THE POTENTIAL FOR RAINWATER HARVESTING

BY

FOR REFERENCE
ONLY

BENEDICTO KAZUZURU

SUBMITTED FOR THE DEGREE OF MASTER OF SCIENCE IN BIOMETRY,
SCHOOL OF APPLIED STATISTICS, UNIVERSITY OF READING
AUGUST 31 2004

DEDICATION

To my Son Yalahita Kazuzuru and my wife Elizabeth Richard Kazuzuru

for their good company

ACKNOWLEDGEMENTS

I would like to thank all my supervisors for their help in my project. These were

Roger stern , Fiona Wood and Ian Dale

I also thank the staff of NRSP for their support and for making available all the materials and reports relating to PT

My great appreciation should go to the Association of Commonwealth Scholarship for sponsoring me through one year period in the M.Sc course

Finally I would like to express my sincere gratitude to the Sokoine University of Agriculture, in particular, the office of the Deputy Vice Chancellor for nominating me for the Scholarship of which I had been eagerly looking for.

ABSTRACT

Rainwater harvest has become one of the strategies by the Tanzanian government in fighting the persistent drought in the country. The government awareness of the importance and use of the rainwater harvest has been facilitated by the joint team of experts from the Sokoine University of Agriculture and the University of Newcastle working to develop a model of the rainwater harvest (RWH) process as an aid to identifying best–bet options

The joint team of experts from the two universities has developed a computer program to support their work on rainwater harvest. This is called PARCHED-THIRST, which is an acronym for, **P**redicting **A**rable **R**esource **C**apture in **H**ostile **E**nvironments **D**uring **T**he **H**arvesting of **I**ncident **R**ainfall in **S**emi-arid **T**ropics). Among its key functions, PARCHED THIRST simulates agrometeorological variables for use as inputs in the simulation of crop growth. This project assesses the simulation of climatic variables, rainfall and temperature in particular. The project looks at the, flexibility of the PT software and the methodology used in the simulation.

The data used for analysis is from Arusha (1970-2000) in Tanzania, which has been supplied within the software. This site has longest records of climatic data among the sites whose data is supplied within the software.

An exploration of the data indicates that, the PT models the means (monthly, weekly, yearly) effectively but underestimate the spread of the data. The analysis of the rainfall data indicates that higher order Markov chains are needed at the site in Tanzania as this leads to an improvement in modelling the spread of the data. For the temperature data, the assumption of first order autoregressive process was found to be valid, but the assumed stationerity of the correlations did not hold.

The flexibility of the PT software has generally been found to be good, following updates by the PT team at Sokoine University over the last three years. However the software needs to be further modified to suit the varied needs of its intended customers.

Table of Content

CHAPTER 1: INTRODUCTION TO THE PROJECT

1.1 Background to the project

There has been quite a number of ways to target poverty alleviation campaigns in developing countries, basing primarily on major means of economic production in those countries. Tanzania being one among the African developing countries has its economy almost entirely relying on agricultural production for the goods it supplies to both domestic and foreign markets. Hence, for most of the time, poverty eradication policies in the country have been targeting the agricultural sector, either directly or indirectly.

The agriculture sector accounts for half of the country's GDP, provides 85% of exports, and employs 80% of the work force. The major staple crops for food consumption include millet, sorghum, maize, rice and cassava, which support the majority of people in both the rural and urban areas. As for the exports crops, the country relies mainly on tea, coffee, cotton, cashew nuts, tobacco and sisal. However, the agricultural sector in Tanzania, just as it is for most of the African sub-Saharan countries is as yet undeveloped, as it depends mainly on rainwater, which is unreliable in some parts of the country. Apparently no more than 4% of the land area is used for crop production owing to unfavourable topographic and climatic conditions. (CIA world fact book, Tanzania 2004)

As a result of the aforesaid, there has been concern on the part of the Tanzanian government to address the poverty issue by looking into ways through which availability of water to the crops can be improved. As irrigation schemes are hard to manage and quite expensive, there remains only one option especially in those areas with no major water bodies and that is **intensive** rainwater harvesting. In one of the Hansard records of July 2001, the Tanzanian prime minister was quoted saying, " the government will strengthen and promote the use of rainwater harvesting technology in both rural and urban areas ". And in another Hansard records of 18 June 2001, one of the Tanzanian members of parliament was quoted saying, " we must do away with a notion that, the drought we face from time to time in many parts of the country are caused by the shortage of rainfall. With good programmes of harvesting rainwater we can avoid droughts even in times or places considered to have low rainfall". All these statements from the country's highest organs indicate their growing awareness and concern about the link between rainwater harvesting, agriculture and poverty eradication in the country.

The concern by the Tanzanian government is undoubtedly reasonable, taking into consideration of the expensiveness of irrigation schemes and the fact that, the dry areas of Tanzania, which are found in the eastern arm of the Great Rift Valley, are somewhat far from major water bodies. The regions that are mainly found in these drylands areas include Dodoma and Singida in the central part of the country, Tabora, Shinyanga and Mwanza in the central north of the country, Arusha and Kilimanjaro in the north east of the county, Dar es Salaam, Zanzibar and Morogoro in the eastern part of the country, Lindi and Mtwara in the southern part of the country. (See appendix 1) Although there is a regular dry season in these regions, there is always a reasonable amount of rainfall to be harvested from the seasonal rainfall. Recognising this, the Tanzanian government has formally incorporated the rainwater harvesting policy as one among its policies to address the persistent drought in those regions. (NRSP News of September 2002)

The growing awareness by the Tanzanian government has further been strengthened and activated by the team of experts at the Sokoine University of Agriculture (SUA) working on soil water management. The group, SWMRG (Soil Water Management Research Group) has conducted research on rainwater harvesting since 1992, with the support of the World Bank via ODA (UK Overseas Development Administration), then DFID Financed Resource Assessment and Farming Systems (RAFS) research programme. (Young 1996)

SWMRG has been working jointly with a team of experts from University of Newcastle and has undertaken extensive research and disseminated a number of findings on rainwater harvesting in Tanzania. Explaining the role of SWMRG in promoting the rainwater harvest to a level of a concern by the Tanzanian government, Hatibu et al., (1996) says, " Technical findings were disseminated through a special issue of the Tanzanian journal of agricultural sciences devoted to rainwater harvesting and the production of a planning guide hand book on rainwater harvesting. Books and pamphlets were also produced in the national language (Swahili) for use by the extension staff and farmers. Training programmes were organised for government extension staff and NGOs who work directly with farmers. All of these activities have played an important role in raising awareness among policy makers."

The project came as a result of the review of DFID's Renewable Natural Resources Research Strategy (RNRRS) in 1993/4, where RAFS was revised and extended to create the new Natural Resources Systems Programme (NRSP). Since then the project fell within the NRSP semi-arid production Systems portfolio (Young 1996). Clarifying how it fell within the hands of SUA and Newcastle, (Young 1996) says, " At about this time, following a World bank assisted review of agricultural research strategy within Tanzania, a National Agricultural Research Master plan was prepared (MOA, 1991). This plan identifies 'soil and water management' as top priority and accordingly Sokoine University of Agriculture (SUA) was invited to provide leadership in establishing a programme of research in soil and water management. The current project was therefore formulated as a collaborative venture between SUA and University of Newcastle."

The group's rainwater harvest project had two main objectives. The first was to "demonstrate viable cropping systems based upon rainwater harvesting techniques" and the second was to "develop a model of the rainwater harvest (RWH) process as an aid to identifying best–bet options" (Gowing 1996) .The joint team of experts from the two universities has gone through the process of developing a computer program to meet the above objectives. This is called PARCHED-THIRST, which is an acronym for **P**redicting **A**rable **R**esource **C**apture in **H**ostile **E**nvironments **D**uring **T**he **H**arvesting of **I**ncident **R**ainfall in **S**emi-arid **T**ropics). We shall refer to PARCHED THIRST as simply **PT.**

In one of their technical reports (Young et al., 1996), the team noted that there is potential for an increase in production during the *Vuli* season (a local name for the rainy season of October to December in Tanzania) due to introduction of viable RWH, whereas there is generally little increase during the *Masika* season (a local name for the main rainy season in Tanzania, which is usually from January to April). The project is now at the final stage, where extension staff and other potential users are being trained on how they can educate the farmers on adoption of a specific rainwater harvesting system for a particular cropping system.

The PT program is a crop simulation model, which incorporates a soil-moisture model, a water runoff model, and a model for simulation of agro meteorological variables, which are rainfall, humidity, temperature, wind, evaporation and radiation.

Although the PT is an established project, with considerable validation efforts by the project team, it has drawn the attention of a number of people, including academicians from various disciplines such as biometricians, agronomist, water engineers, meteorologists and many others. For example, an agronomist would like to know how PT works in comparison with other known crop models. What specific advantages does it have as opposed to other crop models? Does it thoroughly deal with all the limitations encountered in crop models, among which is the problem of accurately measuring the model inputs such as pests and water management (Xinyou Yin et al., 2003).

The University of Reading, through the School of Applied Statistics has also taken keen interest in assessing the performance of the PT programme and has sought to address the relevant features of the programme by suggesting this as a topic for a Biometry MSc thesis. Some key statistical issues that have been raised concern the functionality and the set-up of the whole of the PT project. We have identified three statistical issues of concern in PT as explained here.

- The model permits climatic data to be simulated at the sites where there is none or only partial real climatic data, through the climatic generator, a component of the model. Basically, the climatic generator simulates weather data based on the statistical properties of the historical data. The simulation model relies heavily on a paper written in 1982 by Richardson. To what extent does the model produces realistic data, and can improvements be suggested to the model?

- The model estimates rainfall intensity, which is crucially needed in the runoff model. The reason for such estimation is the fact that, in most areas in Tanzania, there has been no record for short time rainfall intensity (i.e. within five minutes) despite having daily rainfall records. As a consequence, the rainfall intensity in the model is estimated based on the total daily rainfall amount and the duration of the rainfall. Three questions are: how does it work in the model, how sensitive are the conclusions to the different assumptions that are made, and can improvements be suggested?

- One advantage of using the simulation model, rather than real experiments is that results can quickly be obtained for a long sequence of years, say 100 years. Usually there will be many aspects that the farmer can vary, such as the crop grown, the planting date, the strategy of weeding and so on. Designing a simulation experiment is like designing a real experiment, (where there are many factors that can be varied) except the results come more quickly. To what extent does the software encourage a well-designed experimental strategy in the simulation? How can the simulation experiments be designed to achieve good precision? How easily can risks as opposed to just comparing means, be estimated and compared for different strategies?

This project, gives an overview on how the three statistical issues have been addressed in the model. It will look critically and in more detail at the first aspect, that is, how realistic is the generated climatic data.

However, to understand **logically** the processes involved in the PT model, including climatic data simulation, it is important first to understand the physiology of a crop growth and its key requirements as may be reflected in any crop model. As initially said, rainwater is one of the chief requirements for any crop growth, and for this reason, discussions on how to manipulate it for crop production forms the cornerstone of the PT model. In the next section, crop growth in general is discussed, where water requirement by a plant is at the centre of the topic.

1.2 The process of plant growth.

Plant growth is the process by which a plant increases in number and size of leaves and stems. The growth of both plants and animals requires energy. Animals get their energy by digesting the plants they eat. Plants get their energy from the sun through photosynthesis. Photosynthesis is the process where the green pigment in the plant's leaf (chlorophyll) absorbs energy from sunlight and, using this energy, water, and carbon dioxide, produces oxygen and simple sugars. The plant then uses these sugars to make more complex sugars and starches for storage as energy reserves, to make cellulose and hemicellulose for cell walls or, with nitrogen, to make proteins. This process is supplemented by nutrients from the soil such as nitrogen, phosphorus, calcium, magnesium, potassium, and sulphur. How the plant uses its energy depends on the developmental stage of the plant and on environmental conditions.

From this explanation of plant growth it is obvious that the daily materials needed for plant growth consist of light, carbon dioxide, water and mineral nutrients. Carbon dioxide is obtained from the atmosphere through the stomata of the leaves, while water and mineral nutrients, enter the plant through the medium of the roots The PT model is constructed based on those main aspects as its daily inputs.

"The model uses a daily time step for the simulation of crop growth. On each day, the resources of light, water and nutrients are `intercepted' or `extracted' and converted into assimilated dry matter. Depending upon the availability of these resources and the crop's ability to sequester them, its growth is considered as light, water, or nutrient `limited' "(http://www.staff.ncl.ac.uk/m.d.b.young/parch.html)

Undoubtedly, the above-mentioned facts justify the study of water as a key ingredient of a plant growth process. Nevertheless, the effect of water in plant's growth depends on how it is accessible by the plant roots. This normally varies with factors such as soil types, slope, temperature and humidity; in general it is the environmental conditions which determine the overall effect of water in plants, apart from the biological nature of the plants themselves.

In view of the above, it follows that rain water management is an extremely important factor influencing crop growth. Various environmental conditions would need different water management techniques for an effective and beneficial crop growth. The various ways of water management as observed in Tanzania by SWRG, are discussed in the next section (1.3)

1.2 Rain water harvesting system

In some parts of the world rainwater harvesting systems can be traced back to the time of the earliest human settlements. In Tanzania, rainwater harvesting for crop production can be traced back to before formal colonisation at the end of 18th century. Most of the country's recorded history began only during the establishment of the colonial rule in the early 19th century. A good example is the *Matengo* people in the Songea region who have been practicing rainwater harvesting for agriculture through the use of contour terraces in the Matengo highlands of Tanzania since and before independence.

By 1920, the majaluba system (excavated bunded basins) had emerged as one of the major ways of rainwater harvesting in central *Sukuma* land of Tanzania (NRSP new June 2001) and has recently been adopted by most farmers in the area for the growth of rice, maize, and other crops.

In both urban and rural areas, there have been extensive efforts to harvest rainwater. It is a common practice in most of the residential houses in Tanzania to have water reservoir, beneath the roof of a house for collecting rainwater runoff from the roof. The SWRG team found the following in Maswa district in Shinyanga Region, with regards to rainwater harvest:
"About 28,000 households were using external rainwater harvesting systems (macro-catchments) for rice production in excavated bunded basins (majaluba). Rainwater was also being collected for domestic use, and for livestock (benefiting an estimated 24,000 livestock units). Water collection within fields (in situ) was also widely practised in this district (45,000 households), and 1800 households collected run-off from roofs". (NRSP news September 2002)

These practices serve to tell us that rainwater harvesting is not a new idea. However, it appears that the definition of rainwater harvesting differs from one area to another, and depends mainly on the purpose of the harvested water. It is thus necessary to clearly explain, in the content of the PT project, what rainwater harvesting really means when applied to crop production.

Rainwater harvesting is defined as a method for inducing, collecting, storing and conserving local surface runoff for agriculture in arid and semi-arid regions (Boers and Ben-Asher, 1982). Rainfall has four facets. Rainfall induces surface flow on the runoff area. At the lower end of the slope, runoff collects in the basin area, where a major portion infiltrates the soil and is stored in the root zone. After infiltration has ceased, then follows the conservation of the stored soil water (Hatibu et al., 1996). So, in general, rainwater harvest system refers to a system, which collects rainwater from one producing area (RPA) and transfers it to another designated area, usually referred to as a rainfall receiving area (RRA).

The rainwater harvest systems (RWH) can be divided into three categories:(i) within-field soil-water conservation systems (in-situ RWH), (ii) microcatchment RWH, and (iii) the macrocatchment RWH.

1.3.1 In-situ rainwater harvest system

In-situ RWH refers to all techniques of conserving soil water around the crops in a field, without considering the runoff water from any other area.

Hatibu et al., (1996) define the in-situ RWH as water conservation plus a prevention of net runoff from a given cropped area by holding rainwater and prolonging the time for infiltration. According to them this seems to be the very basic requirement of any rainwater harvesting system. There are a number of techniques falling under this category as outlined below

Tillage:

Tillage simply refers to the preparation and cultivation of the land for crops. It seems to be the basic practice by most farmers in Tanzania, that before the planting is done, land is usually cleared and tilled accordingly. Tillage helps to conserve the soil moisture and provides an easy access of water to the plant roots, reducing surface runoff and increasing the infiltration rate of water in the soil. There is no standard way for how the land should be tilled to improve conservation of water, but it is generally agreed that the deeper the tillage the greater is the amount of rainwater to be conserved and the more likely an increase in crop yield. Significant reduction of surface runoff and increase in crop yields have been shown to occur with increased depth of tillage in Hombolo, Central Dodoma (Mahoo et al., 1996)

Contour farming

This technique refers to land tillage in which successive terraces are formed on the slopes of the hills so as to reduce the speed of runoff from the high to the low ground. In Tanzania it is a common practise in highland areas, especially in the southern Matengo highlands where farmers cultivate coffee and other food crops. This practice is sometimes complemented by the use of water barriers such as stones and constructed ridges.

1.3.2 Microcatchment rainwater harvest system

The microcatchment water harvest systems refers to rainwater harvesting techniques of collecting and conserving the runoff water from one RPA to an adjacent RRA, within a farming area. The distance between a RPA and RRA is taken to be from5 to50metres. So the chief distinction between the in situ rainwater harvesting and the microcathment rainwater harvest is the fact that in the former case there is not a distinct producing or receiving area, whereas in the later, there is a very clear distinction between the two areas. The common techniques in the microcathment system includes, strip catchments tillage, pitting, contour bunds, semi-circular bunds and *meskat* –type-systems which are explained below: -

Strip catchments tillage

This involves tilling strips of land along crop rows and leaving appropriate sections of the inter-row space uncultivated so as to release runoff. It is a very common practice in Tanzania for most part of the country especially in the growth of sweet potatoes, though virtually all the crops can be grown using this technique.

Pitting

According to Hatibu et al., (1996) these are small semi-circular pits dug to break the crusted soil surface. Farmyard manure is added in the pits thus permitting the concentration of water and nutrients. Seeds are planted in the middle of the pits. They are used in areas with rainfall of between 350-600 mm. There is no rainwater harvesting system in Tanzania directly comparable to this technique, though it appears that a technique similar to this is common in the banana growing areas of Tanzania. Usually in those areas banana trees are raised in circular holes rather than in semi -circular pits.

1.3.3 Macrocatchment RWH

As opposed to microcatchment systems, this is a system which involves the transfer of rainwater from a distant area to a particular farming area. The distance between the RPA and RRA is relative large, up to a few hundred metres. The system has the advantage of collecting the rainwater runoff found between the two areas. Although the system set-up may look very similar to conventional irrigation schemes, unlike them, the water used is not sourced from any permanent water body and is available during the rainy season only. Some common techniques in the microcatchment RWH includes sheet flow/hill-side system, streambed system and the stream diversion system which are explained below: -.

Sheet flow system

This involves catching and utilizing the water flowing from the hillside areas to the lowlands parts. It is a common system within the rice growing population of Tanzania, where water flowing from the hills is collected in excavated basin without any modification to its course of flow. This practice in Tanzania is commonly known as *majaluba* system, mainly practised in the lake zone.

Stream bed system

It is a kind of rainwater harvesting where water in the bed of a stream is blocked and spread in various direction for an easy absorption by the crops. The blocking of water is done by the use of either permeable stone dams or earth bunds. It is a common practice to be found in Ifakara area, located in the south of Morogoro Region of Tanzania

Stream diversion system

The system involves the diversion of water from its natural ephemeral stream into a cropped area. The diversion is done in such a way that the new system becomes a system of several water structures coming out of the whole and going into several directions of the field.

1.2 Conclusion

The above discussion summarise the motivation for this project, its background and its importance to rainwater harvest practices, particularly to the Tanzanian country. In the next chapter we will discuss the functionality of the PT program as a prerequisite in understanding the simulation of climatic data.

CHAPTER2: THE FUNCTIONALITY OF THE PARCHED THIRST SOFT WARE

2.1 Introduction

Having got a general idea about a PT project as explained in chapter one, it is better then to understand in general how the software function and its key components towards crop growth simulation before we embark on the task of assessing how it simulates climatic data. Its general understanding reveal the climatic simulation of agro meteorological variables as merely one among its key functions towards modelling crop growth.

2.2 The simulation of crop growth in the PT

PT Version2.3 is developed in Microsoft Visual Basic. It works on the assumption that there is always a runoff receiving area (RRA) and runoff producing areas (RPA) called profiles. A profile represents an area of land with the same soil characteristics and slope. Usually a single crop will be grown in a given profile and the same types of crop management techniques such as, weed control, nutrient treatment and soil surface management would be applied to all the plants. The totality of the profiles forms what is called a system, which is ideally supposed to have profiles of the same climatic conditions, and the same altitude. The system should have the same seasons and thus the same sowing criteria. As for the current set-up of PT, a system comprises of at most 20 profiles.

Having defined the site characteristics as well as the profile characteristics of the given site, one can easily run the simulation process for crop growth of four crops, which are maize, sorghum, rice, and millet. The expected output in any profile would be the amount of harvest (yield per ton), amount of water runoff and run-on, crop transpiration, soil evaporation, soil drainage and rainfall. The outputs are given as an annual summary as well as on a daily basis both in graphs and in tables. Additionally one may also have a look on the growth development of a plant's root, stem and leaf (weight/tonnes) as well as weeds transpiration (mm). A demonstration of how to simulate crop growth using the PT program is in appendix 8.

2.3 The main components of the PT

Following the demonstration in appendix 8, one should be able to see more clearly why PT comprises a number of simulation models. The crop growth needs a defined amount of water intake by a plant per specified time. The determination of water intake by a plant in a particular time, necessitate an estimation of the net amount of water in the soil which brings the need of a soil moisture simulation model as well as the water runoff model. The soil evaporation as well as plant transpiration depends on the amount of water vapour in the air and the temperature condition of the day, and hence a need for climatic data generation in areas where such variables are missing.

The PT software essentially comprises of four main simulation components; the crop model component, soil-moisture component, water runoff component; and the climatic data simulation component. The details for each of these components and their roles are outlined below. Unless reference is made, the details in each of these components are based on Young et al., (1996).

2.3.1 The crop model component of PARCHED THIRST

According to Young et al (1996), the PARCHED THIRST comprises of two models, (i) Parch (Bradley and crout, 1994) for the simulation of sorghum, millet and maize and (ii) Oryza_W (woperies et al., 1996) for the simulation of rain fed, lowland rice.

The PARCH part of the software has already been used for crop simulation, what is new therefore is the THIRST component of the software. Stephens and Hess (1999) used the model to extrapolate field results beneficial effects of soil conservation techniques on yields over longer time periods than few years of measurement. The results showed that runoff control and runoff harvesting produce significant yield increases in average years in both long rains and short rains Matthews (2002).

Most crop models including the PT are built to reflect the interactions between the various factors influencing crop growth and development, such as water and nutrient supply, biotic stresses, and the timing of planting and harvesting of the crop in relation to the prevailing environment. Matthews et al., (2002) PT produces quite a number of outputs apart from the crop growth itself. As it can be seen in appendix 8, soil evaporation, soil moisture content, and some climatic conditions of the are among the outputs from the PT simulation process.

2.3.2 Climatic generator

The PT Model needs daily climatic data in the simulation of crop growth, soil-moisture content and the rainwater runoff. These are temperature, rainfall, humidity, wind, and radiation. However, in most developing countries, the daily records are either unavailable or incomplete, or spatially sparse, or contains a large number of missing data. Thus the climatic generator component is meant to simulate weather data in case of the presence of at least one of the above scenarios. Basically the climatic generator component uses the historical data fed in the model. The larger the amount of historical data supplied, the better is the quality of the generated weather data.

Wyseure et al.,(1996) gives the following account on the functionality of the climatic generator component of PARCH-THIRST.

"The climatic generator component of the model uses the statistical properties of the historical weather data and random numbers to stochastically generate synthetic sequences of weather data. Rainfall is the controlling variable, with all other variables, (except wind speed) dependent upon whether a simulated day is wet or dry. Overall seven weather variables are considered by the model"

Rainfall, which is the leading variable, is generated by a two stage processes using a Markov chain method. The method involves the determination of whether a day is wet or dry, and if the day is wet then the rainfall amount is estimated from the gamma distribution.

For temperature, both maximum and minimum are generated by a multivariate process, which involves the generation of residuals about long-term means. The means used and the residuals generated depend upon the wet or dry status of the day Young et al., (1996). The generation of radiation data follows an equivalent procedure.

Relative humidity is sampled from one of the two gamma distributions depending upon the wet or dry status of the day, whereas wind speed is also sampled from the gamma distribution, but not dependent on the wet or dry status of the day.

It is only evaporation, which is not generated, but rather calculated from other generated variables. The approach of calculating the evaporation data is based upon the Penman-Monteth method for reference crop evapotranspiration as presented by Feddes and Lenselink(1994), Young et al., (1996).

2.3.3 The water runoff model component.

In order to generate a specific amount of water runoff, rainfall intensity is needed in the runoff generation model, preferably at short intervals of time. PT uses five-minute intervals. But, in Tanzania, as in most developing countries there are little data records on rainfall intensities. As a consequence, a provision is made in the model to calculate the five-minute rainfall intensities, using a so-called rainfall disaggregator.

The common rainfall records in developing countries are the total daily rainfall (R), rainfall duration (D) and the maximum 30-minute daily rainfall intensities (I_{30}). Using these three common records, the rainfall disaggregator, fits them to an assumed rainfall intensity distribution. Explaining briefly, how it works in the model, Gowing et al., (2002) says",

"PT provides a rainfall disaggregate which generates 5-minute rainfall intensity data from an assumed distribution of rainfall intensities (figure2.1) similar to that proposed by Oron et al (1989). The distribution of rainfall intensity is assumed to have a linear rise in 30 minutes period, followed by a linear decay in 15 minutes and finally an exponential decay through out the time.

This distribution is fitted to rainfall amount, duration and I_{30} , using a Newton Raphson iterative technique on daily basis. Where observed durations and I_{30}s, are not available, regression equations (which can be developed at climatically similar stations) are used to estimate them".

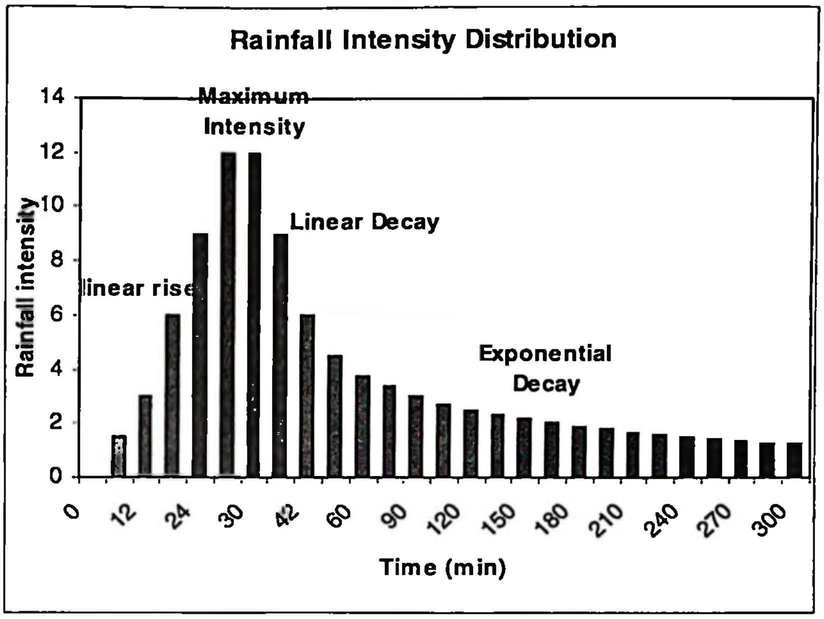


Figure.2.1

Having determined, the rainfall intensity, as described above, water runoff is then calculated as an excess of rainfall intensity to infiltration rate. The infiltration rate is determined by the so called Green and Ampt equation which is given as $fp(t) = K_{fs}\left(1+\frac{S_{av}D_i}{F(t)}\right)$ where $fp(t)$ is the potential infiltration at time t, K_{fs} the field hydraulic conductivity, S_{av} the wetting front suction, D_i the initial water deficit and F(t) the cumulative infiltration at time t.

The excess rainfall intensity is further modified by the depression storage of the soil as well as the surface sealing. Wyseure et al., (2002), explains the role of soil depressions as essentially being to reduce the amount of water runoff from the soil. According to him, in a situation where rainfall intensity exceeds infiltration rate for a prolonged period, the smallest depressions are quickly filled and overland flow begins. On the other hand soil crusting increases surface runoff by reducing rainfall infiltration. Soil crusting most often occurs when rain separates the soil into very small aggregates and individual particles that cement into hard layers at the soil surface when drying occurs rapidly. Vandyke (2000)

In some cases especially with macrocatchment RWH, estimation of runoff routing becomes important, because the runoff water is expected to travel for a quite significant distance eroding some parts of the soil. PT estimates the runoff routing by using a simple hydrograph convolution based on SCS(Soil Conservation Services)-synthetic hydrograph approach. A hydrograph is a graph of water flow rate versus time. In order to estimate this rate of water flow, a shape of flow is assumed after which the peak and volume runoff are determined, which eventually define the hydrograph. It is this type of estimation, which is known as a SCS synthetic approach. The word SCS emphasizes the fact that, the estimation process utilizes the information on Land use and soil types. (http://www.ce.utexas.edu/prof/maidment/gishydro/txdot/litrev.htm#uno)

2.3.4 Soil water model component

The determination of moisture content of the soil is an essential part in the modelling process of any crop growth. The amount of water in the soil will depend upon a number of factors, notably, the amount of runoff, infiltration rate, harvested water, and the soil characteristics of the area under consideration. PT has a provision for determining, the soil moisture content depending on the soil type under the area of study.

2.3.5 Pedotransfer function

This is an inbuilt component meant to facilitate the practical functionality of other models especially the soil-moisture model as well as the water runoff models, which rely heavily on the soil hydraulic properties that are usually hard to measure and establish. This is how Young (2002) explain the importance and role of the Pedotransfer component in PT program.

"The moisture retention curve, the saturated and unsaturated hydraulic conductivity require a soil-physical laboratory. In practice this requires transport of numerous samples over a long distances (several 100km) to carry them to the few well-equipped laboratories. Those measurements are also notorious for failing one or more aspects. Alternatively instantaneous profile methods can be applied in the field, which are major logistic operations and require huge amounts of water in an area where water is scarce. In order to minimise the amount of data collection, there has been a great deal of interest in Pedotransfer functions (PTF's). These equations estimate the hydraulic properties of the soil from easily measurable soil properties (usually **soil texture,**

organic matter content and **bulk density**), which are produced during a conventic soil survey".

2.3 Conclusion

In this chapter we have discussed the key components of the PT, including the clim generator. In the next chapter we shall discuss the process of simulation of clim variables having done a demonstration. Further more we will assess the agreem between the generated data and the historical data using the PT itself and by work out our own calculations.

CHAPTER 3: PARCHED THIRST AS CLIMATIC DATA SIMULATION SOFTWARE

3.1 Introduction

In this chapter we demonstrate how to simulate climatic data using the PT software and how to assess the agreement between the historical data and the PT generated data. It is on the basis of this assessment that we build our motivation to find alternative simulation methods

3.2 How to simulate data using PT software

This explanation of the process of climatic data simulation is based on the PT User Guidebook by Young et al., (1996) as well as on the tutorials included within PT. A user can generate climatic data using PT in three different ways.

1. Using a weather parameter file: A weather parameter file, contains summary of weather parameters necessary for generation of daily climatic data in a year. The file contains monthly or daily summaries of seven climatic variables. The user can either create it based on the existing historical files of data or can use the already created weather parameter files found in PT.

2. Using historical data files: These files contains, the daily climatic records of data. The PT has been supplied with historical weather files for some sites in Tanzania and Uganda. These sites are Same, Arusha, and Morogoro for Tanzania and Soro for Uganda. Alternatively a user has to supply the PT software with his own historical data files, which must be in a format compatible to that of PT

3. Using the rainfall overlay option: This option gives the user an opportunity to generate other agrometeorological variables at a site where only rainfall data is available or at a site where other agro meteorological variables are available, but with some missing data. To

use this option, a user has to first create a number of rainfall only weather files which correspond to the number of years with available rainfall data. Furthermore a user must make sure that the rainfall records contain no missing data by putting a 0 in a place of missing values. Having done so, he can then use these files along with a weather parameter file to generate climatic data.

Of all the options above, it is only the second option, (i.e. historical data files) that gives a user an opportunity to compare the statistical properties of the generated data and the historical data on various dimension of time, such as on daily, weekly, monthly or yearly basis. On the other hand, one can as well make the comparison using a weather parameter file, but the comparison would be limited because the weather parameter files contains only daily or monthly summaries of weather parameters.

In view of the above, in this project we will use only historical data files in simulating agrometeorological variables as well as in assessing the degree of agreement between the two sets of data. Of the sites with historical records of data, it is Arusha, which has the longest record of data (31years), and for this reason, this site will be used for demonstration and to assess the agreement between the historical data and the generated data.

Using the PT software, I simulated 100 years of climatic data, based on the 31 years of Arusha climatic records. Details of the steps, which one can follow to simulate climatic data in PT as I have done, are attached, as appendix 9.To understand the discussion in the next section a reader is advised first to go through appendix 9 from which some important graphics will be discussed. Of particular importance are the figures 3.2.1 (a) (b) and (c) as well as figures 3.2.2 (a) and (b). All are in appendix 9.

3.3 Comparison between historical against simulated data by PT

After a user has gone up to step 8 of the demonstration. (Please see appendix 9 for the details), he has an opportunity to see the comparison between the generated data and the historical data in graphs. As an example rainfall and maximum temperature have been considered and their graphs are found in appendix 9 as figures 3.2.1 (a) (b) and (c) and figures 3.2.2 (a) and (b).

Before I discuss the implication on those five graphs, I would like to stress that the graphs are generated automatically by the PT software after a user has gone up to step 8 of the demonstration and there exist no further details behind those graphs apart from the labels and titles.

Looking on what has been written on the five graphs, there appears to be quite a good match between the historical data and the generated data basing on the mean observations. The agreement of their means should be expected, because the generated data is derived on these very means of the historical data. Therefore what is important to consider is whether there is an agreement between the standard deviations of the two data sets

Two graphs (figure 3.2.1(c) and figure 3.2.2(b)), which compare the standard deviations, indicate that the generated data is consistently less variable over the months than the historical data. This casts a doubt on the reliability of the generated data and call for a further look into how agreement between the two data sets based on standard deviations can be improved.

I have attempted to find out how the means and standard deviations have been worked out in the PT software. The user guidebook by Young (1996) gives no detailed description of how they are calculated. The uncompiled code of the software (see appendix 2), was also of little help. The formulae used for calculating the means and standard deviation were correct, but the difficulties of unravelling the code meant I

0650481

SNAL DATE 31/5/05 SPECIAL COLLECTION

could not see exactly what values were being used. Finally, I exported both the historical and generated data into Instat to work out the standard deviations, but I was unable to get the same numerical results as the ones shown in the figures generated by PT. Nevertheless, my own calculations using the Instat software yield the same conclusion that, the historical data is more variable than the generated data. Below is a comparison of the two data sets basing on my own calculations.

3.4 Comparison of historical data against generated data based on calculations using the Instat software.

3.4.1 Introduction

Before we compare the two sets of data, it is important to make clear that this comparison will also be done based on the graphs of the two data sets. This to me, seems to be a clear way of comparing the agreement between the two data sets. It is surprising that, although within the PT software, the comparison is made based on the graphs, in some of its literature a slightly different approach is reported to check the validity of the PT components, including that of the climatic generator.

