Nonparametric estimation of Rainfall Return levels under Climate Change Scenarios in Legazpi City, Albay

Siegfred Roi L. Codia
MS Statistics Student, UP School of Statistics, UP Diliman

ABSTRACT

In a warming world, dry years will become drier, and wet years will become wetter. With this, it is important that we analyze extreme rainfall events by estimating the probability distribution of rainfall amount, since the shape and tails of these distributions give more impact in analysis instead of simple average of rainfall totals. Analysis of uncertainty in future climate is best approached by probabilistic risk assessment using the concept of return period or recurrence interval of a single day hydrologic event such as extreme daily rainfall that causes floods.

In this paper, we estimate the return levels of extreme rainfall amounts in Legaspi City by employing a non-parametric approach. For future periods, LARS-WG was used to simulate future climate in Legaspi under RCP 4.5 and RCP 8.5 greenhouse gas emission scenarios. In each simulated data, probability distribution of annual maxima was estimated by non-parametric kernel estimation and using cross-validation as the bandwidth selection procedure.

The results revealed that if the emission of greenhouse gas will not be controlled in the next decades, high magnitude single day rainfall amounts will be more frequent in Legaspi in terms of annual occurrence by the end of the century. Single day rainfall magnitudes with 5-year, 25-year, 50-year, and 100-year return periods under the different climate change scenarios are presented, which can then be used by disaster scientists for flood modeling.

1. INTRODUCTION

1.1. Significance and Motivation

In the past half century, the Philippines have been stricken by multiple typhoons that costed property damages and fatalities. According to the Climate Change Vulnerability Index, the Philippines is named one of the Asian countries that face extreme risks from natural disasters. Recently, several natural disasters such as the Taal Volcano eruption, Typhoon Molave (Quinta), and Typhoon Goni (Rolly), Typhoon Vamco (Ulysses) were some of the most destructive disasters in 2020, resulting in human casualties and property destructions.

Historically, Bicol is one of the regions that experience the first destructions caused by typhoons, with it facing the Pacific Ocean. Some of the typhoons that crossed the Bicol Region before the year 2000 are Typhoon Emma in 1967, Typhoons Kelly and Lynn in 1981, Typhoon Babs in 1998, among others.

In Legazpi, the highest rainfall amount recorded by the PAGASA weather station was 484.6 mm, observed on November 3, 1967, during the Typhoon Welming (Emma).

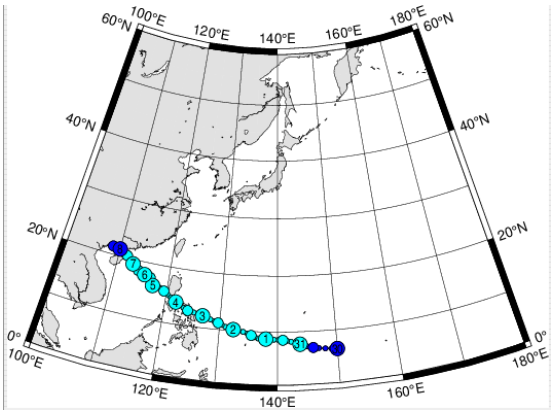


Figure 1. Track map of Typhoon Welming (Emma), October 31 – November 8, 1967
Source: National Institute of Informatics (Japan)

In more recent decades, tropical cyclones that crossed the Bicol Region are Tropical Storm Juaning (international name Nockten) in 2011, Tropical Storm Salome (Haikui) in 2017, and Super Typhoon Rolly (Goni) in 2020, all bringing extreme rainfall amounts in the region.

NOAA data shows that the Legaspi weather station had recorded the highest daily rainfall of 2020 on October 26, 2020, during the onslaught of Quinta (Molave), with 282 mm amount of rainfall. According to a report by DSWD, Bicol Region, among the other PH Regions affected, has the greatest number of people affected by Typhoon Quinta in 2020. The typhoon also caused agricultural damage in Bicol of up to ₱286 million (US$5.9 million). 6,671 of houses were damaged and 243 of them were destroyed.

We are interested in knowing the frequency of occurrence of these extreme rainfall events in the future under different climate change scenarios guided by a goal of building a resilient city. Analysis of uncertainty in future climate is best approached using the probabilistic risk analysis using the concept of return period or recurrence interval of a hydrologic event such as rainfall that causes floods. Hydrologic events for n-years, (e.g., 5-, 25-, 50-, and 100-year) have specific applications in engineering hydrology such as drainage management, flood control, etc.

1.2. Goals and Objectives

The overarching goal of this study is to assess the frequency of extreme daily rainfall events in Legaspi in terms of return periods. This paper mainly applies non-parametric procedures in estimating the return levels.

Specifically, this paper aims:

- To determine 2-year, 50-year, and 100-year precipitation return levels using the historical data and future simulated data.
- To characterize trends of the extreme rainfall magnitude in the 21st century under different climate change scenarios

1.3. Scope and Limitations

This study is focused only on the application of non-parametric estimation of return levels using data for Legaspi weather station. Additionally, limitation of data includes being extracted from an unofficial data source, NOAA's ftp server. Multiple blank and invalid data points were found but were removed and estimated by interpolation methods.

2. RELATED LITERATURE

2.1. Extreme value theory and Return Levels

Analysis of rainfall data strongly depends on its distribution pattern. It has long been a topic of interest in the fields of meteorology in establishing a probability distribution that provides a good fit to daily rainfall. (Sharma and Singh, 2010). The famous statistician R.A Fisher was one of the pioneers in this field by studying the influence of rainfall on the yield of wheat in Rothamsted and showing that instead of total rainfall amount, the distribution of rainfall during a season influences the crop yield. Particularly, statistics of extremes plays a very important role in flood frequency analysis (e.g. Adamowski, 2000).

There are two primary approaches to analyzing extremes of a dataset: block maxima and peaks-over-threshold. The first approach reduces data considerably by taking maxima of long blocks of data, e.g. annual maxima. The second approach analyzes excesses over a high threshold. Coles (2001) wrote a good book on the subject, where he defined concepts used in the topic.

The Generalized Extreme Value (GEV) distribution has theoretical justification for fitting block maxima, while the Generalized Pareto (GP) distribution provides the theory on peaks-over-threshold (POT).

The Generalized Extreme Value Distribution is given by

$$G(x) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-\frac{1}{\xi}}\right\}$$

where $y_+ = \max\{y, 0\}$, $\mu \in \mathbb{R}$, $\sigma > 0$, and $\xi < \infty$. The GEV envelopes three types of distribution function depending on the shape parameter:

I: $(\xi \to 0)$ Gumbel: $\Lambda(x) = e^{-e^{-\left(\frac{x-\mu}{\sigma}\right)}}, \quad x \in \mathbb{R}$

II: $(\xi > 0)$ Fréchet: $\Phi_\xi(x) = \begin{cases} 0 & x \le \mu \\ e^{-\left(\frac{x-\mu}{\sigma}\right)^{-\frac{1}{\xi}}} & x > \mu \end{cases}$

III: $(\xi < 0)$ Weibull: $\Psi_\xi(x) = \begin{cases} e^{-\left|\frac{x-\mu}{\sigma}\right|^{\frac{1}{\xi}}} & x < \mu \\ 1 & x \ge \mu \end{cases}$

In dealing with hydrological extremes, block maxima of rainfall, as described by Chow et al. (1988) or Khaliq et al. (2006), is one of the most natural ways. The annual maxima series (AMS), which consists of one maximum value from each year of record, is used to fit a GEVD.

However, for short historical rainfall series, the use of AMS results to a small sample size which are unsuitable for further analyses. Another approach to modelling hydrological extremes is using the partial duration series (PDS), or sometimes called peaks-over-threshold (POT) as previously mentioned.

For the case of the POT approach, a threshold of exceedance can be approximated by the Generalized Pareto Distribution (GPD). Let X be a random variable and u a high enough threshold. Then the distribution of the exceedance $x-u$ conditional on X exceeding u, $0 \le u < x$, can be approximated as follows

$$G(x-u) = 1 - \left[1+\xi\left(\frac{x-u}{\sigma}\right)\right]_+^{-\frac{1}{\xi}}$$

scale parameter $\sigma > 0$ and shape parameter $\xi \in \mathbb{R}$. Again, the shape parameter determines the type of distribution function (with the same interpretations as the GEV df): heavy tail when $\xi > 0$ (Pareto), upper bound when $\xi < 0$ (Beta), and exponential in the limit as $\xi \to 0$.

It is usually more convenient to interpret these extreme value models in terms of quantiles or return levels, rather than the individual parameter values μ, σ, ξ (Coles, 2001).

Suppose a time series random variable X that is observed every T time intervals and has a cumulative distribution function G_* (commonly an extreme value distribution). Let x_p be the upper p quantile, where $p = 1 - G_*(x_p)$, i.e. $P(X \geq x_p) = p$.

The return period of the value x_p is $\frac{1}{p}$, which can be interpreted as "the number of T periods in which the value x_p is expected to be observed once".

