Data preparation, filtering, data quality, for instance selection, normalization, transformation, feature extraction and selection steps can take considerable amount of processing time.

- Data quality is the process of detecting, correcting or removing the inaccurate records from data;
- Data normalization is the process used to standardize the range of independent variables or features of data;
- Data transformation is the process of converting data from a format to the new format people expect;
- Feature extraction is the process of transforming the input data into a set of features which can very well represent the input data;
- Data reduction is the transformation of numerical data into a corrected, ordered, and simplified form, minimizing the amount of data or reducing the dimensionality of data.

3.3.1. Data Format

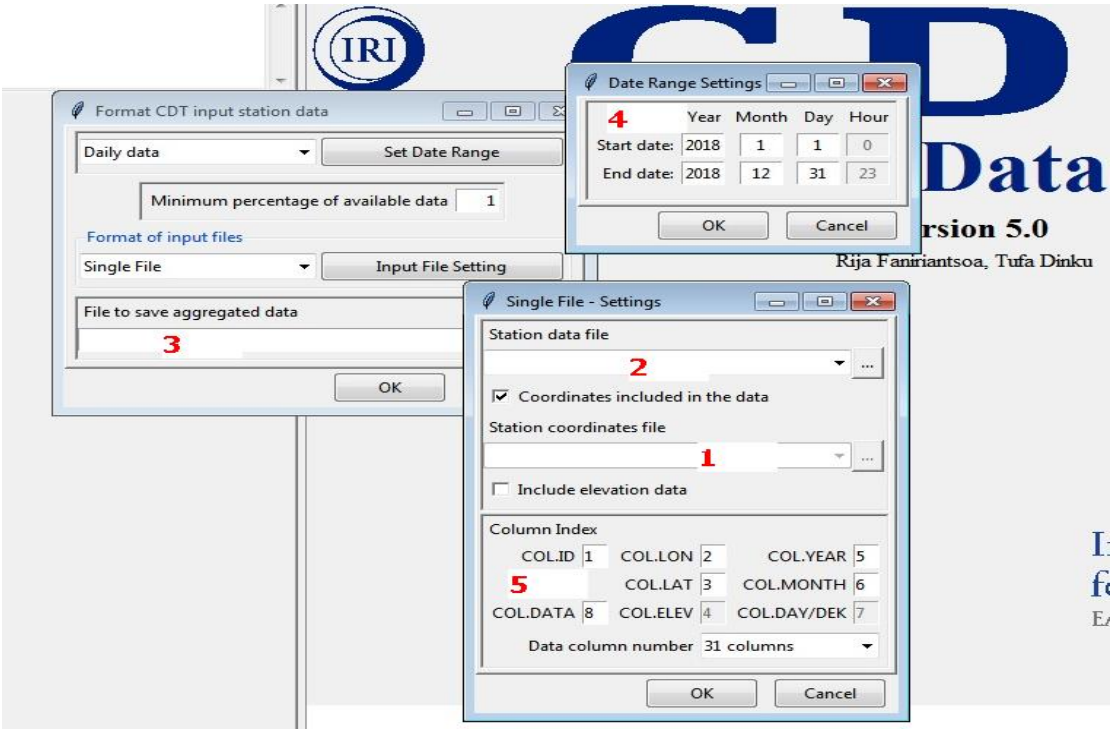
CDT uses one standard data format for all the process, which is shown below. For station daily data format; can do the formatting with other software (e.g. Excel) and then convert to text format, which separations are spaces or tabs. As shown in the table below, from CDT data

format the first line contains the ID of the stations, the second and 3rd lines longitude and latitude. The first column contains dates in the form **yyyymmd (e.g.** 20120101)**,** where **yyyy** is the year, **mm** the month and **d** is the daily or dekade. Missing values are denoted by -99. Input station data files should be in text format (". txt").

GH_ID	YEAR	MO	VAL01	VAL02	VAL03	.....	VAL31
ARABOM11	1988	1	29	25.7	29		
ARABOM11	1988	2	30	31.4	31.4		
ARABOM11	1988	3	32.5	32.6	30		
ARABOM11	1988	4	32.3	33.7	33		
ARABOM11	1988	5	31.5	31.5	30.6		
ARABOM11	1988	6	30.2	29.7	28.5		
ARABOM11	1988	7	30.5	30.6	31.4		
ARABOM11	1988	8	25.5	26.4	26		
ARABOM11	1988	9	27	26	22.3		
.	.	.	.	.	.		
.	.	.	.	.	.		
.	.	.	.	.	.		

Data format of CDT transferred from station data format: CDT has also a function to convert the daily and dekadal observational data to this format. Go to the ***Tools*** menu - ***Format data Temperatures*** and ***Rainfall*** or according to the parameter to be formatted. The following dialog box opens:

GH_ID	LAT	LONG	ELEV	Year	Mon	Dek	1st_Dek	2nd_Dek	3rd_Dek
ARABOM11	8.47	39.83	1630	2017	12	2	-99	0	-99
SHADDI12	8.99	38.79	2330	2017	12	2	-99	0	-99
SHADDI21	9.02	38.75	2386	2017	12	2	-99	0	-99
ARADEL11	7.75	39.90	2480	2017	12	2	-99	0	-99
GOADET11	11.27	37.49	2179	2017	12	2	-99	0	-99
TIADIG11	14.28	39.45	2497	2017	12	2	-99	0	-99
TIADWA11	14.18	38.88	1911	2017	12	2	-99	0	-99
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.



