Estimating Rainfall Erosivity Using Nonlinear Regression Models and the Relationship with El Niño Southern Oscillation in South America

Xiaoxuan Li [1]

[1] *Dept. of Geography and Geoinformation Science, Fairfax, VA 22030-4444*

Abstract

As a key parameter to estimate the erosivity power of rainfall, the rainfall erosivity factor plays an important role in land-use management and sediment control studies. Besides, the variability of rainfall erosivity is also associated with the El Niño-Southern Oscillation (ENSO). In this study, multiple regression models were built to estimate the rainfall erosivity in South America during the period from 1980 to 2017. Then, the relationship between calculated rainfall erosivity and the ENSO index was examined using rainfall erosivity and three ENSO indicators: Multivariate ENSO Index (MEI), Southern Oscillation Index (SOI), and Sea Surface Temperature (SST). The results show that the presented models have the potential to estimate rainfall erosivity accurately, with relatively high correlation coefficients (CC>0.5) between the observed and estimated rainfall erosivity. Furthermore, the results indicate that the relationship between rainfall erosivity and ENSO indicators varies spatially and temporally, especially during El Niño and La Niña years. Compared to normal years, rainfall erosivity is higher during El Niño years and lower during La Niña years across most of the study sites.

Introduction

From sediment degradation to potential agriculture land-use loss, soil erosion issues concern governors, farmers, and researchers all over the world. People seek to predict this important soil factor such that conservationists and decision-makers can adapt and mitigate erosive soil trends under the threat from climate change. Soil erosion is caused by either passive or active factors in the environment, including some human activities like irrigation, deforestation, and other forms of land-use actions that passively increase the risk of soil erosion (Mello et al., 2011). Particularly, divergent rainfall distribution, caused by local and global climatic conditions, is the major active factor that affects soil erosivity. This kind of rainfall force is referred to rainfall erosivity (R), the potential of rainfall to cause soil erosion by raindrop impact and surface wash out when infiltration capacity is exceeded. Rainfall erosivity was initially introduced as the capability of rainfall to cause soil loss from hillslopes by water (Nearing et al., 2017) in the Universal Soil Loss Equation (USLE). The function of USLE can be explained as follows:

$$A = RKLSCP \quad (1)$$

where A is soil loss factor, R is annual rainfall erosivity factor, K is soil erodibility factor, L is slope length factor, S is the slope steepness factor, C is the cover and management factor, and P is the supporting practices factor.

The USLE erosivity equation has been widely used for soil erosivity and conservation planning purposes for decades. Among these parameters in USLE, R is one of the most important factors because it shows a significant ability to disaggregate soil (D'Odorico et al., 2001). It can be calculated as the product of 30-minute rainfall intensity (I) and rainfall energy (E) (Lee et al., 2015). Wischmeier and Smith (1978) also defined R as the annual summation of EI for storms that

produce more than 12.7mm of rainfall. In this paper, the rainfall erosivity factor at different spatial and temporal scales was calculated.

As the climate conditions discussed previously, due to its extremely high sensitivity to precipitation events, rainfall erosivity is commonly used to understand local and global climate change by examining its relationship with ENSO indicators. Any positive or negative correlation between the two factors may have great soil loss implications across different regions with abnormal rainfall erosivity values during El Niño and La Niña years. Recorded by National Oceanic and Atmospheric Administration (NOAA) (Psd [APA], n.d.), the SST, MEI, and SOI are frequently used as strong indicators of ENSO indicators (Lee et al., 2014).

This paper seeks to broaden the spatial-temporal view of the relationship between ENSO events and rainfall erosivity by analyzing the entire continent of South America at different scales. Divided by different climate zones, with their linear, and nonlinear regression models, broad conclusions can be made to explain how ENSO events impact rainfall erosivity across each South America subregion.

Data

Study Area

The study area consists of four subregions based on different climate zones across the South America continent, which have diverse climates and topographic features. Based on Koppen Climate Classifications (Köppen et al., 1954), these climate zones can be classified as (A) tropical climate, (B) moist climate, (C) dry climate, and (D) moist mid-latitude climates with mild winters (Figure 1). These subregions were downloaded from Koppen-Geiger.vu in shapefile format. 21

climate polygons were collected and classified into four major categories for data processing and regression modeling. This study aims to estimate rainfall erosivity based on these climate zones to examine the potential variability of rainfall erosivity affected by the global climate phenomenon.