The coefficient of determination was used by Young et al., (1996) to validate the water runoff model component, crop model component, as well as the climatic generator component. But Young is not alone in doing so, Chineke et al., (1996) also used the coefficient of determination in assessing the validity of his simulation methods. While I am not against this approach, its use should be of secondary consideration and rather the graphs themselves should tell us about how good is the agreement. The reason is that the coefficient of determination explains on average how the two data sets match, but because of similar patterns between the two data sets over time, its value may unnecessarily be exaggerated. (This point will be further clarified in the next subsection 3.4.2 (i))

In this comparison, I use the **"mean monthly rainfall totals"** instead of the "mean monthly daily rainfall" as implied in the PT graph (figure 3.2.1(a)) and **"the standard deviations of the monthly rainfall totals"** instead of the "mean monthly standard deviations of rainfall on rainy days" as implied in the PT graph (figure 3.2.1(b)) My reason for doing so is due to the fact that, rainfall totals are more recognisable and helpful measure of rainfall than the mean monthly daily observations.

For the temperature data I use **"mean monthly maximum daily temperature** " instead of "mean monthly maximum temperature" as implied in the PT graph (figure3.2.2(a)) and **"standard deviation of the mean monthly maximum daily temperature"** different from " mean monthly standard deviation of maximum temperature" as implied in the PT graph (figure3.2.2(b)). My reason is that, I have an impression that, the phrase "mean monthly standard deviation " as used in PT was meant to calculate the variation within years for each month and not between years. The key comparison of the two data sets should be based on their variability between the years and not within years. Furthermore the phrase "mean monthly maximum temperature" does not clearly state whether it is a monthly daily temperature or monthly mean temperature, which is being considered.

To overcome the lack of clarity about the calculations as encountered in PT software, the summary measures are given below: -

1. To obtain mean monthly rainfall totals
   (a) All the daily rainfall observations are summed in each month of a year irrespective, to form 12 totals each year.
   (b) The above process is repeated for all the years; in my case, from 1970 to 2000(31 years) for historical or 100 years for simulated data.
   (c) The totals are then averaged over the years

2. To obtain the standard deviations of the monthly rainfall totals
   (a) The mean monthly total found in 1 above is considered.
   (b) The standard deviation is worked out using the following formulae

$$S.D_r = \sqrt{\frac{1}{n-1}\left(r_i - \overline{r_i}\right)^2}$$

Where r_i is a monthly rainfall total and $\overline{r_i}$ is the mean monthly rainfall total considered in (1a) above.

n is the total number of years which is 31

3, To obtain the mean monthly number of rainy days

(a) All rainy days are counted in each month of a year to form 12 different totals of the number of rainy days. A rainy day is regarded as a day with rainfall value of more than 0.26mm. My choice of the value "0.26" was based on the fact that a threshold value was 0.26 appears to be a default value in PT (refer to figure 3.14 at step 3 of the demonstration). This choice was necessary in order to have a meaningfully comparison between the PT generated data and the historical data.

(b) Step (a) is repeated for all the 31 years. Each month then has 31 totals

(c) The 31 totals are then averaged over the years

4. To obtain the mean monthly daily maximum temperature

i. All the daily temperature observations are summed in each month of a year irrespective of the size of any observations, to form 12 totals each year.

ii. Each total found in (a)above is divided by its respective number of days to form 12 averages (monthly daily temperature)

iii. The two processes above are repeated for all the years

iv. The averages are then averaged over the years.

5. To obtain the standard deviation of the mean monthly maximum temperature

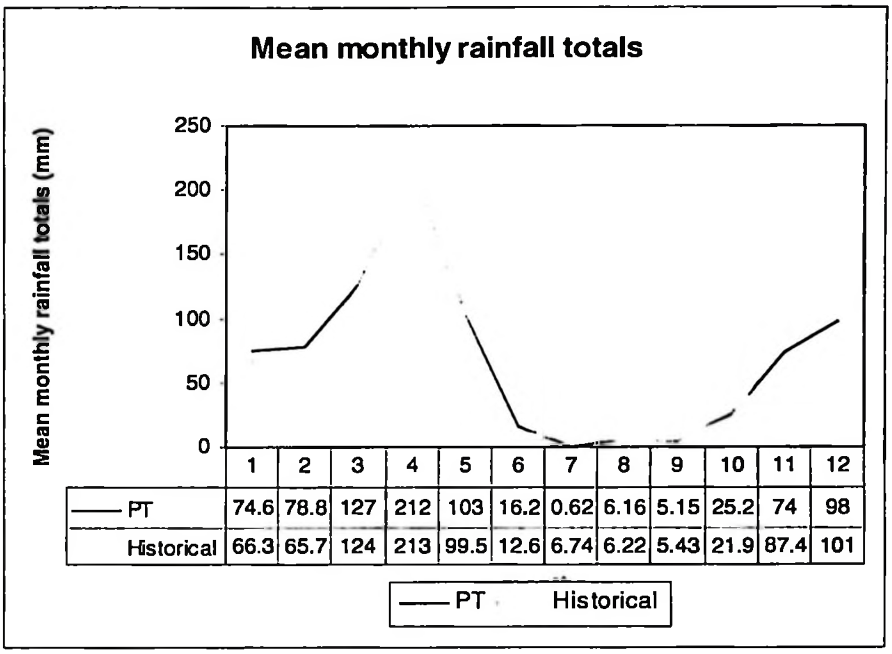
v. The mean monthly daily maximum temperature found in 4(d) above is considered

vi. The standard deviation is worked out using the following formula $S.D_t = \sqrt{\frac{1}{n-1}\left(t_i - \bar{t}_i\right)^2}$ where t_i is a monthly daily maximum temperature and $\bar{t}_i$ is the mean monthly maximum daily temperature found in 4(d) above.

3.4.2 Comparison of the two data sets based on the above-defined summary measures

(i) Rainfall

(a) Based on the mean monthly totals

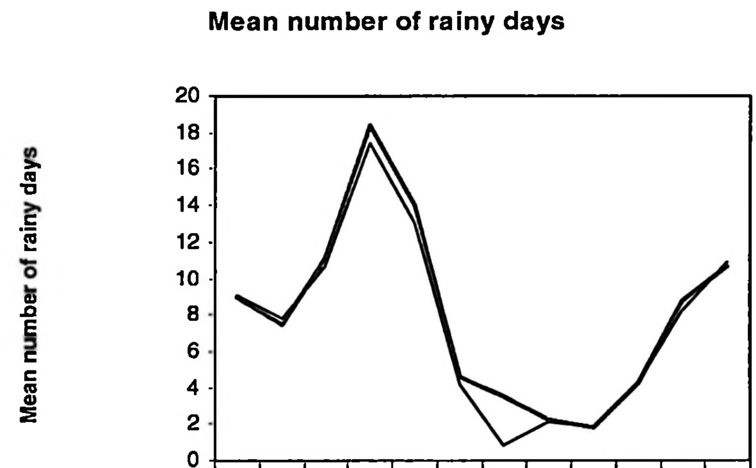


	1	2	3	4	5	6	7	8	9	10	11	12
PT	74.6	78.8	127	212	103	16.2	0.62	6.16	5.15	25.2	74	98
Historical	66.3	65.7	124	213	99.5	12.6	6.74	6.22	5.43	21.9	87.4	101

Figure 3.2.3(a)

The graph above indicates comparison of mean monthly rainfall totals between the PT and the historical data. The agreement between the two data sets is extremely good. This result is the same as the one found using the PT graphs. The message here is the same as before, that the agreement of the means should be expected because the generated data is derived from these very means.

(b) Based on the mean number of rain days



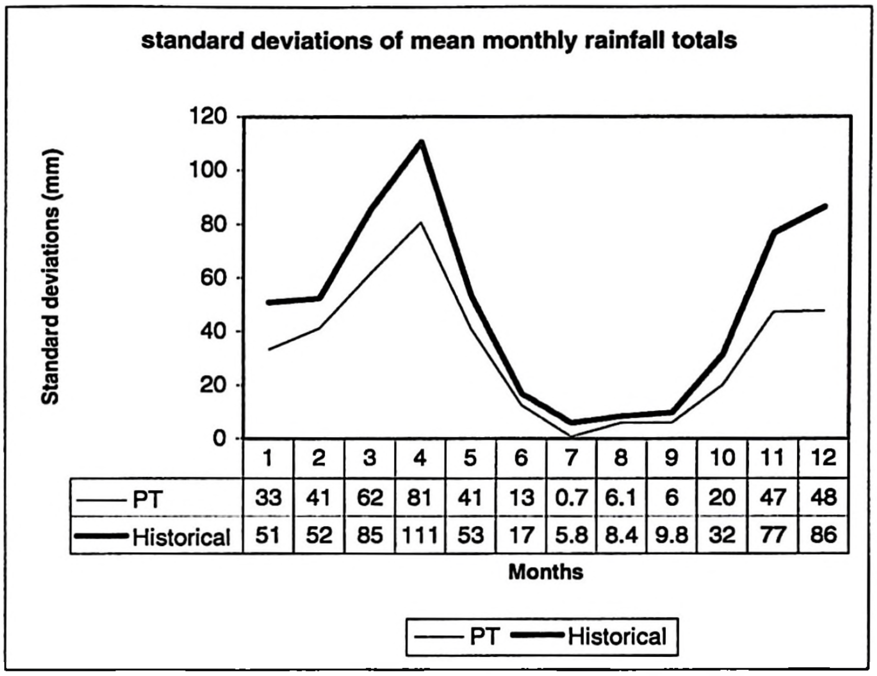
	1	2	3	4	5	6	7	8	9	10	11	12
PT	9	8	11	17	13	4	1	2	2	4	8	11
Historical	9	8	11	18	14	5	4	2	2	4	9	11

Months

Figure 3.2.3(b)

This graph indicates a very good agreement of the two data sets. The message is the same as in (a)

(c) Based on the standard deviations of monthly totals



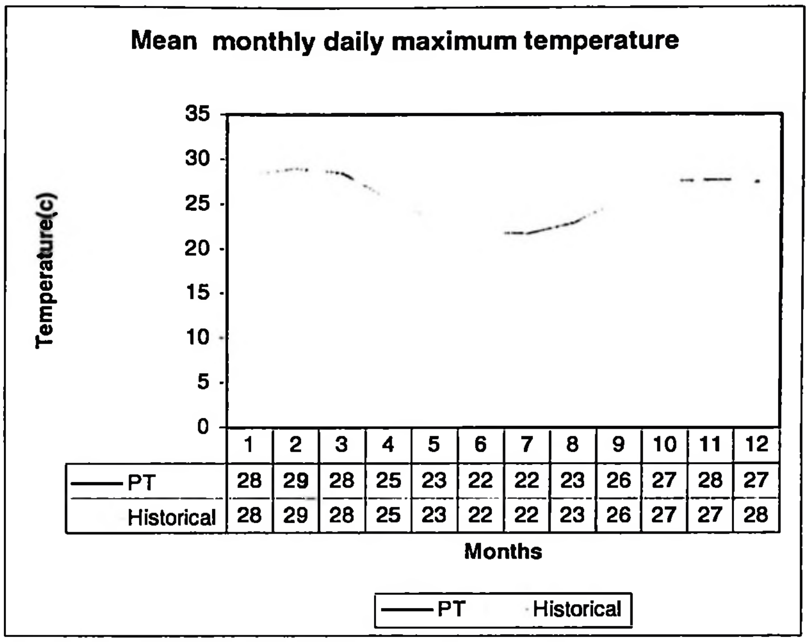
Months	1	2	3	4	5	6	7	8	9	10	11	12
PT	33	41	62	81	41	13	0.7	6.1	6	20	47	48
Historical	51	52	85	111	53	17	5.8	8.4	9.8	32	77	86

Figure 3.2.3(c)

Figure 3.2.3(c) indicates the comparison between the two data sets based on the standard deviations of the monthly rainfall totals. The graph shows that the standard deviations of the historical data are consistently higher than those of the PT generated data over the months. The coefficient of determination between the two data sets is 96%, which may be interpreted as showing a very close agreement, contrary to what has been displayed by the graphs. This clarifies the point raised in section 3.4, that in a situation where the two data sets have the same seasonal variation it is very risk to use the coefficient of determination to assess their agreement.

(ii) Maximum daily Temperature

(a) Based on the mean daily maximum temperature

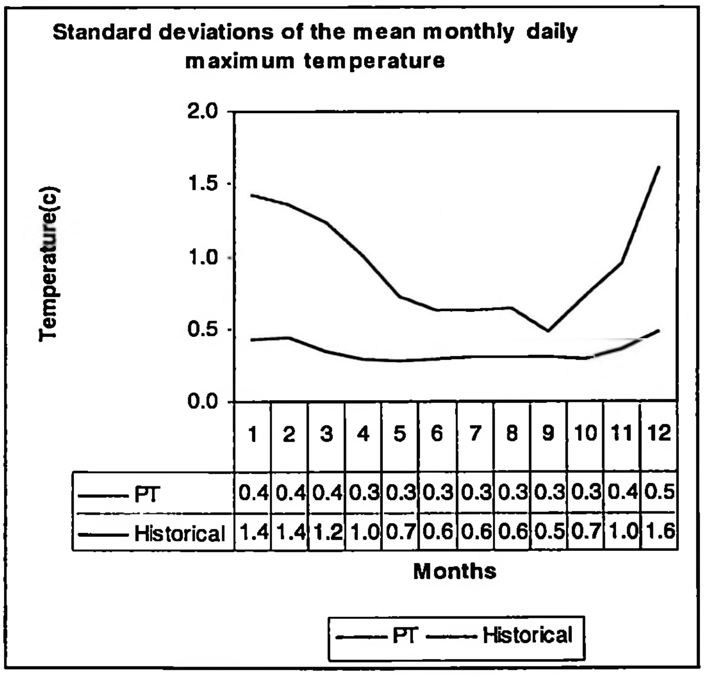


Months	1	2	3	4	5	6	7	8	9	10	11	12
PT	28	29	28	25	23	22	22	23	26	27	28	27
Historical	28	29	28	25	23	22	22	23	26	27	27	28

Figure 3.2.4(a)

The figure above indicates the comparison between the two data sets based on the mean monthly daily temperature. The agreement here is almost perfect, which emphasizes what has been said before regarding the agreement between the two data sets on monthly means.

(b) Based on the standard deviations of the mean monthly daily maximum temperature



Months	1	2	3	4	5	6	7	8	9	10	11	12
PT	0.4	0.4	0.4	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.4	0.5
Historical	1.4	1.4	1.2	1.0	0.7	0.6	0.6	0.6	0.5	0.7	1.0	1.6

Figure 3.2.4(b)

The figure above indicates the comparison between PT generated data and the historical data on standard deviations of mean monthly daily maximum temperature over the months. As the graph indicates, the discrepancy between the two data sets is much higher than with the rainfall data.

At this stage it is clear that the PT model is not properly modelling the spread of the observed data. Figures 3.2.3(c) and 3.2.4 (b), which compares the standard deviations between the two data sets for rainfall and temperature, proves this assertion beyond doubt. To convince the users of the model about its deficiency let us further look at the suitability of the generated data in terms of some events of interest by farmers, and planners. In the next section, a brief description of these events and how they are reflected by the generated data as compared to the historical data is discussed. As an example, the comparison is made using rainfall data only.

3.5 Comparison of the historical data against generated data based on some events of interest by farmers.

An "event" in the sense used here refers simply to a characteristic of interest for which there is a single observed value each year. For example, "total rainfall in March" is an event as it occurs only once in a year. Below are some events of possible interest to farmers:

3.5.1 The 20 and 80 percent points of the distribution of monthly totals

The 20 and 80 percent points are useful to a farmer as they can tell him about the expected monthly total rainfalls.

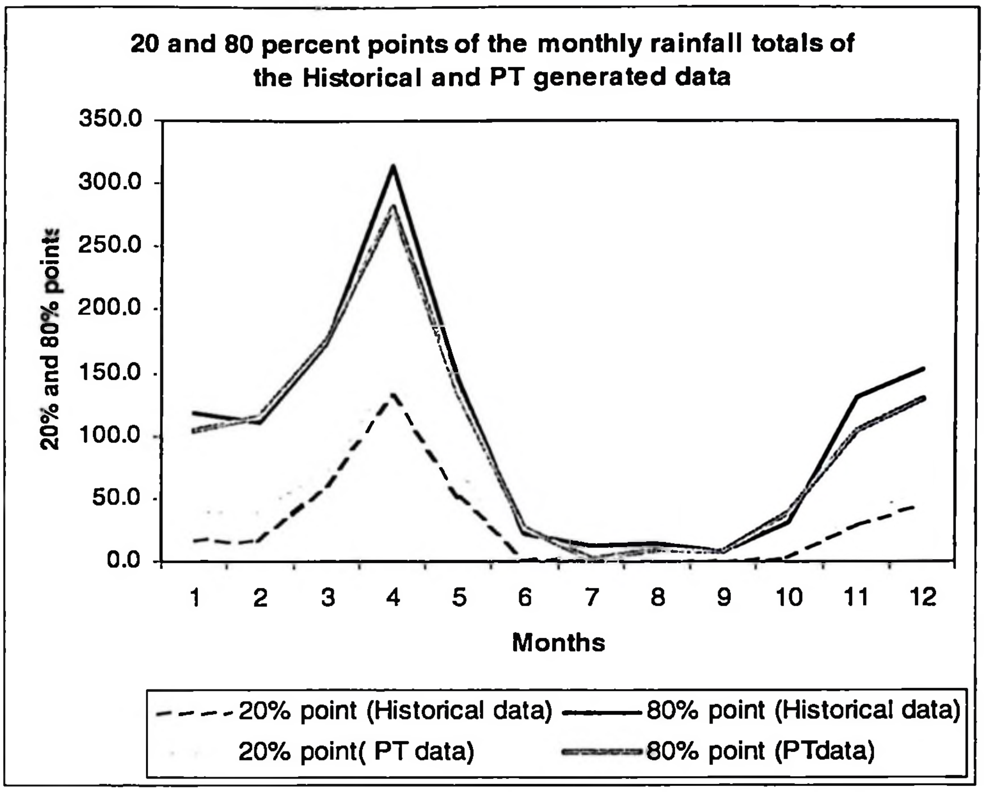


Figure 3.2.5

From the above graph we can see that, the gap between 20 and 80 percent points of the historical data is much higher than that of the PT generated data. The reason is that the 20% point of the PT generated data is consistently higher than that of the Historical data. This is a further confirmation of what has been said before about the difference in variability between the two data sets. A farmer's expectation on monthly total rainfall in a year would be different between PT generated data and historical data.

3.5.2 The distribution of the possible sowing dates

The knowledge on the distribution of the start of rains are important to farmers as they help them to avoid the risk of plant's death as well as in planning the whole of their farming activities. Knowing the start of rain to most farmers in developing countries goes by experience relying mainly on factors such as perceived amount of moisture in the soil and the number of days with consecutive raining. In this aspect the distribution of the start of rains will be determined based on two factors, which are common to Tanzanian farmers.

(a) The earliest planting date as practised by farmers
(b) The total amount of rain within consecutive days of raining

As an example let us consider Maize growing in Tanzania. The planting by most farmers is usually between late December to late February and rarely goes to March. I will regard 1st of January as the earliest planting date for maize in Tanzania. The planting date is defined here as the one when there have been 20 mm total rainfall in three consecutive days.

Table3.1

Date of planting	Historical	PT
10 Percentile	3rd January	1st January
Median	18th January	14th January
90 Percentile	2nd March	9th February
Mean	26th January	17th January
Standard deviation (days)	23	16

The table indicates that the distribution of sowing dates between the two data sets is very different especially in late planting. While with the Historical data late planting would be around 2nd March, with PT it will be around 9th February. Further to this, there is variation of about three-weeks (23days) of a planting date by farmers using historical data while using PT data there is a variation of only two weeks (16days).

3.5.3 Combining the dry spells and sowing dates

Dry spells are events indicating the maximum number of consecutive dry days within a specified time of a year. For example one may consider say a dry spell of 5 days or more within a particular month of a year. Such a consideration is known as "an unconditional dry spell". However if a restriction is made apart from the specified time period, a dry spell is called "a conditional dry spell". For example one may say a dry spell of 5 days or more in June after 2nd of the month. The study on possible dry spells is very important to farmers because it can help them to know which crops are supposed to be planted at a particular time of a year

The event in section 3.5.2, which was on possible sowing dates, was not considered together with risks associated with dry spells. Usually the knowledge on the distribution of dry spells help the farmer to adjust their ideal sowing dates. Below is a comparison, which includes the distribution of 10-day dry spells, one more condition apart from those given in section 3.52. With the inclusion of the 10 days dry spells, we may now state our event as follows:

"The earliest planting date is 1st of January, whereas as the next possible date would be the one, before which there have been 20 mm total rainfall in three consecutive days with no more than a 10 day dry spell within the next 30 days

Table 3.2

Date of planting	Historical	PT
10th Percentile	3rd January	1st January
Median	17th February	15th January
90th Percentile	31st March	15th February
Mean	17th February	21 th January
Standard deviation	35	22

After considering the dry spells, the distribution of the sowing dates between the two data sets has been very different. With the Historical data most farmers will be planting around 17th February whereas with the PT generated data most farmers will be planting around 21th January. Further more the planting date using the historical data is more variable than that derived from the PT generated data.

3.5.4 The distribution of dry spells

Below are the probability distributions of the chances of dry spells of 7 and 10 days conditioned at 24 different possible sowing dates each within the following 30 days.

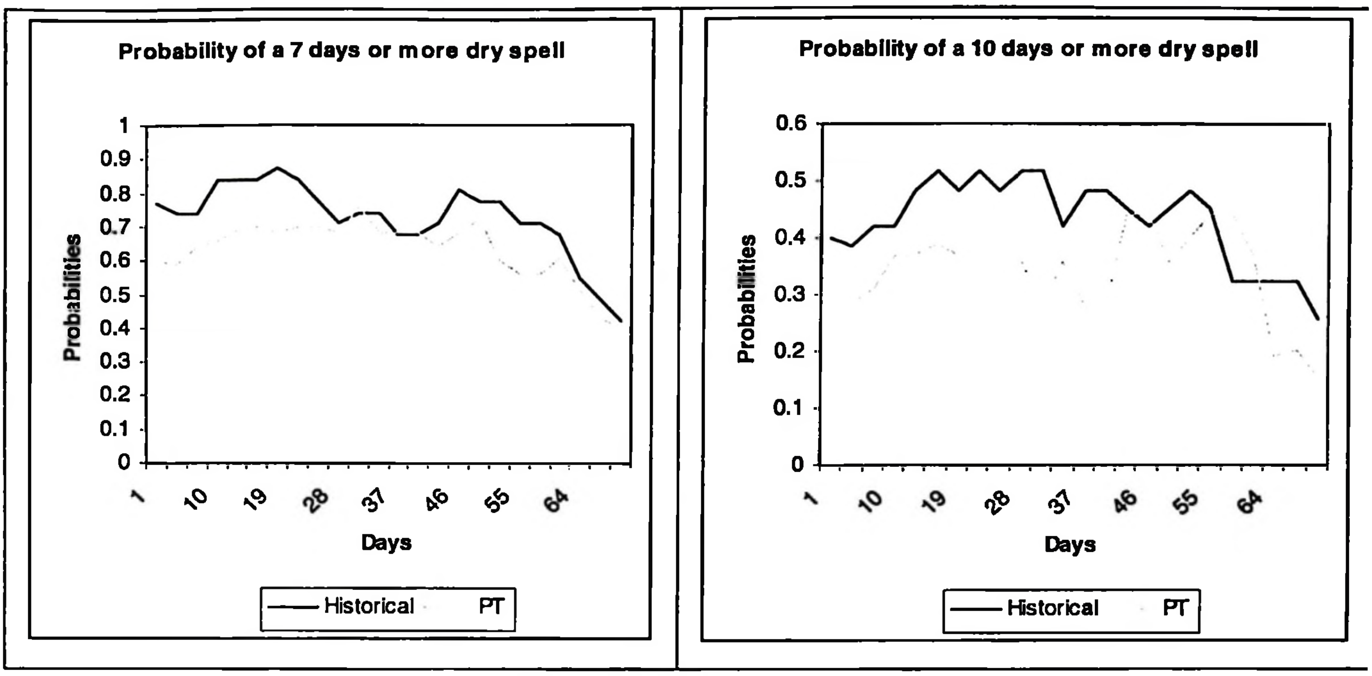


Figure 3.2.6

The distribution of the dry spells indicates that, PT generated data tends to underestimate the risks of the dry spells, more than the historical data.

The discussed events clearly indicate the discrepancy between the two data sets in exploring some important information to the farmers. This is a further justification of finding ways to improve the method of data generation in PT

In order to be able to possibly improve the agreement between the historical data and the generated data, we have to clearly understand the methodology behind data generation in PT software. This is attempted in chapter four. However, it is appropriate, at this stage, to make a comment on the flexibility of the PT software, both in terms of its general use (simulation of crop growth) and in terms of simulation of climatic data. These are discussed in the next section.

3.6 Problems in Using the PT software.

Regarding its general use, the PT is a "user-friend" software. My own concern is on frequent crashing, which often occurs when you misspecify some of the requirements in windows. This causes frustration for the user, as the same processes must be repeated several times whenever a mistake is made.

There are further problems with the set up of the PT software as explained by people who have used it. The development of PT has been funded by NRSP since 1992. NRSP wanted an independent assessment of the PT to help them decide if and how to continue their support of PT. They commissioned the Statistical Services Centre to carry out this review (Dale, 2004).

The review was generally positive about the usability and helpfulness of the software, but indicates various aspects that need improvement to make PT more useful for the intended user groups. In particular the inclusion of some detailed case studies is recommended. Full details on the comments given by Dale are attached as appendix 2 of the project.

The report by Dale did not assess the climatic simulation features of PT, and thus contains no particular problems regarding climatic data simulation. Having, personally used it many times to simulate climatic data I have noted the following frustrating problems which may be encountered by any user, especially a researcher or a student.

- There is no clear description in the user guidebook and in the tutorials on how the summary measures are worked out. The phrase "standard deviation" can sound clear by itself, but how it is applied is very important. For example, does the software work out the variation, within the years for each month, or does it work out the variation between the years for every month?

- Although the software presents the graphics, the numerical values behind those graphics are not readily available. Only simulated data files in a raw form are indicated. The numerical results behind those graphics may only be obtained by creating a parameter file out of the historical data file used in the simulation.

- The graphics shown are not automatically saved in a file, at the end of the simulation process. Once you close the software, there is no way to reproduce those graphics, unless you simulate again, in which case you will have graphics corresponding to a different set of data. The graphs themselves are in a picture format, which is difficult to edit or save.

- The way in which the simulated data is saved is also frustrating. Each year of data is saved in a different file. For analysis purpose you have to copy and paste as many times as there are simulated years. In my case I had to repeat this process 100x 7 times, as I had 100 years of simulation, and 7 variables under consideration. The situation becomes even worse when more than one site is considered.

- The software itself does not have facility to look at descriptive statistics of the historical data. It is only after the simulation, that some descriptions are given along with the simulated data.

- Although, the software works out the comparison of the simulated data and the generated data, because of the lack of details on how they are being calculated, it is then not easy to know whether some of the common statistical practices on particular variables are also considered during their calculations. For example, the value used for the rainfall threshold may vary from one station to another and by one researcher to another. In PT, it is not clear of what is meant by a "rainy day". While the user guide indicates that a user has an opportunity to specify a rainfall threshold, in the simulation process (see figure3.1.4 of step 3) the default value of the threshold is set at 0.26. And in the source code it appears as, this threshold is set to a value of 0 (see appendix2).

- The format of the simulated data e.g. figure 3.9(c) above is not very easy to operate with. There are a lot of columns with missing values, which need not be there. Furthermore, the column contains no titles.

- The comparison of data is based on common summary measures, which are not of practical interest by users of the data in the relevant field. For example, a farmer, or extension officer would be more impressed if the simulated data predicted a similar distribution for the start of rains or distribution of dry spells, as these are important for farming activities, rather than simply seeing that the means and standard deviations are similar.

- The use of rainfall data to simulate other agro meteorological variables involves a long and tedious process of deleting, copying and pasting. If there could be a way of shortening the process, the users would be relieved

- The reason there is an option for using a weather parameter file in simulating data is not explained. The surprising issue is that a weather parameter file itself can be created from a historical data file. My own opinion is that, the creation of a weather parameter file would have been more useful for descriptive statistics than for simulation purpose, the fact which has not been overemphasized.

- There is an option for a user to use a constant random number seed or to use a separate one. But the option does not indicate which number has been used in the simulation process and how many times has it been used. I encountered a problem when I wished to simulate similar climatic data, as I could not know the number seed of the previous simulation.

500316 48072

CHAPTER 4:GENERATION OF DAILY RAINFALL DATA IN PARCHED THIRST SOFTWARE.

4.1 Introduction

In chapter three we have seen two main problems with the PT model;

1. There is a documentation problem. This leads to a lack of clarity on how summary measures are calculated to compare simulated and historical data.

2. The comparison of standard deviations of the mean monthly rainfall totals of the Arusha historical data with the PT generated data indicated that the simulated data did not adequately model the historical data. The same applied to the other variables.

To understand why the two data sets do not match, we need to know how the rainfall data were simulated in PT and the assumptions that are made. This may indicate why the two data sets do not match and also, suggest ways to improve the method of simulation. So in this chapter the objective is to describe the methodology used in the PT model and discuss whether the assumptions are adequate.

In chapter three we saw how to use PT to generate climatic data including rainfall data, using a given set of historical data. The methodology behind rainfall data generation in PT involves three steps

1. Determination of the probability model for the chance of the occurrence of rain in a given day of a year.

2. The determination of the parameters of the probability distribution of the daily rainfall amount for each day of a year.

3. Simulation of daily rainfall amounts.

Step 3 is the implementation of steps 1 and 2 and is a relatively simple and so is described below before giving details of steps 1 and 2.

To obtain the amount of rainfall of a particular day we will have to draw a uniform random number and compare it to the probability of a dry/wet status of the day under consideration. If the number is larger than the probability of having a dry day , we conclude that the particular day is a wet day; otherwise it is a dry day. If the day is wet, the rainfall amount x, is generated from the model of rainfall amount for that particular day of the year. This process is carried out for each day of the year and repeated for all the years of simulation.

4.2 The model for the chance of rain.

The observed data is a set of records that state whether it rained or not on each day of the year. The aim is to use this data to obtain the probability of rain on a given day of the year.

The probability of an event is the ratio of the number of times it has appeared to the number of trials. For example with the Arusha data, January 6 was dry in 20 years and rainy for 11 years out of the 31years. The observed probability is therefore 11/31=0.35 (see appendix 5 for details)

This probability describes the simple case in which the chance of a day being wet does not depend on the state of any previous day. We refer to such a situation as a "zero order Markov chain"

In PT, the chance of rain in a day, is conditioned upon the state of the previous day. This is called a " first order two state Markov chain". The two states are either wet or dry. Similarly when the chance of rain in a day is conditioned upon the state of the last two previous days , it is called a "second order two state Markov chain". So basically the orders ranges from zero, first, second, third and so on. The theory of Markov chain and how it has been applied in PT is explained briefly in the next section.

4.2.1 An overview of a Markov chain process

Markov chain theory is concerned with the determination of the probability for the occurrence of a given event subject to the condition of another antecedent event or events. In the discrete-time case, the process consists of a sequence X_1, X_2, X_3, ... of random variables. The domain of these variables is called the *state space*, with the value of X_t being the state at time *t*. If the distribution of X_{t+1} conditional on past states is a function of X_t alone , we then have $P(X_{t+1} / X_0, X_1, X_2, \ldots, X_t) = P(X_{t+1} / X_t)$ which is said to have the **Markov property**.

With the first order Markov chain as used in PT, the state space for the random variable "rainfall occurrence" are wet(w) and dry(d). The two states are summarised below with their transition matrix P(t) , using the same notations as used by Young et al., (1996), the author of the PT UserGuide book.

$$P(t)=\begin{bmatrix} P_{DD} & P_{DW} \\ P_{WD} & P_{WW} \end{bmatrix}$$

Where

P_{DD}= conditional probability that a dry day is followed by a dry day

P_{DW}= conditional probability that a wet day is followed by a dry day

P_{WD}= conditional probability that a dry day is followed by a wet day

P_{WW}= conditional probability that a wet day is followed by a wet day

Since our interest is on the chance of a day being wet, we shall consider only the conditional probabilities of wet given wet (P_{WW}) and of wet given dry(P_{WD}) Accordingly P_{DW}=1-P_{WW} and P_{DD}=1-P_{WD}. The two conditional probabilities are given by the follwing formulae;

$$p_{WD}(i) = \frac{\sum_{j=1}^{j=n}\{X(i+1,j) = W / X(i,j) = D\}}{\sum_{j=1}^{j=n}\{X(i,j) = D\}}$$

$$p_{WW}(i) = \frac{\sum_{j=1}^{j=n}\{X(i+1,j) = W / X(i,j) = W\}}{\sum_{j=1}^{j=n}\{X(i,j) = W\}}$$

Where X(i,j) is a random variable that a day i in year j is wet or dry

$p_{WD}(i)$ is the probability that if day, i is dry then day i+1 will be wet

$\sum_{j=1}^{j=n}\{X(i+1) = W / X(i,j) = D\}$ is the sum of all the occurences of a wet day following a dry day on the previous day, i in years j=1 to j=n

$\sum_{j=1}^{j=n}\{X(i,j) = D\}$ is the sum of all the occurences of a dry day, i in years j=1 to j=n

The same applies for P_{WW}

4.2.2 The fitting of the probability model

The fitting of the probability model in PT has been done using a five days moving average approach and later interpolated to give daily values. This method unlike the parametric approach method simply smoothes the time series of the observed probabilities to obtain a curve, which describes the observed probabilities. The reason is that, in most case the time series of the observed probabilities will not be smooth. As an example, time series of observed probabilities as well as their smoothed first order probabilities using a 5 days moving average based on data from Arusha site are given in figures 4.1 and 4.2 below: -.