1. Enter the path containing the IDs and coordinates and data of stations or browse the button; the file format is as follows:

GH_ID	LAT	LONG	ELEV	Year	Mon	Dek	1st_Dek	2nd_Dek	3rd_Dek
ARABOM11	8.47	39.83	1630	2017	12	2	-99	0	-99
SHADDI12	8.99	38.79	2330	2017	12	2	-99	0	-99
SHADDI21	9.02	38.75	2386	2017	12	2	-99	0	-99
ARADEL11	7.75	39.90	2480	2017	12	2	-99	0	-99
GOADET11	11.27	37.49	2179	2017	12	2	-99	0	-99
TIADIG11	14.28	39.45	2497	2017	12	2	-99	0	-99
TIADWA11	14.18	38.88	1911	2017	12	2	-99	0	-99
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.

2. Enter the directory and file name containing the daily or dekadal data;
3. Enter the path for saving the output or browse the button;
4. Select the start dates and end date for the input and output data.
5. This section collects information about the files attributes.

- Missing values: Enter the code used in the files to record the missing values (eg -99 or NA), and it is necessary that all the files are consistent.
- Header files stations: check if the files in (2) header. Again, it is necessary that all the files are consistent.
- Coordinates Header files check if the file containing the coordinates has a header.
- Minimum fraction:

After you have entered all the required parameters, you can run the program using the **OK** button. Finally, the output is as the following table in text format.

Stations	BAROBE15	GGARBA15	GOBAHI15	HADIRE15	ILGORE15	**…..**
LON	40.05	37.55783	37.36	41.51	35.533333	
DAILY/LAT	7.133333	6.057167	11.5997	9.36	8.1333	
20120101	22.3	30.6	28.2	31.7	26.8	
20120102	22.2	30.6	27.3	30.6	26.6	
20120103	22.6	30.2	27.2	29.2	26.3	
20120104	22.9	30.2	26.7	29.9	25.4	
20120105	23.9	30.8	27	30.6	26	
20120106	24	31	25.5	29.8	25.3	
20120107	23.4	30.8	27.2	29.7	25.6	
20120108	23	31.1	26.8	31.4	24.9	
20120109	22.9	31.9	28	30.4	25.8	
20120110	22.9	31.8	28.2	30.1	25.7	
20120111	22.3	31.5	27.6	30.8	24.8	
. . .	. . .	. . .	. . .	. . .	. . .	**…..**

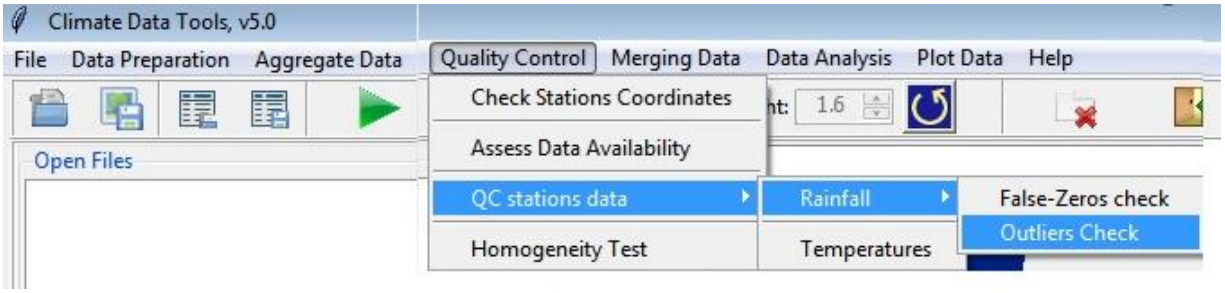
3.3.2. Quality control

The basic concepts for data quality control used in CDT (Climate Data Tools) are focused on outlier detection for the purpose of elimination of data contamination. The most common method used for outlier analysis is standardization of mean square. The reason for standardizing the in fit and outfit mean square statistics is to allow their statistical significance (or p-values). A familiar method to use for this purpose is the Z-score, or standard score (Schulz, 2002).