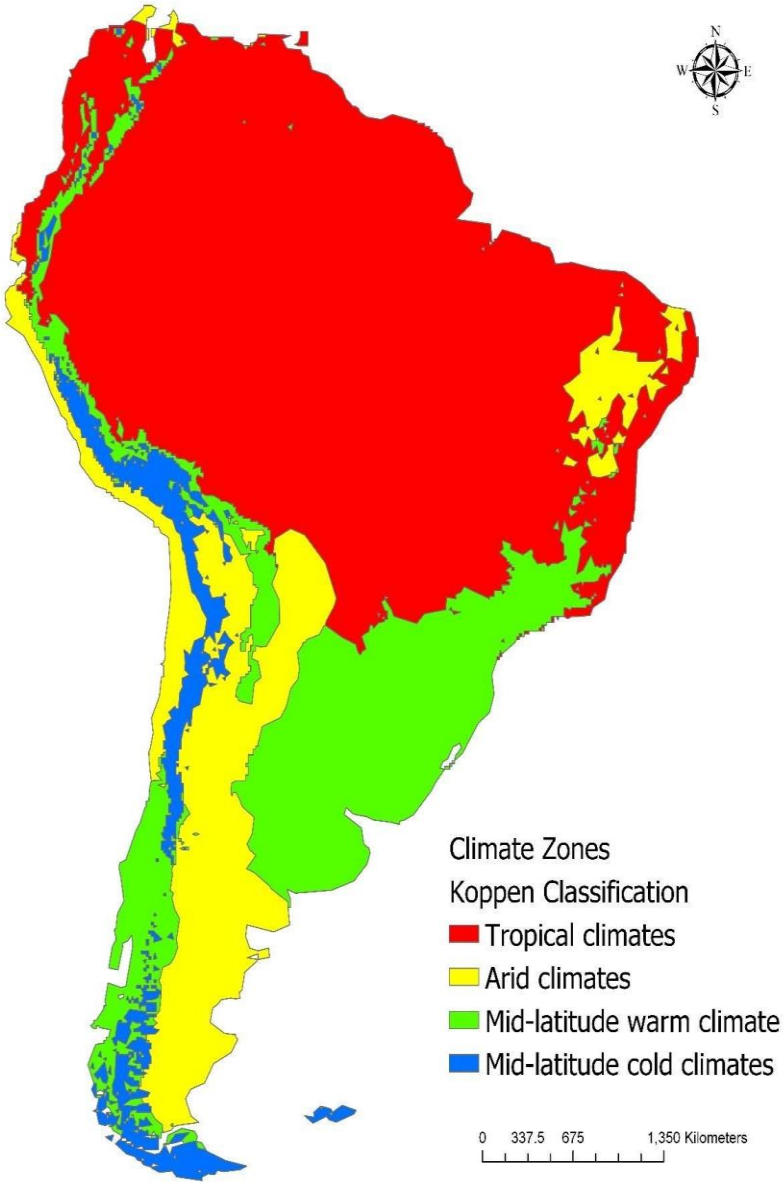

Figure 1. Koppen Climate Zones in Study Sites. In the analytic method section, A, B, C, and D regions represent tropical climates, arid climates, mid-latitude warm climates, and mid-latitude cold climates, respectively.

MERRA-2 precipitation data collection and pre-processing

In this study, MERRA-2 hourly precipitation data, distributed by EARTHDATA, was collected as primary data from 1980 to 2018, with a high temporary resolution of 60 minutes, and a spatial resolution of 0.5 degrees by 0.625 degrees (Table 1). The dataset was well-formatted in NetCDF format, which can be downloaded by calling EARTHDATA API (Earthdata, 2019). In this study, some Python scripts were written to deal with the bulk download process (see support documents for more details). Each NetCDF original file was then converted into raster files in R (version 3.3.4), which contain 24 bands representing 24 hours in each day.