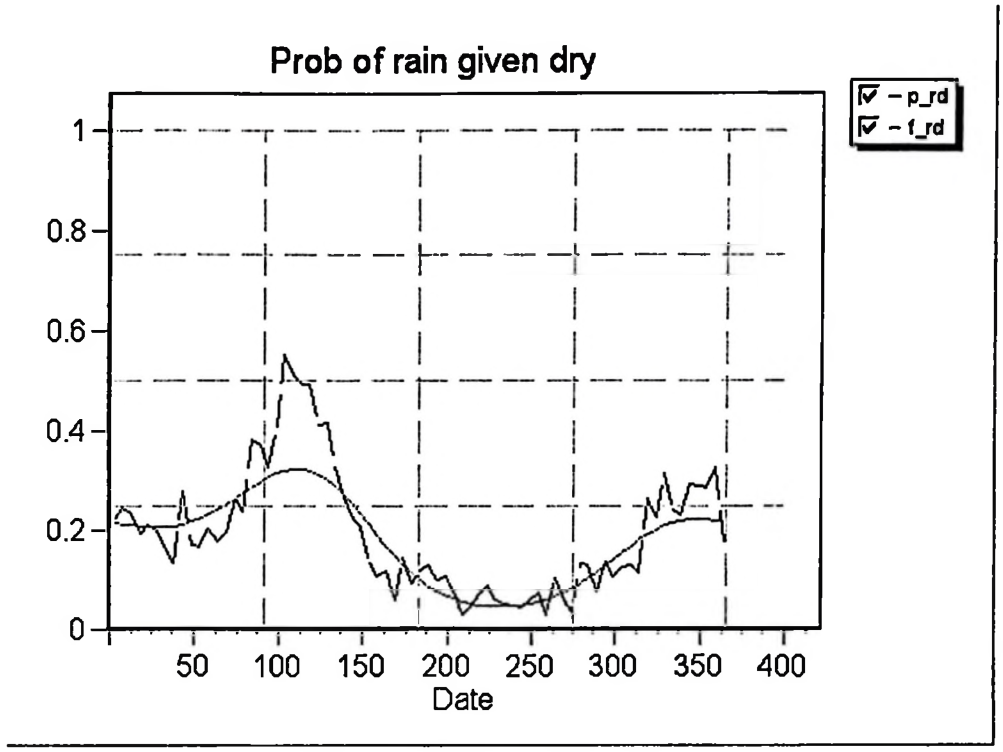


Figure 4.1

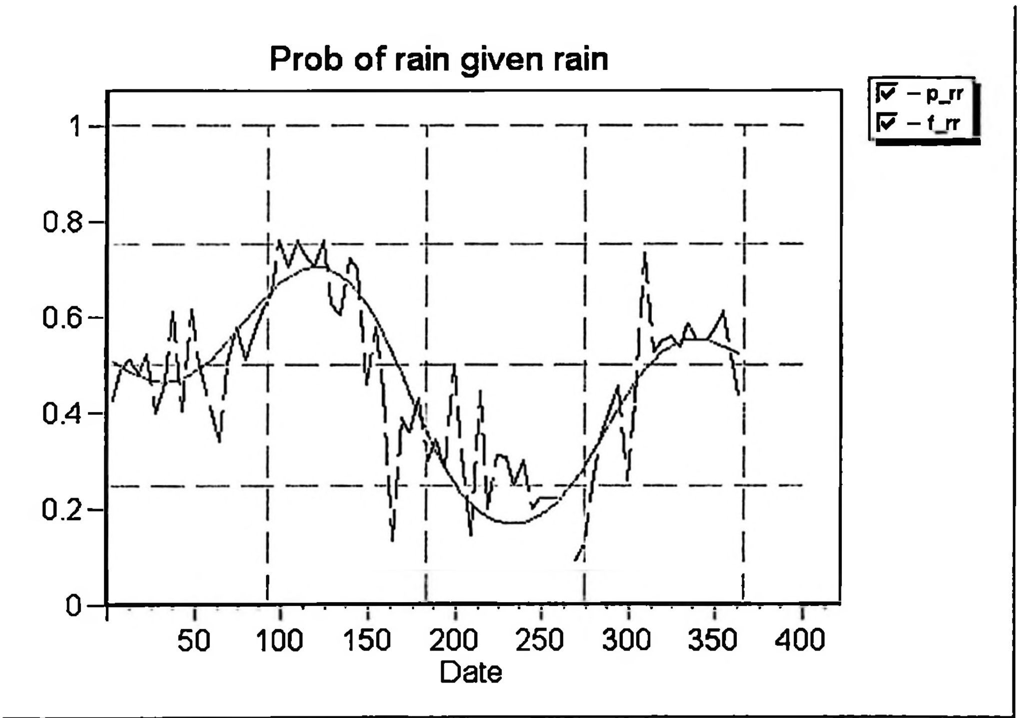


Figure 4.2

The graphs above indicate the observed probabilities at Arusha site in Tanzania. The observations confirm the normal rainfall patterns in Tanzania whereby rainfall is heavy during March to April and during October to December. Furthermore, as it can be seen the observed probabilities are not smooth.

Young et al., (1996); describes the fitting of rainfall probabilities in PT model in the following way

" In most locations and particularly in arid and semi arid areas, the transition matrices of rainfall occurrence will change with time of a year. To fully characterise the system, therefore, a number of years of data are required in which every day of the year has at least one instance of the two transitions (i.e. ww and wd). In many cases the required quantity of data is not available. Even if it is, the resulting time series of each probability will often be 'spiky'.

To account for this, the probabilities need to be smoothed and interpolated. Richardson (1981) used Fourier series to archive this. However, because of the need for **simplicity**, and in order to match **historical patterns** as closely as possible, the PT Climate Generator uses a simple five-day moving average followed by Lagrange interpolation to smooth the data".

There exist a number of alternatives to the use of a moving average. Apart from Richardson (1981), there has been quite an extensive work by a number of statisticians who have used different methods in modelling rainfall probabilities. Their methods are based on parametric approach using the modern statistical theory of "generalised linear model" which has been in place before the PT model was formulated. Stern et al., (1982), Chineke(1993) and Chandra et al., (1986) provide just a few examples of how rainfall probabilities can be modelled using modern statistical methods approach. The use of a "generalised linear model" will be discussed in the next chapter.

4.3 Modelling the rainfall amounts.

The rainfall amount X, being a positive quantity, is assumed to be a random variable following the gamma distribution, with parameters, α and β where

$f(x) = \beta^{\alpha} x^{\alpha-1} e^{-\beta\alpha,}$ x. >0. Where α is the shape parameter, which is assumed to be constant through out the year.

The mean and variance of the distribution are respectively, α/β and α/β^2. When α=1, the distribution is an exponential distribution, which is a special case of the gamma distribution.

The use of the gamma distribution in PT seems to be synonymous with the practice by most of the statisticians when modelling the rainfall amount. Stern (1979) also used the gamma distribution in modelling rainfall amount at Samaru Nigeria

The advantages upon using the gamma distribution is due to its closeness to two famous distributions in statistics, the exponential and the normal distribution. When α=1, the distribution becomes an exponential distribution and when α is 5, the distribution approaches the normal distribution. This means that, the distribution can be fitted to different frequency distributions of the rainfall amounts.

Estimation of the two parameters α and β is done by using the method of maximum likelihoods, basing on the historical data. These estimates are usually complex and are computationally, time consuming.

The, computationally easy method for estimation would have been the, method of moments in which $\hat{\alpha} = 4m_2^3 / m_3^2$ and $\hat{\beta} = \frac{1}{2} m_3 / m_2$ where m_2 and m_3 are the second and the third moments from the mean. However, estimates from the "methods of moments" are less precise than those from maximum likelihood estimates.

Hence the decision to use the maximum lilkihood estimates is appropriate. The maximum Likelihood estimates in PT are given approximately with the following formulae. Young et al., (1996)

$$\hat{\alpha} \approx Y^{-1}(17.79728 + 11.968477Y + Y^2)(8.898919 + 9.059950Y + 0.99775373Y^2)$$

where $Y = In(\overline{x}) - \left(\frac{1}{n} \sum_{i=1}^{i=n} In(x_i) \right) = \ln \left(\frac{arithmeticmean}{geometricmean} \right)$

and $\hat{\beta} = \frac{\overline{x}}{\hat{\alpha}}$

In PT, estimation process is done by considering the whole of monthly observations as a single point and later interpolated using the Lagrange method to give the daily values of

α and β. This process is done in an attempt to attain accuracy, which may not be attained if each of the daily historical observation is considered as a single point.

4.4 Statistical issues of interest in the generation of rainfall data using the PT

Following the discussions in the previous sections, we have some statistical issues to raise with regards to the methodology used in PT

The PT uses first order Markov chain to model the chances for occurrence of rainfall. The question of interest would be, does the assumption of the first order Markov chain, really holds true, in areas under the study? Given that it is true that Markov chain theory is applicable, could there be a possibility that, other higher orders fits better to the chance of rain than the first order? More generally, could it be that, the Markov chain model is as such not applicable in the given sites?
It is a common practice in statistics to use the theory of generalised linear model when modelling the proportions as in this case. But in PT, simply a 5 days moving average, is used. What can be the improvement, when the generalised linear model theory is used?

The questions raised above are addressed again in chapter five, when an alternative simulation method is considered for rainfall data.

CHAPTER 5:THE METHODOLOGY FOR ANALYSIS OF RAINFALL DATA.

5.1 Introduction

In chapter 4 we saw that PT modelled the probability of rain assuming a first order Markov chain using a five day moving average approach. In chapter 3 we demonstrated that the resulting model were not adequate. At the end of chapter 4 we described some of those areas in PT methodology, which need further investigation. In this chapter we look at the generalised linear model theory, which is briefly reviewed in the following sections. The review is based on Collett (1991), pages 53-73 and on lecture notes.

5.2 Introduction to generalised linear model theory

The generalised linear model theory is an extension of the general linear model theory. One of the key assumptions of the general linear model is that the response variable needs to follow the normal distribution. This assumption does not always hold with some response variables. The generalised linear model theory is used in fitting linear models, and so is useful when the assumption of normality is not appropriate. It applies to response variables, which fall under the family of the exponential distributions of which the normal distribution is one. Some other distributions which belong to this family, are the binomial distribution and the gamma distribution. In this study the response variables are chances of rain in a day and the amount of rainfall in a day, following under the binomial and gamma distributions respectively. Therefore the use of the generalised linear model theory is justifiable.

5.3 The Model for the chances of rain

In fitting a generalised linear model to the chances of rain we first transform the response variable, which is in the range of (0,1) to a range $(-\infty,\infty)$. There are three different types of transformations that can be made. These are logistic, probit and complementary log log transformations. The logistic transformation is the most used because of its computational efficiency.

5.3.1 The logistic regression

For logistic regression this transformation is a logistic transformation of success probability, p_i, which is the probability of having rainy in a day. This transformation is given by the general expression:

$\text{Logit}(p_k) = \log\left(\frac{p_j}{1-pj}\right) = \sum_j \hat{\beta}_j x_{ji}$ where $\sum_j \hat{\beta}_j x_{ji} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_j x_{ji}$ In my case, $\sum_j \hat{\beta}_j x_{ji} = g_i(t) = a_{i0} + \sum_{k=1}^{m} [a_{ik} \sin(kt') + b_{ik} \cos(kt')]$, because I will use Fouries series to smooth the proportions of rainy days.

Having estimated the value of the linear systematic component of the model denoted by $\hat{\eta} = g_i(t) = a_{i0} + \sum_{k=1}^{m} [a_{ik} \sin(kt') + b_{ik} \cos(kt')]$, we can get the estimated fitted success probabilities given by $\hat{p}_k = \frac{e^{\hat{\eta}}}{1+e^{\eta}}$.

The use of generalized linear model theory is a common approach by a number of statisticians. According to Stern (1982), the use of this theory in modelling the rainfall probabilities as a function of time (days) is motivated by the following reasons.

1. The observations x are from a distribution which is a member of the exponential family. The normal distribution, the gamma and the binomial distributions, which can be used, are both members of the exponential distribution.

2. The mean number of x, i.e. μ, can be related to the explanatory variables by a link function h, so that μ=h (TB), where T is a matrix of " explanatory variables " and β is the vector of unknown parameters. The parameters β are then estimated by solving the least-square equations $T'T\hat{\beta} = Tx$,

   The function g (t), which defines the expected number of wet days in a given season, can take a number of forms ranging from constant function, quadratics, and higher order polynomial up to Fourier series. As a matter of fact, g (t) rarely assumes a constant function. Stern (1982) gives the following account regarding the choice of g (t).

", Probabilities that vary continuously with time are more attractive. The simplest are polynomials; $g_i(t) = \sum_{k=0}^{m} a_{ik} t^k$. Fourier series may also be used, in which case $g_i(t) = a_{i0} + \sum_{k=1}^{m} [a_{ik} \sin(kt') + b_{ik} \cos(kt')]$, where $t' = \pi(t-183)/183$."

5.3.2 Test for the goodness of fit of a model.

After a model is fitted, it is usually the practice to assess whether, it is has effectively described the observed data or not. The effectiveness of the model in describing the data is referred to as its goodness –of- fit. On the other hand when the model is not effectively describing the observed data is termed as lack of fit. To assess the goodness of fit of a model we have to compare the fitted response values against the observed response values. There are several methods, which can be used to assess goodness-of-fit of a model, the most common one being the use of the "deviance" which is discussed below.

5.3.3 Deviance

Deviance is defined as D=-2log($\hat{L}_c / \hat{L}_f$)=-2(log $\hat{L}_c$ -log $\hat{L}_f$)

Where $\hat{L}_c$ is the maximised log-likelihood under a current model.(fitted model)

$\hat{L}_f$ is the maximised log-likelihood of an alternative baseline model, know as full model. Upon substitution of the appropriate livelihoods, we finally get

$$D = 2\sum_k \left[y_k \log\left(\frac{y_k}{\hat{y}_k}\right) + (n_k - y_k)\log\left(\frac{n_k - y_k}{n_k - \hat{y}_k}\right)\right]$$

Where

n_k is the observed binomial total at the kth day of a year

y_k is the observed number of rainy days

$\hat{y}_k$ is the estimated number of rainy days

As it can be seen, the deviance basically compares the observed value y_i
and the corresponding fitted value y_i. The deviance statistic follows under chi-squire distribution with (m-p) degrees of freedom where m is the number of binary observations and p is the number of unknown parameters in the current model. To determine whether a model is satisfactory we compare the value of the deviance with the percentage points of the chi-squared on (m-p) degrees of freedom. If the value of the deviance is larger than the chi-square percentage points then the model is judged to be unsatisfactory.

5.3.4 Comparing linear logistic models

When one model contains terms that are additional to those in another, the two models are said to be **nested**. The deviance can as well be used to assess the difference between two nested models. This is done by looking at the change in deviance due to an addition of extra term to another model. In my case I will use this concept for two purposes.

1.To determine the number of harmonics to be used in a model.

For example, the model $g_i(t) = a_0 + a_1 \sin(t) + a_1 \cos(t)$, which has only one harmonic, is nested in the model $g_i(t) = a_0 + a_1 \sin(t) + a_1 \cos(t) + a_2 \sin(t) + a_2 \cos(t)$, which has two harmonics. When the two models are fitted to the same set of data, the difference in their deviances compared to the corresponding percentage chi-value at an appropriate degrees of freedom will tell whether the second harmonic is really needed or not. This is idea will be used to determine the number of harmonics for a given probability model.

2.To compare the zero, first and second order Markov chain probability models

The response values are proportions of the three orders of the Markov chains, which depend on the same explanatory variables, i.e. days of a year. The models for the three orders are not nested. In order to use the deviance we have to find a way of making some models nested to others, by introducing some factors (dummy variable).

To clearly understand this point considers the table below.

Table: 5.1

Date	rdd	rdr	rrd	rrr
1	2	2	2	1
2	3	1	1	1
3	5	2	2	1
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
366	3	1	2	0

The table indicates observed frequencies on the four states of rain of a second order Markov chain basing on Arusha historical data (1970-2000)

Notice that:

- rdd is a number of times a day is wet given the two previous days were both dry days
- rdr is a number of times a day is wet given the previous day was a dry day preceded by wet day.
- rrd is a number of times a day is wet given the previous day was a wet day preceded by a dry day
- rrr is a is a number of times a day is wet given the last two days were both wet days

(Please refer to appendix 7, for other details associated with each of the given column)

The table indicates that, the model for the chance of rain can be explained using four different curves f_rdd, f_rdr, f_rrd and f_rrr associated with the four states of a second order Markov chain.

We can now introduce factors "first" to make the first order model nested to the zero order model and the factor "second" to make the second order model nested to the first order model by appending the four columns of table5.1 in the order rdd, rdr, rrd, rrr as shown in column six of table5.2

In table 5.2 below, factor "second" has got four levels, which means the whole set of observations can be explained using four curves. On the other hand the factor "first" has got two levels, which are dry, and rain. With the level "dry" we assume that, f_rdd and f_rdr are explained by one curve whereas with the level " rain" we assume that, f_rrd and f_rrr are one curve as well. In other words the whole set of observations can be explained using two curves only.

The factor "zero" has only one level "rain" meaning that only one curve can explain the whole set of observations.

Table: 5.2

Number	Date	Zero	First	Second	Response
1	1	rain	dry	rdd	2
2	2	rain	dry	rdd	3
3	3	rain	dry	rdd	5
.	.	rain		.	.
.	.	rain		.	.
.	.	rain		.	.
366	366	rain	dry	rdd	3
337	1	rain	dry	rdr	2
338	2	rain	dry	rdr	1
339	3	rain	dry	rdr	2
.	.			.	.
.	.			.	.
.	.			.	.
732	366	rain	dry	rdr	1
735	1	rain	rain	rrd	2
736	2	rain	rain	rrd	1
737	3	rain	rain	rrd	2
.	.		.	.	.
.	.		.	.	.
.	.		.	.	.
1098	366	rain	rain	rrd	2
1099	1	rain	rain	rrr	1
1100	2	rain	rain	rrr	1
1101	3	rain	rain	rrr	1
.	.		.	.	.
.	.		.	.	.
.	.		.	.	.
1464	366	rain	rain	rrr	0

We can now make comparison of the three Markov chain models through the following steps: -

1. Fit the zero order model by considering

$$g_i(t) = a_{i0} + \sum_{k=1}^{m}[a_{ik}\sin(kt') + b_{ik}\cos(kt')],$$

2. Fit the model $g_i(t) = first + a_{i0} + \sum_{k=1}^{m}[a_{ik}\sin(kt') + b_{ik}\cos(kt')]$, and compare the change in deviance. This will tell whether a first order model (two curves) is an improvement to a zero order model. When the fitting was done, five harmonics were found to be appropriate in this model.

3. If the first order model is an improvement to the zero order model, we then fit the following model,

   $g_i(t) = second + a_{i0} + \sum_{k=1}^{m}[a_{ik}\sin(kt') + b_{ik}\cos(kt')]$, and compare the change in deviances. This will tell whether the second order model (four curves) is an improvement to the first order. Again when fitting was done only **five** harmonics were appropriate.

4. Alternatively we will consider fitting each of the models in 1 and 2 above with an interaction term between the factor and the harmonics. When this was done only the **first harmonic** proved to be significant in both cases. Therefore we did not need other higher order interactions.

All these processes will be done automatically with the use of Genstat macros attached as appendix 4 of the project. The results of the analysis are presented in chapter six

5.4 The model for the rainfall amounts

As explained before, the rainfall amount will be fitted to the gamma distribution since the gamma distribution is a member of an exponential family of distributions, the generalised linear model theory will also be used. All other procedures, which applied to the determination of a probability model, will also apply. However this time there will be no comparison of regression curves because the order of amount that will be considered is one and that is zero order.

The gamma distribution of X, for rainfall amount will be given by

$f(x)=(k/\mu(t))^k x^{k-1} \exp[-kx/\mu(t)]/\Gamma(k)$ where $E(X'(t))=\mu(t)$ and the time dependence will be taken to be of the form $\log(\mu(t))=g(t)$ where is g(t) is a function linear in unknown parameters. k is the shape parameter of the distribution.

5.5 Problems for investigation

It would be appropriate at this stage to again summarise the problems under investigation of whose results will be reported in the next chapter. These are: -

1 Do higher order Markov chains fit to Tanzanian data than the first order?
2. If 1 is true how is the order of Markov chain reflected in terms of the standard deviations of the generated data.

3. Again if 1 is true, what is its implication on the standard deviations of the generated data against those of the historical data?

4. Is the use of modern statistical methods (generalized linear model theory + Fourier series smoothing) in fitting the model for the chances of rain and the model for the daily rainfall amount of any improvement to the Model?

5. Again If 1 is true, how then does it improve the use of the model in predicting some important events of interest to the farmers? The events are specifically

    a. 20, and 80 percent points of the monthly total rainfall distribution
    b. The distribution of the possible sowing dates
    c. The distribution of the possible sowing dates when the dry spells is considered.

    d. The distributions of the chances of dry spells within the given periods of a year (refer to chapter three)

5.5.1 How to investigate

In order to investigate the problems above, the following needs to be done

Question one

1. Comparison of regressions curves for the zero, first and second order Markov chains using the theory of generalised linear models. From the comparison of the curves we will know whether the first order Markov chain is better than the zero order and also whether the second order markov chain is better than the first.

Question 2, 3, 4 and 5

Before answering question two to five the following need to be done.

1. Fit the zero, first and second order probability models

2. Fit the zero order model for the mean daily rainfall amounts using the, Zero, first and second order probability models.

3. Simulate daily rainfall amounts using (Instant+v3.04 program) for the zero, first and second order probability models with a common random number seed

Having done the three steps above,
The answer to the second question will be obtained by comparing the Standard deviations of the mean monthly totals of the Zero, first, and second order generated data relative to those of the historical data.

The answer to the third question will be obtained by comparing the standard deviations of the mean monthly totals of the second order generated data against those of the PT generated data relative to those of the historical data.

The answer to the fourth question will be obtained by comparing the standard deviations of the mean monthly totals of the PT generated data against those of the first order generated data relative to those of the historical data.

As for the last question, the listed events will be explored from PT generated data and second order generated data and compared relative to the historical data.

CHAPTER 6:RESULTS ON THE ANALYSIS OF RAINFALL DATA

6.1 Comparison of higher order Markov chain probability models to the first order model

Using the method described in section 5.2.4, zero order Markov chain was fitted to the data with 5 harmonics. This model was extended to test whether a first order or second order Markov chain was more appropriate. The results of the comparison are shown in table 6.1 below:-

Table 6.1: Changes in deviances from the zero order to second order

	Zero order	First order		Second order	
	1 curve	2 parallel curves	2 separate curves	4 parallel curves	4separate curves
Deviance	2323.	1607	1596.97	1522.7	1515.13
Degrees of freedom	1347	1346	1344	1344	1338
Reduction in deviance		716.6	10	84.21*	7.6

* Change in deviance from two parallel curves

From the above table it is clear that,

(a) The first order probability model is preferable to the zero order models because the change in deviance of 716 is significantly higher than the corresponding percentage chi value at 1 d.f which is 3.84.This first order model corresponds to the model fitted in PT.

(b) The second order probability model fits better than the first order model because the change in deviance of 84.2 is significantly higher than the corresponding percentage chi value at 2 d.f which is 6. This confirms the fact that a higher order Markov chain model may be needed.

These two facts can as well be reflected in figure 6.1 and 6.2 below: -

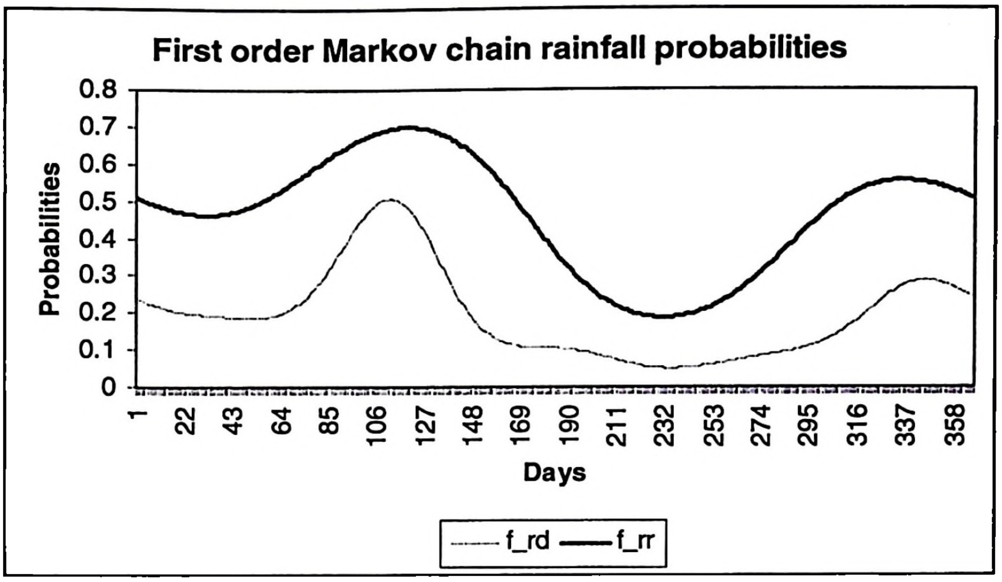


Figure 6.1

The two curves above for the first order probabilities are quite far apart (parallel). This is indicates the need for a first order probability model to the zero order model.

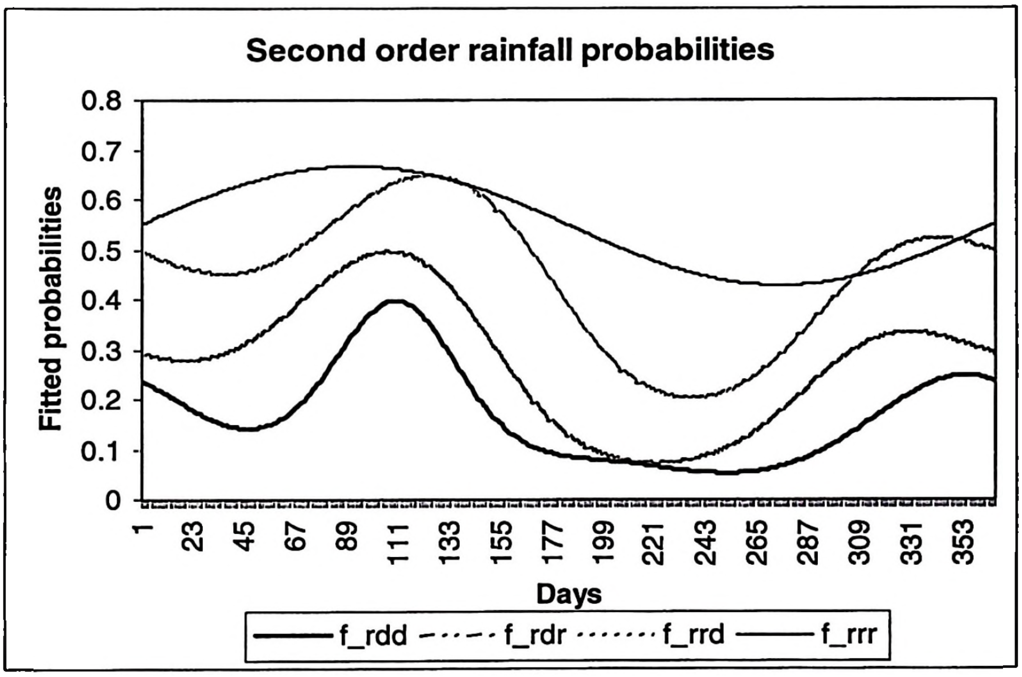


Figure 6.2

The four curves above for the four states of the second order Markov chain appears not to be parallel. Probably there was a technical problem during the fitting. Further investigation might be needed.

Having determined that, a second order Markov chain model is required, further modelling was carried out to determine the appropriate number of harmonics for the zero, first and second order probability models. The reason for doing so was to compare the standard deviations of the zero, first and second order generated data. My hypothesis as explained in section 5.5 is that the standard deviations of the generated data improves with the order of Markov chain. These results are attached as an appendix 10 of the project.

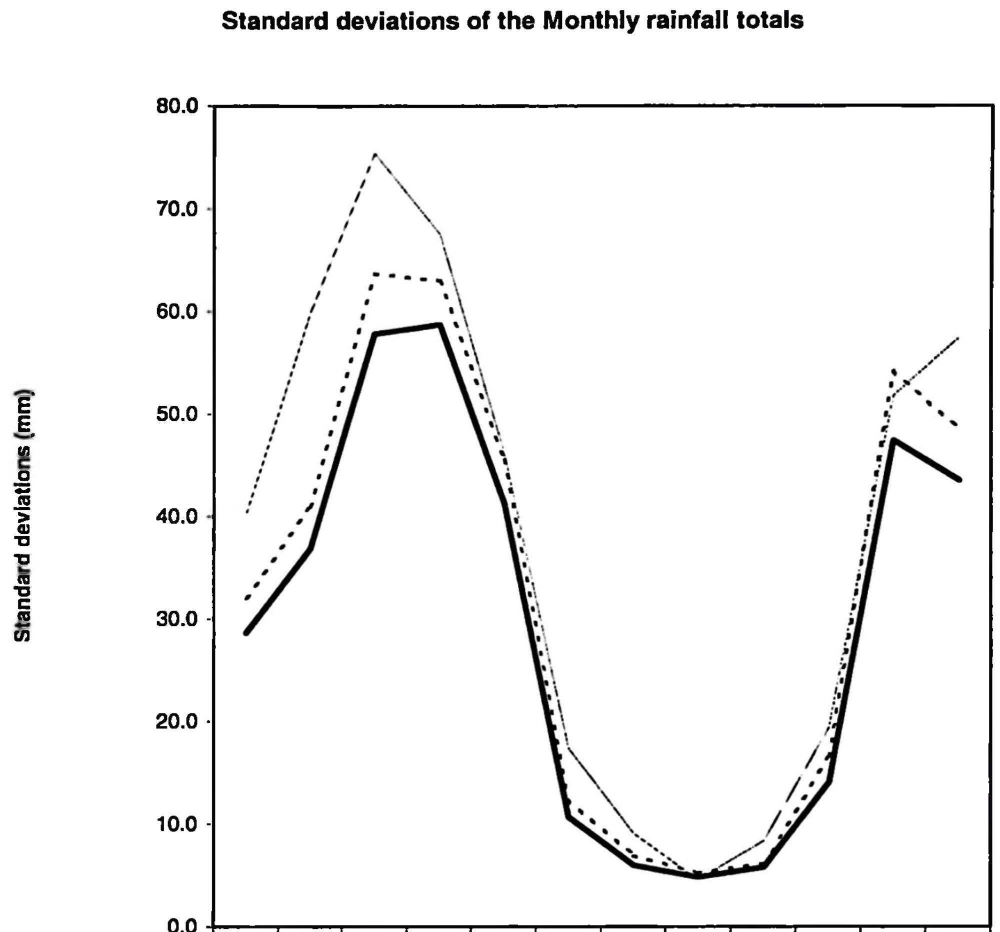
6.2 Using the fitted probability models and the rainfall amounts models

As I wished to compare the standard deviations of the zero, first and second order generated data, I had to simulate the rainfall amounts The zero order daily rainfall amounts for 100 years for each of the zero, first and second orders Markov chains were simulated based on their respective probability models and the fitted gamma distributions, using a random number seed of 4434 in Instat (V3.04). The objective of using a common random number seed was to have a meaningfully comparison among the three orders of Markov chain. The results of the comparison are reported in the next subsection.

Further to this comparison, I will also provide answers to other questions raised in section 5.5

6.2.1 Comparison of the standard deviations with the three orders of Markov

chain



	1	2	3	4	5	6	7	8	9	10	11	12
Zero	28.6	36.9	57.8	58.7	41.3	10.7	6.0	4.8	5.8	14.1	47.4	43.5
First	32.1	41.0	63.7	63.0	45.5	12.4	6.9	5.2	6.2	16.6	54.2	48.5
Second	40.2	60.0	75.4	67.5	46.1	17.4	9.1	4.5	8.4	19.4	51.9	57.5

Figure 6.4

From figure 6.4, the comparison of the three orders indicates that;

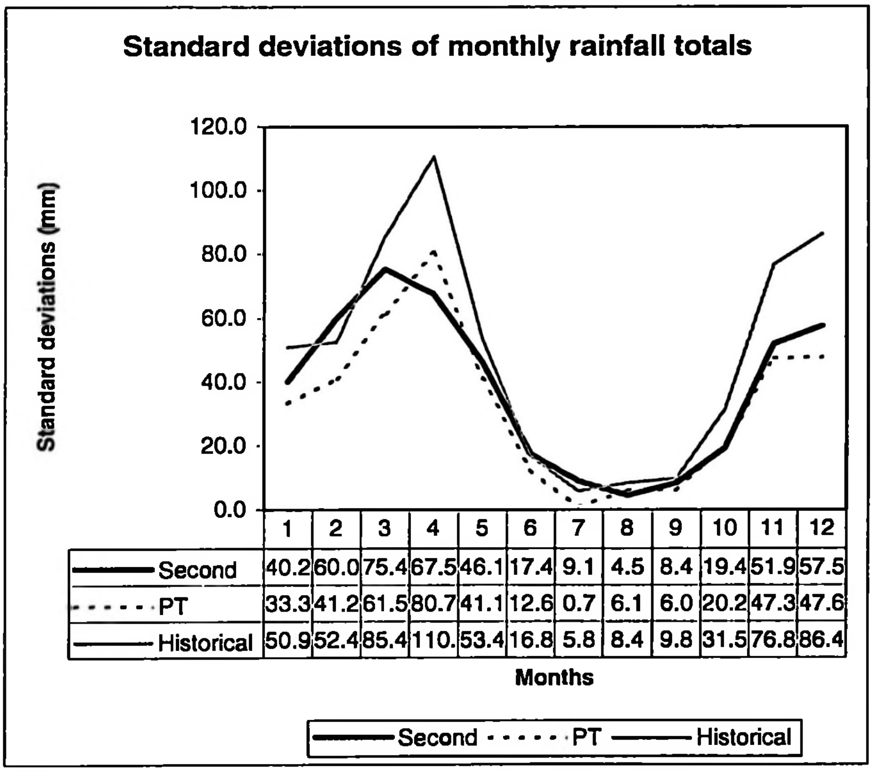
(i) Standard deviations of the first order-generated data are consistently higher than those of the zero order generated data.

(ii) Standard deviations of the second order-generated data are also consistently higher than those of the first order generated data except in August.

(iii) The model by the second order seems to be not very good when compared to the zero and first order. The peak in its curve is seen to be on March instead of April which is the month with heaviest rain in Tanzania. The weakness is noted and probably it is because of what we saw in figure 6.2 of the previous section. Further investigation is needed

This is a further indication that higher order models may be needed at this site. That is, there is a possibility that, as we move from lower order to higher order we get a more realistic model for the rainfall probabilities.

6.2.2 The implication of the higher order Markov chain model on standard deviations of the generated data against those of the historical data.



	1	2	3	4	5	6	7	8	9	10	11	12
Second	40.2	60.0	75.4	67.5	46.1	17.4	9.1	4.5	8.4	19.4	51.9	57.5
PT	33.3	41.2	61.5	80.7	41.1	12.6	0.7	6.1	6.0	20.2	47.3	47.6
Historical	50.9	52.4	85.4	110.	53.4	16.8	5.8	8.4	9.8	31.5	76.8	86.4

Figure 6.5

The figure above shows a comparison of the standard deviations of the PT generated data against those of the Second order-generated data.

The standard deviations of the second order Markov chain generated data are consistently higher than those of the PT generated data in all the months except in April and August. There seems to be a slight improvement of the standard deviations when the second order Markov chain model is used instead of the PT generated data. This slight improvement could be attributed to a number of reasons, which need further investigation. One of the most obvious reason could be the fact that, we stopped at the second order without knowing whether other orders are needed. Possibly third order can do better than this.

6.2.3 Comparison of the PT generated data against the first order-generated data basing on the standard deviations.

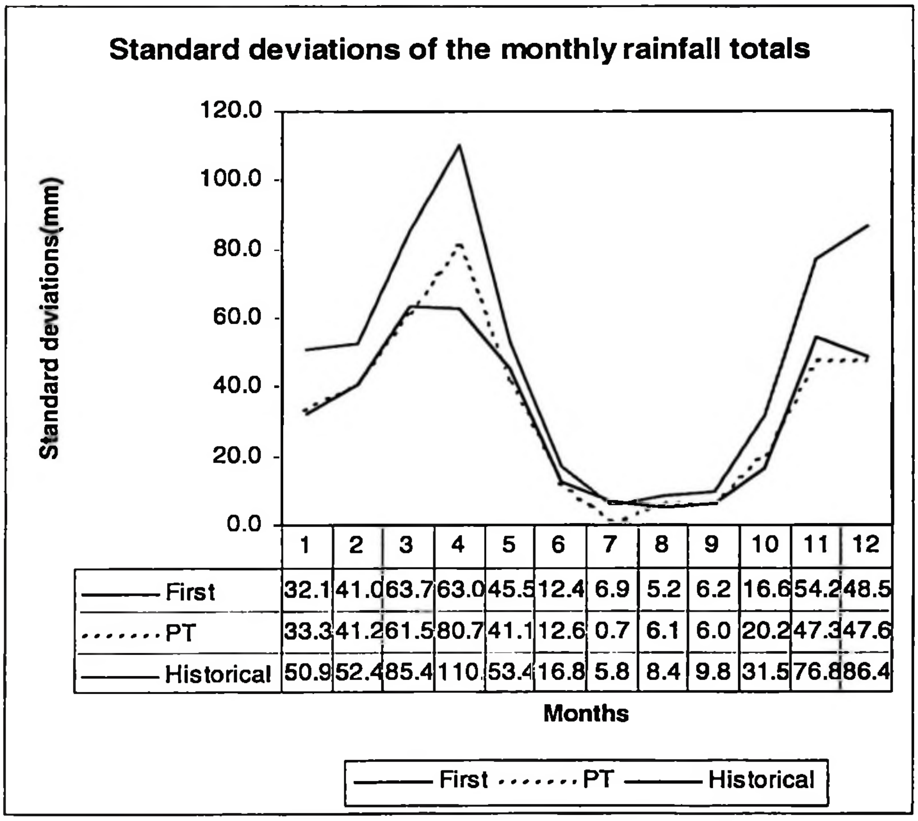


Figure 6.6

The figure above shows a comparison between the first order-generated data and the PT generated data on standard deviations. The aim is to test the performance of the smoothing technique in PT against "the use of the modern statistical theory of generalised liner model" The standard deviations of the first order Markov chain generated data are more or less the same as those of the PT generated data. As it can be seen their graphs frequently cross each other over time. This fact can further more be seen by looking at the values of the two data sets from the attached table. Their values are consistently very close except on April. This fact is a message that the PT non-parametric approach is in no way less good than the modern theory of generalised linear model. The reason as to why the PT model is not modelling well the historical data may be attributed more to other factors than the smoothing technique.