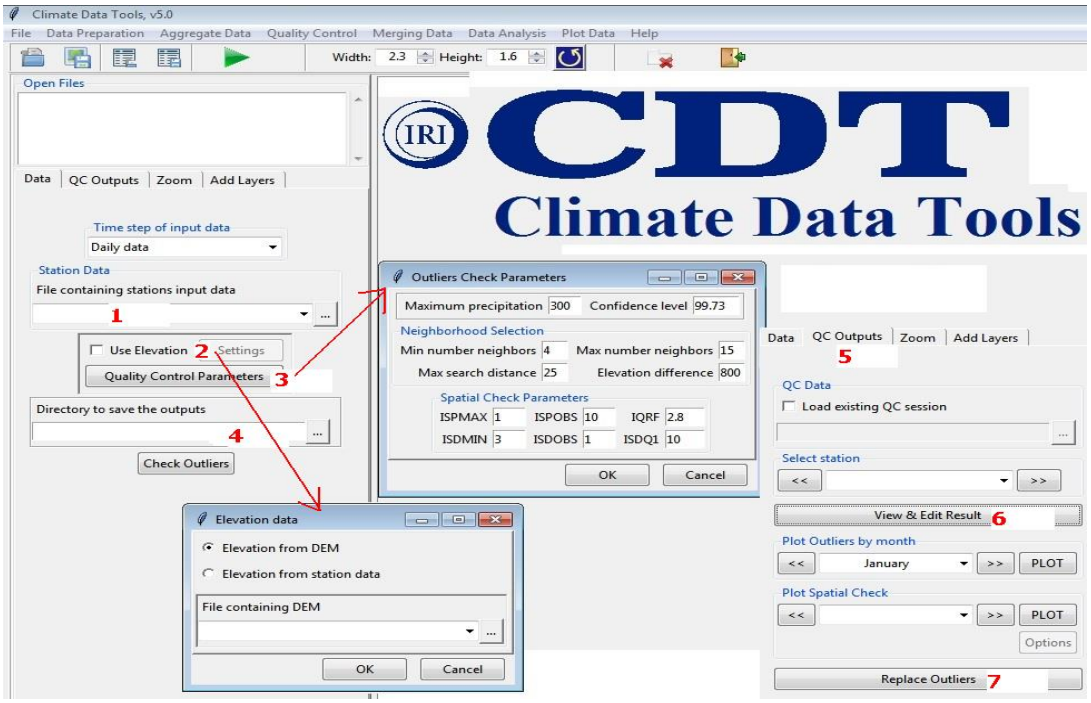
A general formula for converting a variable X, to the standard score Z, is:

$$Z(X) = \frac{X - \mu_X}{\sigma_X} \qquad (1)$$

Before performing the outlier test, negative values and the values greater than 300 mm for daily, 1000 mm for dekadal and 3000 mm for monthly rainfall are removed by replacing by missing values. Then, outliers are detected for each station for each month. In CDT, the values detected as upper outliers are assigned to a flag. The quality control is performed using a regression method and the Fisher test. To start quality control, go to the Quality Control menu then quality control stations data followed by rainfall or temperatures is depending the variable you wish to check. The following dialog box will open:



For instance, select outliers check to do ensure the measured rainfall amount from station. To do this quality control the following dialog box will open:



1. Enter the path of the file containing the station data for quality control. This is the file containing the daily or decadal values in the format specified above.
2. Enter the file name of containing DEM
3. Settings parameters required for quality control

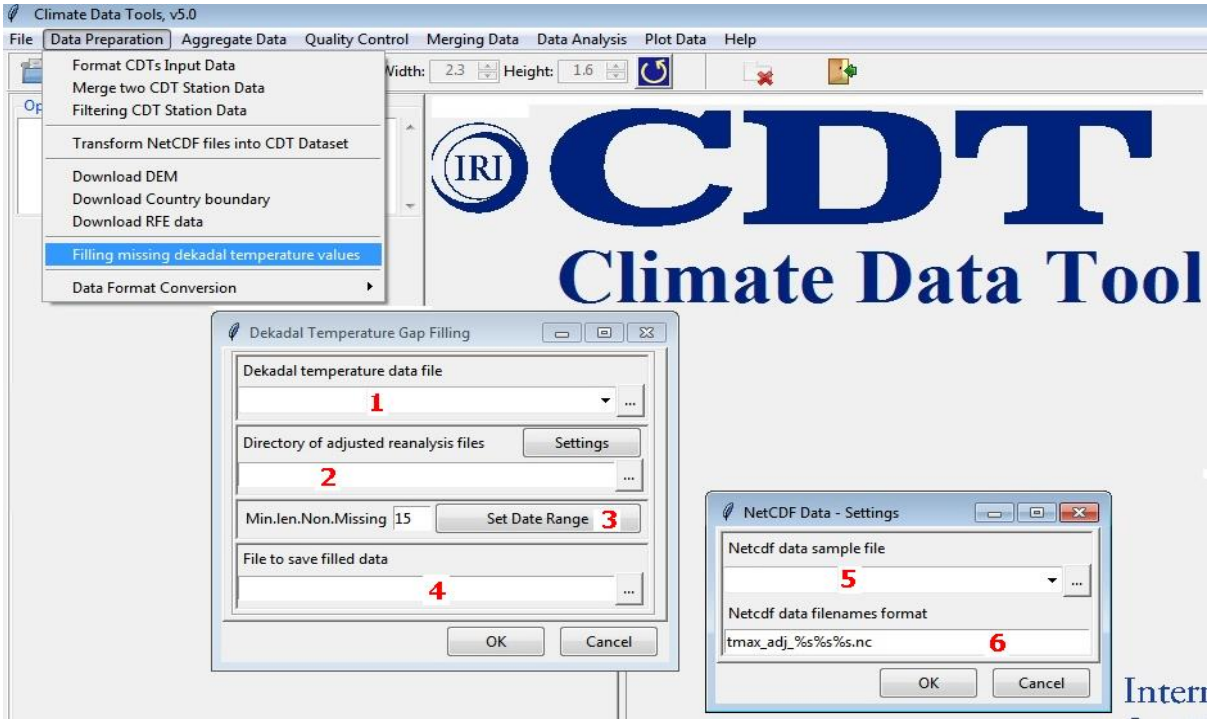
   - Confidence level for test (%) field enters the confidence level for the percentage test.
   - The minimum numbers of stations for the spatial check enter in the field the minimum number of stations to perform the spatial verification.
   - Maximum distance of neighboring stations (km): Enter here the maximum distance for selecting neighboring stations, stations outside this distance will not considered.