ENSO indicators

ENSO indicators represent the backbone of this work's capability to analyze the relationship between rainfall erosivity and the ENSO phenomenon. Three ENSO indicators were collected in this study: SST, SOI, and MEI. Due to the complex format of the ENSO index, a couple of Python programs were developed to clean and transform these unformatted index files to CSV tables (see support documents). Specifically, the ENSO index was converted to a time-series matrix that can be used to compare with calculated rainfall erosivity. These ENSO indicators are essential to represent the dynamic trend of global climate phenomena over the past climate period such that we may have a better understanding of soil erosivity power concerning climate change.

Table 1. Primary datasets and secondary datasets.

Index	Data	Resolution	Type	Source
1	MERRA-2 precipitation data	Spatial: 0.5 x 0.625, Temporal: hourly	NetCDF	EARTHDATA
2	Köppen climate classifications		Shapefile	Koppen-Geiger.vu
3	MEI	Monthly	array	NOAA
4	SOI	Monthly	array	NOAA
5	SST	Monthly	array	NOAA

Methods

Rainfall erosivity calculation

To calculate monthly rainfall erosivity, the precipitation raster files were aggregated and calculated in R to calculate monthly rainfall erosivity. Specifically, rainfall erosivity was calculated using the precipitation values extracted from MERRA-2 precipitation data. There are two critical and high-correlated factors considered in rainfall erosivity calculation: rainfall amount and rainfall intensity. The rainfall erosivity calculation approach has been presented by Wischmeier and Smith (1978) and Renard et al. (1997) as:

$$R = \frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{m_j}(EI_{60})_k \quad (2)$$

where R is the mean annual rainfall erosivity (MJ mm ha^{-1} h^{-1} yr^{-1}), n is the number of years of data, m_j is the number of erosive events in the j year and EI_{60} is the rainfall erosivity index of a storm k. The event's rainfall erosivity index EI_{60} is defined as:

$$EI_{60} = I_{60}(\sum_{r=1}^{m} e_r v_r) \quad (3)$$

Where e_r is the unit rainfall energy (MJ ha^{-1}) and v_r is the rainfall depth (mm) during a period r. *I60* is the maximum rainfall intensity during a 60 min period of the rainfall event (mm h^{-1}).

$$e_r = 0.29(1 - 0.072e^{-0.05i_r}) \quad (4)$$

where i_r is the rainfall intensity during the period (mm h^{-1}).

This calculated rainfall erosivity, along with original precipitation file, will be converted to different formats (points, raster, tables) and processed in Excel (version 2019), GraphPad Prism (version 8.3.0), ArcGIS Pro, R, and Python (version 3.6.2) for further regression modeling. Before rainfall erosivity and precipitation data were implemented into regression models, some processes

were conducted in these tools to clean the data and format the data into software recognized format. About 70% of the total precipitation data was then randomly selected in ArcGIS Pro and R to run regression models and estimate rainfall erosivity. The rest of the rainfall data was used to validate these estimated rainfall erosivity data. More details about data collection, data processing, and regression analysis procedures can be found in Figure 2.