6.2.4 How does the second order Markov chain improve the use of the model in predicting some important events of interest to the farmers.

The events are defined in the same way as in section 3.2.2 of chapter three

1. **The 20% and 80% points of the monthly rainfall totals.**

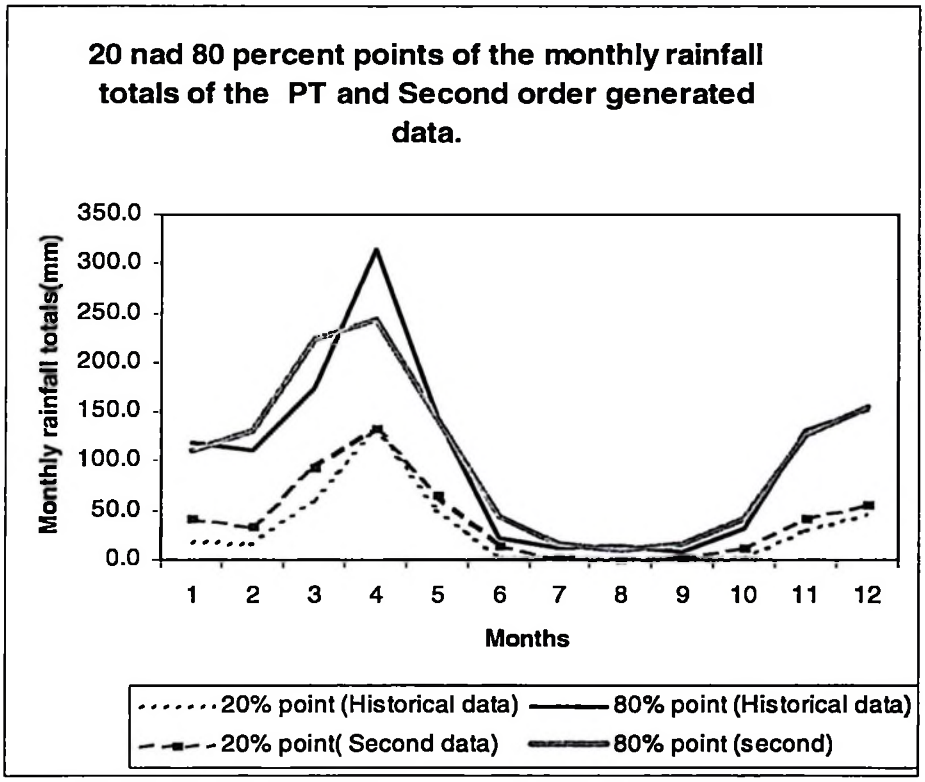


Figure 6.7

From the above graph we can see that, the gap between 20 and 80 percent points of the historical data is much higher than that of the Second order generated data As a matter of fact this event is not well predicted by the second order-generated data.

2. The distribution of the possible sowing dates

(Using the same criteria as in chapter three)

Table 6.8

Date of planting	Historical	PT	SECOND
10 Percentile	3rd January	1st January	1st January
Median	18th January	14th January	13th January
90 Percentile	2nd March	9th February	14th February
Mean	26th January	17th January	19th January
Standard deviation	23	16	19

The table indicates that there is a similar distribution of sowing dates over the years. There appears to be a very slight advantage of the second order-generated data over the PT generated data.

3. Combining the dry spells and sowing dates

(Using the same criteria as in chapter three)

Table 6.9

Date of planting	Historical	PT	SECOND
10th Percentile	3rd January	1st January	2nd January
Median	17th February	15th January	6 February
90th Percentile	31st March	15th February	15th March
Mean	17th February	21 th January	7th February
Standard deviation	35	22	28

After considering the 10 days dry spells, the distribution of the sowing dates basing on the second order-generated data is closer to that of the historical data than to the PT generated data. With the historical data most farmers will be planting around 17 February whereas with the second order generated data most farmers will be planting around 7th February, a gap of only 10 days period.

4. The distribution of the dry spells

Below are the probability distributions of the chances of dry spells of 7 and 10 days at 24 different possible sowing dates beginning from 1 of January.

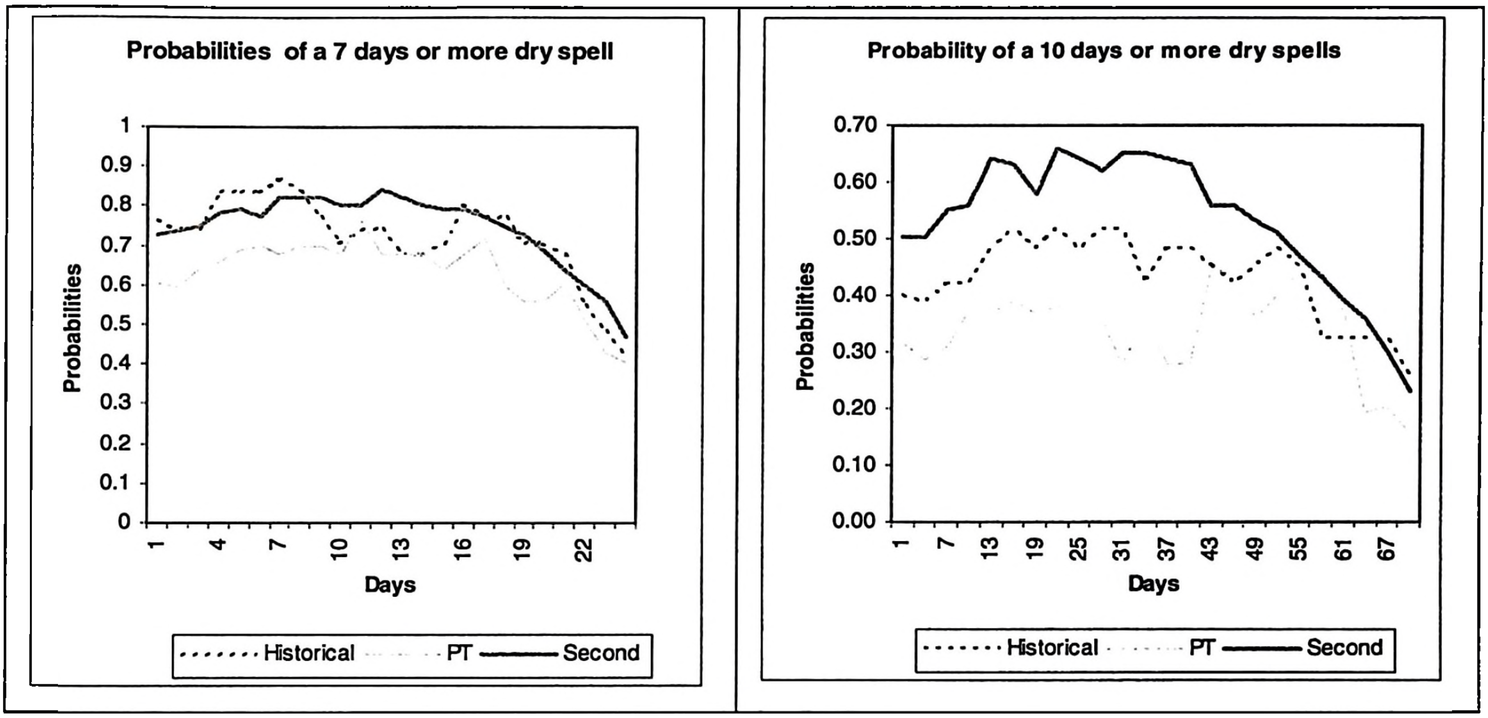


Figure 6.8

The graphs indicate that the second order generated data and the historical data give the same risks of dry spells of a 7 days or more where as with a 10 days dry spell the second order generated data overestimates the risks of dry spell. This situation is relatively better compared to the PT generated data, which underestimates the risks for **both** a 7 days and a 10 days dry spell.

CHAPTER 7:GENERATION OF MAXIMUM TEMPERATURE, MINIMUM TEMPERATURE AND RADIATION IN PT

7.1 Introduction

In chapter four we looked at the methodology of rainfall data generation in PT, to understand whether the assumptions behind it were adequate or not. In this chapter we look at the methodology used in PT for generation of maximum temperature, minimum temperature and radiation. Again the description of the methodology is based on explanations by Young et al., (1996).

7.2 The model for the three variables

According to Young et al.,(1996) the three variables are assumed to be dependent upon each other and can be determined through a stochastic process. In modelling the three variables, a multivariate distribution of the three of them is assumed.

The three time series are usually not stationery and need to be smoothed to apply conventional statistical methods, which deal with stationery time series. In PT, the daily observations for each variable are smoothed by subtracting each of the observation by the appropriate daily mean and divided by the standard deviation as shown below:-The notation used is the same as used by Young et al .,(1996)

$$\chi_{p,i}(j)=\frac{X_{p,i}(j)-\overline{X}_i^0(j)}{\sigma_i^0(j)} \quad \text{or} \quad \chi_{p,i}(j)=\frac{X_{p,i}(j)-\overline{X}_i^1(j)}{\sigma_i^1(j)}$$

Where $\chi_{p,i}(j)$ = Residual component of variable j on day i, year p

$\overline{X}_i^0(j)$ = Mean of variable j on dry day i

$\overline{X}_i^1(j)$ = Mean of variable j on wet day i

$\sigma_i^0(j)$ = Standard deviation of variable j on dry day i

$\sigma_i^1(j)$ = Standard deviation of variable j on wet day i

$X_{p,i}(j)$ =Value of variable j on day i, year p.

7.3 Generation process

Having obtained the historical mean and standard deviation for each variable on each day of a year, the generation process involves, the generation of residuals which are then multiplied by their respective standard deviations and added to their appropriate means. For example, if e_i is a residual in day i of the year, for a variable x_i, then the value of this variable in that particular day is $\overline{x_i} + e_i * \sigma_{x_i}$ where $\overline{x_i}$ and σ_{x_i} are the mean and standard deviation of the variable in that day. The mean and the standard deviation to be used would depend on the wet/dry status of that particular day. If the day is dry, the mean as well as the standard deviation for dry days will be used and the vice versa.

Young et al.,(1996) explain that, the residuals for the three variables can be generated through a first order autoregressive linear process given as $\chi_{p,i}(j) = A\chi_{p,i-1}(j) + B\varepsilon_{p,i}(j)$

where $\chi_{p,i}(j)$=(3x1) matrix for day i, year p whose elements are residuals of maximum temperature (j=1), minimum temperature (j=2) and radiation (j=3)

$\chi_{p,i-1}(j)$ = The same as $\chi_{p,i}(j)$ except for day i-1

$\varepsilon_{p,i}(j)$ = (3x1) matrix of independent random varieties that are normally distributed (mean=0, variance=1).

A = (3x3) matrix

B = (3x3) matrix

$\sigma_i^0(j)$ = Standard deviation of variable j on dry day i

$\sigma_i^1(j)$ = Standard deviation of variable j on wet day i

$X_{p,i}(j)$ =Value of variable j on day i, year p.

7.3 Generation process

Having obtained the historical mean and standard deviation for each variable on each day of a year, the generation process involves, the generation of residuals which are then multiplied by their respective standard deviations and added to their appropriate means. For example, if e_i is a residual in day i of the year, for a variable x_i, then the value of this variable in that particular day is $\overline{x_i} + e_i * \sigma_{x_i}$ where $\overline{x_i}$ and σ_{x_i} are the mean and standard deviation of the variable in that day. The mean and the standard deviation to be used would depend on the wet/dry status of that particular day. If the day is dry, the mean as well as the standard deviation for dry days will be used and the vice versa.

Young et al.,(1996) explain that, the residuals for the three variables can be generated through a first order autoregressive linear process given as $\chi_{p,i}(j) = A\chi_{p,i-1}(j) + B\varepsilon_{p,i}(j)$

where $\chi_{p,i}(j)$=(3x1) matrix for day i, year p whose elements are residuals of maximum temperature (j=1), minimum temperature (j=2) and radiation (j=3)

$\chi_{p,i-1}(j)$ = The same as $\chi_{p,i}(j)$ except for day i-1

$\varepsilon_{p,i}(j)$ = (3x1) matrix of independent random varieties that are normally distributed (mean=0, variance=1).

A = (3x3) matrix

B = (3x3) matrix

The least square estimations of the parameters A and B, leads to $A = M_1 M_0^{-1}$ and $BB^T = M_0 - M_1 M_0^{-1} M_1^T$ where M_1 = lag 1 covariance matrix and M_0 = lag 0 covariance matrix. With the assumption that, the $\chi_{p,i-1}(j)$ series have a variance of 1 and the mean of 0 we have

M_0 given as
$$\begin{bmatrix} 1 & \rho_0(1,2) & \rho_0(1,3) \\ \rho_0(2,1) & 1 & \rho_0(2,3) \\ \rho_0(3,1) & \rho_0(3,2) & 1 \end{bmatrix}$$

And M_1 given as
$$\begin{bmatrix} 1 & \rho_1(1,2) & \rho_1(1,3) \\ \rho_1(2,1) & 1 & \rho_1(2,3) \\ \rho_1(3,1) & \rho_1(3,2) & 1 \end{bmatrix}$$

The equation "$\chi_{p,i}(j) = A\chi_{p,i-1}(j) + B\varepsilon_{p,i}(j)$" as given by Young et al., (1996) is an example of an autoregressive process. The theory on autoregressive process is briefly discussed in the next section.

7.4 An overview of an autoregressive model.

(This overview is based on Chapter three of Chatfield (1992))

The word auto refers to the dependence of errors in different time periods. In most time series, the successive observations are autocorreleted. This can be discovered by looking at the correlation among the errors in different time periods.

The autocorrelation between ε_t and ε_{t-k}, also known as serial correlation is given as $\rho_k = \frac{E(\varepsilon_t \varepsilon_{t-k})}{E(\varepsilon^2{}_{t-k})}$ where ε_t and ε_{t-k}, are errors in time t and time t-k respectively. If the processes are assumed to be stationery as is the case with PT, then, ρ_k simply becomes $\frac{E(\varepsilon_t \varepsilon_{t-k})}{E(\varepsilon^2{}_t)} = \rho_{-k}$. i.e $E(\varepsilon_t \varepsilon_{t-1}) = E(\varepsilon_t^2)$.

The reason that we study the autocorrelation of errors is because we want to ensure the fulfilment of one of the basic assumptions in the theory of ordinary least square estimation. The least square estimation which is often adapted in time series analysis require the errors in successive periods to be uncorrelated, an assumption which is rarely met.

Nevertheless, the estimation process can often take place irrespective of whether there is autocorrelation or not. The point though is, how should the linear autoregressive process be treated for a particular time series?

Generally a process $\{x_t\}$ is said to be an autoregressive process of order p if $X_t = \alpha_1 X_{t-1} + \ldots + \alpha_p X_{t-p} + Z_t$. This is like a multiple linear regression model, but X_t is regressed not on independent variables but on past values of X_t; hence the prefix 'auto'.

So for a first order we have p=1 in which case $X_t = \alpha_1 X_{t-1} + Z_t$ where as for a second order p=2 and $X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + Z_t$.

In PT a first order linear autoregressive process has been adapted, and this assumes that, If ε_t and ε_{t-1}, are the errors in time t and t-1 , then $\varepsilon_t = \rho\varepsilon_{t-1} + \eta_t$. The aim is to define a new set of errors in time t such that $\eta_t = \rho\varepsilon_t - \varepsilon_{t-1}$. With such a transformation we have a fulfilment of the assumption of independence of errors i.e $E(\eta_t \eta_{t-k}) = 0$, given that $0 < \rho < 1$. The other assumptions, which also hold true, are:

(i) $E(\eta_t) = 0$,

(ii) $\text{var}(\eta_t) = \sigma_\eta^2 = E(\eta_t^2)$

(iii) $E(\eta_t \eta_{t-k}) = 0$.

An estimate of ρ is given by the usual least square estimates as $\rho_1 = \dfrac{E(\varepsilon_t \varepsilon_{t-1})}{E(\varepsilon^2{}_{t-1})} = \dfrac{E(\varepsilon_t \varepsilon_{t-1})}{E(\varepsilon^2{}_t)}$ = cov(t, t-1)/var(t). It can generally be shown that, if we consider lag k, the autocorrelation between ε_t and ε_{t-k}, is given as $\rho_k = \dfrac{E(\varepsilon_t \varepsilon_{t-k})}{E(\varepsilon^2{}_{t-k})} = \dfrac{E(\varepsilon_t \varepsilon_{t-k})}{E(\varepsilon^2{}_t)} = \rho_1{}^k$ where ρ_1 is the autocorrelation between ε_t and ε_{t-1} (lag 1)

With the second order autoregressive process, the coefficients a_1 and a_2 can be estimated from the equation $X_t = a_1 X_{t-1} + ... + a_p X_{t-p} + Z_t$ by using the method of least square estimation. This leads to $a_1 = \rho_1(1-\rho_2)/(1-\rho_1^2)$ and $a_2 = \rho_1(\rho_2 - \rho_1^2)/(1-\rho_1^2)$ where ρ_1 and ρ_2 are the lag 1 and lag 0 first order autocorrelations. Notice that, first order autoregressive process assumes that $\rho_2 = \rho_1^2$. This means an expression for a_2 above, reduces to zero if the errors are to follow the first order autoregressive process.

7.4.1 Computation of ρ_1 an ρ_2

We had earlier given an expression for ρ_k as $\frac{E(\varepsilon_t \varepsilon_{t-k})}{E(\varepsilon^2_{t-k})}$.

This expression can be rewritten as follows for an ease of computation:

$$\rho_k = \left[\sum_{i=1}^{n} \left(x_{it} - \overline{x_t}\right)\left(x_{it-k} - \overline{x_{t-k}}\right)\right] \Big/ n s_t s_{t-k} \quad \text{Where}$$

n is the number of paired observations. (n=number of years of study)

x_{it} is the i[th] observation in time t where i=1 ,2, ..n and t=k, k+1, ...366

$\overline{x_{it}}$ is the mean of the n observations in time t.

$$s_t = \sqrt{\frac{1}{n-1}\left(x_{it} - \overline{x_{it}}\right)^2}$$

From the given formulae, it follows that, ρ_1 and ρ_2 will be given as

$$\hat{\rho}_1 = \left[\sum_{i=1}^{n} \left(x_{it} - \overline{x_t}\right)\left(x_{it-1} - \overline{x_{t-1}}\right)\right] \Big/ n s_t s_{t-1}$$

$$\hat{\rho}_2 = \left[\sum_{i=1}^{n} \left(x_{it} - \overline{x_t}\right)\left(x_{it-2} - \overline{x_{t-2}}\right)\right] \Big/ n s_t s_{t-2}$$

The formulae given above are assuming a non-stationery correlation structure over years contrary to what has been assumed in the PT methodology.

7.5 Statistical issues of interest in the generation of the three variables

Following the discussion on the methodology used in PT for generation of the three variables, we will have the following major statistical issues of investigation.

1. The assumption that residuals follow the first order linear autoregressive process. It is well known that, residuals in time series can take a number of stochastic process such as moving average or a mixture of the two Chatfield (1992). What is the justification on the use of a first order autoregressive process? Why not a second order autoregressive process?

3. The assumption that residuals in successive time periods are stationery.

CHAPTER 8:EXPLORING THE CORRELATION STRUCTURES IN TEMPERATURE DATA

8.1 Introduction

In chapter seven we have seen that the Richardson (1982) method as explained by Young et el., (1996) assumed the following:

1. Successive observations on the variables maximum temperature, minimum temperature and radiation have a first order linear autoregressive correlations structure.
2. The autoregressive residuals are stationary through years.

In this chapter we aim to explore the validity of those assumptions based on Arusha data (1970-2000) for maximum and minimum temperature. The Radiation data had a lot of missing values and is therefore omitted from the analysis. The statistics are calculated using the formulae given in chapter seven

8.2 Maximum temperature

8.2.1 Daily maximum temperature against time.

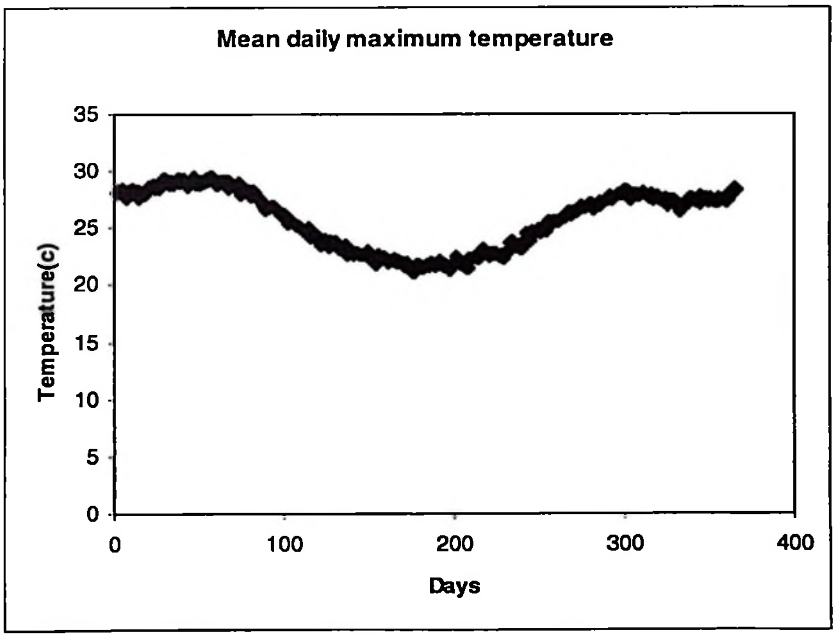


Figure 8.1

The figure above indicates the usual distribution of temperature in Tanzania whereby June is the coldest month.

8.2.2 Daily Variance of maximum temperature against time.

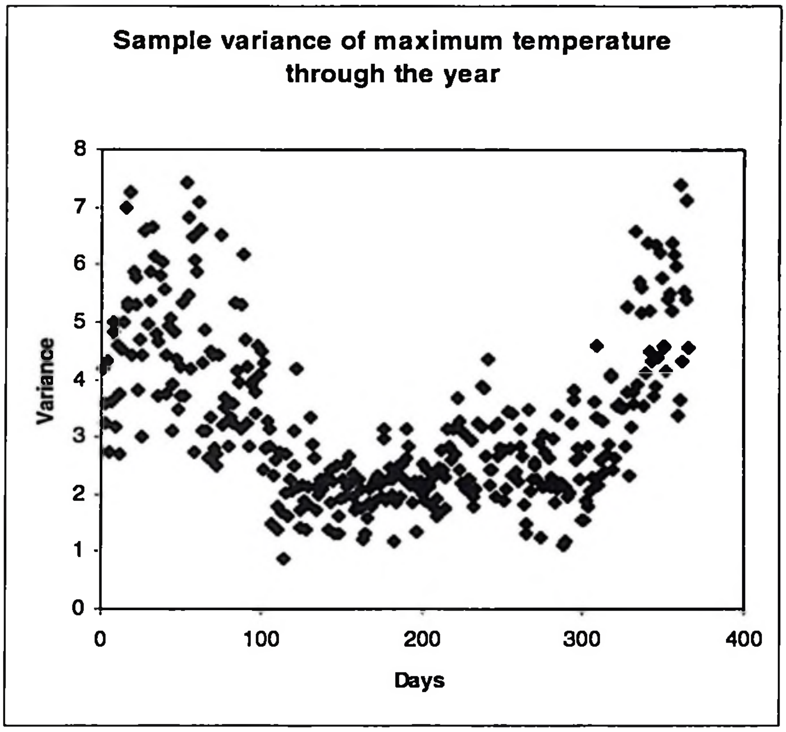


Figure 8.2

The figure indicates the distribution of variance of daily maximum temperature through out the year. The variance is less smooth than temperature over time. There is an extremely high variation in temperature between January and March as well as between October and December.

First order serial correlation

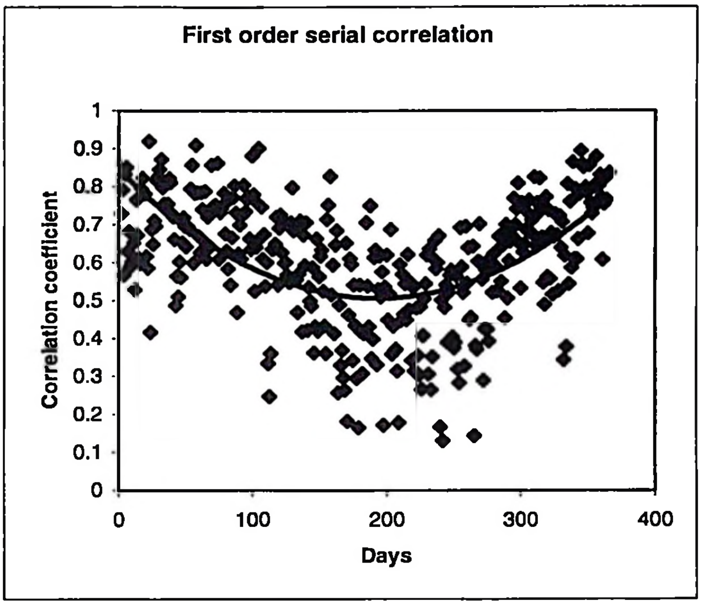


Figure 8.3

The graph above shows the first order autoregressive errors on maximum temperature through out the year. The residuals are certainly non-zero and non-stationery.

8.2.3 Second order serial correlation

(i) Graph of a1 against time

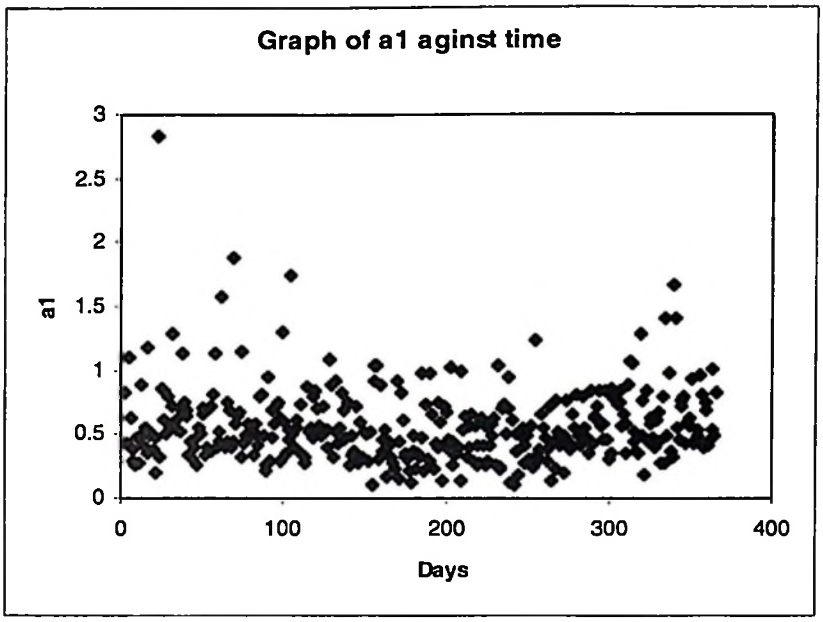


Figure 8.4(i)

The graph above indicates that most of the times the values of a1 are non-zero.

(ii) Graph of a2 against time

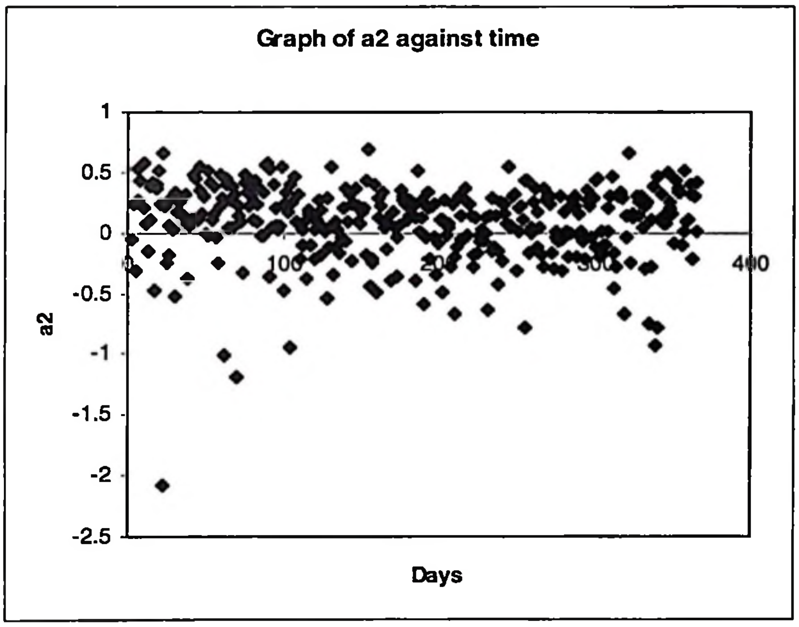


Figure 8.4(ii)

The graph above indicates that most of the time the values of **a2** are around zero. (i.e. between –0.5 and 0.5) These results justify the assumption of a first order linear autoregressive process in modelling data for maximum temperature.

8.2 Minimum temperature

8.3.1 Mean daily minimum temperature

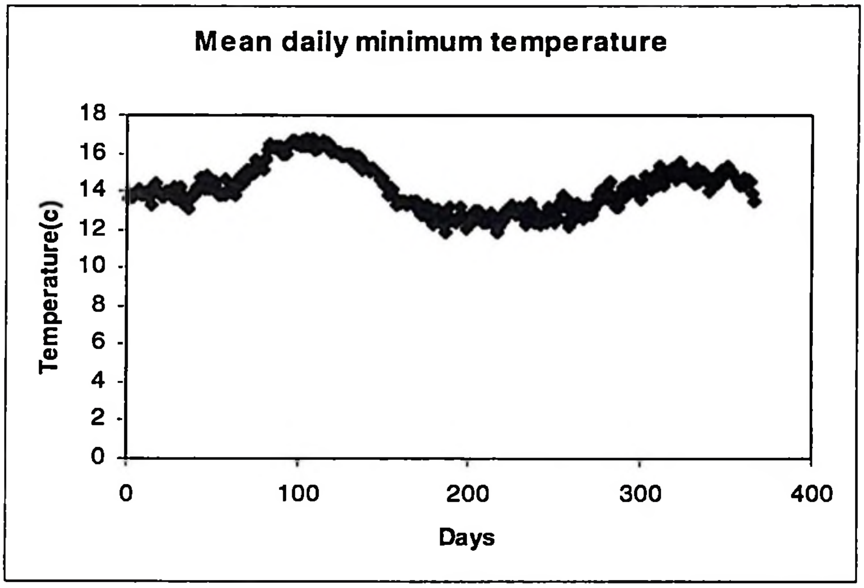


Figure 8.5

The graph above indicates the usual minimum temperature in Tanzania where June is the coldest month.

8.3.2 Daily Variance of minimum temperature against time.

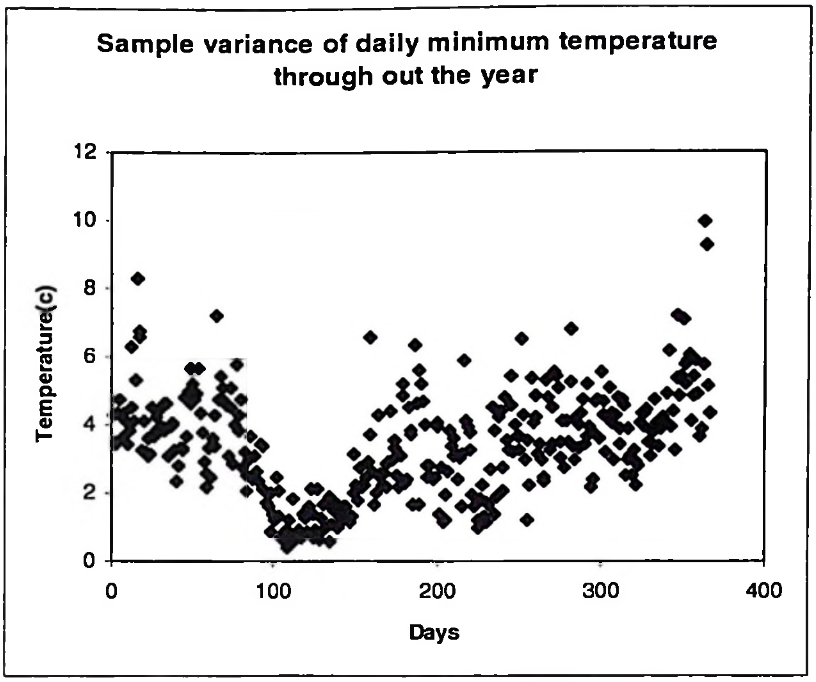


Figure 8.6

The figure indicates the distribution of daily minimum temperature variance through out the year. Like the case with the maximum temperature, variance is less smooth than mean daily minimum temperature.

8.3.3 First order serial correlation

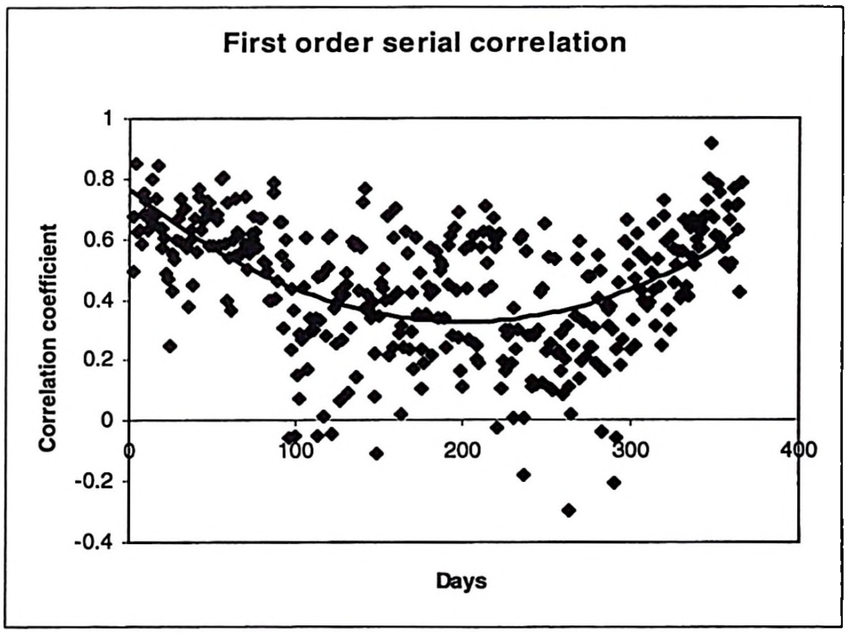


Figure 8.7

The graph above shows the first order autoregressive errors on minimum temperature through the year. The correlations are certainly non-zero and non-stationery.