4. Enter the file name and path to save the results of quality control.
5. Select quality control outputs station to plot;
6. To view and edit the results;
7. Clicking replace outliers button;

After you have entered all the required parameters, you can run the program using the **Replace Outliers** button. Moreover, do the same techniques for False-Zeros check.

3.3.3. Filling missing Values

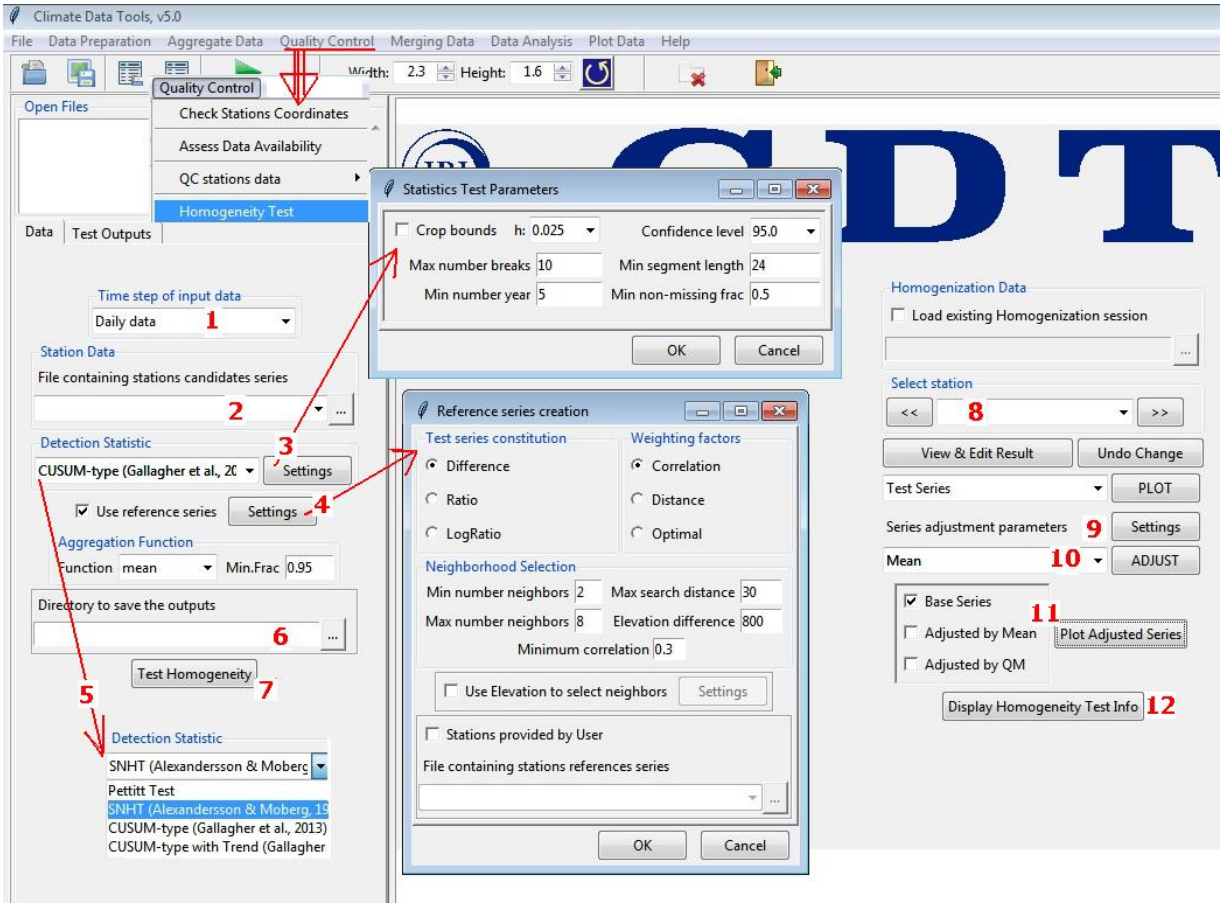
To start filling missing values, go to the Data Preparation menu then filling missing values followed by dekadal temperatures gap filling is depending the variable you wish to check. Filling missing is used to fill missing data for Temperatures. For precipitation, two methods are available; one uses satellite rainfall estimate (RFE) data and the other uses data from nearby stations. The following dialog box will open:



1. Enter here the path to the file containing the observations;
2. Directory containing reanalysis temperature data;
3. Set data range from starting to end;
4. Enter here the file path for saving the results (filled data);
5. Enter the sample file of netCDF;
6. Define the format for the name of netCDF files. For example, maximum temperature "tmax_adj_%s%s%s.nc" where the first %s is for year, second %s is for months and %s for day.