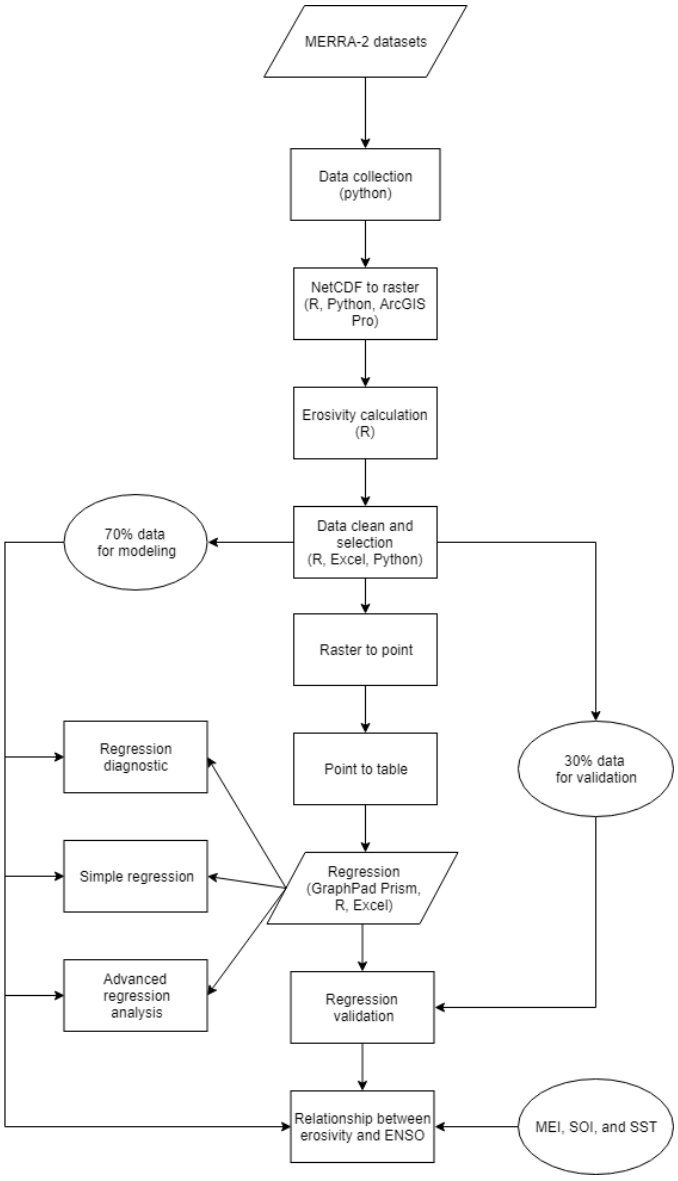

Figure 2. Workflow for estimating rainfall erosivity and examining the relationship between rainfall erosivity and ENSO indicators.

Regression

A variety of R packages were used for developing the best-fit model and displaying resulting data and matrix. These tools were included but not limited to: “gvlma”, “psych”, “glmulti”, “MASS”, “rsq”, “ggplots2”, “ggmisc”, “car”, “lmtest”, “plotrix”, and “BBmisc”. Especially, “glmulti” was used to automatically find all possible regression models with specific responses and explanatory variables. Besides, GraphPad Prism is another major regression model generator in this study. Multiple nonlinear regression models including polynomial regression model, power function model, complex power function model, were developed in GraphPad Prism. It is noted that the polynomial function applied performs as follows:

$$Y_i = \beta_0 + X_i + \beta_1 {X_i}^2 + \beta_2 {X_i}^3 + \in_i \quad (5)$$

Where Y_i devotes rainfall erosivity, X_i represents rainfall observations, β_i equals the corresponding constant for each X_i and $\in_i$ is error term.

For the power function, there are two formats used in this study to build a nonlinear regression model, which are showing in equation 5 and 6:

$$Y_i = a{X_i}^b + \in_i \quad (6)$$

$$Y_i = a{X_i}^b + c{X_i}^d \in_i \quad (7)$$

Where Y_i devotes rainfall erosivity, X_i represents rainfall observations, $\in_i$ is error term and a,b,c, and d are all parameters.

These models generated from R and GraphPad Prism were evaluated and selected based on their R^2 and AIC values. Detailed procedures can be found in Figure 3. Those models who have bigger R^2 and smaller AIC were selected as the best-fit model in each region, which was then used to derive erosivity from each of the climate zones. To further evaluate these models from different

perspectives, regression diagnostic assumptions, Durbin Watson test, Pearson's correlation, and Taylor diagrams were included in this study to test the goodness of each model. The resulting plots can be found in the following sections.