8.3.4 Second order serial correlation

(i) Graph of a1 against time

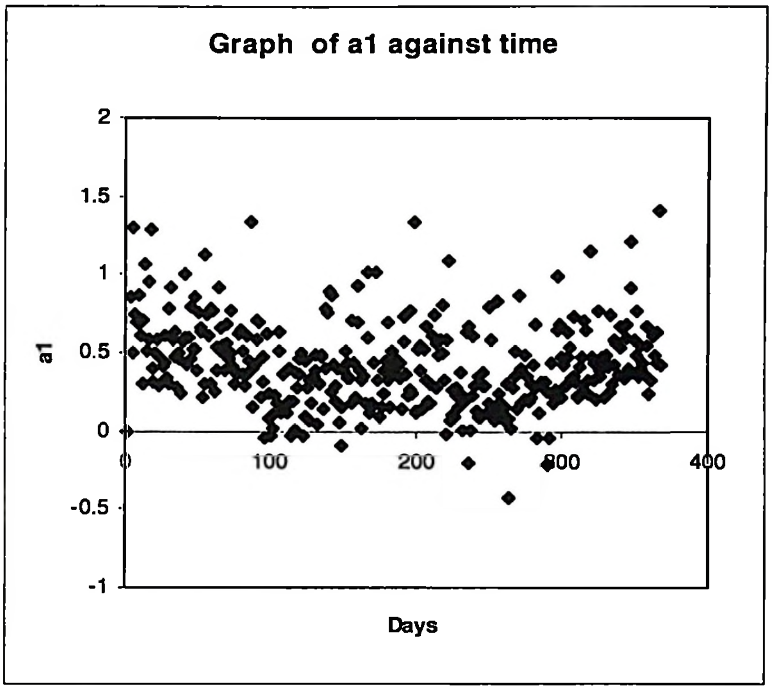


Figure 8.8(i)

The graph above indicates that most of the time the values of a1 are non zero

(ii) Graph of a2 against time

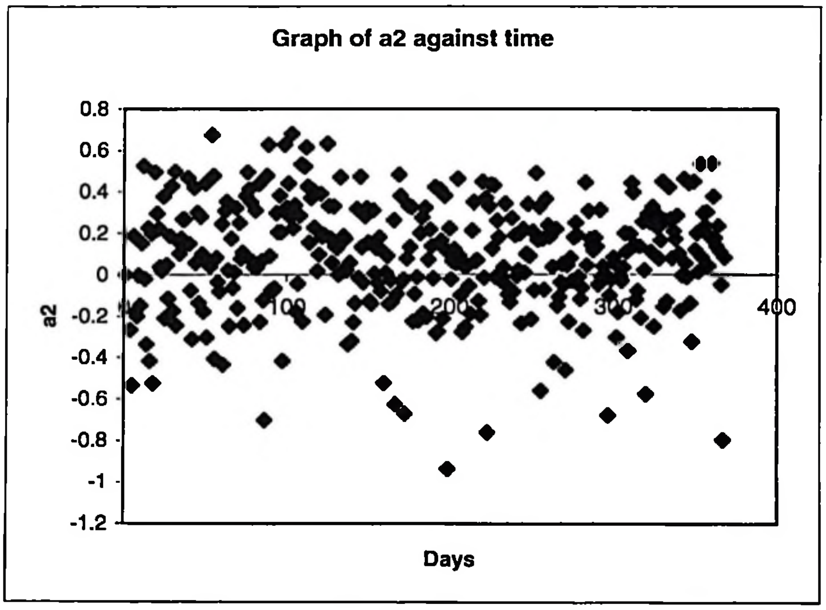


Figure 8.8(ii)

The graph above indicates that, most of the time the values of a2 are around zero
These results justify the assumption of a first order autoregressive process used in PT when modelling data on minimum temperature.

8.4 Conclusion

The temperature data from Arusha (1970-2000) indicates that, an assumption of first order autoregressive process on errors as used in PT is quite justifiable. The only problem is on the assumption of errors being stationery. Further study is needed in this area

CHAPTER 9:DISCUSSIONS, RECOMMENDATIONS AND FURTHER WORK

The results on analysis of rainfall data confirm that, higher orders Markov chain fit the chances of rain at Arusha site and leads to an improvement in the generated data. There is a possibility that, this is also the case for other stations in Tanzania. Hence there is a need to improve the PT program by allowing for other higher order models

The method of modelling temperature data as described in chapter seven makes two main assumptions. First is the assumption of residuals being stationery over years and second an assumption that residuals follow under first order autoregressive process. The assumption of first order autoregressive process has been found to be valid for data from Arusha, but the stationary assumptions is questionable. I would propose a modelling approach in which the errors are allowed to vary with time. Coe (1981) provides a good example on how to allow for such variation.

Owing to what has been found on climatic data simulation, it would be as well important later to look into some other statistical issues raised in chapter one. These were how well is the estimation of rainfall intensity and how does the software encourages the design of a strategic experiment in the simulation.

As introduced in the project, the estimation of rainfall intensity is very crucial in determining, water runoff from one profile to another. In PT, rainfall intensity distribution is assumed to be similar to that proposed by Oron et al (1989). The basic assumption is that rainfall intensity in Tanzania follows such a distribution, irrespective of a region under consideration. It would be important to carefully examine the assumptions underlying such a distribution and see whether they apply to all the stations in Tanzania. Suggestions on how to improve the estimation of rainfall intensity may then follow after such a study.

As discussed in chapter two, the PT model simulates cop growth while considering a number of factors such as weed control and soil surface management. This is essentially a design of a real experiment, only that the results come out so quickly. To what extent does the software encourage a well-designed experimental strategy in the simulation? How can the simulation experiments be designed to achieve good precision? How easily can risks as opposed to just comparing means, be estimated and compared for different strategies? The issue of a good simulation process has not been addressed in PT. This is evidenced by the fact that, in the simulation process, there is no indication of the used number and the number of times it has been used. Further investigation is needed in this aspect to assure the quality of the conclusions from the experiments done by the PT project team

The modelling of the agrometeorological variables, including rainfall does not take into consideration of possible climatic changes in Tanzania. However, taking into consideration, of the on going droughts in some parts of the country, it would be more wise later to consider the climatic change when modelling the climatic data. This consideration becomes even more important when the planners are thinking of long-term plans such as 10 years or more

As seen from chapter three, using the PT model in simulating crop growth as well as climatic data is a frustrating process. The problems range from documentation problem to an ease of using the generated data for research purposes. To identify some more problems, I propose that, the PT model be practically implemented in teaching the students. The students would simply use the data generated by the PT model in their relevant subjects such as "experimental designs" The concerned instructor would then have the students opinions about how the model works.

REFERENCES

Chandra, M., Oguntuase, O., 1986:Measurements of global solar radiation in Nigeria. Nigeria building and Road Research Institute Report No.7

Chatfield (1989), The analysis of time series an introduction

Chineke T.C and Jagtap .S.S and. Aina. J.I (1999):

Applicability of a weather simulation model based on observed daily meteorological data in humid Tropical climate. Theoretical Applied Climatology. 64, 15-25

Collett, D. (1991) Modelling Binary Data, Chapman and Hall, London

Dale, I.C. (2004) Evaluation of the PARCHED-THIRST version 2.3 model. Report to NRSP.

Matthews and William Stephens (2002) Crop soil simulation models, Applications in developing countries CABI Publishing Wallinford UK

Ndhokero Juma MSc in Biometry 2002

STERN.R.D and COE .R, Fitting models to daily Rainfall Data, Journal of Applied Meteorology, Vol 12, No 7, July 1982

Vandyke John (2004) Integrated Crop Management, IC-492 (18)

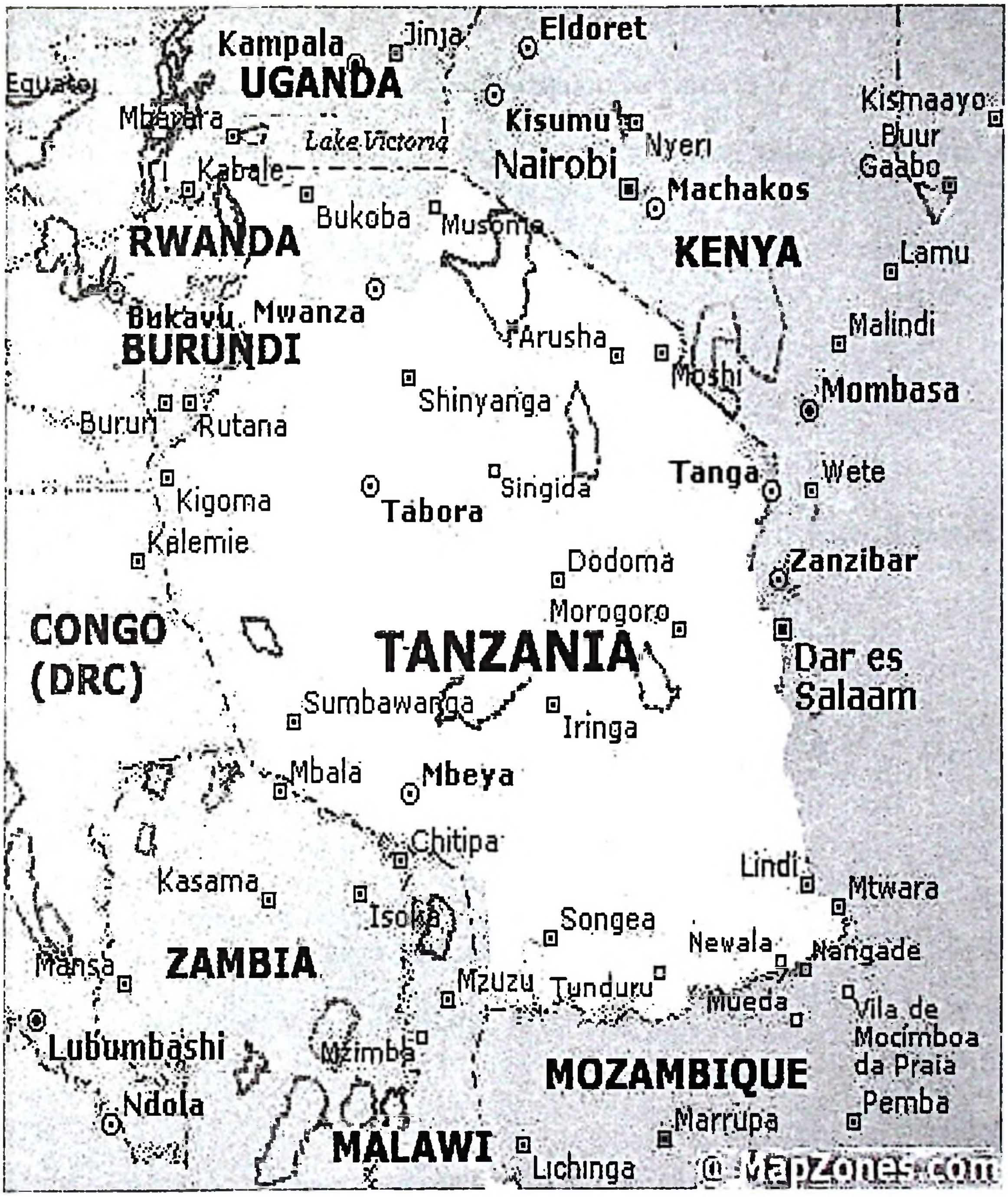
Young M.D and Gowing J.W, (1996). The PARCHED THIRST, User Guide, Version 1.0

Young M.D, Gowing J.W, Wyseure G.C.L (2002) /Agricultural water management 55, 121-140

Xinyou Yin[*,a], Piet Stam[b], Martin J. Kropff[a] and Ad H. C. M. Schapendonk[c] (2003)
Agronomy Journal 95:90-98 (2003)

APPENDIX 1

THE MAP OF TANZANIA INDICATING THE COUNTY'S REGIONS



Source: http://www.maps.com/reference/geoshelf/factbook/tanzania.html#Geo

APPENDIX 2

MICROSOFT VISUAL BASIC PROGRAM FOR CALCULTING THE SUMMARY MEASURES FOR RAINFALL

```
If OutRainX(MMonth%) > 0 Then
        OutRainMean(MMonth%) = OutRainX(MMonth%) /
OutRainN(MMonth%)
        OutRainSD(MMonth%) = Sqr((1 / (OutRainN(MMonth%) - 1)) *
(OutRainX2(MMonth%) - (OutRainX(MMonth%)) ^ 2 /
OutRainN(MMonth%)))
    Else
        OutRainMean(MMonth%) = 0
        OutRainSD(MMonth%) = 0
    End If
    If OutRainyDayN(MMonth%) > 1 Then

        OutRainyDayMean(MMonth%) = OutRainyDayX(MMonth%) /
OutRainyDayN(MMonth%)
        OutRainyDaySD(MMonth%) = Sqr((1 / (OutRainyDayN(MMonth%)
- 1)) * (OutRainyDayX2(MMonth%) - (OutRainyDayX(MMonth%)) ^ 2 /
OutRainyDayN(MMonth%)))
        OutMeanNoOfRDays(MMonth%) = OutRainyDayN(MMonth%) /
SNoOfYears%
        '10/04/97
        OutSDNoOfRDays(MMonth%) = Sqr((1 / (SNoOfYears% - 1)) *
(OutRainyDaysX2(MMonth%) - (OutRainyDaysSum(MMonth%) ^ 2 /
SNoOfYears%)))
    Else
        OutRainyDayMean(MMonth%) = 0
        OutRainyDaySD(MMonth%) = 0
        OutMeanNoOfRDays(MMonth%) = 0
        '10/04/97
        OutSDNoOfRDays(MMonth%) = 0
    End If
```

APPENDIX 3

Evaluation of the Parched-Thirst v2.3 model
NRSP short-term contract PD110

Report by
Ian Dale, Statistical Services Centre, The University of Reading
6-9 February 2004

Terms of Reference

E-mail request: "NRSP wants an independent assessment of the current version of the PT2 model, essentially to determine how useful, useable etc it is for users and to identify what would be needed to make it such if it's not already."

Contract letter, dated 6 Feb 2004:
To carry out an evaluation of the model PT v2.3 to determine the extent to which the current product meets specified target user needs.
To provide specification of any improvements needed to the model to meet the needs of the specified users.
To suggest the process and method by which improvements in 2 (above) could be achieved, as a basis for NRSP decision-making on further support to the model's development and promotion with users.

Background Materials
A web search located more than thirty pages where Parched-Thirst (PT) was mentioned. I visited each site briefly and to put PT in context I downloaded all that seemed relevant. As well as this downloaded material and the PT software itself, I had access to:

Details from the NARSIS database on the web about NRSP projects
NRSP Final Technical Report on Project R7949
Consultant's Report on the PT model (Peter Muraya)
Assorted e-mails from different authors concerning PT

Questions and aspects considered

What is PT for? Does it meet its claimed objectives?
Who are the intended users? Is it appropriate for them? Does it meet their needs?
Dissemination: What information on PT is readily available?
Downloading issues.
Installation issues.
Help sub-system.
Tutorials. Who are they for, how useful are they...?
Supplied background information, publications etc.
Using the system.
File locations, structure, etc.
Using your own crop/climate data.

1. What is PTv2 for? Does it meet its claimed objectives?

According to the help sub-system, the model has the potential to be used for:

Modelling rain-water harvesting (RWH) macro-systems or micro-systems
Adding value to experimental fieldwork by extrapolating results over larger temporal and spatial extents
Strategic planning to save resource and man-power when used in conjunction with other tools such as GIS etc.
A study aid to emphasise important aspects of a RWH system and identify the effects of changes in the properties of soil, climate, crops etc.

According to the Tutorial, PTv2.3 is a planning, research and teaching tool.
As a planning tool:
Testing various RWH and rain-fed system scenarios ... in implementation of new investment and in improving existing projects.
Using the climatic generator component of the model to predict what will happen in ten years or so in agricultural planning.

As a research tool:

Testing out a proposed field experiment in order to predict the likely range of response to the proposed inputs.
Extrapolating the results of one or two years' field experiments to different weather conditions and considering the long-term variability of the yields.
Extrapolating the results of the field experiments to different soil conditions, different planting density and different soil-water management practices.

As a teaching or educational tool:

Studies of rainfall-runoff relationship.
Allowing student to carry out 'field experiments' in the classroom and consider a wide range of treatments and treatments interactions such as weeding treatments (days after emergence), planting density experiments, planting date experiments (treatments), soil variability experiments (texture, structure, drainage, nutrients).

Does PTv2 meet its claimed objectives?

It seems to do so quite well; it's not really possible to gauge this thoroughly in a rapid appraisal, without better evidence of its actual use in the intended working environments.

In terms of the software functionality, PTv2.3 can certainly do the jobs it claims to do, with some limitations, mainly to do with ease-of-use and ease-of-adaptation to users' own needs. These are outlined in the sections that follow.

2. Who are the users? Is PT appropriate? Does it meet their needs?

The intended ultimate beneficiaries - of improved RWH – are "the poor in semi-arid areas" (from the NARSIS database record: purpose of project R8088).
My prior assumption was that the potential user of PT would be an agricultural / hydrological researcher in an institute in a semi-arid country. A reading of the documentation suggests that this is too narrow a view.
According to NARSIS, users are "clients (researchers, non-researchers, GO, NGO, PVO)."

According to the Tutorial, the intended users of PT are "agricultural planners and advisers."
According to the help sub-system, the model will be of value to a great number of potential users including:
Planners (agricultural, governmental)
Researchers (water management, crops, livestock)
Extension agents / staff
Students (Undergraduate or postgraduate)

There is a lack of clarity in the documents about the identification of exactly who would use PT, and under what circumstances. The above-mentioned users are rather vague groupings, so specific examples of PT users (in e.g. Tanzania) would make things clearer.

The users of primary interest to NRSP are likely to be agricultural/hydrological planners and advisers within government, i.e. those whose analyses inform decisions about resource management that directly affect farmers' livelihoods. The related group of extension staff are likewise involved in influencing activities at the field and community levels.

Students as a group may well learn from looking at the PT software, but this is of less interest to NRSP since students' work in the short and medium term has little direct effect on the implementation of RWH. If this is not so, evidence and examples are needed.

Researchers are similarly of secondary relevance, unless their work directly influences the implementation of improved RWH practices. Again, evidence and examples are needed. If there is a clear route by which RWH research activities in agricultural/educational institutes are translated into adoption of new practices by farming communities, then this ought to be clearly demonstrated in the PT documentation. Currently it is not.

Is PT appropriate for these users?

For the sake of discussion, I have considered users at three classes of computing ability:

Expert / skilled computer user (researcher), able to modify models
Supported / competent computer user, follows instructions, can solve problems
Non-technical / unskilled computer user, baffled by most software errors

Researchers would typically be in the first or second class, as would some governmental staff, e.g. computer-literate graduates in central agricultural/planning offices. University students typically fall into the third class, though increasingly are in the second class. Extension agents I would expect to find mostly in the third class.

Researchers could use PT in its current form, as could some planners in government – though some case studies of its actual use by such people would be more convincing of this.

Agricultural extension service staff would find PT overly complex, though could probably learn to operate the system after several training sessions, for a set of pre-specified tasks.

A similar comment applies to its use by university students: it would take some time to bring a student to a level where he or she could use the model effectively, so it would be a tool more appropriate to be used on multiple assignments over several weeks, or for a dissertation project.

Using PT for teaching university students is likely to be a good way to test and hone its usability and stability. If the program is appropriately modified as a result of this testing, trainees in agricultural extension offices will then find it easier to use.

3. Dissemination / What information is readily available?

Assorted website materials; mainly from CLUWRR in Newcastle. (Somewhat outdated; some pictures not accessible.) A list of these pages can be provided if required. Information formerly available (in 2001) via Damion Young's web pages were no longer accessible from the Newcastle site.

A site in Tanzania was identified but not accessed (site not working).

4. Downloading

The only downloadable version that I located was PTv2.1 from the CLUWRR website. The Tanzanian site at Sokoine University of Agriculture (www.suanet.ac.tz) that was revealed by the search was not responding (on 5 February 2004).

A copy was requested from Tanzania by email and telephone, and was promptly sent as an e-mail attachment. A follow-up message identified the current download page. This quick response indicates that the help-desk is functioning.

This aspect can be improved to give potential users a better opportunity to examine PT. Mirror sites could be hosted at e.g. HTS, or at CLUWRR in Newcastle, or perhaps in Nairobi (ICRAF) where Nutu Hatibu, the former leader of the Sokoine team, is now based.

5. Installation

The installation is a setup file that uses the commercial "Wyse" installation software. It does not follow current Microsoft guidelines for installation, which are to use an "msi" package rather than a executable setup program. The potential problem is that installing PTv2 causes several files in the system folder to be overwritten: this is not recommended (though other programs do it, as it used to be acceptable practice).

Accepting that, the installation process ran smoothly on both Win-NT and Win-ME (for info: administrator privileges are needed).

Check: the version number on the "About" window that is picked up from the internal version information is still set to v2.2.

6. Help sub-system

The Help sub-system is impressive, and seems complete. Some pictures (e.g. formulas) that were missing from the help in v2.2 have been inserted.

It is written using "compiled HTML help" files which is a sensible choice: it means that the source files are available as HTML and therefore can be made accessible via the web – which in turn makes the rapid provision of updates possible.

A minor but useful improvement would be to show a suitable title for the help system in the Title Bar – when both the Tutorial and Help windows are open, it is not currently possible to tell which is which.

A more significant improvement would be to add a context-sensitive 'help' button to every dialogue box and form: this puts a link to the relevant help at the point where it would be most useful, and would enhance the ease-of-learning of the package considerably.

Several of the currently-available help buttons that are present jump to the one of the "PT Wizard profiles screen" help topics, instead of to a relevant topic.

7. Tutorials

(Except where stated otherwise, comments refer to v2.3.) The tutorial has been revised and extended somewhat from v2.1, and now includes more detail about using the climate generator. As with the help sub-system, it is written using "compiled HTML help" files. The same comment about use of the Title Bar made about the Help system applies here.

Note: in Tutorial v2.1, various pictures are missing (e.g. cover page, some formulae). Some pictures used in the tutorial do not exactly match the descriptive text (e.g. figs. 2, 3, 5). There are differences between the menus and their description (e.g. para 2 above fig. 6).

For teaching/learning purposes, an easily printable (pdf) version of the tutorials would be useful, in addition to the compiled HTML help versions.

The major component missing from the tutorials is a clear exposition of how a real user – say, an agricultural extension officer (AEO) – would take PT, adapt it for his or her locality, and use it to run through a set of scenarios to help decide what RWH operations could be used, what parameters (e.g. catchment size relative to planted area) would be appropriate, etc. This is clearly what the PT model is intended for; but some worked examples – taken from real-life use of PT – are necessary in the tutorials.

The main thing added to the v2.3 tutorial concerns how to use the 'simulation' or 'data generation' facilities in PT to create long series of climatic data from short records. While this is useful in some circumstances, it is peripheral to the main issue of enabling planners (and AEOs) to use their actual data records to rapidly compare alternative RWH scenarios.

My impression is that the tutorials have been developed further for the purpose of teaching university students about modelling and simulation, rather than to help planners in the field.

8. Supplied background information, publications, etc.

Readme, pdf files etc. – none were available from the installed system. Publication lists were included in the Help and Tutorials, but none of the listed documents were provided as pdf or otherwise screen-readable and printable versions. This is easily corrected, and should be, if PT is to be a more accessible and self-contained package.

The associated background documentation should certainly be made available from the PT website(s), if it is not actually included in printable form within the installation package.

9. Using the system

PTv2.3 uses a simplified Windows multiple-document interface (MDI): effectively it's a "single-document interface". It has a Menu Bar similar in style to that in popular Windows programs such as Word or Excel. This has advantages: users learn to expect a program to behave in certain ways, which then shortens the time they need to become familiar with a new program. PTv2.3 has improved in this area over its predecessors, and appeared stable and robust in general use.

However, when I tried loading two instances of the package simultaneously (which naïve users often do accidentally, e.g. by double-clicking twice on the PT icon), the system ran out of memory. This circumstance should be checked for and the second attempt to run the program should be ignored.

I also crashed the software by attempting to access a "simulation summary" before I had run a simulation and generated any output. This could be avoided by disabling ("greying") menu options until their use is valid.

Another concern I have is to do with the use of "modal" forms: these are dialogue and other boxes that appear during the normal use of the program and that cannot be resized or minimized; in other words, they demand immediate attention from the user. The start-up screen is an example; there are some others.

This is not only bad Windows programming, it is also very confusing for unskilled computer users. It can be avoided by following standard Windows programming practices, which include advice on using appropriate dialogue boxes [see, for example, http://msdn.microsoft.com/library/default.asp?url=/library/en-us/dnw2kcli/html/W2Kcli.asp].

On some dialogues is an "Autosize" button: its purpose is unclear and undocumented.

10. File locations, structure etc.

All installed files are put into subfolders below the "Program Files" folder, which is fine in general. Some Windows system files are updated/modified without warning (no longer an accepted practice – see Microsoft documentation, as above).

However, user's data and other files should not be added to subfolders of the "Program Files" folder, but instead to the user's own writeable folders. This is because - on Windows 2000/XP and later systems, and often on earlier (Win-NT/98/ME) systems – normal, unprivileged users (i.e. not "administrators" or "power users" who would carry out the installation) are not able to write to system areas such as "Program Files".

There are standard ways from within a program to identify the appropriate areas to use for storing users' files, and defaults should be identified and created (if necessary) at installation [see Appendix in Microsoft document, as above].

The format and structure of the weather data files are still cumbersome and awkward. An advance would be to improve possibilities for reading different formats of climatic data. A step in the right direction has been to add a facility for reading Clicom files, though this basically converts a specific set of Clicom files into PT-format files. [It is worth looking at the use of ODBC connections to read external databases directly.]

11. Putting in your own crop/climate data

On the whole, the facilities in PT for modifying 'system' and 'profile' information are flexible and clearly explained. What is unclear is how a researcher might adapt PT to use a crop other than those predefined (rice, maize) or adapt it to use a different crop growth model. This area also therefore needs more documentation.

Conclusions and summary of recommendations

The Sokoine team is to be commended for the progress made in improving the robustness and stability of PT under different Windows versions. They have also usefully added to the help and tutorial materials. As with all software, there is of course room to improve further.

How far does PT v2.3 meet specified target user needs? (usefulness, usability etc.)
Moderately well overall; in some areas very well. It depends on who is the 'target user'.
What improvements are needed to the model to meet the needs of the specified users?
The main general change I would recommend is to focus less on the university/teaching aspects, and much more on the planning/application target groups.

On the evidence I had available, I was unable to tell whether in fact PT is used in 'live' situations other than teaching, in a university setting. The question is: are the PT development team actually using PT in planning activities with agricultural/hydrological officers? The most important requirement is therefore the writing-up of some case studies and worked examples of the use of PT to assist real-live RWH planning and implementation.

If no such cases can be identified, it throws doubt on the practical benefits of the project.

For planning, the point is to compare different RWH scenarios (across a number of years), not to look in detail at the growth of a crop through a season. How is this done? Within the program, can different scenarios be easily compared? (This depends on the file formats and output.) Is it easy enough already, or could it improve? Demonstrations and examples are needed.

Some specific recommendations:

Make the necessary changes to the program as noted in this report, and as arise during the use of the program in teaching situations, for reliability and usability, particularly for the more important target user-groups, e.g. agricultural extension officers.

Carry out general checking, to ensure consistency between help/tutorial text and the program itself.

Capture all related documents and publications in electronic form, and make readily available (within the installation bundle, and/or via the PT website).
Produce, make available and maintain a list of edits planned under the current project.
Produce, make available and maintain a wish-list of desirable improvements – these are more long-term objectives, and could be added to by people other than the current team.

Generally:
The PT development team should concentrate on producing a solid product, with fully documented source code, and a 'hand-over manual' describing the software structure and code in such a way that a different programming team could relatively easily continue to develop the package. Electronic versions of the code and documents could be housed with NRSP, and perhaps also elsewhere.

What process and method could achieve the necessary improvements? (Suggestions may assist NRSP decision-making on further support to the PT's development and promotion.)

In terms of the software development, a clear and detailed statement of what is expected by end of the project would help matters. The recommendations above could form part of the statement.

The current funding process seems to be working quite well at maintaining a steady rate of improvements to the PT package. The direction and focus of the work needs to be brought more in line with the intentions of the donors. Emphasising the need for some detailed case studies would assist this realignment.

If there is any possibility of an extra input of funds, it could perhaps best be directed towards helping the write-up of RWH case studies that use the PT package. These examples would also provide clear evidence that the PT project has been successful.

Issues raised but unanswered.

What are the current development team's plans? What happens when the funding stops?
How are they proposing to continue / maintain PT after March 2005?
What happens to the Help-desk after March 2005?
Will there be web support? How will this be done? Where will it be located?
What other sources of funds or sponsorships are or might become available?
Use? Users? Ownership?

Ian Dale
Statistical Services Centre,
The University of Reading, UK.
9th February, 2004

APPENDIX 4

I: GENSTAT MACROS FOR COMPARISON OF RERESSION CURVES

```
set [diag=fault]
variate [val=(1...366)4]date
factor
[lev=4;lab=!t(dry2,dry1,rain1,rain2);val=366(1...4)]second
factor [lev=2;lab=!t(dry,rain);val=732(1...2)]first
factor
[lev=3;lab=!t(dry2,dry1,rain);val=366(1...2),732(3)]halfway
append [dry] ddd,ddr,drd,drr
append [rain] rdd,rdr,rrd,rrr
calc total=dry+rain

"Now generate the sine and cosine terms, ready for the fitting"
scalar const
calc const=2*c('pi')/366
for i=1...8
     calc sin[i]=sin(i*const*date)
     calc cos[i]=cos(i*const*date)
endfor

"Adding of weights so the fitted values are given for each day,
even if there is no data"
calc wt=(total>0)  "This gives a weight of zero if there are no
dry days"
calc total=total*(total>0)+(total.eq.0)  "Now give a denominator
of at least 1 always"

"General Model for rain with 5 harmonics."
MODEL [DISTRIBUTION=binomial; LINK=logit; DISPERSION=1;wei=wt]
rain; NBINOMIAL=total
FITindividually [PRINT=model,summary,estimates,acc;
CONSTANT=estimate; FPROB=yes; TPROB=yes; FACT=9]\

sin[1]+cos[1]+sin[2]+cos[2]+sin[3]+cos[3]+sin[4]+cos[4]+sin[5]+c
os[5]+first +first*(sin[1]+cos[1])

"General Model for rain with 5 harmonics."
MODEL [DISTRIBUTION=binomial; LINK=logit; DISPERSION=1;wei=wt]
rain; NBINOMIAL=total
FITindividually [PRINT=model,summary,estimates,acc;
CONSTANT=estimate; FPROB=yes; TPROB=yes; FACT=9]\
```

```
sin[1]+cos[1]+sin[2]+cos[2]+sin[3]+cos[3]+sin[4]+cos[4]+sin[5]+c
os[5]+second +second*(sin[1]+cos[1])



MODEL [DISTRIBUTION=binomial; LINK=logit; DISPERSION=1;wei=wt]
rain; NBINOMIAL=total
FITindividually [PRINT=model,summary,estimates,acc;
CONSTANT=estimate; FPROB=yes; TPROB=yes; FACT=9]\

sin[1]+cos[1]+sin[2]+cos[2]+sin[3]+cos[3]+sin[4]+cos[4]+sin[5]+c
os[5]+halfway+halfway*(sin[1]+cos[1]+sin[2]+cos[2])+sin[3]+cos[3
])




RCHECK [GRAPHICS=high] residual; fitted
RDISPLAY [PRINT=fitted]

FITindividually [PRINT=model,summary,estimates,acc;
CONSTANT=estimate; FPROB=yes; TPROB=yes; FACT=9]\
  second*(sin[1]+cos[1]+sin[2]+cos[2]+sin[3]+cos[3])



RKEEP FITTEDVALUES=fit
calc fit=fit/total

PEN [RESET=yes] 1,2,3,4; METHOD=line,line; JOIN=ascending;
SYMBOL=0; LINESTYLE=1,1;colour=1,2,3,4;th=2,2,2,2
dgraph fit;date;pen=first halfway second
```

II: GENSTAT MACROS FOR ZERO ORDER PROBABILITY MODEL

```
"Simple program to take first order Markov chain for the chance
of rain and zero order for
amounts. Then does the sort of analysis that is alternatively
handled by Instat, but properly"

set [diag=fault]

variate [val=1...366]Date

"These columns are not needed for the fitting, but are given to
be consistent with\
the calculations done in Instat"
calc p_r=rain/tot

"Put the data into a spreadsheet. This is optional too"
%wspread Date,tr,lr,dry,rain,tot,p_r

"Now plot the conditional and overall proportions"
XAXIS [RESET=yes] WINDOW=1; LPOSITION=outside;
LDIRECTION=parallel; MPOSITION=outside;\
 ARROWHEAD=omit; ACTION=display
YAXIS [RESET=yes] WINDOW=1; LPOSITION=outside;
LDIRECTION=perpendicular; MPOSITION=outside;\
  ARROWHEAD=omit; ACTION=display
 PEN [RESET=yes] 1,2; METHOD=line,line; JOIN=ascending;
SYMBOL=0; LINESTYLE=1;colour=1,2


"Generate the sine and cosine terms, ready for the fitting"
scalar const
calc const=2*c('pi')/366
for i=1...6
     calc sin[i]=sin(i*const*Date)
     calc cos[i]=cos(i*const*Date)
endfor

  "Adding weights so the fitted values are given for each day,
even if there is no data"
calc wtprev=(tot>0)  "This gives a weight of zero if there are
no tot days"
calc tot=tot*(tot>0)+(tot.eq.0)  "Now give a denominator of at
least 1 always"
```

```
"General Model for rain given no status of the previous day."
MODEL [DISTRIBUTION=binomial; LINK=logit;
DISPERSION=1;wei=wtprev] rain; NBINOMIAL=tot
FITindividually [PRINT=m,s,e,a; FPROB=yes; TPROB=yes]\

sin[1]+cos[1]+sin[2]+cos[2]+sin[3]+cos[3]+sin[4]+cos[4]+sin[5]+c
os[5]

RKEEP FITTEDVALUES=frain
calc frain=frain/tot
dgraph f_rr,p_rr;Date


"Now fit the amounts"

calc mrain=tr/rain "This will give a warning when there is 0/0."
calc meanrain=tr/(rain+(rain.eq.0))
calc mrain=meanrain+(rain.eq.0)
"mrain is set to 1, when there is no rain on that day.  That
means the log link is ok.
  The fitting is not affected, because the weight is set to
zero."

"Estimate the shape parameter"
calc dev=lr/(rain+(rain.eq.0))
calc dev=-2*(dev-log(mrain))
calc temp=dev*(rain-1)
scalar k
calc k=sum(dev>0)
print k
calc k=0.5*sum(temp)/(sum(rain)-k)
if k<0.5772
  calc k=(.5000876+.1648852*k-.0544274*k*k)/k
else
  calc
k=(8.898919+9.05995*k+.9775373*k*k)/(k*(17.79728+11.968477*k+k*k
))
endif
print k
calc temp=1
calc kest=k*temp

"General Model for rainfall amounts with gamma model and 2
harmonics."
MODEL [DISTRIBUTION=gamma; LINK=log;wei=rain] mrain
FITindividually [PRINT=m,s,e,a; FPROB=yes; TPROB=yes]\
  sin[1]+cos[1]+sin[2]+cos[2]+sin[3]+cos[3]
```

```
RKEEP FITTEDVALUES=fm
PEN [RESET=yes] 1,2; METHOD=line,line; JOIN=ascending; SYMBOL=0;
LINESTYLE=1;colour=1,2;th=4,1
 DGRAPH Y=fm,mrain; X=Date
  PEN [RESET=yes] 1,2

"Put the fitted model into a spreadsheet"
%wspread Date,f_rd,f_rr,fm,kest,frain

"This fitted model is now in the form that can be read back into
Instat for the
next stage of using the results"
```