Related with return period is the return level x_T which is the expected amount to be exceeded once every T interval. Mathematically, it can be defined as $x_T = G_*^{-1}\left(1 - \frac{1}{T}\right)$.

For example, if we take the GEVD as the cdf G_*, the $\frac{1}{p}$ period return level is obtained as

$$z_p = \begin{cases} \mu + \frac{\sigma}{\xi}\left[y_p^{\xi} - 1\right] & \text{for } \xi \neq 0 \\ \mu + \sigma \ln y_p & \text{for } \xi = 0 \end{cases}$$

where $y_p = -1/\ln(1-p)$

For the GPD, the formulation of m-observation return level is as follows

$$x_m = u + \frac{\sigma}{\xi}[(m\varsigma_u)^{\xi} - 1)]$$

From these concepts, estimating the return level requires estimating the distribution function G_* with $\hat{G}_*$.

Recent advancements also introduced new procedures in analyzing extremes in hydrology, such as applying bootstrap methods in fitting general extreme value models on rainfall data written in the works of Saeb (2014) and Gilleland (2020). While MLE is perfectly valid estimator of the parameters in an EV distribution, alternatives such as profile likelihood (Gilleland and Katz, 2016) and bootstrap methods are appealing for constructing confidence intervals especially for the case that return levels exceed the temporal range of the data (e.g. 100-year return level with only 20 years of data)

Furthermore, since uncertainties in precipitation can be caused by other climate variables, Kim et.al (2022) proposed a novel procedure for extreme value modeling by using large scale climate indices as covariates of extreme rainfall quantiles, which is relevant given that there are proofs that nonstationary climate due to climate change influence the estimated return levels (Cooley, 2013; Alipour and Leal, 2019).

2.2. Nonparametric estimation

Parametric estimation requires certain assumptions such as knowing the distribution of the data. If we choose a parametric model that is not of appropriate form, then there is a danger of reaching incorrect conclusion. Current flood frequency analysis methods assume that that flood observations come from known pdf, but in hydrological context, the population distribution is not known exactly (Kim and Heo, 2001).

Recently, nonparametric density methods have gained popularity in different fields of science, including hydrology. These methods have the advantage of not requiring assumptions about the distribution of the population of interest (Faucher, et.al, 2001).

Let X be a continuous random variable, with probability density function (pdf) f and cumulative distribution function (cdf) F. Given a sample of data $x_1, x_2, x_3, \ldots, x_n$, the kernel estimator or the Parzen-Rosenblatt estimator (Parzen, 1962) of the density f is defined by

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

where K is the kernel (a nonnegative function), and $h > 0$ is bandwidth parameter that determines the degree of smoothing. Using the relationship between the pdf and cdf, the kernel estimator of the distribution function is given by

$$\hat{F}_h(x) = \int_{-\infty}^{x} \hat{f}_h(t)\, dt = \frac{1}{n}\sum_{i=1}^{n} H\left(\frac{x - x_i}{h}\right)$$

where $H(u) = \int_{-\infty}^{x} K(t)dt$

Other well-known nonparametric estimator of the distribution function is the empirical distribution (Reiss, 1981)

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \le x\}} = \frac{\text{number of } x_i \text{ less than or equal to x}}{n}$$

Since Parzen and Rosenblatt's study, multiple investigations about the kernel density estimators has been done (Scott, 1979; Bowman, 1985; Terell and Scott, 1985; Silverman, 1986; Terell, 1990). However, these theoretical results are asymptotic in nature and will require large samples, hence raising questions on their performance for small samples. Bandwidth selection methods were also discussed by Marron (1989), Sheater (1992), Jones et.al. (1992), and Park and Turlach (1992).

Annual maxima and minima have also been estimated via nonparametric estimators. From the context of kernel estimator, we can use the kernel distribution function to estimate percentiles to a given probability of exceedance. Yakowitz (1983) and Adamowski and Feluch (1983) introduced the kernel method in hydrology. We still use the definition of T period return level

$$\hat{x}_T = \hat{G}_*^{-1}\left(1 - \frac{1}{T}\right)$$

where $\hat{G}_*$ is the estimated distribution function using a kernel estimator.

However, in hydrology context, bandwidth selection methods for the density estimators have been improperly used. Faucher et al. (2001) highlighted this issue then proposed a new bandwidth estimator derived from cumulative distribution function and the properties of the quantile estimator instead of the density function. Kim and Heo (2002) emphasized that while many studies have utilized bandwidth estimators, literature still fail to determine which bandwidth selection method is the best.

Other than the use of kernel estimators for rainfall data, bootstrap methodologies were also introduced to estimate distributions, motivated by availability of short-time series (Holešovský et.al., 2016) or uncertainties in rainfall due to spatial interpolation (Zhang et.al 2017)

Quintela-del-Rio (2011) further pointed out the common mistakes in dealing with statistical problems in hydrology but nonetheless supported the use nonparametric methods in dealing with annual maxima flood series, given that the only perceived disadvantage of nonparametric methods so far is the higher demand in computational time.

Other than the estimation of the distribution, kernel estimators are also used in interpolating missing hydrological data points. Lee and Kang (2015) concluded that the kernel approaches provide higher quality of interpolation than K^{th} nearest neighbor regression approach.

2.3. Climate Change and the Representative Concentration Pathways

In climate change projections, the Representative Concentration Pathways (RCP) from the International Panel for Climate Change (IPCC) 5th Assessment Report (AR5) are used to describe the future world based on different driving forces: human activities, policies, and greenhouse gases (GHG) emissions.

The RCPs represent the range of GHG emissions; they include a stringent mitigation scenario (RCP 2.6), two intermediate scenarios (RCP 4.5 and RCP 6.0), and one scenario with very high GHG emissions, or also called the "Business-as-usual scenario" (RCP8.5).

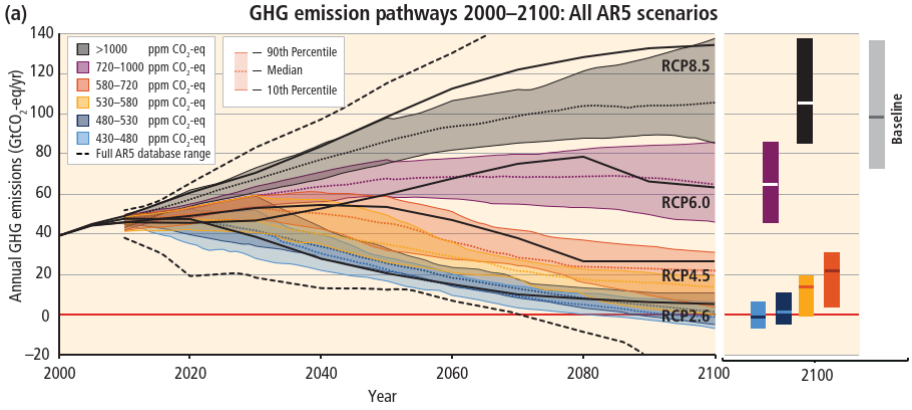


Figure 2. GHG emission pathways (IPCC, Climate Change 2014)

When it comes to precipitation magnitudes, according to AR5, changes in precipitation in a warming world is not uniform. Areas that are on the high latitudes and equatorial Pacific regions (such as the Philippines) are likely to experience an increase in annual mean precipitation by the end of the 21st century under the RCP 8.5 scenario, while a decrease is projected over mid-latitude and subtropical dry regions.

It is also projected that extreme precipitation events are very likely to be more intense and frequent over most mid-latitude land masses and over wet tropical regions by the end of the century as global mean surface temperature increases.

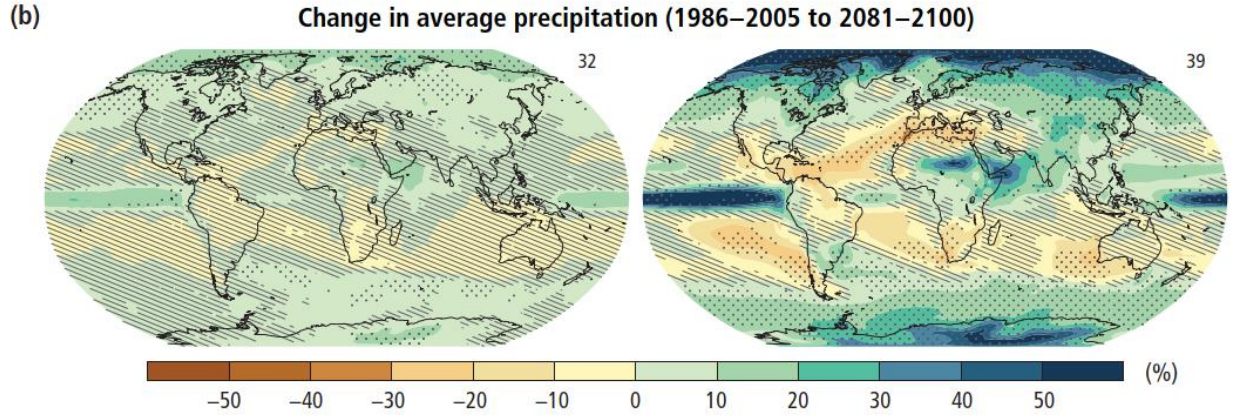


Figure 3. Changes in average precipitation (IPCC, Climate Change 2014)

In the Philippines, DOST-PAGASA and Manila Observatory released a report on Philippine Climate Extremes for 2020, in a goal to present information on historical and projected annual climate extremes in the country. Precipitation extremes have a distinct spatial variability. In Albay, where Legaspi City is located, the total rainfall amount on extremely wet days tends to be higher by the end of the century, especially under RCP 8.5.