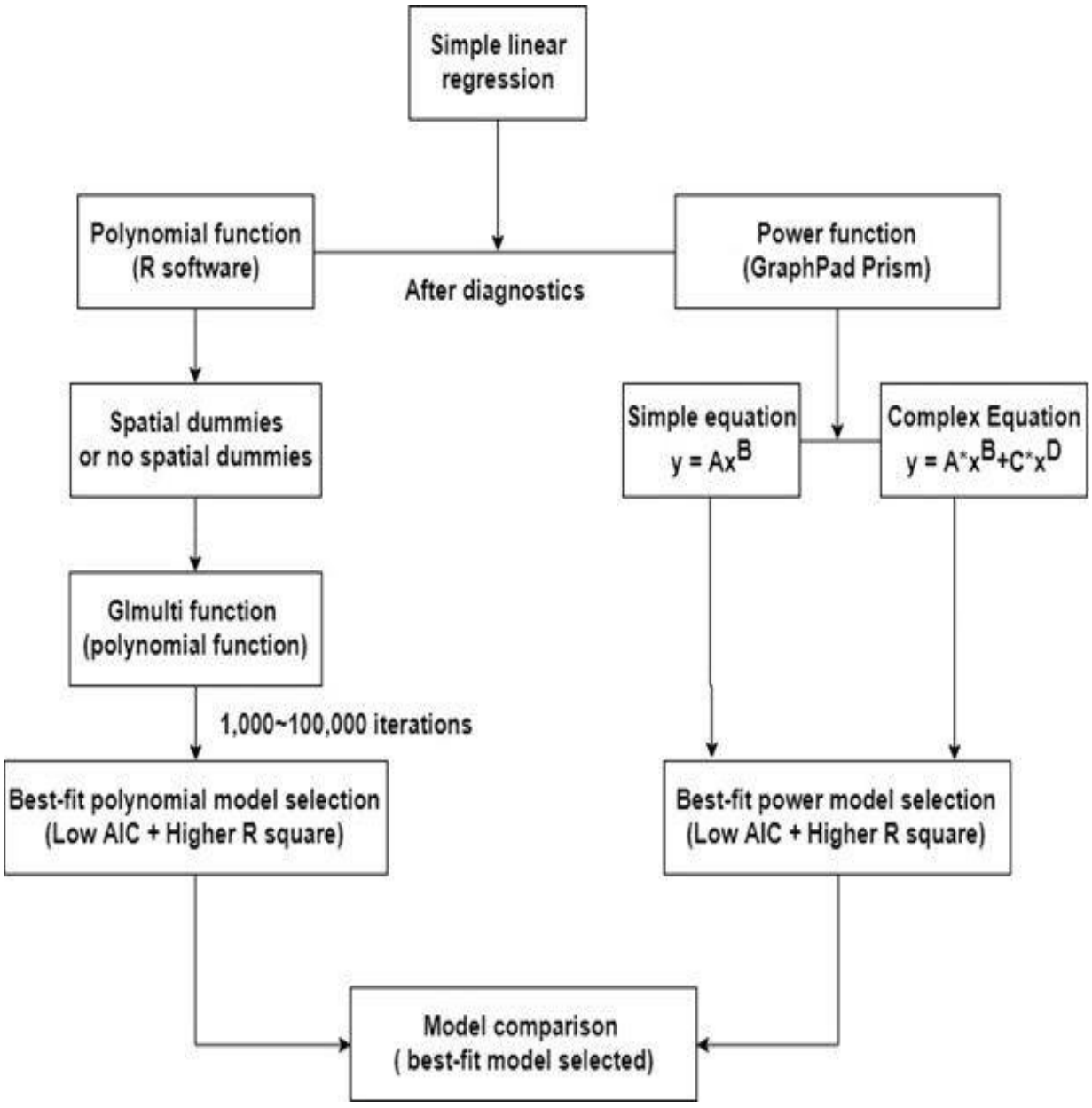

Figure 3. Advanced regression analysis procedure.

Result

Original data interpretation

The original rainfall dataset and rainfall erosivity derived from rainfall observations were displayed and checked if there is any linear relationship between these two variables. The results showing in Figure 4 demonstrate that rainfall and rainfall erosivity variables have moderate relationships, with correlation coefficients between 0.6 and 0.71 in each climate zones.