III: GENSTAT MACROS FOR FIRST ORDER PROBABILITY MODEL

```
"Simple program to take first order Markov chain for the chance
of rain and zero order for
amounts. Then does the sort of analysis that is alternatively
handled by Instat, but properly"

set [diag=fault]

variate [val=1...366]Date

calc _d=dd+rd "_d is number of days it was dry yesterday"
calc _r=dr+rr "_r is number of days it was rainy yesterday"

"These columns are not needed for the fitting, but are given to
be consistent with\
the calculations done in Instat"
calc p_rd=rd/_d
& p_rr = rr/_r
& dry=dd+dr & rain=rd+rr & tot=dry+rain

"Put the data into a spreadsheet. This is optional too"
%wspread Date,dd,dr,rd,rr,tr,lr,_d,_r,p_rd,p_rr,

"Generate the sine and cosine terms, ready for the fitting"
scalar const
calc const=2*c('pi')/366
for i=1...6
calc sin[i]=sin(i*const*Date)
calc cos[i]=cos(i*const*Date)
endfor

"Adding weights so the fitted values are given for each day,
even if there is no data"
calc wtdry=(_d>0)  "This gives a weight of zero if there are no
dry days"
calc _d=_d*(_d>0)+(_d.eq.0)  "Now give a denominator of at least
1 always"

"General Model for rain given dry with 5 harmonics."
MODEL [DISTRIBUTION=binomial; LINK=logit;
DISPERSION=1;wei=wtdry] rd; NBINOMIAL=_d
FITindividually [PRINT=m,s,e,a; FPROB=yes; TPROB=yes]\
```

```
sin[1]+cos[1]+sin[2]+cos[2]+sin[3]+cos[3]+sin[4]+cos[4]+sin[5]+c
os[5]

RKEEP FITTEDVALUES=f_rd

calc wtrain=(_r>0)
calc _r=_r*(_r>0)+(_r.eq.0)

"General Model for rain given rain with 4 harmonics."
MODEL [DISTRIBUTION=binomial; LINK=logit;
DISPERSION=1;wei=wtrain] rr; NBINOMIAL=_r
FITindividually [PRINT=m,s,e,a; FPROB=yes; TPROB=yes]\
 sin[1]+cos[1]+sin[2]+cos[2]+sin[3]+cos[3]+sin[4]+cos[4]

RKEEP FITTEDVALUES=f_rr

"Now fit the amounts"

calc mrain=tr/rain "This will give a warning when there is 0/0."
calc meanrain=tr/(rain+(rain.eq.0))
calc mrain=meanrain+(rain.eq.0)
"mrain is set to 1, when there is no rain on that day.  That
means the log link is ok.
  The fitting is not affected, because the weight is set to
zero."

"Estimate the shape parameter"
calc dev=lr/(rain+(rain.eq.0))
calc dev=-2*(dev-log(mrain))
calc temp=dev*(rain-1)
scalar k
calc k=sum(dev>0)
print k
calc k=0.5*sum(temp)/(sum(rain)-k)
if k<0.5772
  calc k=(.5000876+.1648852*k-.0544274*k*k)/k
else
  calc
k=(8.898919+9.05995*k+.9775373*k*k)/(k*(17.79728+11.968477*k+k*k
))
endif
print k
calc temp=1
calc kest=k*temp
```

```
"General Model for rainfall amounts with gamma model and 3
harmonics."
MODEL [DISTRIBUTION=gamma; LINK=log;wei=rain] mrain
FITindividually [PRINT=m,s,e,a; FPROB=yes; TPROB=yes]\
  sin[1]+cos[1]+sin[2]+cos[2]+sin[3]+cos[3]

RKEEP FITTEDVALUES=fm
PEN [RESET=yes] 1,2; METHOD=line,line; JOIN=ascending; SYMBOL=0;
LINESTYLE=1;colour=1,2;th=4,1
 DGRAPH Y=fm,mrain; X=Date
  PEN [RESET=yes] 1,2

"Put the fitted model into a spreadsheet"
%wspread Date,f_rd,f_rr,fm,kest

"This fitted model is now in the form that can be read back into
Instat for the
next stage of using the results"


"Simple program to take first order Markov chain for the chance
of rain and zero order for
amounts. Then does the sort of analysis that is alternatively
handled by Instat, but properly"

set [diag=fault]

variate [val=1...366]Date

calc _dd=ddd+rdd  " _dd is number of days it was dry in two
consecutive days"
calc _dr=ddr+rdr  "   _dr is number of days it was dry preceded
by  rainy "
calc _rd=drd+rrd  "   _rd is number of days it was rainy
preceded by a dry day"
calc _rr=rrr+drr   "    _rr is number of days it was rainy
preceded by  rainy "
calc rain=rdd+rdr+rrd+rrr
calc tot= rdd+rdr+rrd+rrr+ddd+ddr+drd+drr
"These columns are not needed for the fitting, but are given to
be consistent with
the calculations done in Instat"
calc p_rdd=rdd/_dd
& p_rdr = rdr/_dr
& p_rrd = rrd/_rd
& p_rrr=rrr/_rr
```

```
"Put the data into a spreadsheet. This is optional too"
%wspread
Date,ddd,ddr,drd,drr,rdd,rdr,rrd,rrr,tr,lr,_dd,_dr,_rd,_rr,p_rdd
,p_rdr,p_rrd,p_rrr,

"Now plot the conditional and overall proportions"
XAXIS [RESET=yes] WINDOW=1; LPOSITION=outside;
LDIRECTION=parallel; MPOSITION=outside;\
 ARROWHEAD=omit; ACTION=display
YAXIS [RESET=yes] WINDOW=1; LPOSITION=outside;
LDIRECTION=perpendicular; MPOSITION=outside;\
  ARROWHEAD=omit; ACTION=display
 PEN [RESET=yes] 1,2; METHOD=line,line; JOIN=ascending;
SYMBOL=0; LINESTYLE=1;colour=1,2

"Graph of second order"
 DGRAPH Y=p_rdd,p_rdr,p_rrd,p_rrr; X=Date

"Generate the sine and cosine terms, ready for the fitting"
scalar const
calc const=2*c('pi')/366
for i=1...6
      calc sin[i]=sin(i*const*Date)
      calc cos[i]=cos(i*const*Date)
endfor

"Adding weights so the fitted values are given for each day,
even if there is no data"
calc wtdd=(_dd>0)  "This gives a weight of zero if there are no
dry days"
calc _dd=_dd*(_dd>0)+(_dd.eq.0)  "Now give a denominator of at
least 1 always"
```

IV: GENSTAT MACROS FOR FIRST ORDER PROBABILITY MODEL

```
"General Model for rain , with four harmonics given it was dry
in the last two days"
MODEL [DISTRIBUTION=binomial; LINK=logit; DISPERSION=1;wei=wtdd]
rdd; NBINOMIAL=_dd
FITindividually [PRINT=m,s,e,a; FPROB=yes; TPROB=yes]\
  sin[1]+cos[1]+sin[2]+cos[2]+sin[3]+cos[3]+sin[4]+cos[4]
```

```
RKEEP FITTEDVALUES=f_rdd
calc f_rdd=f_rdd/_dd

calc wtdr=(_dr>0)
calc _dr=_dr*(_dr>0)+(_dr.eq.0)

"General Model for rain  with three harmonics given that,
previous day was dry preceded by rain"
MODEL [DISTRIBUTION=binomial; LINK=logit; DISPERSION=1;wei=wtdr]
rdr; NBINOMIAL=_dr
FITindividually [PRINT=m,s,e,a; FPROB=yes; TPROB=yes]\
  sin[1]+cos[1]+sin[2]+cos[2]+sin[3]+cos[3]

RKEEP FITTEDVALUES=f_rdr
calc f_rdr=f_rdr/_dr
dgraph f_rdr,p_rdr;Date

 calc wtrd=(_rd>0)
calc _rd=_rd*(_rd>0)+(_rd.eq.0)

"General Model for rain  with two harmonics given that, previous
was rain preceded by dry"
MODEL [DISTRIBUTION=binomial; LINK=logit; DISPERSION=1;wei=wtrd]
rrd; NBINOMIAL=_rd
FITindividually [PRINT=m,s,e,a; FPROB=yes; TPROB=yes]\
  sin[1]+cos[1]+sin[2]+cos[2]

RKEEP FITTEDVALUES=f_rrd
calc f_rrd=f_rrd/_rd
dgraph f_rrd,p_rrd;Date

calc wtrd=(_rr>0)
calc _rr=_rr*(_rd>0)+(_rr.eq.0)

"General Model for rain  with one harmonic given that, previous
was  rain preceded by dry"
MODEL [DISTRIBUTION=binomial; LINK=logit; DISPERSION=1;wei=wtrd]
rrr; NBINOMIAL=_rr
FITindividually [PRINT=m,s,e,a; FPROB=yes; TPROB=yes]\
 sin[1]+cos[1]

RKEEP FITTEDVALUES=f_rrr
calc f_rrr=f_rrr/_rr
dgraph f_rrr,rrr;Date

dgraph f_rrr,f_rrd,f_rdr,f_rdd;Date
```

```
"Now fit the amounts"

calc mrain=tr/rain "This will give a warning when there is 0/0."
calc meanrain=tr/(rain+(rain.eq.0))
calc mrain=meanrain+(rain.eq.0)
"mrain is set to 1, when there is no rain on that day.  That
means the log link is ok.
  The fitting is not affected, because the weight is set to
zero."

calc dev=lr/(rain+(rain.eq.0))
calc dev=-2*(dev-log(mrain))
calc temp=dev*(rain-1)
scalar k
calc k=sum(dev>0)
print k
calc k=0.5*sum(temp)/(sum(rain)-k)
if k<0.5772
  calc k=(.5000876+.1648852*k-.0544274*k*k)/k
else
  calc
k=(8.898919+9.05995*k+.9775373*k*k)/(k*(17.79728+11.968477*k+k*k
))
endif
print k
calc temp=1
calc kest=k*temp

"General Model for rainfall amounts with gamma model and 2
harmonics."
MODEL [DISTRIBUTION=gamma; LINK=log;wei=rain] mrain
FITindividually [PRINT=m,s,e,a; FPROB=yes; TPROB=yes]\
  sin[1]+cos[1]+sin[2]+cos[2]+sin[3]+cos[3]

RKEEP FITTEDVALUES=fm
PEN [RESET=yes] 1,2; METHOD=line,line; JOIN=ascending; SYMBOL=0;
LINESTYLE=1;colour=1,2;th=4,1
 DGRAPH Y=fm,mrain; X=Date
  PEN [RESET=yes] 1,2

"Put the fitted model into a spreadsheet"
%wspread Date,f_rdd,f_rdr,f_rrd,f_rrr,fm,kest

"This fitted model is now in the form that can be read back into
Instat for the
next stage of using the results"
```

APPENDIX 5

Zero order probabilities					
Date	fm	Dry	Rain	Tot	Pr
1	7.69	23	8	31	0.26
2	7.68	24	7	31	0.23
3	7.66	21	10	31	0.32
4	7.65	22	9	31	0.29
5	7.64	22	9	31	0.29
6	7.62	20	11	31	0.35
7	7.61	20	11	31	0.35
8	7.59	23	8	31	0.26
9	7.57	21	10	31	0.32
10	7.56	21	10	31	0.32
11	7.54	21	10	31	0.32
12	7.52	23	8	31	0.26
13	7.50	23	8	31	0.26
14	7.48	20	11	31	0.35
15	7.46	19	12	31	0.39
16	7.44	18	13	31	0.42
17	7.42	19	12	31	0.39
18	7.40	23	8	31	0.26
19	7.37	24	7	31	0.23
20	7.35	26	5	31	0.16
21	7.33	24	7	31	0.23
22	7.31	23	8	31	0.26
23	7.29	18	13	31	0.42
24	7.27	22	9	31	0.29
25	7.26	22	9	31	0.29
26	7.24	22	9	31	0.29
27	7.23	22	9	31	0.29
28	7.22	24	7	31	0.23
29	7.21	25	6	31	0.19
30	7.20	23	8	31	0.26
31	7.20	23	8	31	0.26
32	7.20	25	6	31	0.19
33	7.20	25	6	31	0.19
34	7.21	20	11	31	0.35
35	7.22	25	6	31	0.19
36	7.24	22	9	31	0.29
37	7.26	25	6	31	0.19
38	7.28	25	6	31	0.19
39	7.31	22	9	31	0.29
40	7.35	23	8	31	0.26
41	7.39	22	9	31	0.29
42	7.43	21	10	31	0.32
43	7.49	19	12	31	0.39
44	7.54	23	8	31	0.26
45	7.61	21	10	31	0.32

342	8.52	17	12	29	0.41
343	8.45	15	14	29	0.48
344	8.39	20	9	29	0.31
345	8.34	20	9	29	0.31
346	8.28	16	13	29	0.45
347	8.23	18	11	29	0.38
348	8.18	19	10	29	0.34
349	8.13	16	13	29	0.45
350	8.09	18	11	29	0.38
351	8.04	16	13	29	0.45
352	8.01	15	14	29	0.48
353	7.97	15	14	29	0.48
354	7.94	16	13	29	0.45
355	7.91	19	10	29	0.34
356	7.88	18	11	29	0.38
357	7.86	17	12	29	0.41
358	7.83	22	7	29	0.24
359	7.81	17	12	29	0.41
360	7.79	15	14	29	0.48
361	7.77	17	12	29	0.41
362	7.76	20	9	29	0.31
363	7.74	20	9	29	0.31
364	7.73	25	4	29	0.14
365	7.71	22	7	29	0.24
366	7.70	23	6	29	0.21

Note that :

fm is the mean daily amount of rainfall as estimated from Gamma distribution

Tot is the total number of days in the given historical years. The number of days is the same as the total number of years, which are 31. Where Tot value <31 means there was a missing value in one of those years

Rain is a rain day

Dry is a dry day

Pr is the probability of a day being wet = rain/Tot

APPENDIX 6

First order Probabilities									
Date	dd	dr	rd	rr	Amount	_d	_r	p_rd	p_rr
1	19	3	4	3	23.0	23	6	0.17	0.50
2	18	6	5	2	52.4	23	8	0.22	0.25
3	17	4	7	3	95.7	24	7	0.29	0.43
4	16	6	5	4	35.2	21	10	0.24	0.40
5	18	4	4	5	81.3	22	9	0.18	0.56
6	15	5	7	4	83.5	22	9	0.32	0.44
7	16	4	4	7	67.1	20	11	0.20	0.64
8	16	7	4	4	32.3	20	11	0.20	0.36
9	18	3	5	5	92.9	23	8	0.22	0.63
10	15	6	6	4	79.9	21	10	0.29	0.40
11	15	6	6	4	114.0	21	10	0.29	0.40
12	16	7	5	3	57.6	21	10	0.24	0.30
13	19	4	4	4	79.2	23	8	0.17	0.50
14	17	3	6	5	69.3	23	8	0.26	0.63
15	16	3	4	8	84.5	20	11	0.20	0.73
16	15	3	4	9	81.0	19	12	0.21	0.75
17	14	5	4	8	58.5	18	13	0.22	0.62
18	13	10	6	2	35.2	19	12	0.32	0.17
19	20	4	3	4	49.1	23	8	0.13	0.50
20	21	5	3	2	13.9	24	7	0.13	0.29
21	23	1	3	4	61.3	26	5	0.12	0.80
22	19	4	5	3	57.7	24	7	0.21	0.43
23	15	3	8	5	57.4	23	8	0.35	0.63
24	14	8	4	5	36.6	18	13	0.22	0.38
25	18	4	4	5	62.9	22	9	0.18	0.56
26	16	6	6	3	49.2	22	9	0.27	0.33
27	18	4	4	5	59.4	22	9	0.18	0.56
28	17	7	5	2	57.7	22	9	0.23	0.22
29	21	4	3	3	45.2	24	7	0.13	0.43
30	20	3	5	3	80.6	25	6	0.20	0.50
31	20	3	3	5	97.3	23	8	0.13	0.63
32	19	6	4	2	32.3	23	8	0.17	0.25
33	22	3	3	3	23.3	25	6	0.12	0.50
34	19	1	6	5	55.8	25	6	0.24	0.83
35	17	8	3	3	101.6	20	11	0.15	0.27
36	19	3	6	3	77.1	25	6	0.24	0.50
37	21	4	1	5	44.8	22	9	0.05	0.56
38	23	2	2	4	44.1	25	6	0.08	0.67
39	20	2	5	4	52.7	25	6	0.20	0.67
40	20	3	2	6	34.9	22	9	0.09	0.67
41	17	5	6	3	102.0	23	8	0.26	0.38
42	16	5	6	4	69.9	22	9	0.27	0.44
43	13	6	8	4	81.3	21	10	0.38	0.40
44	16	7	3	5	71.9	19	12	0.16	0.42

45	16	5	7	3	84.7	23	8	0.30	0.38
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
342	12	5	5	7	57.1	17	12	0.29	0.58
343	10	5	7	7	106.3	17	12	0.41	0.58
344	12	8	3	6	60.4	15	14	0.20	0.43
345	16	4	4	5	45.5	20	9	0.20	0.56
346	13	3	7	6	56.8	20	9	0.35	0.67
347	11	7	5	6	82.5	16	13	0.31	0.46
348	15	4	3	7	99.3	18	11	0.17	0.64
349	11	5	8	5	132.8	19	10	0.42	0.50
350	12	6	4	7	84.9	16	13	0.25	0.54
351	13	3	5	8	97.6	18	11	0.28	0.73
352	13	2	3	11	125.3	16	13	0.19	0.85
353	8	7	7	7	135.4	15	14	0.47	0.50
354	10	6	5	8	119.3	15	14	0.33	0.57
355	13	6	3	7	110.1	16	13	0.19	0.54
356	14	4	5	6	71.3	19	10	0.26	0.60
357	12	5	6	6	44.0	18	11	0.33	0.55
358	15	7	2	5	51.6	17	12	0.12	0.42
359	15	2	7	5	53.1	22	7	0.32	0.71
360	10	5	7	7	84.4	17	12	0.41	0.58
361	8	9	7	5	177.6	15	14	0.47	0.36
362	14	6	3	6	85.1	17	12	0.18	0.50
363	17	3	3	6	76.2	20	9	0.15	0.67
364	18	7	2	2	13.4	20	9	0.10	0.22
365	19	3	6	1	67.5	25	4	0.24	0.25
366	18	5	4	2	94.0	22	7	0.18	0.29

Note that:

dd is the number of times a day was dry preceded by a dry day
dr is the number of times a day was dry preceded by a wet day.
rd is the number of times a day was wet preceded by a dry day
rr is the number of times a day was wet preceded by a wet day
_d is the total number of occurrences of dry days in the previous day
_r is the total number of occurrences of wet days in the previous day .
p_rd is the probability of rain given the previous day was a dry
p_rdr is the probability of rain given the previous day was a rain day

Example

Let us Consider day 1 of a year. Using the formulae given in chapter four we can work out the probabilities associated with day 1 as follows

1.Probability of having rain in day 1 given that a previous day was a dry day (P_rd)

Where $p_{WD}(i) = \frac{\sum_{j=1}^{j=n} \{X(i+1, j) = W / X(i, j) = D\}}{\sum_{j=1}^{j=n} \{X(i, j) = D\}}$ = rd/_d = 4/23=0.17

2.Probability of having rain in day 1 given that a previous day was a wet day (P_rr)

We have $p_{WW}(i) = \frac{\sum_{j=1}^{j=n} \{X(i+1, j) = W / X(i, j) = W\}}{\sum_{j=1}^{j=n} \{X(i, j) = W\}}$ = rr/_r =3/6=0.50

APPENDIX 7

Second order Probabilities

Date	ddd	ddr	drd	drr	rdd	rdr	rrd	rrr	tr	_dd	_dr	_rd	_rr	p_rdd	p_rdr	p_rrd	p_rrr
1	16	3	2	1	2	2	2	1	23.0	18	5	4	2	0.11	0.40	0.50	0.50
2	16	2	3	2	3	1	1	1	52.4	19	3	4	3	0.16	0.33	0.25	0.33
3	13	4	3	1	5	2	2	1	95.7	18	6	5	2	0.28	0.33	0.40	0.50
4	14	2	5	1	3	2	2	2	35.2	17	4	7	3	0.18	0.50	0.29	0.67
5	13	5	1	3	3	1	4	1	81.3	16	6	5	4	0.19	0.17	0.80	0.25
6	14	1	3	2	4	3	1	3	83.5	18	4	4	5	0.22	0.75	0.25	0.60
7	12	4	2	2	3	1	5	2	67.1	15	5	7	4	0.20	0.20	0.71	0.50
8	14	2	3	4	2	2	1	3	32.3	16	4	4	7	0.13	0.50	0.25	0.43
9	12	6	1	2	4	1	3	2	92.9	16	7	4	4	0.25	0.14	0.75	0.50
10	13	2	3	3	5	1	2	2	79.9	18	3	5	5	0.28	0.33	0.40	0.40
11	11	4	3	3	4	2	3	1	114.0	15	6	6	4	0.27	0.33	0.50	0.25
12	11	5	4	3	4	1	2	1	57.6	15	6	6	4	0.27	0.17	0.33	0.25
13	13	6	4	0	3	1	1	3	79.2	16	7	5	3	0.19	0.14	0.20	1.00
14	14	3	2	1	5	1	2	3	69.3	19	4	4	4	0.26	0.25	0.50	0.75
15	14	2	2	1	3	1	4	4	84.5	17	3	6	5	0.18	0.33	0.67	0.80
16	14	1	0	3	2	2	4	5	81.0	16	3	4	8	0.13	0.67	1.00	0.63
17	11	3	2	3	4	0	2	6	58.5	15	3	4	9	0.27	0.00	0.50	0.67
18	11	2	4	6	3	3	0	2	35.2	14	5	4	8	0.21	0.60	0.00	0.25
19	13	7	4	0	0	3	2	2	49.1	13	10	6	2	0.00	0.30	0.33	1.00
20	18	3	2	3	2	1	1	1	13.9	20	4	3	4	0.10	0.25	0.33	0.25
21	20	3	0	1	1	2	3	1	61.3	21	5	3	2	0.05	0.40	1.00	0.50
22	18	1	2	2	5	0	1	2	57.7	23	1	3	4	0.22	0.00	0.33	0.50
23	13	2	1	2	6	2	4	1	57.4	19	4	5	3	0.32	0.50	0.80	0.33
24	12	2	5	3	3	1	3	2	36.6	15	3	8	5	0.20	0.33	0.38	0.40
25	11	7	2	2	3	1	2	3	62.9	14	8	4	5	0.21	0.13	0.50	0.60
26	13	3	3	3	5	1	1	2	49.2	18	4	4	5	0.28	0.25	0.25	0.40
27	13	5	3	1	3	1	3	2	59.4	16	6	6	3	0.19	0.17	0.50	0.67
28	13	4	2	5	5	0	2	0	57.7	18	4	4	5	0.28	0.00	0.50	0.00
29	17	4	3	1	0	3	2	1	45.2	17	7	5	2	0.00	0.43	0.40	0.50
30	18	2	1	2	3	2	2	1	80.6	21	4	3	3	0.14	0.50	0.67	0.33
31	18	2	2	1	2	1	3	2	97.3	20	3	5	3	0.10	0.33	0.60	0.67
32	16	3	2	4	4	0	1	1	32.3	20	3	3	5	0.20	0.00	0.33	0.20
33	17	5	3	0	2	1	1	2	23.3	19	6	4	2	0.11	0.17	0.25	1.00
34	17	2	1	0	5	1	2	3	55.8	22	3	3	3	0.23	0.33	0.67	1.00
35	17	0	4	4	2	1	2	1	101.6	19	1	6	5	0.11	1.00	0.33	0.20
36	15	4	1	2	2	4	2	1	77.1	17	8	3	3	0.12	0.50	0.67	0.33
37	19	2	3	1	0	1	3	2	44.8	19	3	6	3	0.00	0.33	0.50	0.67
38	19	4	0	2	2	0	1	3	44.1	21	4	1	5	0.10	0.00	1.00	0.60
39	19	1	1	1	4	1	1	3	52.7	23	2	2	4	0.17	0.50	0.50	0.75
40	19	1	1	2	1	1	4	2	34.9	20	2	5	4	0.05	0.50	0.80	0.50
41	17	0	2	3	3	3	0	3	102.0	20	3	2	6	0.15	1.00	0.00	0.50
42	12	4	3	2	5	1	3	1	69.9	17	5	6	3	0.29	0.20	0.50	0.33
43	11	2	4	2	5	3	2	2	81.3	16	5	6	4	0.31	0.60	0.33	0.50
44	11	5	4	3	2	1	4	1	71.9	13	6	8	4	0.15	0.17	0.50	0.25
45	10	6	2	3	6	1	1	2	84.7	16	7	3	5	0.38	0.14	0.33	0.40

.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
347	9	2	3	4	4	1	4	2	82.5	13	3	7	6	0.31	0.33	0.57	0.33
348	10	5	2	2	1	2	3	4	99.3	11	7	5	6	0.09	0.29	0.60	0.67
349	8	3	1	4	7	1	2	3	132.8	15	4	3	7	0.47	0.25	0.67	0.43
350	8	4	3	3	3	1	5	2	84.9	11	5	8	5	0.27	0.20	0.63	0.40
351	9	4	2	1	3	2	2	6	97.6	12	6	4	7	0.25	0.33	0.50	0.86
352	11	2	0	2	2	1	5	6	125.3	13	3	5	8	0.15	0.33	1.00	0.75
353	6	2	2	5	7	0	1	6	135.4	13	2	3	11	0.54	0.00	0.33	0.55
354	5	5	4	2	3	2	3	5	119.3	8	7	7	7	0.38	0.29	0.43	0.71
355	9	4	3	3	1	2	2	5	110.1	10	6	5	8	0.10	0.33	0.40	0.63
356	10	4	2	2	3	2	1	5	71.3	13	6	3	7	0.23	0.33	0.33	0.71
357	10	2	3	2	4	2	2	4	44.0	14	4	5	6	0.29	0.50	0.40	0.67
358	12	3	5	2	0	2	1	4	51.6	12	5	6	6	0.00	0.40	0.17	0.67
359	10	5	1	1	5	2	1	4	53.1	15	7	2	5	0.33	0.29	0.50	0.80
360	10	0	3	2	5	2	4	3	84.4	15	2	7	5	0.33	1.00	0.57	0.60
361	4	4	5	4	6	1	2	3	177.6	10	5	7	7	0.60	0.20	0.29	0.43
362	7	7	5	1	1	2	2	4	85.1	8	9	7	5	0.13	0.22	0.29	0.80
363	11	6	0	3	3	0	3	3	76.2	14	6	3	6	0.21	0.00	1.00	0.50
364	16	2	3	4	1	1	0	2	13.4	17	3	3	6	0.06	0.33	0.00	0.33
365	14	5	1	2	4	2	1	0	67.5	18	7	2	2	0.22	0.29	0.50	0.00
366	16	2	4	1	3	1	2	0	94.0	19	3	6	1	0.16	0.33	0.33	0.00

Note that:

ddd is a number of times a day is dry given the two previous days were both dry days
ddr is a number of times a day is dry given the previous day was a dry day preceded by a wet day.
drd is a number of times a day is dry given the previous day was a wet day preceded by a dry day
drr is a number of times a day is dry given the last two days were both wet days
rdd is a number of times a day is wet given the two previous days were both dry days
rdr is a number of times a day is wet given the previous day was a dry day preceded by wet day.
rrd is a number of times a day is wet given the previous day was a wet day preceded by a dry day
rrr is a is a number of times a day is wet given the last two days were both wet days
_dd is the total number of occurrences of dry days given the previous day was dry.
_dr is the total number of occurrences of dry days given the previous day was wet
_rd is the total number of occurrences of wet days given the previous day was dry .
_rr is the total number of occurrences of wet days given the previous day was wet
p_rdd is the probability of rain given the two previous days were both dry days
p_rdr is the probability of rain given the previous day was dry preceded by wet day.
p_rrd is the probability of rain given the previous day was wet preceded by dry day
p_rrr is the probability of rain given that the last two days were both wet days

Example

Let us consider day 1 of the year.We can work out the probabilities associated with day 1 as follows

Probability of rain given the two previous days were both dry days

p_rdd = rdd/_dd = 2/18=0.11

2. Probability of rain given the previous day was dry preceded by wet day

p_rdr = rdr/_dr = 2/5=0.4

Probability of rain given the previous day was wet preceded by dry day
p_rrd = rrd/_rd = 2/4=0.5
Probability of rain given that the last two days were both wet days
p_rrr = rrr/_rr = 1/2 = 0.5

APPENDIX 8

2.1.1 Demonstration on how one can simulate crop growth using PT

Having opened the software, the user would be faced with a screens as shown below: -

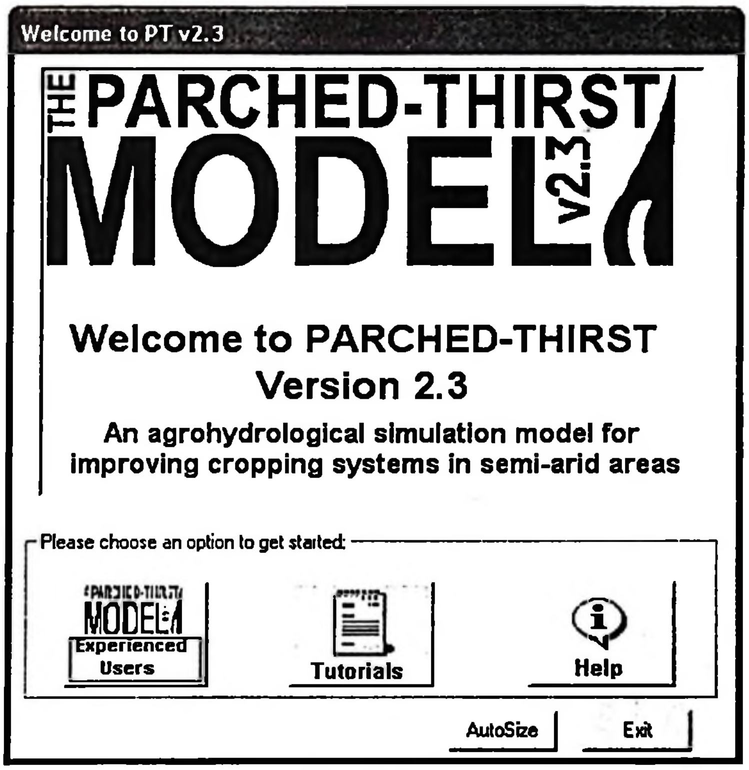


Figure 2.1

There are three options as they can be seen from the screen. Assuming that, the user is already experienced with the software, he should, then click on the experienced users option, otherwise he should go for the tutorials. By clicking in this option, a user would be brought into the PT main window as it appears below:-

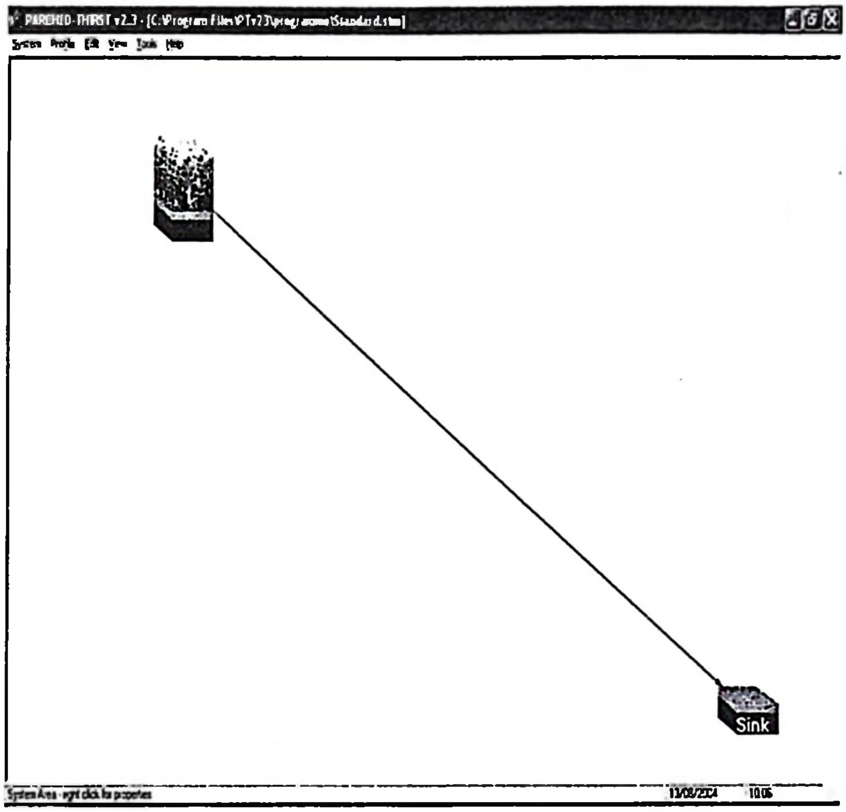


Figure 2.2

3. The screen above shows a single profile with a sink which is meant to be the final recipient of water runoff from the profile. For a simulation to take place, at least two profiles are needed. Assuming that, a user is going to use two profiles only, he should click in the profile option from the window's menu to make a new profile, and a screen would appear with two profiles numbered 1 and 2. Later, the user should right click on the second profile icon to activate it, and then join it to the first profile by an arrow to form a complete farming system as shown below.