3. METHODOLOGY

3.1. Data

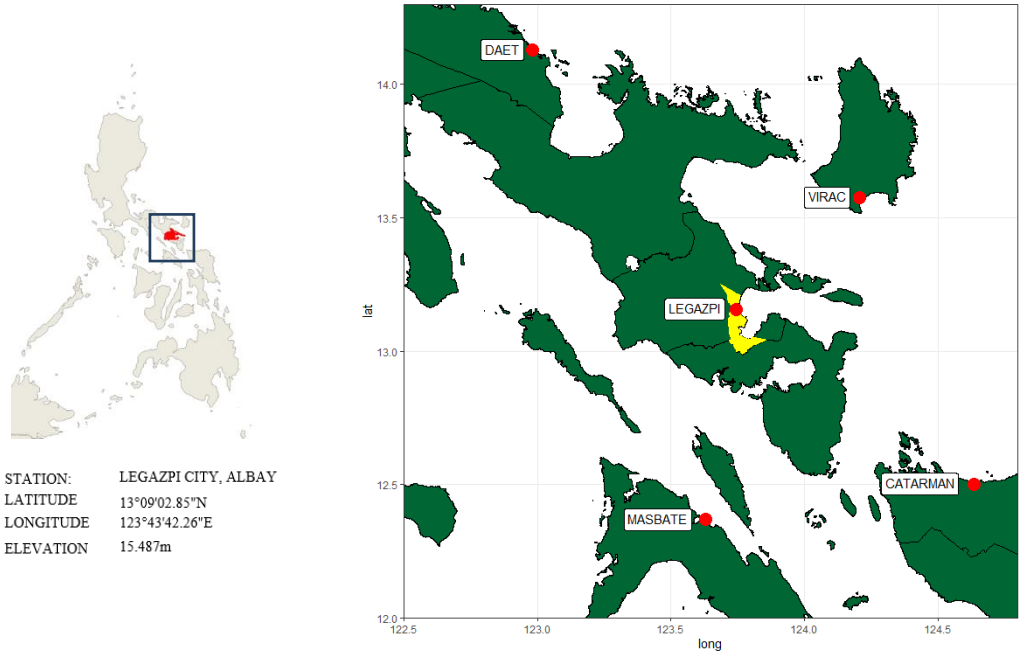


Figure 4. Location of Legazpi weather station

The daily rainfall and temperature data were extracted from NOAA through ftp://ftp.ncdc.noaa.gov/pub/data/gsod/. However, note that this data is not officially from PAGASA and there are some blank or erroneous data points. For the solar radiation, hourly data is extracted from SOLCAST API Toolkit, then the average of the hourly data will be used to represent the solar radiation for a specific day. Table 1 summarizes the definition of the variables, units used, year range of the data available, and the sources.

Table 1. Data definition and Sources

Variable	Definition	Units		Year range	Source
		Original	Converted		
PRCP	Total precipitation (rain and/or melted snow) reported during the day	Inches	millimeter	1973 – 2020	NOAA
TEMP	Mean temperature for the day in degrees Fahrenheit to tenths.	Fahrenheit	Celsius		
MIN	Minimum temperature reported during the day	Fahrenheit	Celsius		
MAX	Maximum temperature reported during the day	Fahrenheit	Celsius		
GHI	Global Horizontal Irradiance: The total irradiance received on a horizontal surface. It is the sum of direct and diffuse irradiance components received on a horizontal surface.	W/m^2	$\frac{MJ}{m^2}/day$	2007-2020	Solcast

3.1.1. Removing invalid data points

PAGASA's official report on climatological extremes shows that within 1903 to 2020, Legazpi weather station has recorded minimum and maximum temperatures at 13.9℃ and 37.7℃ respectively, and the greatest daily rainfall amount recorded is 484.6 mm. These will serve as the upper and lower bounds of the data. Values that are outside these are removed.

3.1.2. Filling the missing datapoints

Rainfall and Mean Temperature

Now, since there are missing data points, we perform imputation and estimation procedures for those missing days. For now, we will just perform a simple linear interpolation for the precipitation `PRCP` and daily mean temperature `TEMP` using the `na_interpolation` function in the `imputeTS` package in R.

Max and Min Temperature

For the Max and Min Temperatures, a different imputation approach will be used due to the constraint $min < mean < max$. Simple interpolation on max and min data will sometimes violate this constraint.

The values of the maximum and minimum temperatures move along with the mean temperature. We can model the daily max and min based on their distance from the daily mean temperature. For this, we will obtain two other variables, the daily temperature ranges in terms of the differences $T_{max-mean}$ and $T_{mean-min}$, estimate the missing differences $\hat{T}_{max-mean}$ and $\hat{T}_{mean-min}$ by linear interpolation, then recompute the values:

$$\hat{T}_{max} = T_{mean} + \hat{T}_{max-mean}$$

$$\hat{T}_{min} = T_{mean} - \hat{T}_{mean-min}$$

Solar Radiation

For the Solcast dataset of Solar Radiation variable, imputation will not work if we estimate data before 2007. A Generalized Linear Model with gaussian family with logarithmic link based on temperature range (Max-Min) and monthly seasonality can be used to estimate the daily solar radiation for the years before 2007. The log link ensures that the estimated value of solar radiation will always be positive.

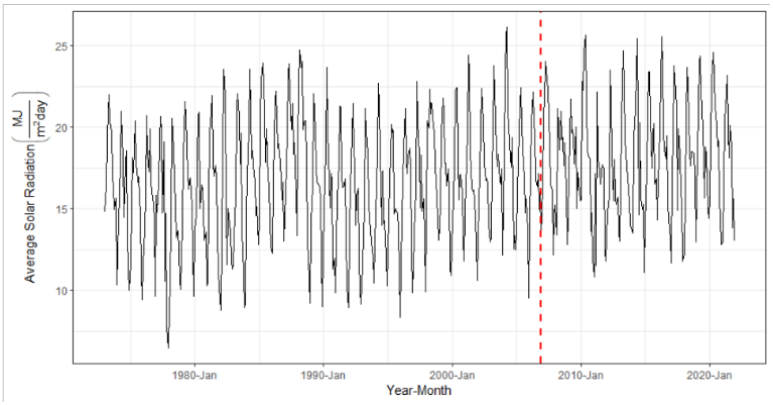


Figure 5. Generated Historical Monthly Mean Radiation using the estimated model

3.2. Non-parametric Estimation of Return Levels

Analysis of uncertainty in future climate in best approached using the probabilistic risk analysis using the concept of return period or recurrence interval of a hydrologic event such as rainfall that causes floods.

In estimating return levels, the annual maxima will be used, and then a distribution function must be estimated using this data. We can estimate the distribution function using non-parametric methods, specifically by a kernel estimation procedure. The kernel estimator for the probability density function f is given by

$$\hat{f}_h(x) = \frac{1}{n}\sum_{j=1}^{n} K_h(x - x_j)$$

where $K_h(u) = \frac{1}{h}K\left(\frac{u}{h}\right)$ with K kernel function and h is the bandwidth parameter.

From this, kernel estimator for the cumulative distribution function F can be constructed:

$$\hat{F}_h(x) = \int_{-\infty}^{x} \hat{f}_h(x)dx = \frac{1}{n}\sum_{j=1}^{n} H\left(\frac{x - x_j}{h}\right)$$

where $H(x) = \int_{-\infty}^{x} K(t)dt$

For this procedure, two choices will be made, the kernel function K and the bandwidth h.

The `kerdiest` package in R provides the 4 kernels. Epanechnikov, Normal, Biweight, and Triweight. For this study, we will only use the Epanechnikov kernel, since the selection of the kernel is of less importance as different functions produce good results (Quintala-del-Rio, Estevez-Perez, 2012).

On the other hand, bandwidth selection is more crucial because the bandwidth influence the estimator: if the bandwidth is small, we will obtain undersmoothed estimator, with high variability. On the contrary, large bandwidth will yield to a very smooth estimator farther from the function we are trying to estimate.

For the bandwidth selection procedure, cross validation by Bowman et.al (1998) will be used as recommended by Quintala-del-Rio (2012). For this method, we will select the bandwidth h that minimizes the function

$$\mathrm{CV}(h) = \frac{1}{n}\sum_{i=1}^{n} \int \left(I(x - x_i) - F_{-i}(x)\right)^2 w(x)\, dx$$

where $I(x - x_i) = \begin{cases} 1 & \text{if } x - x_i \geq 0 \\ 0 & \text{o.w} \end{cases}$ and $F_{-i}(x) = \frac{1}{n}\sum_{j \neq i} H\left(\frac{x - x_j}{h}\right)$

The `CVbw` and `rl` functions of the `kerdiest` package in R will be used to compute the bandwidth and return levels respectively.