3.3.4. Homogeneity Test

The homogenization method used here is that of RHtests. Go to the Quality Control menu Homogeneity Test and change point detection to detect any breaks in the series. The following dialog box will come into view.



1. Select the time step of input data (Daily, Dekadal, Monthly…);
2. Enter here the path to the file containing the data observations or browsing situ from the button to right of the text box;
3. Select station test parameters i.e. maximum number of breaks, minimum number of year, confidence level and etc. it is set parameters for homogenization procedures;
4. Reference series creation box is used to set parameters to create the reference series;
5. Select the method to test homogeneity from detection statistics text box;
6. Enter here the file path for saving the results;

7. Clicking Test Homogeneity button and open the test outputs dialog box will come into view;
8. Select station to view and edit the output result;
9. Set the series adjustment parameters: minimum adjustment in month and segment to adjust;
10. Select none or mean or quartile matching;
11. Click plot adjust series button;
12. Click display homogeneity test info button to finalize the test

After execution, directories with ID stations are created in the directory you selected above.

3.4. Data Blending (Merging) over a long period and Updating rainfall Data

3.4.1. Merging data over a long period

To perform the merging over a long period, go to the menu Merging Data - Merging then Rainfall and Temperatures by the variable to merge. Then, next select Merging Rainfall and temperature. The following dialog box for merging rainfall will come into view. The procedures are the same as merging with RFE or MODIS/Reanalysis, but instead of choosing a single date, choose the date of beginning and conclusion of the merger.

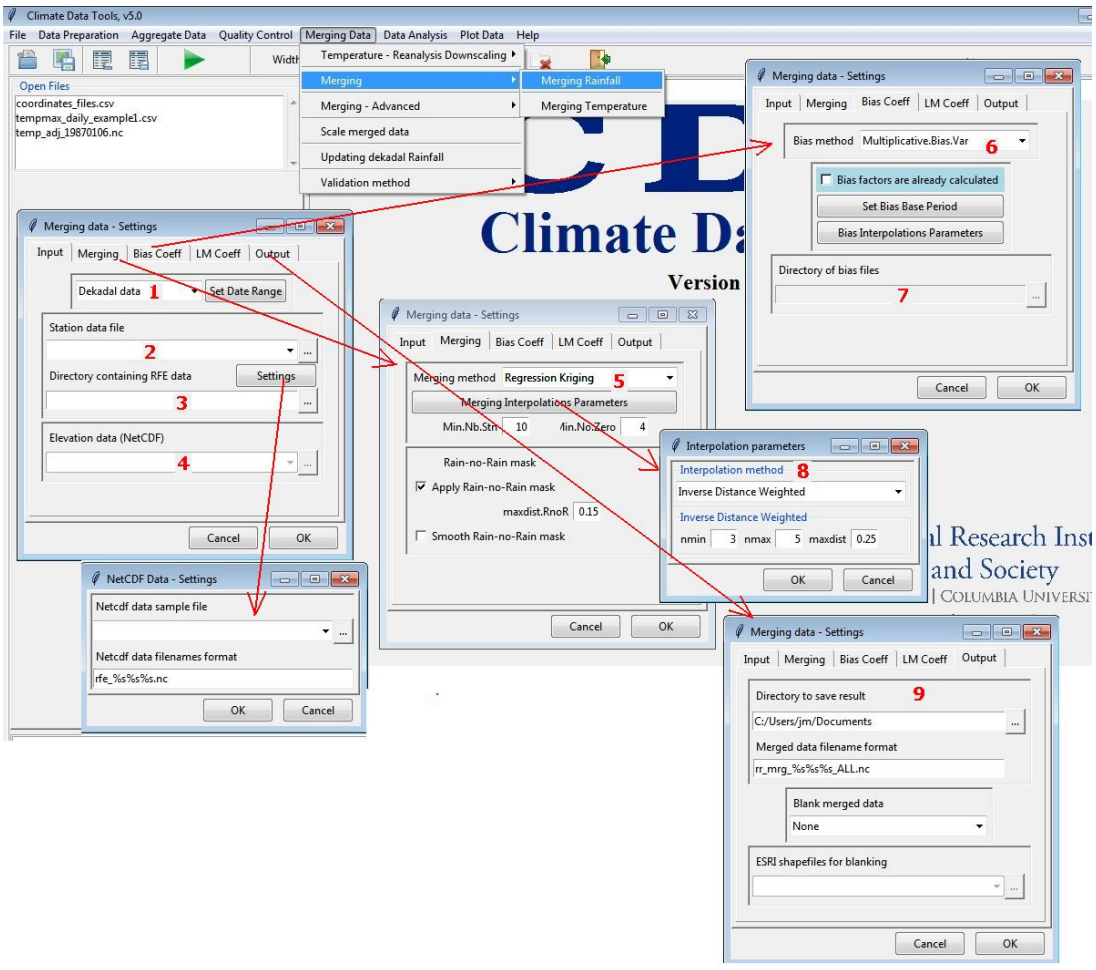
Basis of the TAMSAT approach

The physical basis of the TAMSAT method is that rainfall in Africa primarily riginates from deep convective clouds that penetrate into the upper levels of the troposphere and thus have cold cloud tops. The method therefore assumes that cloud tops colder than a certain threshold temperature produce rain, while cloud tops warmer than this threshold don't rain. From this assumption it also follows that the longer the time the cloud is below this rain/no rain temperature threshold, the more rain it produces. Thus, the length of time a cloud is below the temperature threshold (cold cloud duration; CCD) can be linearly related tothe rainfall amount, R using:

$$R = \begin{cases} a_0 + a_1 CCD & CCD > 0 \\ 0 & CCD = 0 \end{cases}$$

Where, R is in mm and a0 and a1are the calibration parameters. The optimal threshold temperature and calibration parameters in this linear relationship are found by calibrating CCD derived from geostationary TIR measurements to a historic rain gauge archive (Tarnavsky et al., 2014). It is important to note that the calibration is conducted for dekadal rainfall and CCD measurements. However, TAMSAT also produces daily rainfall estimates by temporally disaggregating the dekadal rainfall estimates using the daily CCD (Maidment et al., 2012).Because convective rainfall characteristics over Africa have marked seasonal and regional variability, the optimal temperature threshold and CCD-rainfall relationship described above varies both spatially and temporally.

Tropical Applications of Meteorology Using Satellite Data and Ground-Based Observations (TAMSAT) rainfall monitoring products have been extended to provide spatially contiguous rainfall estimates across Africa. This has been achieved through a new, climatology-based calibration, which varies in both space and time. As a result, cumulative estimates of rainfall are now issued at the end of each 10-day period (dekade) at 4-km spatial resolution with pan-African coverage. The utility of the products for decision making is improved by the routine provision of validation reports, for which the 10-day (dekadal) TAMSAT rainfall estimates are compared with independent gauge observations. This paper describes the methodology by which the TAMSAT method has been applied to generate the pan-African rainfall monitoring products. It is demonstrated through comparison with gauge measurements that the method provides skillful estimates, although with a systematic dry bias. This study illustrates TAMSAT's value as a complementary method of estimating rainfall through examples of successful operational application.



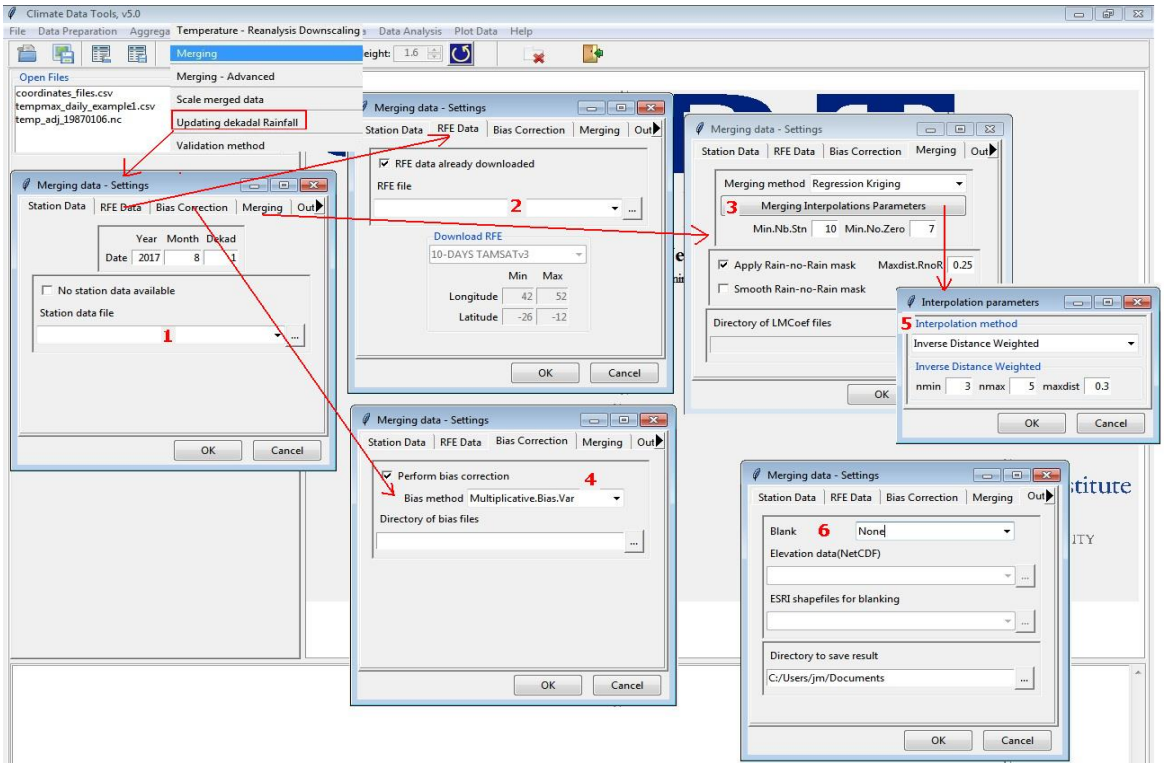
1. Select the time step of input data (Daily, Dekadal, Monthly…);
2. Enter here the path to the file containing the data observations;
3. Enter here the path to the file containing Rfe;
4. Enter here the path to the file containing DEM;
5. Select merging method;
6. Enter bias coefficient method;
7. Enter here the path to the file containing bias coefficient files;
8. Select the merging interpolation parameters methods;
    1. Number Min: Minimum number of stations used to interpolate a point. If the number of stations reached;
    2. Max Number: Maximum number of stations used to interpolate a point. The interpolation is performed with the number of the nearest stations.
    3. Max Distance: Maximum distance (in km) use.