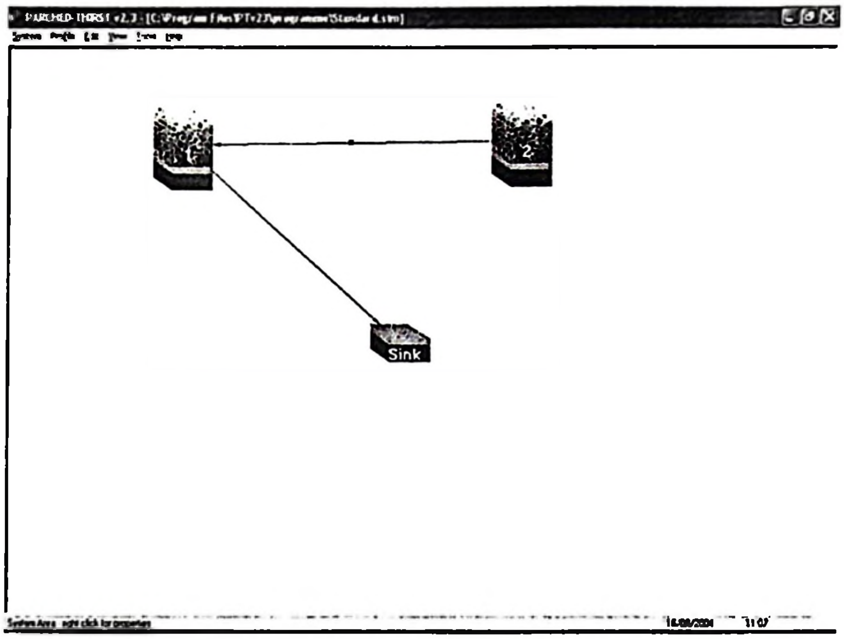


Figure 2.3

4. The above screen indicates two profiles, 1 and 2, with profile2 joined to profile 1.This means that, profile 1 is the water receiving profile whereas profile 2 is the water producing profile. At this stage, a user needs to characterise the properties of the profiles and of the whole system . To characterise the system, a user should click in the system option of the window's menu and choose system properties option. A window will appear. The system window will include four menu bars for four main characteristics of a system. These are, time, sowing , site and weather conditions of the system. By clicking in each of those a user will have an opportunity to fill in some system properties. Below is a screen showing the system window with summary of the four main characteristics of a system

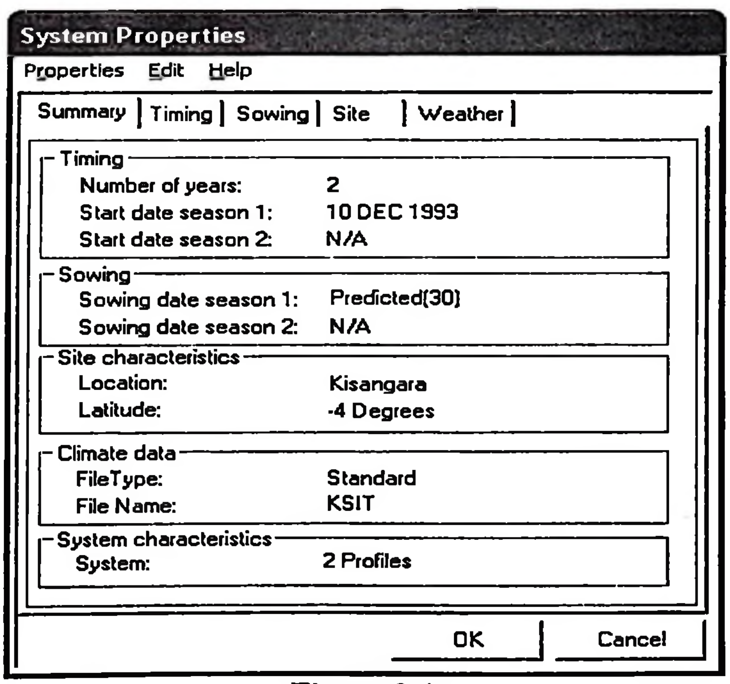


Figure 2.4

5. The above screen indicates that, the system is found at Kisangara site located at 4 degrees south of equator. The climatic conditions of the system will therefore be determined by the site's climatic condition fed in the PT program with a name KSIT.The screen indicates that, the system is simulating crop growth for two different years in only one season of a year commencing from 10 December 1993. By clicking OK the user will be brought back into the previous screen (figure 2.3)

As for the profile properties, the user should right click a profile's icon and fill in the profile's properties in the displayed window. The profile's window will include four menu bars for four main characteristics of a profile. These are, type of crop grown, weeds management practice, soil surface management for water control and soil properties. In addition there will be a menu for the general properties of a profile such as profile size, slope, and others. Thus by clicking in each of those menu a user will be able to fill in some properties to characterise the profile. As an example properties of profile 2 are shown in the screen below:-

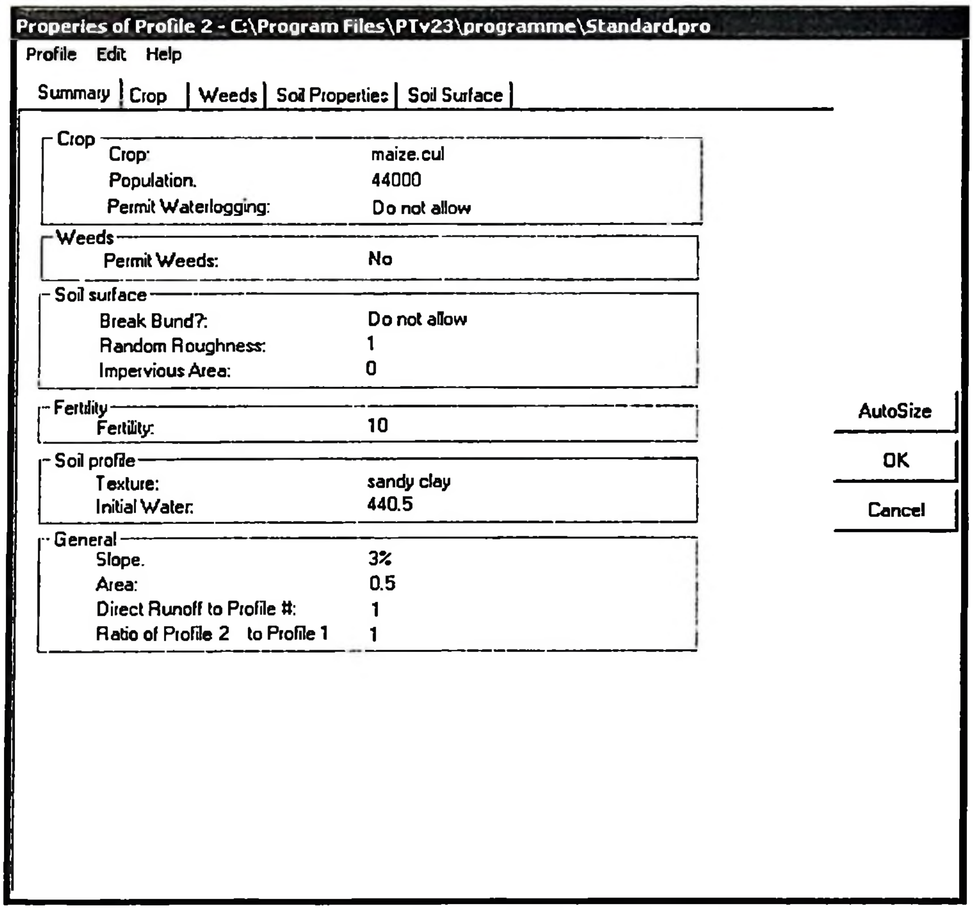


Figure 2.5

The above screen indicates properties of profile 2. The screen indicates that, about 44,000 maize plants are being grown in the profile. Also no particular techniques are applied to control weeds. Further more, the profile is shown to be of sandy clay, with an area of about 0.5 hector and a slope of 3%.Again by clicking OK he will be brought back into the previous screen (figure4)

Having specified both, the system properties and the profile's properties as explained before, the user would now be facing PT main window and should now go into the window's menu in the system bar and click run option, to simulate the crop growth. A screen would appear asking him whether he wants to see the process of the simulation, if he opts for yes then another screen will appear asking him specify the speed of the simulation which is either, slow, medium or fast. The two screens are shown below:-

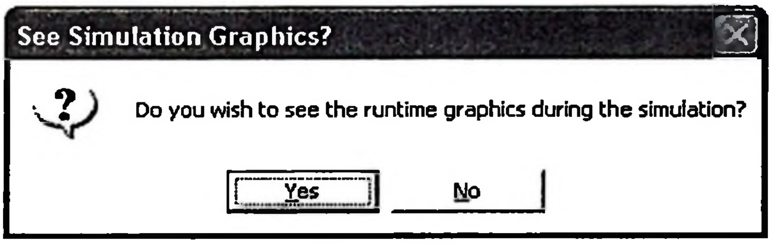


Figure 2.6

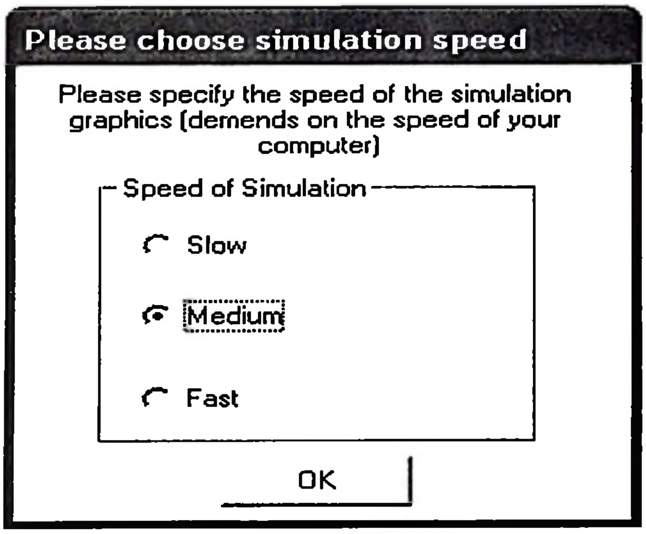


Figure 2.7

By clicking OK in the above screen simulation will run within few seconds and user would be seeing how the process goes on, as shown in the screen below.

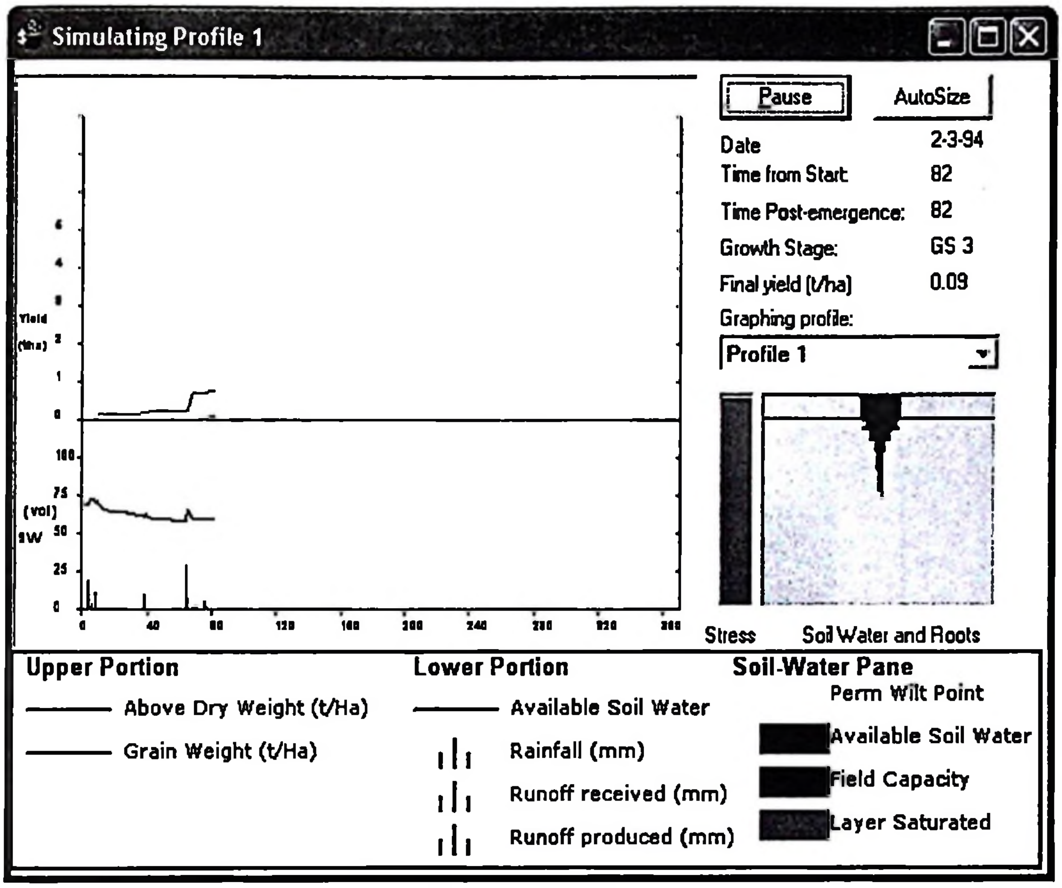


Figure 2.8 (a)

At the end of the simulation process summary of results will be immediately displayed in graphics as shown in the next two screens (figures 2(10)&2(11). The first screen(figure10) indicates the final output at the end of the season in each of the two years. Seven outputs for the system as a whole can are considered, these are rainfall, runoff, yield, evapotransporation (ET), evaporation and drainage The second screen (figure11) gives the user an opportunity to view, outputs by profile and on daily basis. It also gives more outputs than the first, where, the crop's and weed's physical development (weight in tonnes) is given over time. To see the output of one variable to another, one needs to simply click in the relevant icon as given in the two screens.

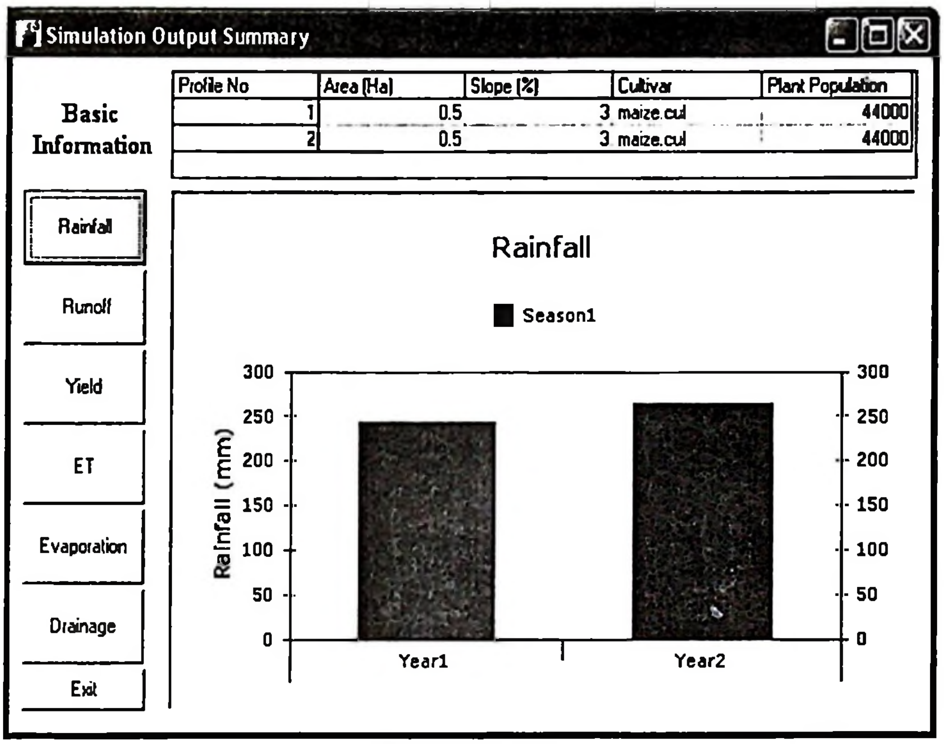


Figure 2.8(b)

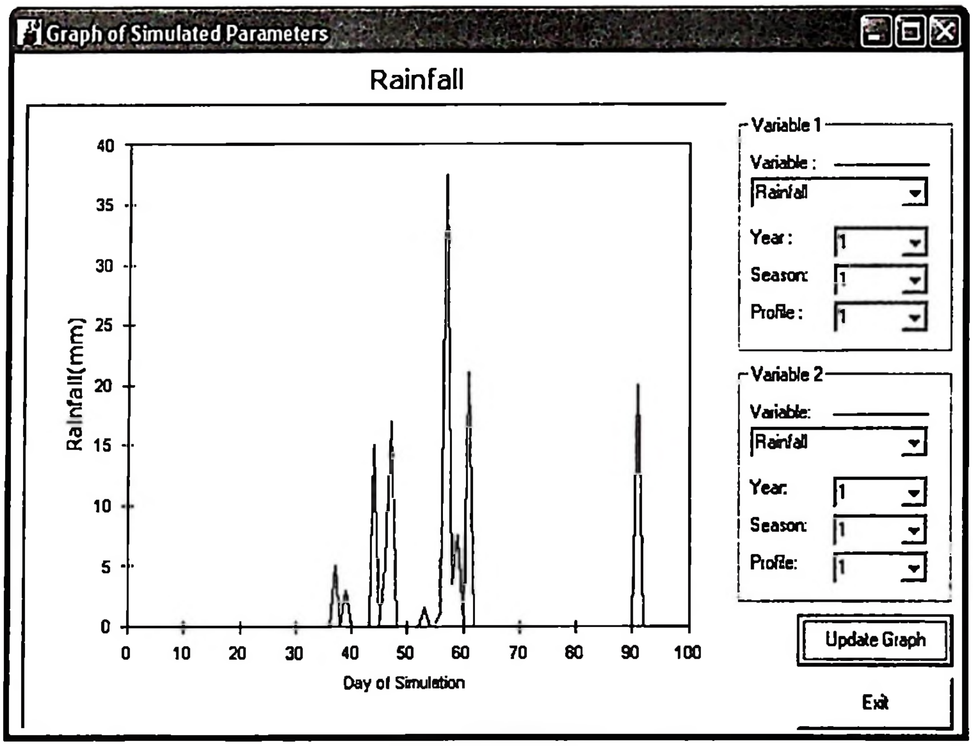


Figure 2.8(c)

APPENDIX 9

How to simulate climatic data using the PT program

Having opened the software, the user would be faced with a screen as
Shown below: -

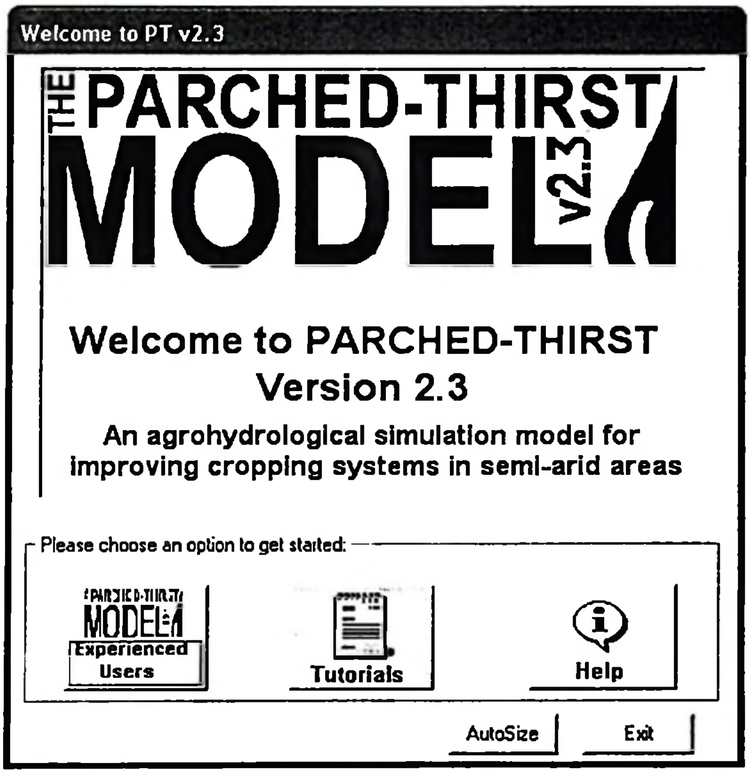


Figure 3.1.1

There are three options as they can be seen from the above screen. Assuming that, the user is already experienced with the software, he should then click on the Experienced Users option, otherwise he should go for the Tutorials. By clicking in this option, the following screen appears.

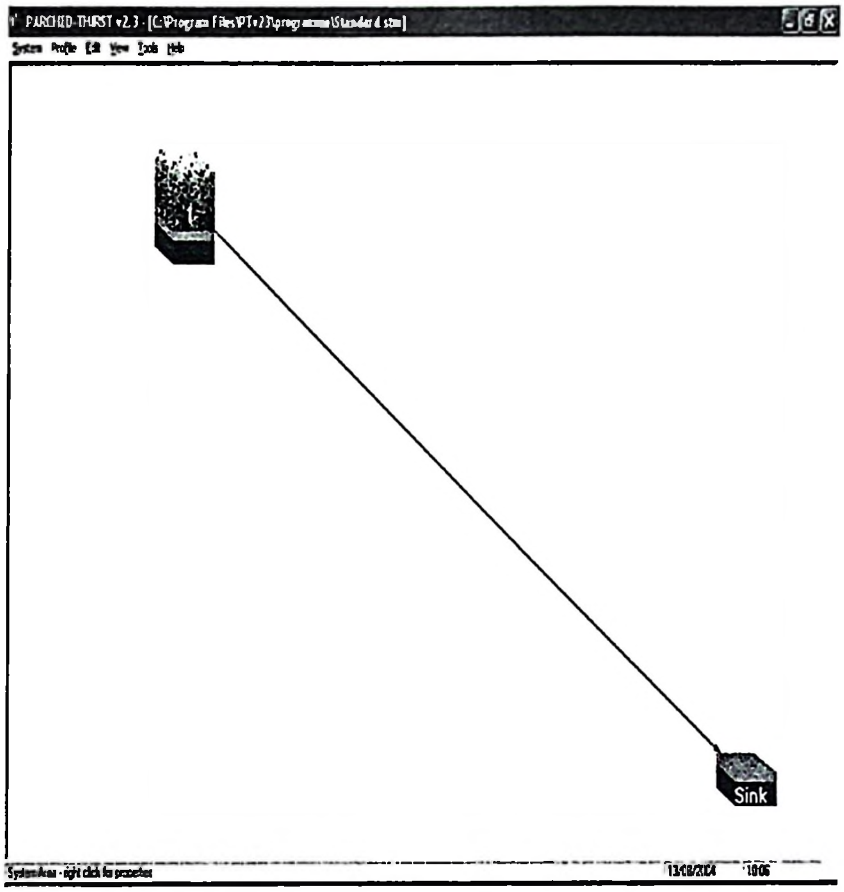


Figure 3.1.2

3. The user then clicks in the tools option on the menu bar for a climatic generator option. The following screen, giving the user an option to simulate data using historical records appears.

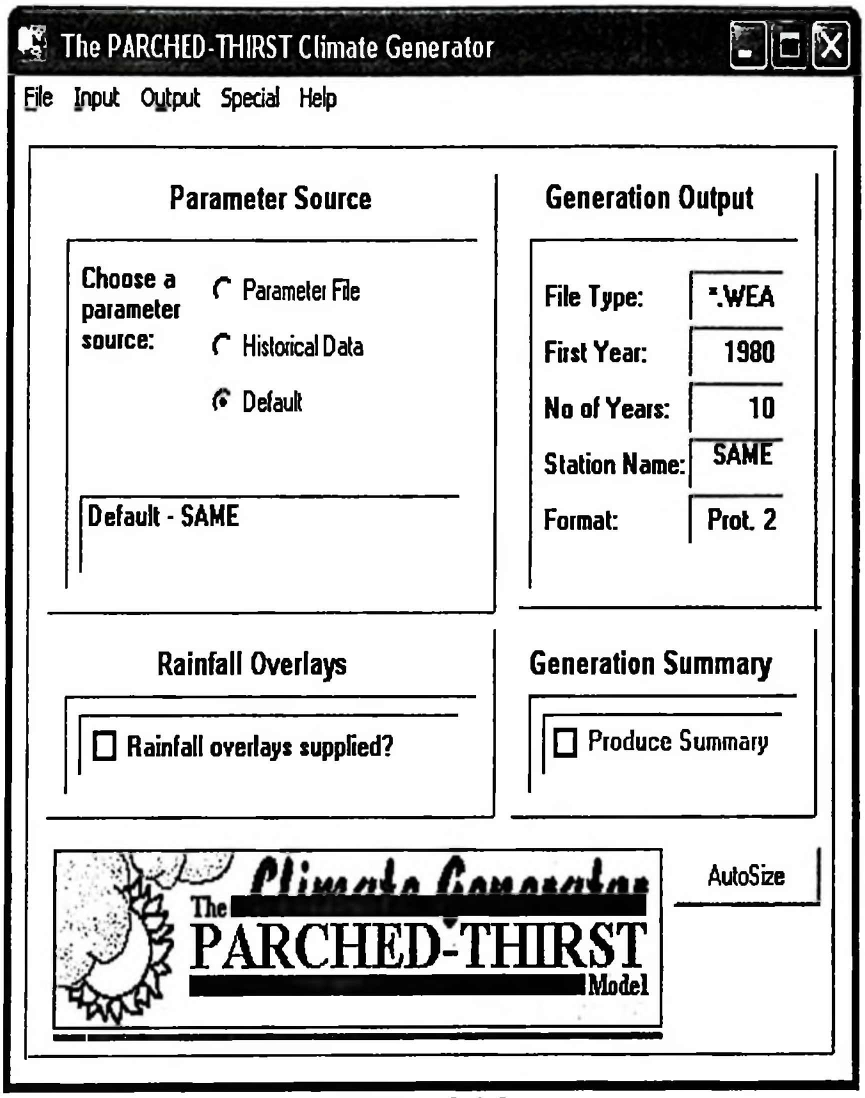


Figure 3.1.3

By clicking on the Historical Data option, the following screen appears, giving the user an option to select a file to be used as historical data.

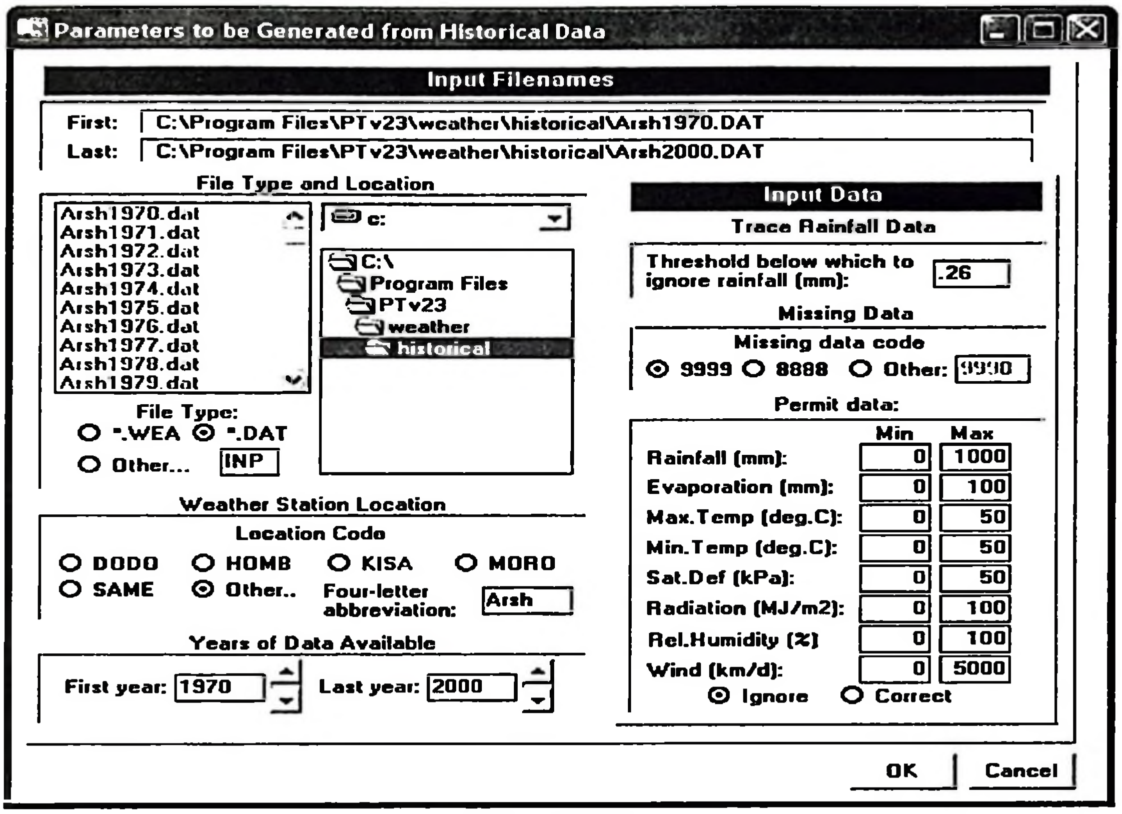


Figure 3.1.4

The screen above, indicates that 31 years (1970-2000) of Arusha climatic records have been selected to be used as a basis for the simulation process.
The user could have as well opted for other sites of climatic records such as Dodoma, Hombolo, Kisa, Same and Morogoro as shown in the above screen. The screen indicates other useful information, such as, the rainfall threshold record, which is taken to be 0.26, and the code for the Missing data, which is 9999.

Having selected the site and the number of years as described above, the user would click the OK option and returns to the earlier screen as shown below: -

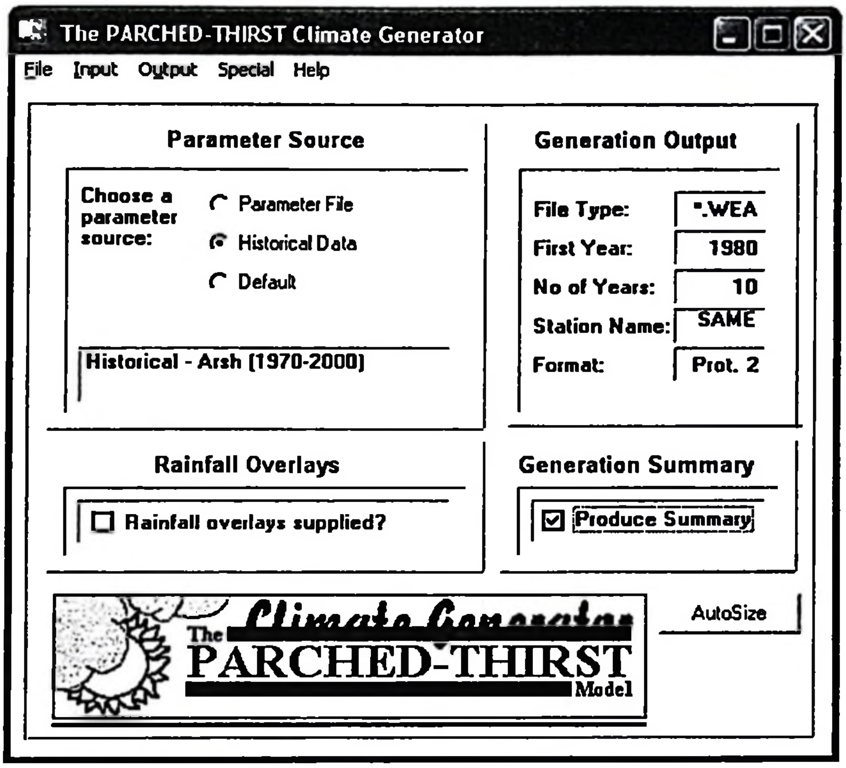


Figure 3.1.5

In this screen the user should now click the option to " Produce Summary," as shown above. This option gives him an opportunity later to view the comparison of statistical properties of the historical data against those of the simulated data.

6 The user should then click in an Output option of the menu bar so as to specify the file name and the location for the generated data as well as the number of years of simulation. In so doing, the following screen will appear

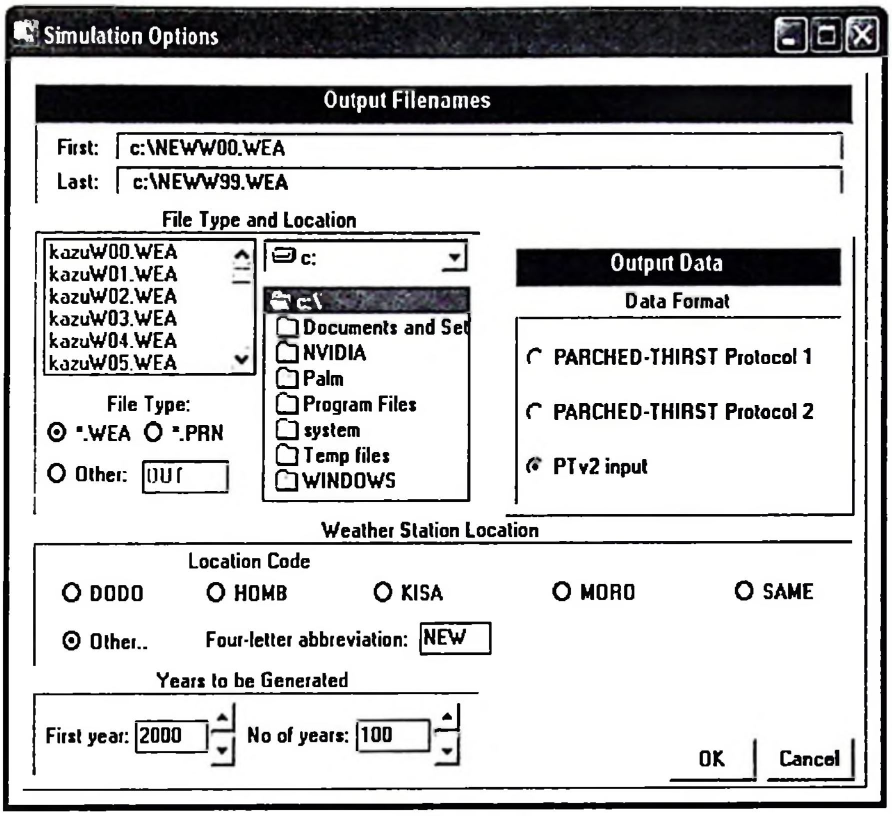


Figure 3.1.6

The above screen indicates that, 100 years of climatic data would be simulated and given the site name "NEW". The generated files would be located on the C: drive

The user should then click the OK option and return to the following screen

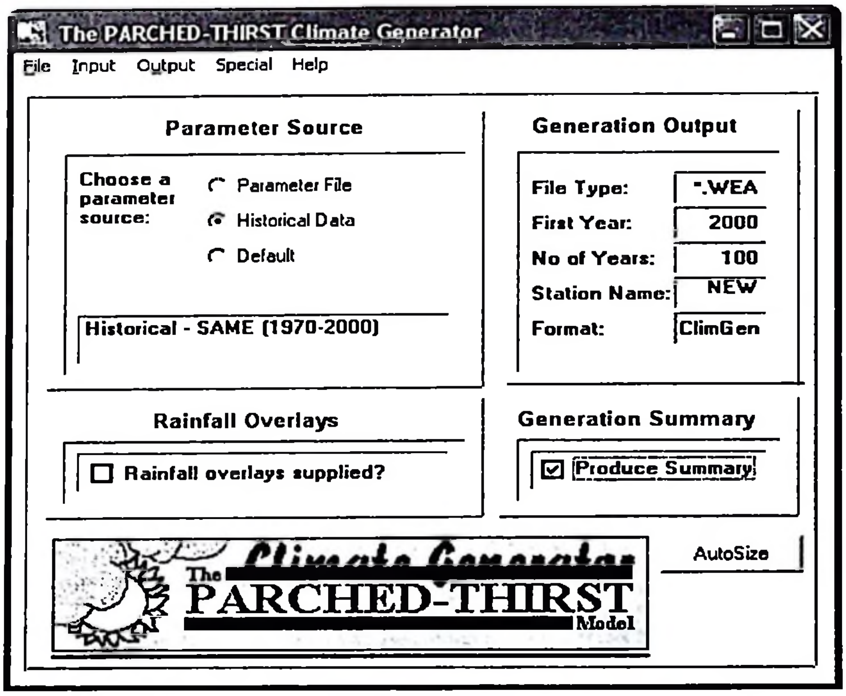


Figure 3.1.7

In this screen, the user should click the Output option again and go into an option for Evaporation. He will then be required to enter details on the latitude and altitude of the chosen site, as shown below:-

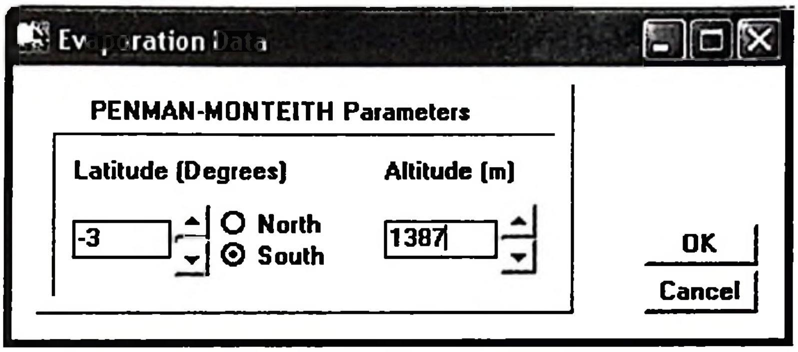


Figure 3.1.8

The screen above indicates the latitude and altitude for Arusha Town in Tanzania

8. Having entered the details on the altitude and latitude of the chosen site, the user can finally run the simulation of data by clicking on the Run option in the File menu. In so doing he will generate 100 years of data records for the variables rainfall, maximum temperature, minimum temperature, radiation, relative humidity, wind and evaporation. The hundred years of data are kept in a hundred separate files named with numbers 00, 01, 02,...99. Below are the hundred files shown as well as the format structure of a file.

The hundred data files.