3.3. Daily Data under Climate Change Scenarios

Long Ashton Research Station Weather Generator (LARS-WG) is a type of weather generator that simulates time series at a single site under different scenarios and different time periods (Semenov, 2012). LARS-WG version 6 can now generate future daily weather data at different climate change scenarios based on the IPCC 5th Assessment Report. More information about LARS-WG can be found on Appendix B.

LARS-WG assumes that observed climate is stationary. If there are any trends in the observed data, they need to be removed, especially on annual total precipitation. Rainfall annual total amounts show an upward linear trend from years 1973-2021. Outliers from the years 1982-1988 may have affected this trend. A Mann-Kendall test proves that the data is non-stationary with p-value = 0.002941.

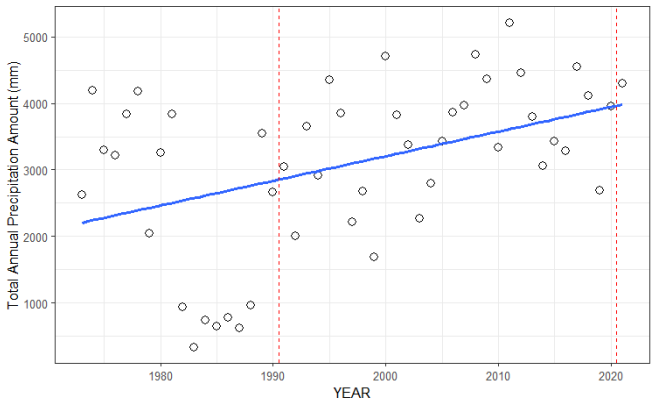


Figure 6. Annual Precipitation Totals from 1973 to 2021 in Legazpi City

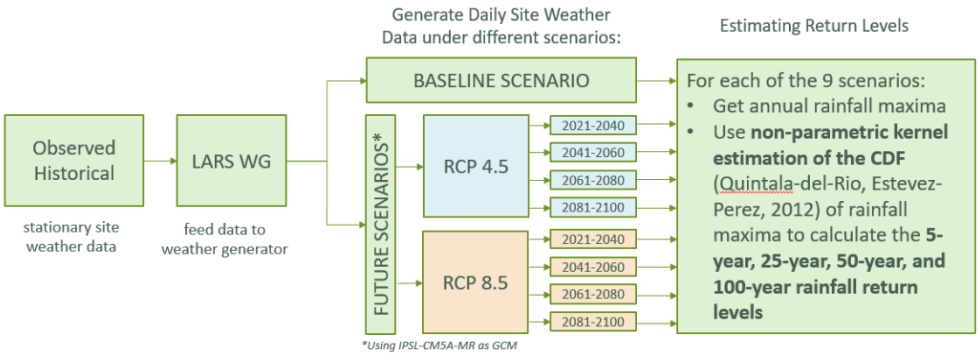
We cut the data, focusing on years 1991 to 2020. The Mann-Kendall test now does not reject the null that the data is stationary, with p-value = 0.080391. The data for the years 1991 to 2020 may serve as the historical data and can now be stored in LARS-WG for site analysis.

Simulation experiments will be conducted for the scenarios RCP 4.5 and RCP 8.5, and for the period 2021-2040, 2041-2060, 2061-2080, and 2081-2100. 100 years' worth of daily data that follow the properties of the historical data will be generated for each scenario and periods.

For this paper, we will use the available IPSL-CM5A-MR GCM from the LARS Weather Generator to simulate daily data at different scenarios. For the rationale of the GCM choice, Ruan et al. (2018) analyzed 34 CMIP5-GCMs for precipitation over the Lower Mekong basin, Southeast Asia. Among their top 5 GCMs, IPSL-CM5A-MR is the only one included in LARS-WG.

By using a weather generator, a reanalysis dataset can be used as quasi-observations to generate climate data in lieu of actual observations. The generated data will be used for the non-parametric estimation of the return levels.

3.4. Methodological Framework



4. RESULTS AND DISCUSSION

4.1. Historical and Generated Baseline Data

Maximum single day rainfall ever recorded within the time frame of the historical data was on Nov 7, 2017, with 437.642 mm of rainfall while the average daily annual maxima is 209.84 mm.

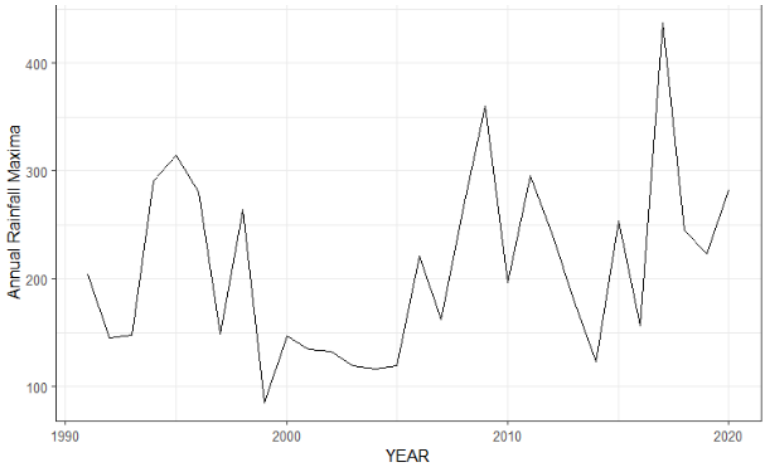


Figure 7. Annual Rainfall Maxima time series from 1991 to 2020

After storing the historical data to the weather generator, LARS-WG outputs a Kolmogorov-Smirnov (KS) test to show that this generated synthetic baseline data follows the same daily rain distribution as the observed historical daily data:

Table 2. KS test for daily rain distributions

Low p-value (<0.05) indicates that generated climate is unlikely to be the same as the observed climate

Month	Effective n	KS Statistic	p-value
January	11.5	0.221	0.572
February	11.5	0.320	0.153
March	11.5	0.206	0.661
April	11.5	0.126	0.988
May	11.5	0.082	1.000
June	11.5	0.054	1.000
July	11.5	0.146	0.952
August	11.5	0.129	0.985
September	11.5	0.158	0.913
October	11.5	0.053	1.000
November	11.5	0.243	0.449
December	11.5	0.147	0.949

Annual Rainfall Maxima were also reproduced well by LARS-WG. A KS test still shows that the annual maxima for the historical and generated 100 years' worth of baseline data follow the same distribution (p-value = 0.5965).

We estimate the density function using the Rosenblatt-Parzen kernel estimator and cross-validation bandwidth selection by Bowman et.al with Normal kernel function.

The obtained value for the bandwidth using the historical data is 37.37 while the bandwidth for the generated baseline data is 20.39. The following graphs are the density curves and distribution functions of the historical and synthetic data:

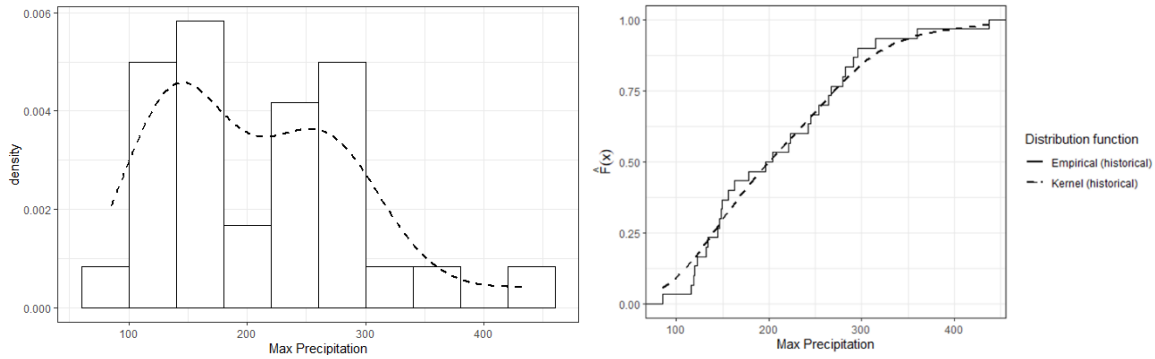


Figure 8. Kernel estimated Density function and Cumulative Distribution function using historical data (1991-2020)

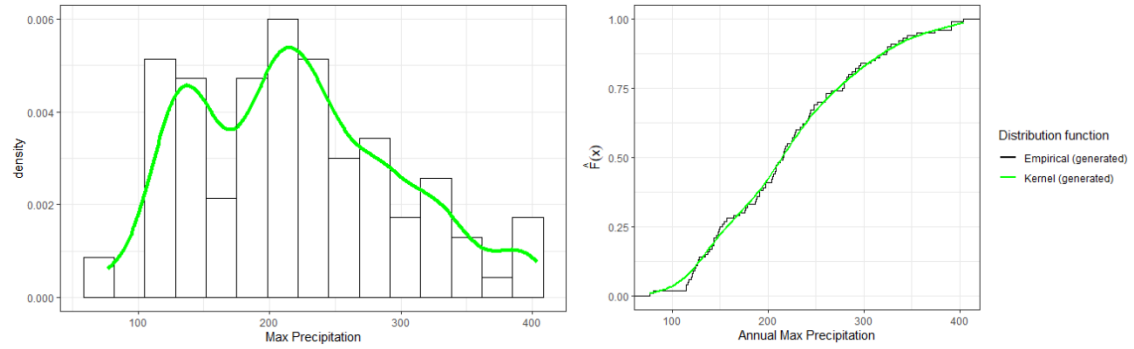


Figure 9. Kernel estimated Density function and Cumulative Distribution function using synthetic baseline data

In analyzing extreme scenarios in terms of return levels, it is interesting to note that the behavior of the tail of the distribution is the most important, specifically the 80th quantile (5-year return period event) and above up to 99th quantile (100-year return period event).