9. Merging data output setting; directory to save the result and file name formats;

The results are saved in the folder you selected in (9). For temperatures, go to the menu Merging Data and Merging next select Merging Rainfall and temperature. The dialog box for merging temperature will come into view and the same as the rainfall merging working procedure.

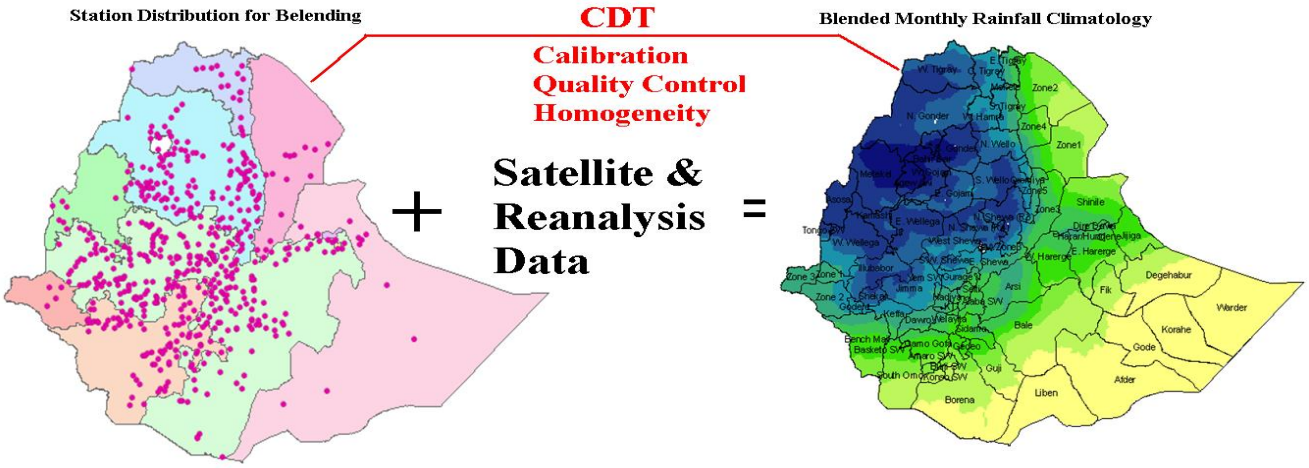
3.4.2. Updating Rainfall Data

Climate information and statistics, based on many types of surface measurements, are produced on local gridded points to national scales. Climate models are used to attribute causes of past climate change that are seen within the observations. The updating rainfall data set is limited to operational stations. There are about 90 operational stations, which report on daily basis by means of communication. Accordingly, NMA uses the operational stations for monitoring activities. Thus, merging these stations with satellite estimates is for improving NMA's monitoring capabilities. To do updating go to the menu Merging Data then updating dekadal Rainfall, then, next select Merging Rainfall and temperature. The following dialog box will come into view.



1. Enter here the path to the file containing the data observations;
2. Enter here the path to the file containing Rfe;
3. Select the merging interpolation parameters methods;
4. Select and enter directory of bias coefficient method files;
5. Select the merging interpolation parameters methods and set the Number Min, the Max Number and the Max Distance
6. Select blank option i.e. none or use DEM or use ESRI shape file and lastly, enter here the directory path to save the result.

To end with, station distribution does not satisfy WMO's minimum requirements even during the best years and in our countries stations are unevenly distributed with most stations located in cities and towns along the main roads. Therefore, as the result, major outputs of the merged data are over 30/50-years of climate time series for every 4km grid across each country and Now data available where there are no stations. Then obtain out product of 4km by 4km merged data at 68,000 gridded points over Ethiopia in numerical values.



As there are no satellite temperature estimates going back 30 years, reanalysis data are used as a proxy. Reanalysis products are climate data generated by systematically combining climate observations (analyses) with climate model forecasts using data assimilation schemes and climate models. The Japanese 55-year Reanalysis (JRA55)4 is used to generate a gridded temperature time series for the period 1961-2016. This product has a coarse spatial resolution of about 50km. Thus, the reanalysis data are downscaled to 4km spatial resolution using station observations and elevation maps The approach thus combines the spatial information from the proxies with the accuracy from point station measurements. The final products are datasets with 30 or more years of rainfall and temperature time-series data at a daily and dekadal time scale for every 4km grid across a country. While the quality of the final products is inevitably dependent on the number, spatial distribution, and quality of the station observations, the result is a great improvement on what was previously. With rainfall, satellite estimates are available for use as a reliable proxy.