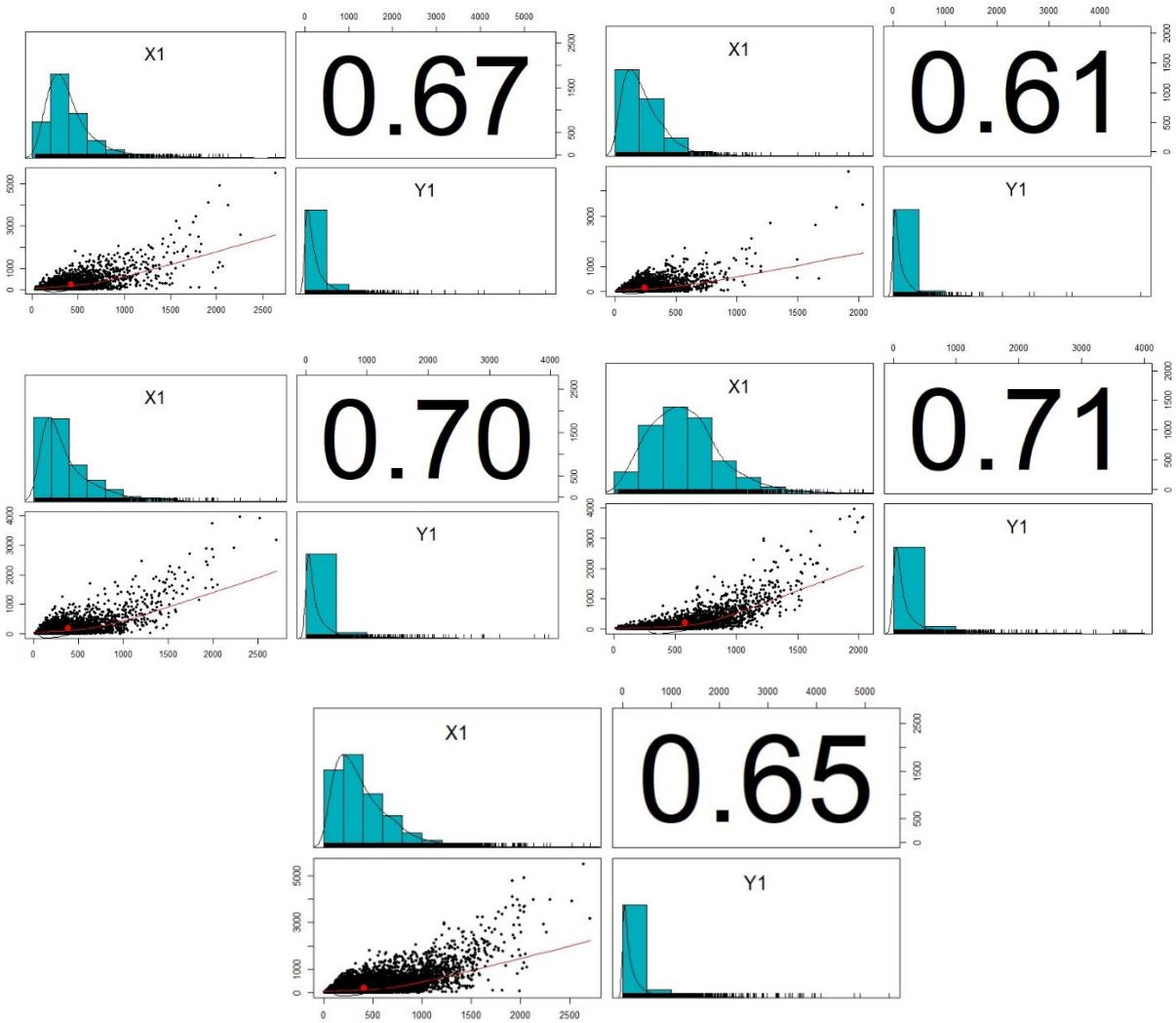

Figure 4. Scatter plots of original data in A (upper left), B (upper right), C (lower left), D (lower right) subregions and the whole South America (bottom) zones.

Regression modeling

To predict rainfall erosivity accurately, multiple regression models including linear regression model, polynomial regression model (quadratic order), power function regression model, and complex power function regression model, were generated in R, GraphPad Prism, and Excel (Figure 5). Besides, the specific regression equation, R^2, and AIC can be found in Table 2. To select the best-fit model for each climate zone, those regression models who have higher R^2 and lower AIC were considered. The results (Table 2) show that complex power regression models generated from GraphPad Prism and Excel are the best choices for regions A, B, and global, while Glmulti regression models created from R "Glmulti" package are better for region C and D.