After estimating the distributions, the T-period return levels are now as follows:

Table 3. Estimated return levels using non-parametric approach

Return Period	Historical	Baseline (generated)
5-year	285.62	290.22
25-year	386.64	375.92
50-year	431.07	396.64
100-year	437.64	404.10

For the next parts, future scenarios are generated using the GCM provided by LARS-WG, and the estimation procedure is the same (non-parametric estimation of return levels).

4.2. Future Scenarios

4.2.1. RCP 4.5

Figure 10. Kernel estimated Density function of future annual rainfall maxima under RCP 4.5 scenario

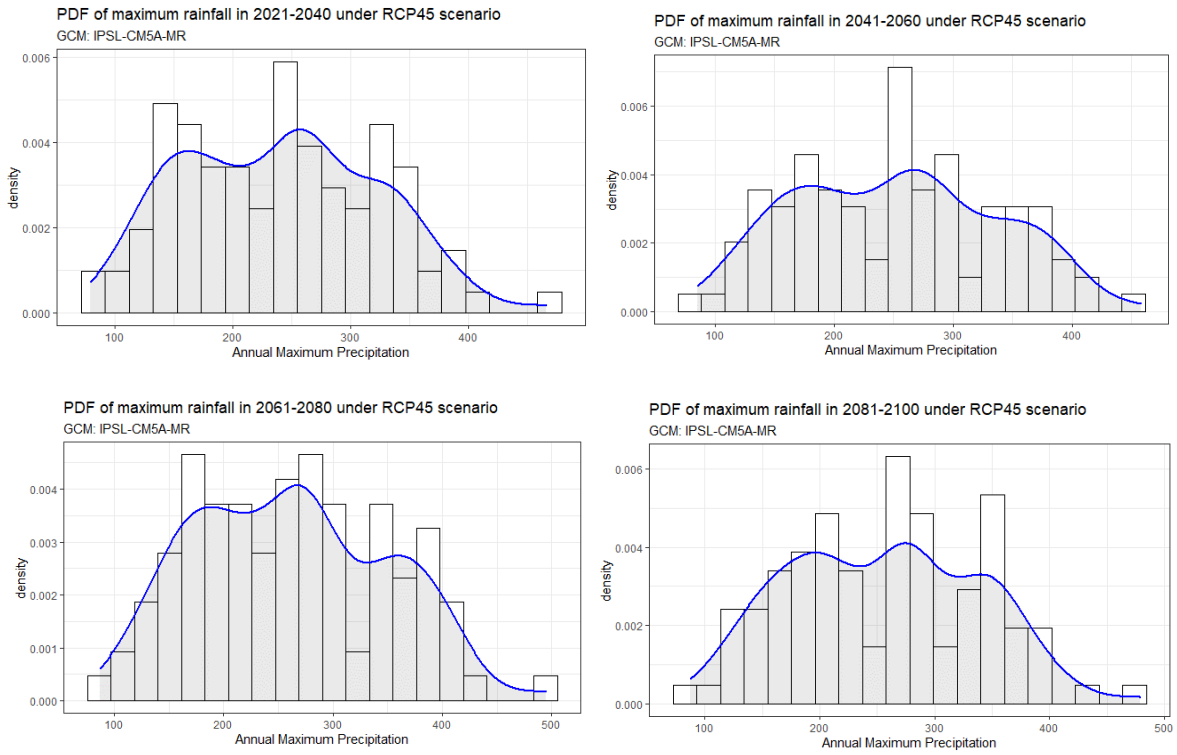
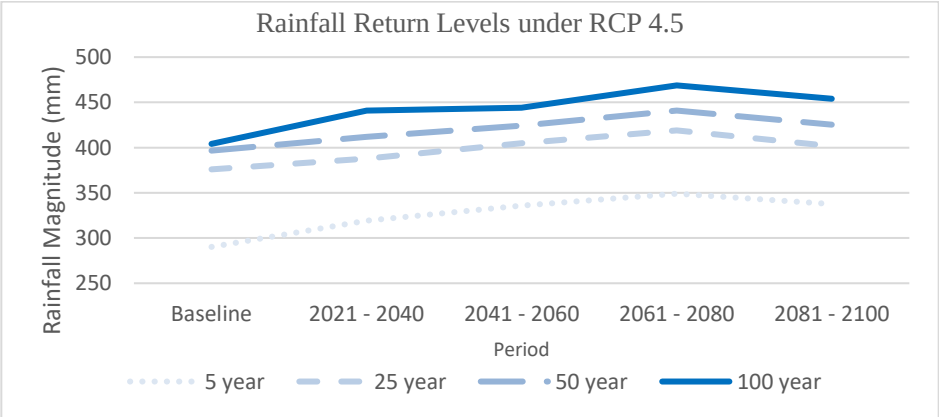


Table 4. Estimated rainfall return levels in Legaspi under RCP 4.5 using non-parametric approach

Return Period	Baseline	2021-2040	2041-2060	2061-2080	2081-2100
5-year	290.22	319.09	335.67	349.05	337.48
25-year	375.92	387.80	404.92	418.99	401.77
50-year	396.64	411.92	424.37	440.92	425.23
100-year	404.10	440.99	443.91	468.77	453.88

Figure 10. Trend of return levels in Legaspi under RCP 4.5



Under RCP 4.5, magnitude of extreme events will peak at 2061-2080 and a projected downward trend by the end of the century.

4.2.2. RCP 8.5

Figure 11. Kernel estimated density functions of future annual rainfall maxima under RCP 8.5 scenario

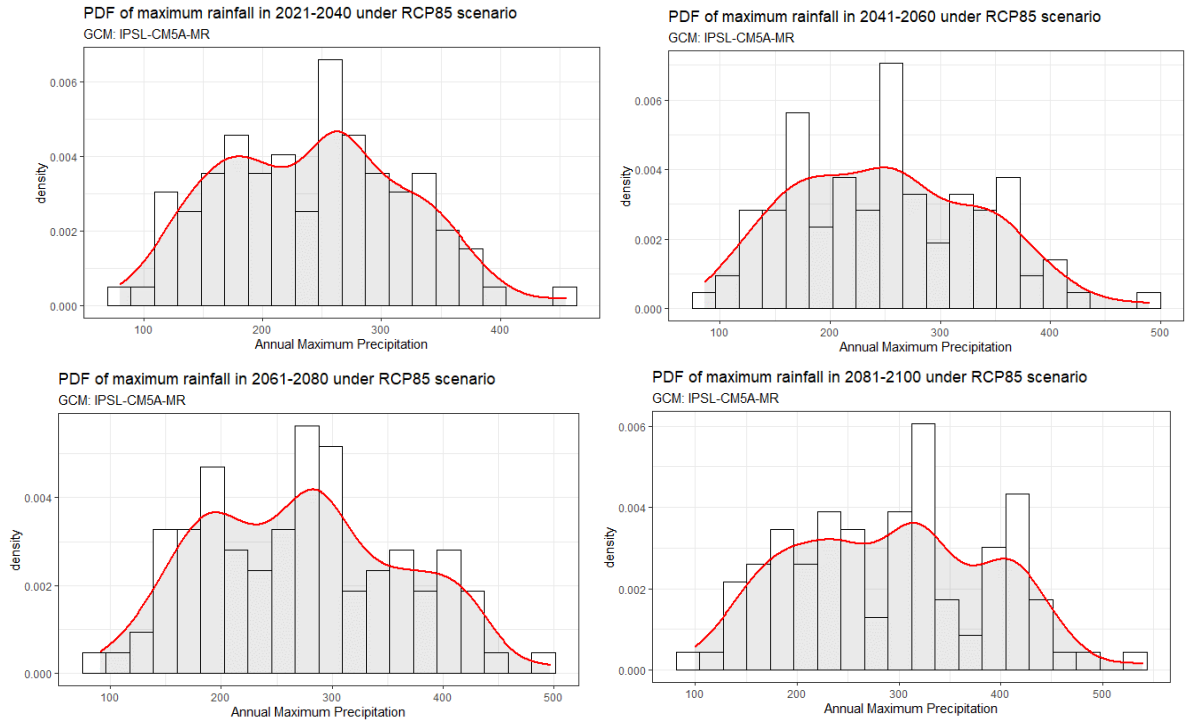
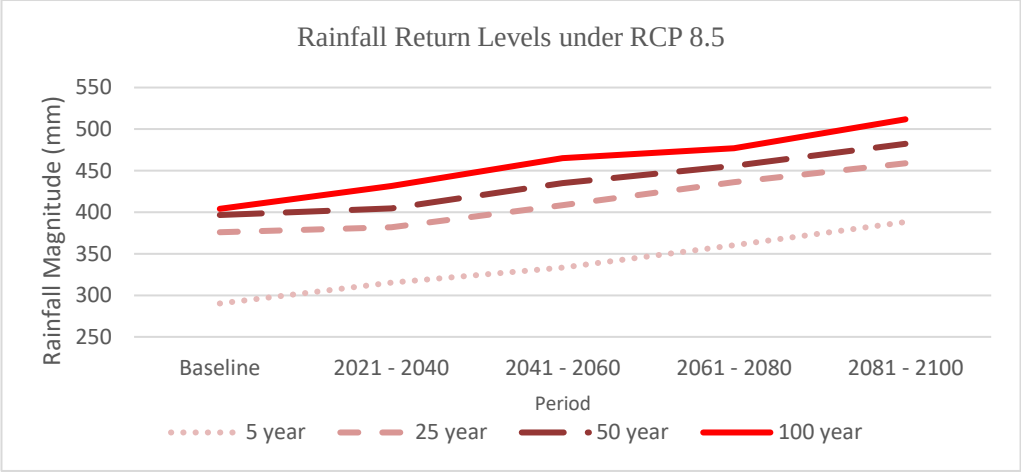