(To save space, only some of them have been shown)

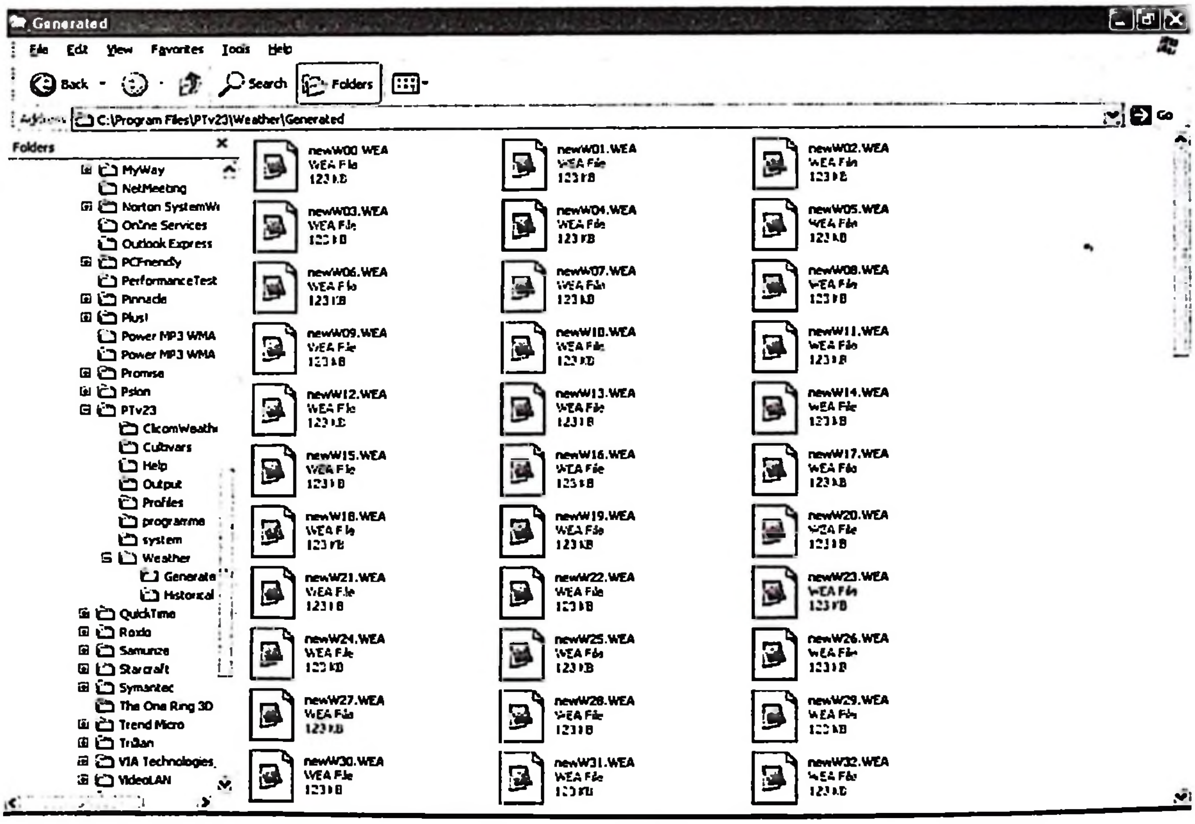
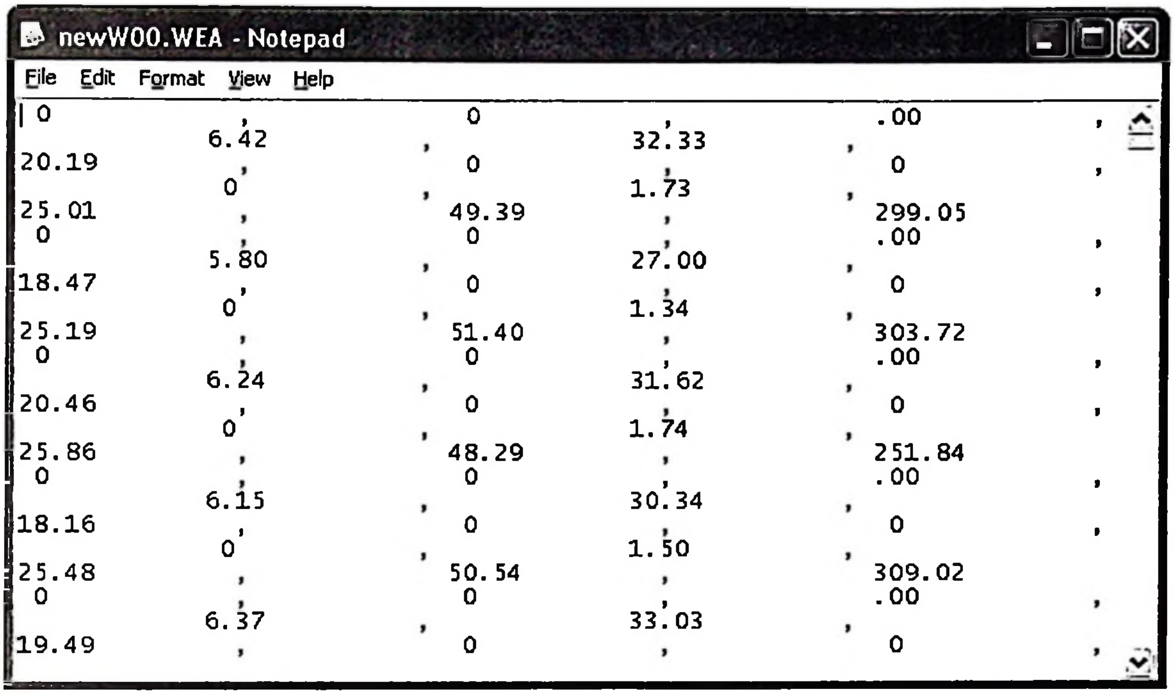


Figure 3.1.9 (a)

The format structure of a file when opened in Notepad
(Note that, lines are wrapped)



```
newW00.WEA - Notepad
File  Edit  Format  View  Help
0            ,          0          ,          .00      ,
       6.42        ,          32.33      ,
20.19        ,          0          ,          0        ,
       0           ,          1.73       ,
25.01        ,          49.39      ,          299.05
0            ,          0          ,          .00      ,
       5.80        ,          27.00      ,
18.47        ,          0          ,          0        ,
       0           ,          1.34       ,
25.19        ,          51.40      ,          303.72
0            ,          0          ,          .00      ,
       6.24        ,          31.62      ,
20.46        ,          0          ,          0        ,
       0           ,          1.74       ,
25.86        ,          48.29      ,          251.84
0            ,          0          ,          .00      ,
       6.15        ,          30.34      ,
18.16        ,          0          ,          0        ,
       0           ,          1.50       ,
25.48        ,          50.54      ,          309.02
0            ,          0          ,          .00      ,
       6.37        ,          33.03      ,
19.49        ,          0          ,          0        ,
```

Figure 3.1.9(b)

The format structure of a file when opened in Microsoft excel
(Note that, only some of the columns are appearing)
Figure 3.1.9(c)

20.19	0	0	0	1.73	25.01	49.39	299.05
18.47	0	0	0	1.34	25.19	51.4	303.72
20.46	0	0	0	1.74	25.86	48.29	251.84
18.16	0	0	0	1.5	25.48	50.54	309.02
19.49	0	0	0	1.89	26.83	44.67	233.35
19.6	0	0	0	1.75	25.42	43.76	307.08
16.96	0	0	0	1.14	25.43	60.35	284.24
18.27	0	0	0	1.89	26.9	40.12	284.32
19.3	0	0	0	1.88	27	41.77	256.96
19.63	0	0	0	2.08	26.22	40.28	283.76
16.49	0	0	0	1.24	25.89	50.16	264.86
19.19	0	0	0	1.92	26.86	40.99	250.31
19.63	0	0	0	1.74	26.12	49.81	295.63
19.13	0	0	0	1.89	26.9	40.07	253.91
18.46	0	0	0	1.76	24.87	42.75	343.61

Further to this output, a screen showing comparison of statistical properties between historical and generated data will appear. The comparison is made using graphs of both historical and simulated data. As an example, rainfall and maximum temperature are used below to make a comparison of historical data against the generated data.

Comparison between historical against simulated data by PT
(i) Rainfall
(a) Based on mean daily observations

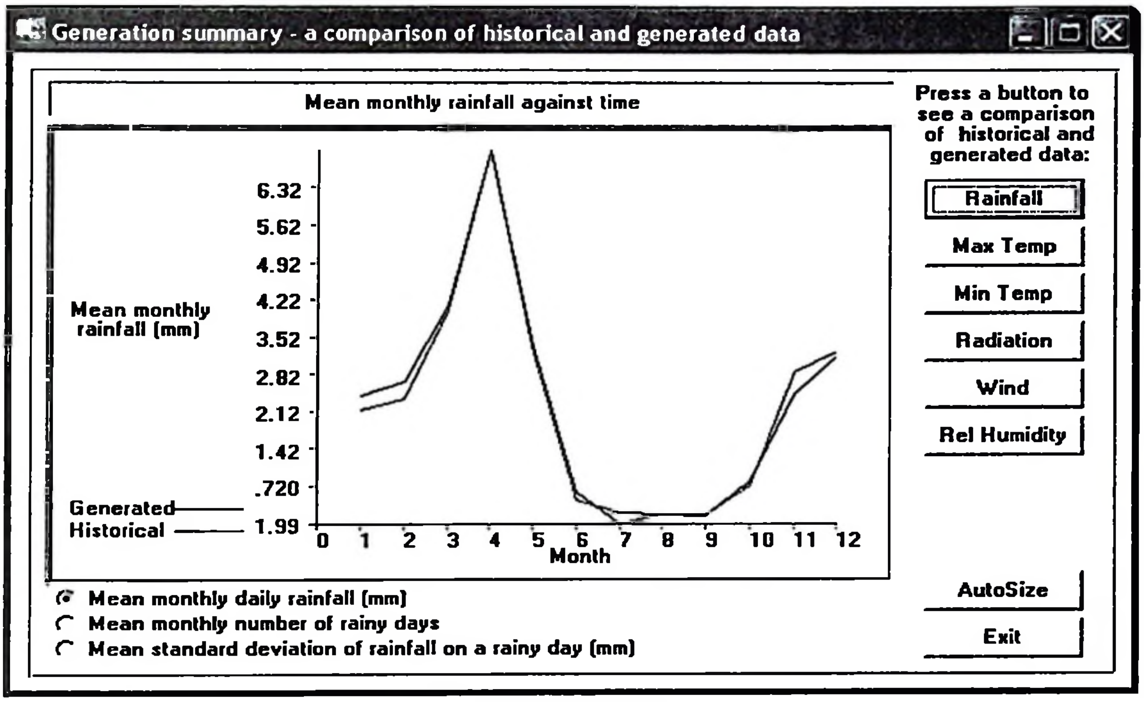


Figure3.2.1(a)

(b) Based on mean number of rainy days

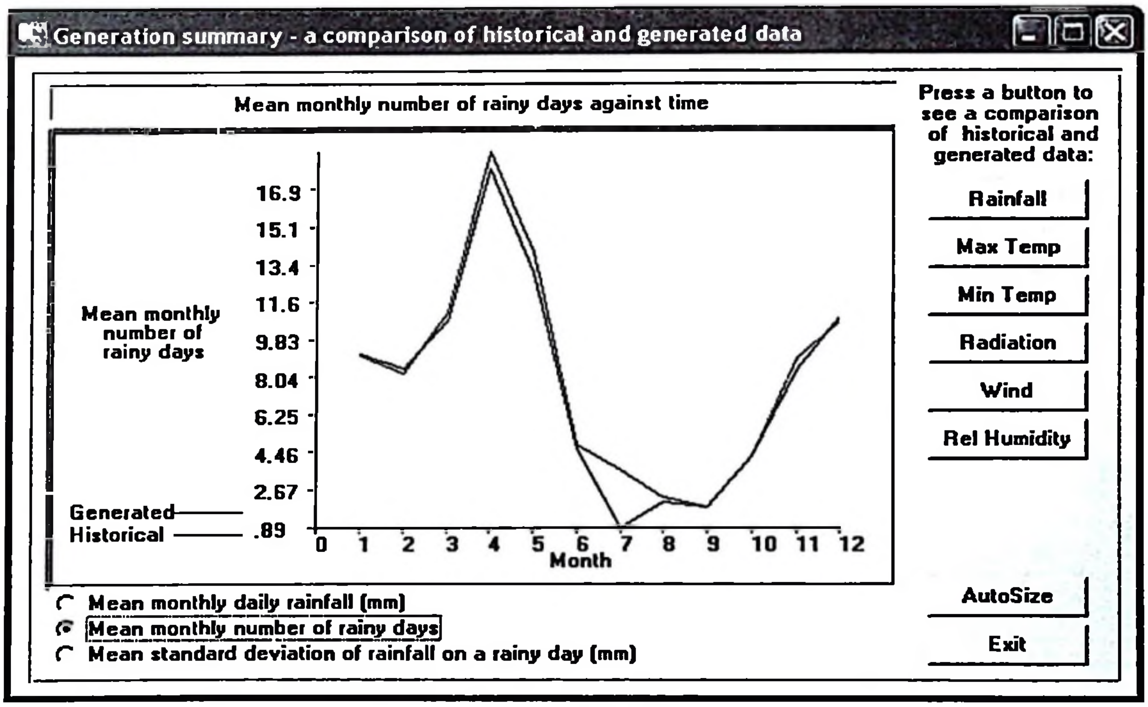


Figure3.2.1 (b)

(c) Based on the standard deviations

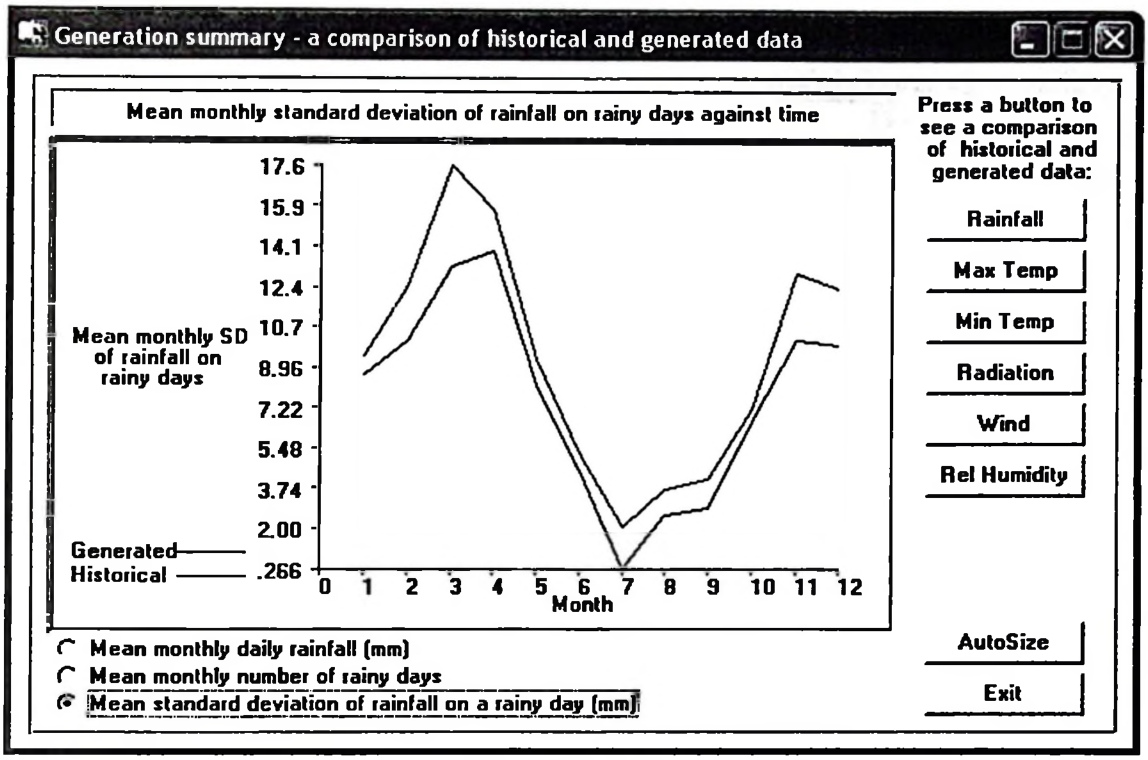


Figure3.2.1(c)

Maximum Temperature
Based on mean monthly maximum temperature

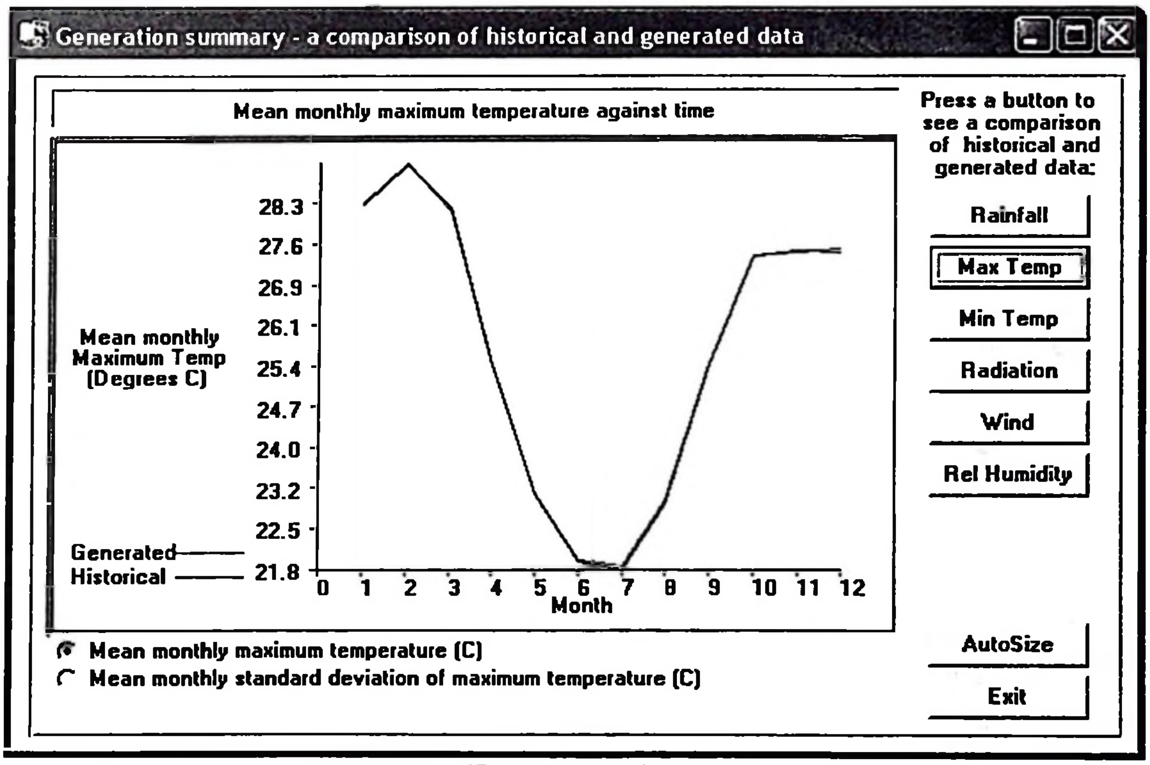


Figure3.2.2 (a)

(c) Based on mean monthly standard deviation of maximum temperature

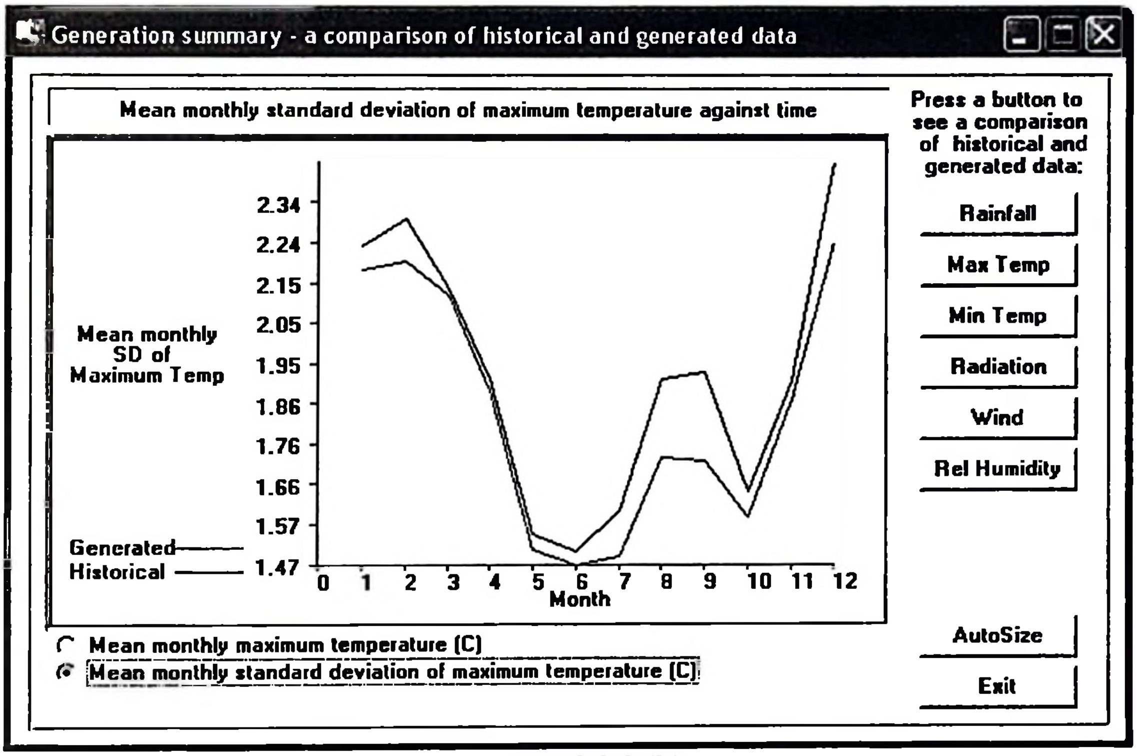


Figure3.2.2 (b)

APPENDIX 10

6.2 Fitting rainfall probability models.

6.2.1 Fitting the zero order

Five harmonics were found to be appropriate and the estimates of the parameters are show in table 6.2 below: -

Table6.2

: Estimates of parameters of for a zero order model					
	Estimate	s.e.	t(*)	t pr.	Antilog of estimate
Constant	-1.174	0.026	-45.44	<.001	0.309
sin[1]	0.949	0.038	25.18	<.001	2.58
cos[1]	0.447	0.035	12.63	<.001	1.563
sin[2]	- 0.669	0.037	-18.30	<.001	0.512
cos[2]	-0.085	0.037	-2.34	<.019	0.918
sin[3]	-0.273	0.036	-7.57	<.001	0.761
cos[3]	0.222	0.037	6.02	<.001	1.248
sin[4]	0.179	0.036	4.98	<.001	1.196
cos[4]	-0.029	0.036	-0.80	<.425	0.972
sin[5]	-0.091	0.034	-2.63	<.008	0.913
cos[5]	-0.107	0.034	-3.12	<.002	0.898

6.2.2 Fitting the first order probability models

Five harmonic were found to be appropriate for the probability model of rain given dry whereas a four harmonics were appropriate for the probability of rain given rain. Their estimates are shown in (i) and (ii) below: -

The probability of rain given dry status in the previous day

Table6.3 (i)

Estimates of parameters					
					Antilog
	Estimate	s.e.	t(*)	t pr.	of estimate
Constant	-1.605	0.033	-49.16	<.001	0.2009
sin[1]	0.806	0.047	17.13	<.001	2.239
cos[1]	0.424	0.045	9.36	<.001	1.528
sin[2]	-0.533	0.047	-11.38	<.001	0.5868
cos[2]	-0.106	0.046	-2.32	<0.020	0.8996
sin[3]	-0.239	0.046	-5.23	<.001	0.7875
cos[3]	0.211	0.047	4.52	<.001	1.234
sin[4]	0.137	0.046	2.99	<0.003	1.147
cos[4]	0.060	0.046	1.31	<0.189	1.062
sin[5]	-0.102	0.045	-2.26	<0.024	0.9032
cos[5]	-0.130	0.045	-2.88	<0.004	0.8781

The probability of rain given rain status in the previous day

Table 6.3(ii)

Estimates of parameters					
	Estimate	s.e.	t(*)	t pr.	Antilog of Estimate
Constant	-0.1856	0.0511	-3.63	<.001	0.8306
sin[1]	0.6760	0.0776	8.71	<.001	1.966
cos[1]	0.2373	0.0660	3.59	<.001	1.268
sin[2]	-0.5036	0.0709	-7.10	<.001	0.6044
cos[2]	0.0304	0.0722	0.42	<0.674	1.031
sin[3]	-0.2603	0.0668	-3.90	<.001	0.7708
cos[3]	0.0639	0.0688	0.93	<0.353	1.066
sin[4]	0.2080	0.0619	3.36	<.001	1.231
cos[4]	-0.1544	0.0609	-2.53	<0.011	0.8569

6.2.3 Fitting the second order probability models

Four harmonics were found to be relevant for the probability of rain given that the two previous days were both dry days whereas as two harmonics were appropriate for the probability of rain given the previous day was dry proceeded by rain day.

The probability of rain given the previous day was rainy preceded by dry was fitted to two harmonics where as the probability of rain given that, the last two days were both rain was fitted to only one harmonic.

Estimates of parameters for the fitted models are shown in the tables below:-

(i) The probability of rain given that, previous day was a dry day preceded by a dry day.

Table.6.4 (i)

Estimates of parameters					
	Estimate	s.e.	t(*)	t pr.	Antilog of estimate
Constant	-1.7501	0.0374	-46.78	<.001	0.1738
sin[1]	0.7335	0.0548	13.39	<.001	2.082
cos[1]	0.3860	0.0513	7.53	<.001	1.471
sin[2]	-0.4933	0.0529	-9.33	<.001	0.6106
cos[2]	-0.0481	0.0530	-0.91	<0.364	0.9530
sin[3]	-0.2474	0.0523	-4.73	<.001	0.7808
cos[3]	0.2120	0.0533	3.97	<.001	1.236
sin[4]	0.1054	0.0526	2.00	<0.045	1.111
cos[4]	0.0276	0.0523	0.53	<0.597	1.028

(ii) The probability of rain given that, previous day was a dry day preceded by rain

Table.6.4 (ii)

Estimates of parameters					
	Estimate	s.e.	t(*)	t pr.	Antilog of estimate
Constant	-1.1043	0.0813	-13.59	<.001	0.3314
sin[1]	0.723	0.109	6.64	<.001	2.060
cos[1]	0.511	0.109	4.67	<.001	1.666
sin[2]	-0.542	0.101	-5.35	<.001	0.5813
cos[2]	-0.2916	0.0941	-3.10	<0.002	0.7471

(iii) probability of rain given that, previous day was a rain day preceded by a dry day.

Table 6.4(iii)

Estimates of parameters					
	Estimate	s.e.	t(*)	t pr.	Antilog of Estimate
Constant	-0.2494	0.0642	-3.88	<.001	0.7793
sin[1]	0.6170	0.0914	6.75	<.001	1.853
cos[1]	0.2189	0.0866	2.53	<0.011	1.245
sin[2]	-0.5129	0.0865	-5.93	<.001	0.5988
cos[2]	0.0106	0.0825	0.13	<0.898	1.011

(iv) The probability of rain given that, previous day was a rain day preceded by a rain day.

Table 6.4 (iv)

Figure11 (iv) Estimates of parameters					
	Estimate	s.e.	t(*)	t pr.	Antilog of estimate
Constant	0.2019	0.0647	3.12	<0.002	1.224
Sin[1]	0.4926	0.0910	5.41	<.001	1.637
Cos [1]	0.0000	0.0852	0.00	1.000	1.000

Having determined the probability models for the three orders of Markov chain, their corresponding models for rainfall amounts were fitted. The results are reported in the next section.

6.3 Fitting the mean daily rainfall amounts to the gamma distribution

When the zero order rainfall amounts were fitted to the gamma distribution, it was found that, five harmonics were appropriate for the zero and first orders probability models while the second order needed only three harmonics. The estimates of the parameters in each case are shown in (i), (ii) and (iii) below: -

Zero order

Table 6.5

Estimates of parameters					
	Estimate	s.e.	t(342)	t pr.	Estimate
Constant	1.6209	0.033	49.84	<.001	5.058
sin[1]	0.5267	0.048	11.08	<.001	1.693
cos[1]	0.6793	0.044	15.30	<.001	1.972
sin[2]	-0.2965	0.046	-6.46	< .001	0.743
cos[2]	-0.2585	0.046	-5.62	<.001	0.772
sin[3]	-0.1790	0.046	-3.87	<.001	0.836
cos[3]	0.0629	0.046	1.38	<.169	1.065
sin[4]	0.0269	0.045	0.60	<.552	1.027
cos[4]	-0.1398	0.043	-3.24	<.001	0.870
sin[5]	0.0792	0.041	1.96	<.051	1.082
cos[5]	0.0766	0.041	1.86	<.063	1.080

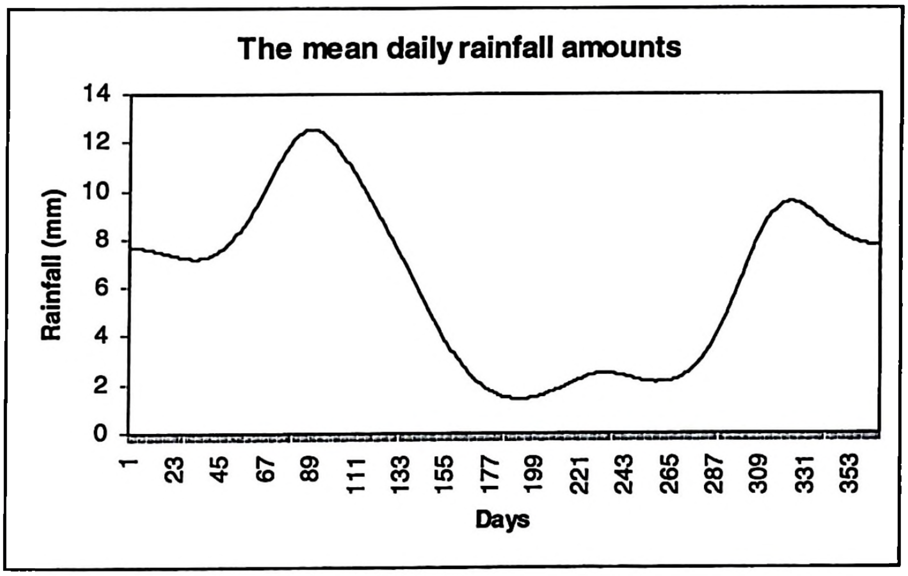


Figure 6.3 (a)

First order

Table 6.6

Figure12 (ii): Estimates of parameters					
	Estimate	s.e.	t(342)	t pr.	Antilog of Estimate
Constant	1.6211	0.0325	49.92	<.001	5.059
sin[1]	0.5267	0.0475	11.10	<.001	1.693
cos[1]	0.6799	0.0443	15.33	<.001	1.974
sin[2]	-0.2965	0.0458	-6.47	<.001	0.743
cos[2]	-0.2579	0.0459	-5.62	<.001	0.773
sin[3]	-0.1789	0.0462	-3.88	<.001	0.836
cos[3]	0.0634	0.0456	1.39	<.165	1.065
sin[4]	0.0272	0.0450	0.60	<.547	1.028
cos[4]	-0.1392	0.0431	-3.23	<.001	0.870
sin[5]	0.0792	0.0404	1.96	<.051	1.082
cos[5]	0.0773	0.0410	1.88	<.061	1.080

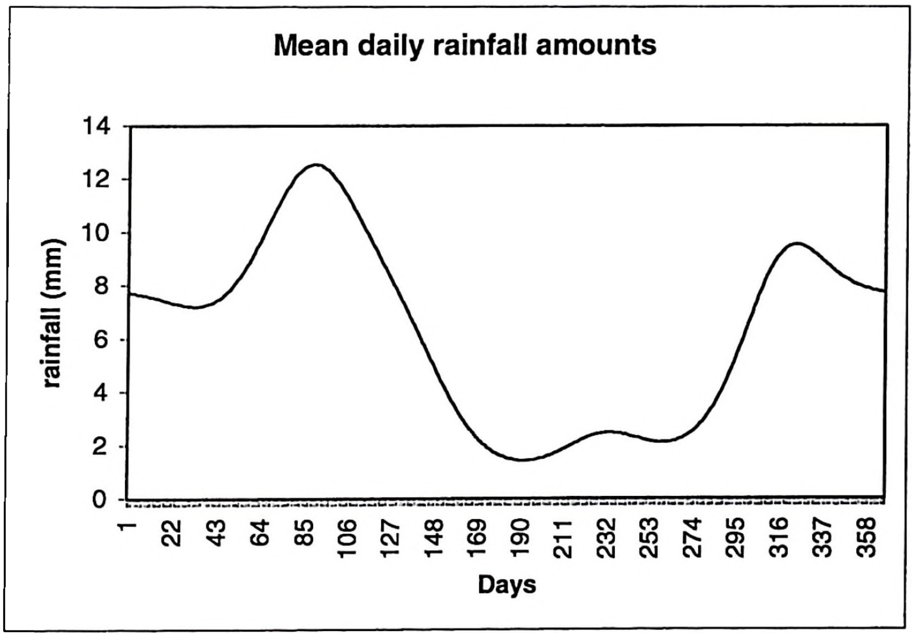


Figure 6.3(b)

Second order

Table 6.7

Estimates of parameters					
	Estimate	s.e.	t(346)	t pr.	Antilog of estimate
Constant	1.6321	0.0328	49.77	<.001	5.114
sin[1]	0.5112	0.0480	10.66	<.001	1.667
cos[1]	0.6704	0.0447	14.98	<.001	1.955
sin[2]	-0.3236	0.0444	-7.29	<.001	0.724
cos[2]	-0.2698	0.0447	-6.04	<.001	0.764
sin[3]	-0.1152	0.0405	-2.85	<.005	0.891
cos[3]	0.0283	0.0419	0.68	<.499	1.029

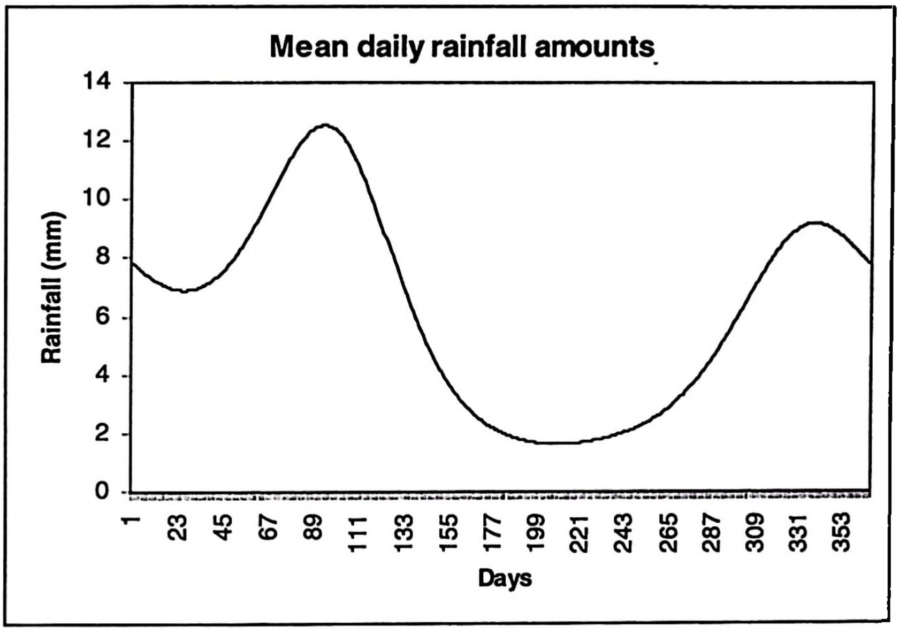


Figure 6.3(c)

SPE
S 619
.W38
K3