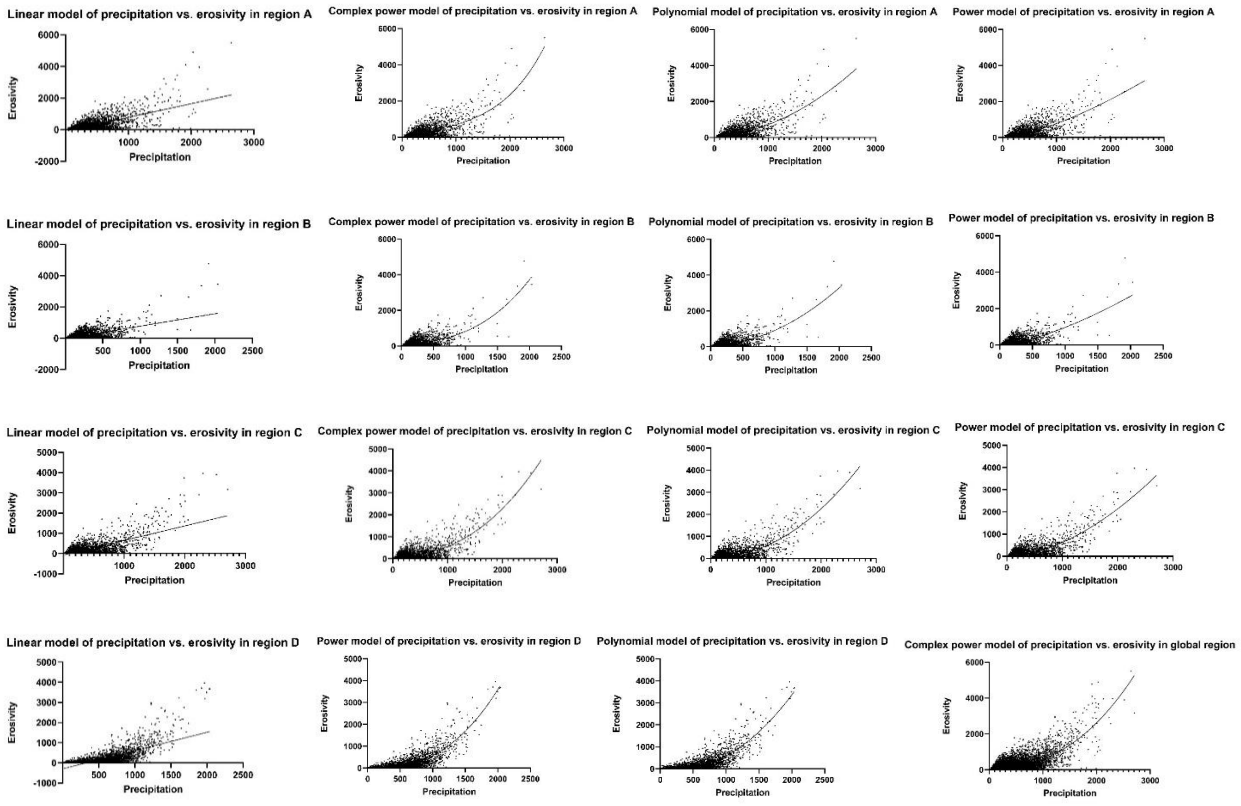

Figure 5. Multiple regression models for region A (first row from left to right), B (second row from left to right), C (third row from left to right), D (fourth row from left to the second right), and Global (lower right): linear regression model, polynomial regression model, power function model, and complex power model.

Table 2. Best-fit model selection table, corresponding R^2, and AIC values.