Table 5. Estimated rainfall return levels in Legaspi under RCP 8.5 using non-parametric approach

Return Period	Baseline	2021-2040	2041-2060	2061-2080	2081-2100
5-year	290.22	315.26	333.51	360.35	388.23
25-year	375.92	381.98	408.50	436.14	458.88
50-year	396.64	404.44	435.15	455.68	482.31
100-year	404.10	431.47	464.96	476.83	511.80

Figure 12. Trend of return levels in Legaspi under RCP 8.5



Under RCP 8.5 or the Business-as-usual greenhouse gas emission, extreme events will bring increased magnitudes. For instance, if previously, a 5-year event brings 290 mm of rain, by the end of the century, a 5-year event will bring 388 mm of rainfall.

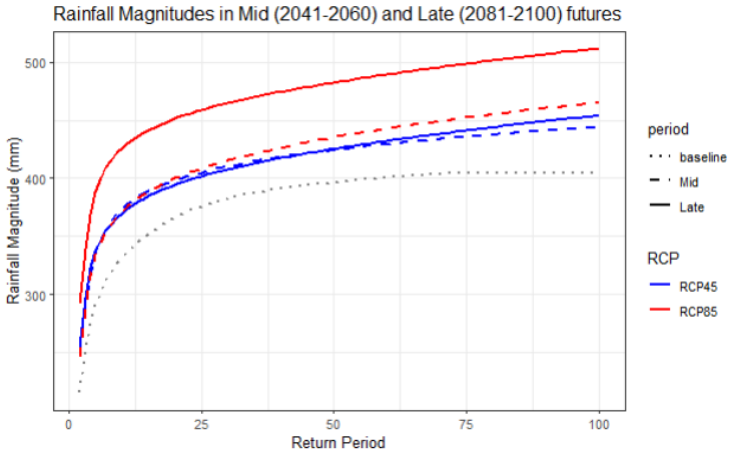
5. SUMMARY: Comparison of RCP 4.5 and RCP 8.5

For brevity, we focus on Mid (2041-2060) and Late (2081-2080) future periods in the 21st century in comparing the effects on rainfall magnitude under the two RCP scenarios. From the table and graph below, it can be observed that that rainfall magnitudes under RCP 8.5 are higher than under RCP 4.5 at different return periods.

Table 5. Estimated rainfall return levels in Legaspi in the Mid and Late futures

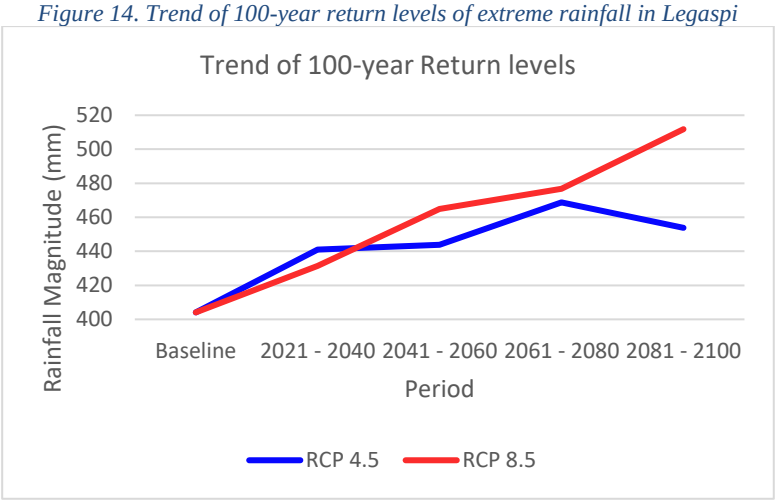
Return period	Baseline	Mid (2041-2060)		Late (2081-2100)	
		RCP 4.5	RCP 8.5	RCP 4.5	RCP 8.5
5-year	290.22	335.67	333.51	337.48	388.23
25-year	375.92	404.92	408.50	401.77	458.88
50-year	396.64	424.37	435.15	425.23	482.31
100-year	404.10	443.91	464.96	453.88	511.80

Figure 13. Return level plots of extreme rainfall in Legaspi



For example, if previously, a 400 mm of rainfall is a at least a 60-year event, by the end of the century, this amount will not be as rare as before. It will be a 24-year event under RCP 4.5 and a 6-year event under RCP 8.5.

Figure 14. Trend of 100-year return levels of extreme rainfall in Legaspi



Extreme events will bring increased rainfall magnitudes in the coming decades, and high rainfall amounts will be more frequent in terms of annual occurrence. While an increase in magnitude of rainfall extremes can also be observed even under the controlled GHG emission scenario (RCP 4.5), difference in effects is still significant compared to the Business-as-usual scenario (RCP 8.5), especially on the late century period.

6. CONCLUSION AND RECOMMENDATIONS

Magnitudes of extreme rainfalls in Legaspi

By the end of the century, it is expected that extreme rainfalls will bring higher magnitude in Legaspi. The estimated return levels in the coming decades will be higher under the Business-as-usual scenario (RCP 8.5) than the mitigated scenario (RCP 4.5). This also means that previous rare events or longer return period (low probability of exceedance) such as a 400 mm rainfall will be more frequent (higher probability of exceedance) if global emission of greenhouse gases are not controlled at a certain level. If mitigation of GHG emissions cannot be done, the resiliency in the city can be improved by preparing for single day high rainfall amounts in terms of planning water flow to avoid flash floods.

The values presented can be used by disaster scientists in flood modelling and mapping for risk assessment.

Assessment of Non-parametric Estimation of Return Levels

In studying extreme annual daily maxima of rainfall amounts, using non-parametric estimation of the distributions have a benefit. For one, the event of daily annual rainfall maximum at higher levels (around 400 mm for the historical data) are more frequent than the events of recording annual maxima between 350-390, but peaks on the frequency can also be observed at around 200 mm-250 mm rainfall values. This can cause multi-modality, as seen in the histograms of rainfalls under the historical, baseline, and different climate scenarios, which the known probability distributions fail to address. Considering multimodality in the center part of the probability distribution is important if we want to address shorter return period events. There is great utility in modelling lower scenarios such as 5, 10, and 25-year rainfall return periods as these are the type of events that people tend to experience more often.

However, non-parametric estimation of extremes also has limitations. Method discussed on this paper will fail to extrapolate more extreme events especially when we want to estimate 100 year, 200-year, 500-year and above return levels given that data is short. Return levels are bounded by the range of the data, unlike if we use a known probability distribution such as Generalized Extreme Value, Pearson type III, Exponential distribution, etc., the asymptotic tails of these distributions fix the issue of extrapolating rarer events. Estimating higher return periods is relevant in planning major infrastructures such as dams that are expected to endure rainfalls for at least 100 years.

For future studies, it is recommended to further assess non-parametric method of estimating return levels by analyzing variance and bias and formally comparing against the parametric methods.