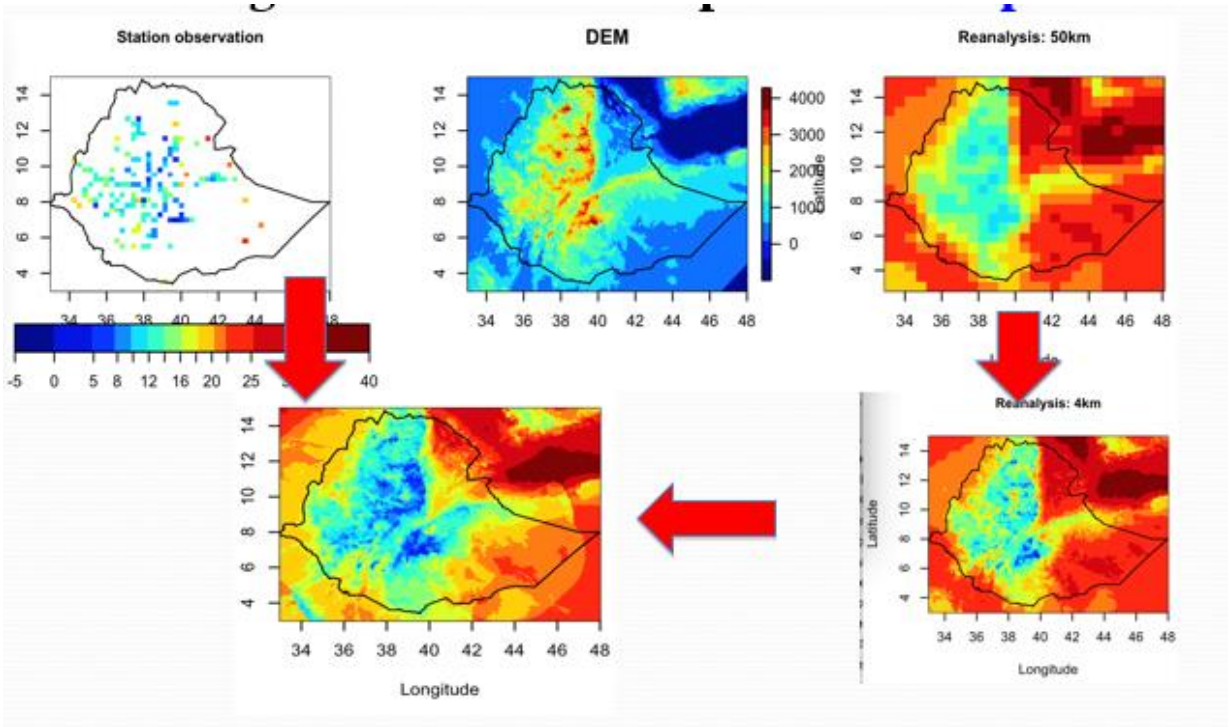


Figure 3: Station measurements of minimum temperature (top left), JRA55 reanalysis data at its original resolution of 50km (top right), bias corrected and reanalysis downscaled to 4km (bottom left) and merged station-reanalysis (bottom right) over Ethiopia.

However, there are no reliable temperature estimates that go back 30 years; as a result, climate model reanalysis products are used as the proxy. One of the challenges in using reanalysis products is their coarse spatial resolution. Thus, they need to be downscaled to 4km resolution. Figure provides an example of a combined minimum temperature map along with the individual station inputs from Ethiopia. Even the raw coarse resolution reanalysis field exhibits a spatial structure similar to that of station measurements, except that the reanalysis overestimates temperature values over some parts of the country. Downscaling and merging with station observations improve the data progressively. The merged product is very close to station measurements and provides high-quality data even where there is no station or data around.

4. REFERENCE

Grimes et al., 1999 D.I.F. Grimes, E. Pardo Igúzquiza and R. Bonifacio, Optimal areal rainfall estimation using rain gauges and satellite data, J. Hydrology **222** (1999), pp. 93–108.

Hengl T, Heuvelink GBM, Rossiter DG. 2007. About regression kriging: From equations to case studies. Comput. Geosci. 33:1301–1315.

http://www.rproject

International Research Institute. https://github.com/rijaf-iri.

Müller, H., 2007: Bayesian transgaussian kriging. Proc. 15th European

Thorne et al., 2001 V. Thorne, P. Coakley, D. Grimes and G. Dugdale, Comparsion of TAMSAT and CPC rainfall estimates with rainfall, for southern Africa, Int. J. Remote Sens. 22 (2001) (10), pp. 1951–1974.

Tufa Dinku. 2016. Enhancing National Climate Service initiative (ENACTS). International Research Institute for Climate and Society (IRI). Columbia

Tufa Dinku, Kinfe Hailemariam, Ross Maidment, Elena Tarnavsky and Stephen Connor.2014. Combined use of satellite estimates and rain gauge observations to generate high-quality historical rainfall time series over Ethiopia. Royal Meteorological Society.UK.

Wang, M., Bailey, S.W.,2001. Correction of sun glint contamination on the Sea WiFS ocean and atmosphere products. Applied. Opt. 40 (27), 4790e4798.

Young Statisticians Meeting, Castro Urdiales, Spain,Bernoulli Society, 5 pp.