7. REFERENCES

7.1. Publications

Alipour, S.M, Leal, J. (2019). Return levels uncertainty under effect of climate change

Chow, V.T., Maidment, D.R., and Mays, L.W., (1988). Applied hydrology. 1st ed. New York: McGraw-Hill. ISBN-13: 978-0070108103

Coles, Stuart. (2001). An introduction to statistical modeling of extreme values. *Springer series in statistics*

Cooley, D. (2013). Return Periods and Return Levels Under Climate Change

Faucher D, Rasmussen PF, Bobée B. (2001). A distribution function based bandwidth selection method for kernel quantile estimation. Journal of Hydrology 250: 1–11.

Fisher R.A. (1924) The influence of the rainfall on the yield of wheat at Rothamsted. Philosophical transaction of the Royal Society of London; Series B, Vol. 213.

Gilleland, E. (2020). Bootstrap Methods for Statistical Inference. Part II: Extreme-Value Analysis

Gilleland, E. and R. W. Katz, (2016): extRemes 2.0: An extreme value analysis package in R. J. Stat. Software, 72, 1–39, https://doi.org/10.18637/jss.v072.i08.

Holešovský, J., Fusek, M., Blachut, V., and Michálek, J. (2016). Comparison of precipitation extremes estimation using parametric and nonparametric methods. Hydrological Sciences Journal, 61:13, 2376-2386, DOI: 10.1080/02626667.2015.1111517

Khaliq, M. N., T. B. M. J. Ouarda, J.-C. Ondo, P. Gachon, and B. Bobée, (2006): Frequency analysis of a sequence of dependent and/or non-stationary hydro-meteorological observations: A review. J. Hydrol., 329, 534–552, doi:10.1016/j.jhydrol.2006.03.004.

Kim K, Heo J. (2002). Comparative study of flood quantiles estimation by nonparametric models. Journal of Hydrology 260: 176– 193.

Kim, H., Kim, T., Shin, J., Heo, J. (2022). "Improvement of Extreme Value Modeling for Extreme Rainfall Using Large-Scale Climate Modes and Considering Model Uncertainty" Water 14, no. 3: 478. https://doi.org/10.3390/w14030478

Lee, H., Kang, K. (2015). Interpolation of Missing Precipitation Data Using Kernel Estimations for Hydrologic Modelling. Hindawi Publishing Corporation. Advances in Meteorology. Volume 2015, Article ID 935868. http://dx.doi.org/10.1155/2015/935868

Mearns, L. O., R. W. Katz, and S. H. Schneider, (1984): Extreme high-temperature events: Changes in their probabilities with changes in mean temperature. J. Climate Appl. Meteor., 23, 1601–1613, doi:10.1175/1520-0450(1984)023,1601:EHTECI.2.0.CO;2.

Olsen, J. R., Lambert, J. H., & Haimes, Y. Y. J. R. A. (1998). Risk of extreme events under nonstationary conditions. 18(4), 497-510.

Parzen E. (1962). On estimation of a probability density function and mode. Annals of Mathematical Statistics 32: 1065– 1076.

Quintela-del-Rio, A. (2011). On bandwidth selection for nonparametric estimation in flood frequency analysis

Reiss RD. (1981). Nonparametric estimation of smooth distribution functions. Scandinavian Journal of Statistics 8: 116–119.

Ruan, Y., Yao, Z., Wang, R. & Liu, Z. (2018) Ranking of CMIP5 GCM skills in simulating observed precipitation over the Lower Mekong Basin, using an improved score-based method. Water 10, 1868.

Saeb, A. 2014. General Extreme Value Modeling and Application of Bootstrap on Rainfall Data - A Case Study

Sharma, M.A., Singh, J.B. 2010. Use of Probability Distribution in Rainfall Analysis. New York Science Journal,

Schnarr, E. and Trzaka, S. (2014) A Review of Downscaling Methods for Climate Change Projections. Tetra Tech ARD.

Zhang, A., Shi, H., Li, T., Fu, X. 2017, A bootstrap method to estimate the influence of rainfall spatial uncertainty in hydrological simulations. Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2017-273

7.2. Reports

IPCC (2014): Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, R.K. Pachauri and L.A. Meyer (eds.)]. IPCC, Geneva, Switzerland, 151 pp.

DOST-PAGASA, Manila Observatory, and Ateneo de Manila University (2021). Philippine Climate Extremes Report 2020: Observed and Projected Climate Extremes in the Philippines to Support Informed Decisions on Climate Change Adaptation and Risk Management. Philippine Atmospheric, Geophysical and Astronomical Services Administration, Quezon City, Philippines.

DSWD DROMIC Report #17 on Typhoon "QUINTA". Accessed through https://reliefweb.int/report/philippines/dswd-dromic-report-17-typhoon-quinta-09-december-2020-6pm

7.3. Data Sources

- DOST PAGASA. Climate Extremes and Climate Normals tables.
  Accessed through www.pagasa.dost.gov.ph/climate/climate-data
- NOAA. Daily Rainfall and Temperatures in Legaspi Station.
  Accessed through ftp://ftp.ncdc.noaa.gov/pub/data/gsod/
- SOLCAST API Toolkit. Historical record of solar radiation data. SOLCAST API Toolkit.
  Accessed through https://toolkit.solcast.com.au/historical/time-series/order

7.4. Softwares

- Semenov, M.A. (2020) Long Ashton Research Station Weather Generator 6.0 (LARS-WG 6). Rothamsted Research, UK. URL https://sites.google.com/view/lars-wg/
- R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

APPENDIX A: Important Concepts

This appendix provides definitions of some concepts introduced in this paper. These provides context about the processes in analyzing climate change. Definitions are extracted from *A Review of Downscaling Methods For Climate Change Projections* by Tzsaka and Schnarr (2014)

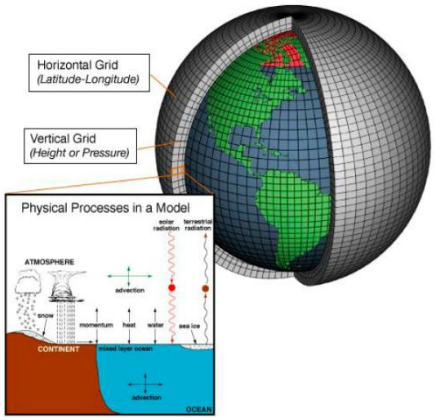
1. GCM

General or global circulation models (GCMs) are the simplified representation of the earth's climate system simulating the physical processes of atmosphere and ocean.

A GCM is composed of many grid cells with estimated climates via mathematical equations that describe atmospheric, oceanic, and biotic processes, interactions, and feedbacks. Each modeled grid cell is homogenous (i.e., within the cell there is one value for a given variable).

GCMs can also provide quantitative estimates of future climate change that are valid at the global and continental scale and over long periods by changing important parameters (such as greenhouse gas emissions and solar radiation).

GCM Concept



Source: National Oceanic and Atmospheric Administration (NOAA), 2012

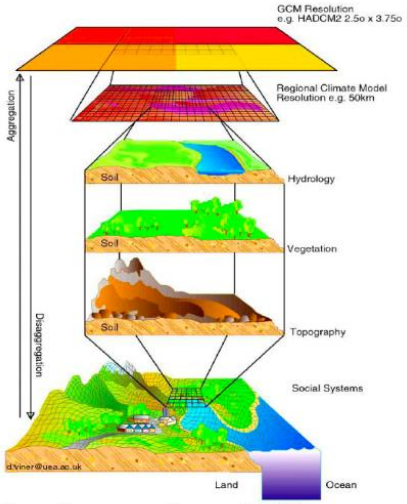
2. Downscaling

Although GCMs are valuable predictive tools, most GCMs have spatial resolution of approximately 200 km (~2° × 2°) which makes it hard to consider fine-scale heterogeneity due to local climate variability and complexity of topography.

To overcome such a limitation, downscaling techniques can be performed to translate a coarse horizontal resolution to a finer resolution while considering regional and local climate variability.

Two main types of downscaling are Dynamical and Statistical Downscaling.

Downscaling Concept



Many of the processes that control local climate, e.g., topography, vegetation, and hydrology, are not included in coarse-resolution GCMs. The development of statistical relationships between the local and large scales may include some of these processes implicitly.
Source: Viner, 2012

a. Dynamical Downscaling

Relies on the use of a Regional Climate Model (RCM), similar to a GCM in principles but with higher resolution.

RCMs take large-scale information supplied by GCM output, then incorporates more complex topography, the land-sea contrast, surface heterogeneities, and detailed descriptions of physical processes.

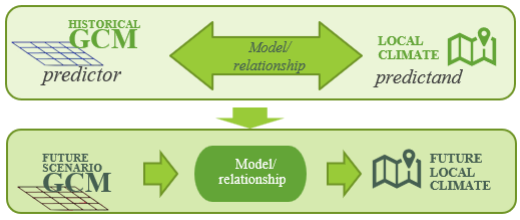
b. Statistical Downscaling

Relies on establishing relationships between large-scale atmospheric variable and local observed climate variables, such as those observed at weather stations.

Once a relationship has been determined and validated, future atmospheric variables that GCMs project are used to predict future local climate variables.

The use of weather generators (such as LARS-WG which is used in this study) is a type of statistical downscaling.

Conceptual framework of Statistical Downscaling

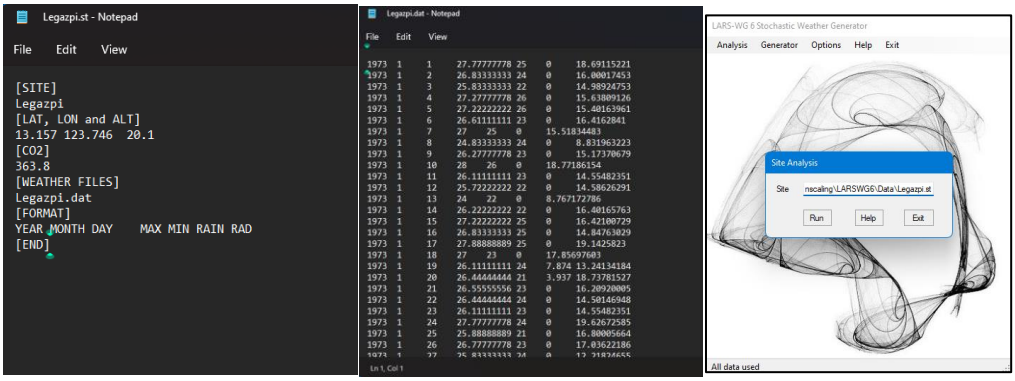


APPENDIX B: LARS-WG

Long Ashton Research Station Weather Generator (LARS-WG) is a stochastic weather generator and a computationally inexpensive downscaling tool to generate local scale climate scenarios based on global or regional climate models for impact assessments of climate change (Semenov, 2020). LARS-WG has been used in more than 75 countries for research and education.

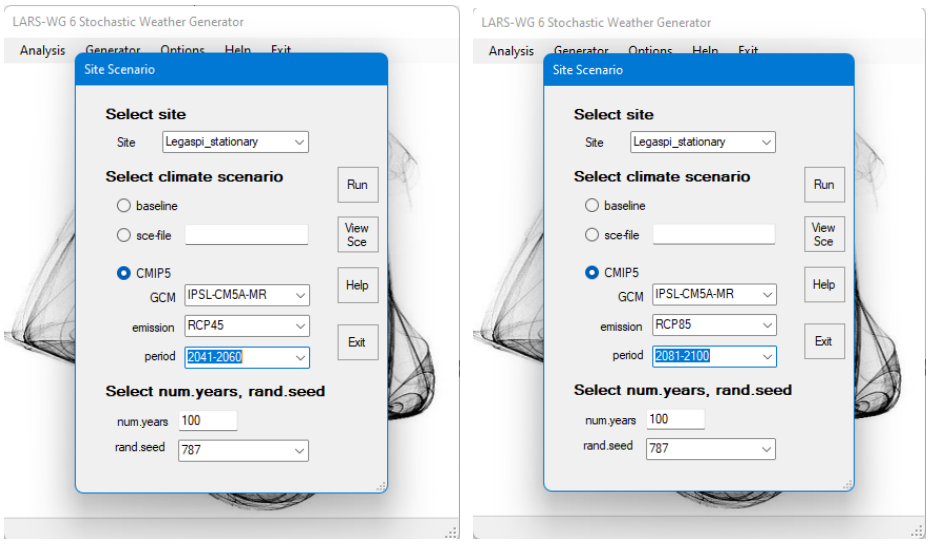
It utilizes semi-empirical distributions for the lengths of wet and dry day series, daily precipitation, and daily solar radiation. Synthetic weather data is then generated by combining these statistical characteristics with a scenario file that contains information about changes in the climate variables. The semi-empirical distribution $Emp = \{a_0, a_i;\ h_i, i = 1, \ldots, 10\}$ is a histogram with ten intervals, $[a_i - 1, a_i)$,where $a_i - 1 < a_i$, and h_i denotes the number of events from the observed data in the i^{th} interval. Random values from the semi-empirical distributions are chosen by first selecting one of the intervals (using the proportion of events in each interval as the selection probability), and then selecting a value within that interval from the uniform distribution.

The current version 6.0 of LARS-WG incorporates climate projections from the CMIP5 ensemble used in the IPCC 5th Assessment Report. LARS-WG has been well validated in diverse climates around the world.



Data to be stored to Long Ashton Research Station Weather Generator (LARS-WG)

(1)..st file: structure of the data; (2)*.dat file: data to be used; and (3) sample interface of LARS-WG*



Sample interface of LARS-WG when generating 100 years' worth of daily weather using IPSL-CM5A-MR

(1) Under RCP 4.5 centered on 2050; (2) Under RCP 8.5 centered on 2090

APPENDIX C: R Codes

The following are the important R codes used in this paper. Only samples are presented. Full codes, functions, and data will be accessible in the author's GitHub page (https://github.com/slcodia) in the future.

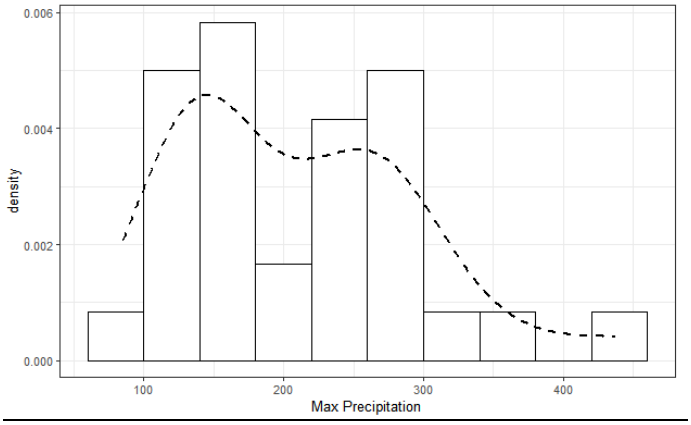
Getting bandwidth using cross-validation bandwidth by Bowman, et.al (1998)

```
bw.CV <- kerdiest::CVbw(vec_data = legaspi_max$MAX.PRCP)
bw.CV$bw
```

```
[1] 37.37771
```

Plotting Histogram and PDF

```
hist.pdf <- ggplot(legaspi_max, aes(x = MAX.PRCP))+
    geom_histogram(aes(y = ..density..),
                   binwidth = 40,fill="white", colour = "black")+
    stat_density(bw = bw.CV$bw, kernel = "gaussian",
                 geom="line",linetype = "dashed",
                 size = 1)+
    theme_bw()+xlab("Max Precipitation")
hist.pdf
```



5-year, 25-year, 50-year, and 100-year Return Levels using Non-parametric Kernel Estimation

```
kerdiest::rl(type_kernel = "n",
             vec_data = legaspi_max$MAX.PRCP,
             bw = bw.CV$bw,
             T = c(5,25,50,100))
```

```
[1] 285.6196 386.6361 431.0655 437.6393
```

Return Level plot

```
# Preparing Data
rl.kerdiest <- data.frame(x = c(2:100),
               baseline_baseline = rl(type_kernel = "n",
                      vec_data = base.max$MAX.PRCP,
                      T  = c(2:100),
                      bw = kerdiest::CVbw(vec_data = base.max$MAX.PRCP)$bw),
               RCP45_Mid = rl(type_kernel = "n",
                      vec_data = RCP45_mid$MAX.PRCP,
                      T  = c(2:100),
                      bw = kerdiest::CVbw(vec_data =RCP45_mid$MAX.PRCP)$bw),
               RCP45_Late = rl(type_kernel = "n",
                      vec_data = RCP45_late$MAX.PRCP,
                      T  = c(2:100),
                      bw = kerdiest::CVbw(vec_data = RCP45_late$MAX.PRCP)$bw),
               RCP85_Mid = rl(type_kernel = "n",
                      vec_data = RCP85_mid$MAX.PRCP,
                      T  = c(2:100),
                      bw = kerdiest::CVbw(vec_data = RCP85_mid$MAX.PRCP)$bw),
               RCP85_Late = rl(type_kernel = "n",
                      vec_data = RCP85_late$MAX.PRCP,
                      T  = c(2:100),
                      bw = kerdiest::CVbw(vec_data = RCP85_late$MAX.PRCP)$bw)
                      )
rl.kerd <- reshape2::melt(rl.kerdiest, id = 'x', value.name = "return
level")%>%splitstackshape::cSplit(splitCols = 'variable', sep="_",
type.convert=FALSE)
colnames(rl.kerd) <- c('x', 'return level', 'RCP', 'period')

# Actual Plot
ggplot(rl.kerd, aes(x=x, y = `return level`, color = RCP, linetype =
period))+
    geom_line(size = 1)+
    scale_color_manual(values = c("baseline" = "green",
                                  "RCP45" = "blue",
                                  "RCP85" = "red"))+
    scale_linetype_manual(values = c("baseline" = "dotted",
                                     "Mid" = "dashed",
                                     "Late" = "solid"))+
    theme_bw()+
    xlab("Return Period")+
    ylab("Rainfall Magnitude (mm)")+
    ggtitle("Rainfall Magnitudes in Mid (2041-2060) and Late (2081-2100)
futures")
```