Area	Models (Glmulti R, excel solver build-in function, and GraphPad Prism)	Goodness
A	$Linear\ Y = -100.8 + 0.8698 * X$	$R^2 = 0.4426$ $AIC = 42339.82$
	$Quad\ Polynomial\ Y = 65.96 + 0.1824X + 0.00047X^2$	$R^2 = 0.4887$ $AIC = 42090.87$
	$Power\ Y = 0.03422X^{1.451}$	$R^2 = 0.4755$ $AIC = 42140.26$
	$Complex\ Power\ Y = 0.2707X^{1.124} + 9.074 \times 10^{-12}X^{4.248}$	$R^2 = 0.4941$ $AIC = 42049.5$
	$Glmulti\ Y = 13.64 + 0.503X + 1.703 \times 10^{-7}X^3$	$R^2 = 0.4929$ $AIC = 42060$
B	$Linear\ Y = -25.22 + 0.7959X$	$R^2 = 0.3689$ $AIC = 40179$
	$Quad\ Polynomial\ Y = 75.56 + 0.08185X + 0.00076X^2$	$R^2 = 0.4636$ $AIC = 39691$
	$Power\ Y = 0.03762X^{1.47}$	$R^2 = 0.4353$ $AIC = 39845.83$
	$Complex\ Power\ Y = 5.32X^{0.6046} + 2.788 \times 10^{-6}X^{2.744}$	$R^2 = 0.473$ $AIC = 39638.76$
	$Glmulti\ Y = 41.94 + 0.4357X + 3.496 \times 10^{-7}X^3$	$R^2 = 0.472$ $AIC = 39650$
C	$Linear\ Y = -70.91 + 0.718 * X$	$R^2 = 0.4954$ $AIC = 41256.17$
	$Quad\ Polynomial\ Y = 109.1 - 0.152X + 0.000612X^2$	$R^2 = 0.6266$ $AIC = 40352.77$
	$Power\ Y = 0.002949X^{1.775}$	$R^2 = 0.6098$ $AIC = 40485.43$
	$Complex\ Power\ Y = 3.98 \times 10^{-5}X^{2.341} + 33.93 \times 10^{-6}X^{0.196}$	$R^2 = 0.6284$ $AIC = 40348.92$
	$Glmulti\ Y = 86.67 + 3.972 \times 10^{-4}X^2 + 7.364 \times 10^{-8}X^3$	$R^2 = 0.6273$ $AIC = 40347.55$
D	$Linear\ Y = 10.61 + 0.01634 * X$	$R^2 = 0.5067$ $AIC = 41905.84$
	$Quad\ Polynomial\ Y = 174.4 - 0.7599X + 0.001176X^2$	$R^2 = 0.7028$ $AIC = 40386.18$
	$Power\ Y = 6.74 \times 10^{-6}X^{2.644}$	$R^2 = 0.7043$ $AIC = 40368.43$
	$Glmulti\ Y = 83.39 - 0.2487X + 4.334 \times 10^{-4}X^2 + 2.941 \times 10^{-7}X^3$	$R^2 = 0.7066$ $AIC = 40347.54$
G	$Complex\ Power\ Y = 2.799 \times 10^{-5}X^{2.407} + 37.62X^{0.1808}$	$R^2 = 0.5468$ $AIC = 164228$
	$Glmulti\ Y = 88.39 + 3.971 \times 10^{-4}X + 1.151 \times 10^{-7}X^3$	$R^2 = 0.5174$ $AIC = 164981$

Then, these best-fit models were tested using the Durbin Watson method to check if there is any autocorrelation among the variables in each model. The results show that there is no obvious autocorrelation identified in all selected models for subregions in South America because the Durbin Watson values are around 2. Furthermore, these best-fit models were applied to estimate the rainfall erosivity using validation rainfall data. After calculating the Pearson's correlation between the estimated rainfall erosivity and the rainfall data for validation, it was found that all values are greater than 0.5, meaning that there is a relatively high correlation between the estimated rainfall erosivity and the rainfall data for validation.

Table 3. Goodness criteria for best-fit models: R^2, AIC, Durbin Watson tests, and Pearson's correlation between estimated and observed data.

Area	R^2	AIC	Durbin Watson	Pearson's correlation with validation data
A	0.4941	42049.5	1.9979	0.7099
B	0.473	39638.76	2.031	0.5097
C	0.6273	40347.55	2.0006	0.6104
D	0.7066	40347.54	2.0063	0.8685
G	0.5468	164228	1.9275	0.7051

Model diagnostics

After generating those best-fit models in the previous step, the regression assumptions were tested in R to evaluate these models in terms of error normality, homoscedasticity of variance, and residuals independence (Figure 6).

Homoscedasticity of residuals or equal variance: in the upper left plots of all regions' models, the red line is approximately horizontal at zero indicating that the disturbances or residuals are homoscedastic.

Normality of residuals: in the upper right plot of all regions' models, we can assume the normality of residuals because all points fall nearly along the reference line.

Outliers and leverage: all models are good and pass this assumption because the red smoothed line is close to the mid-line. Also, most values have Cook's distance smaller than 5, indicating that these existing outliers do not have much influence on regression models.

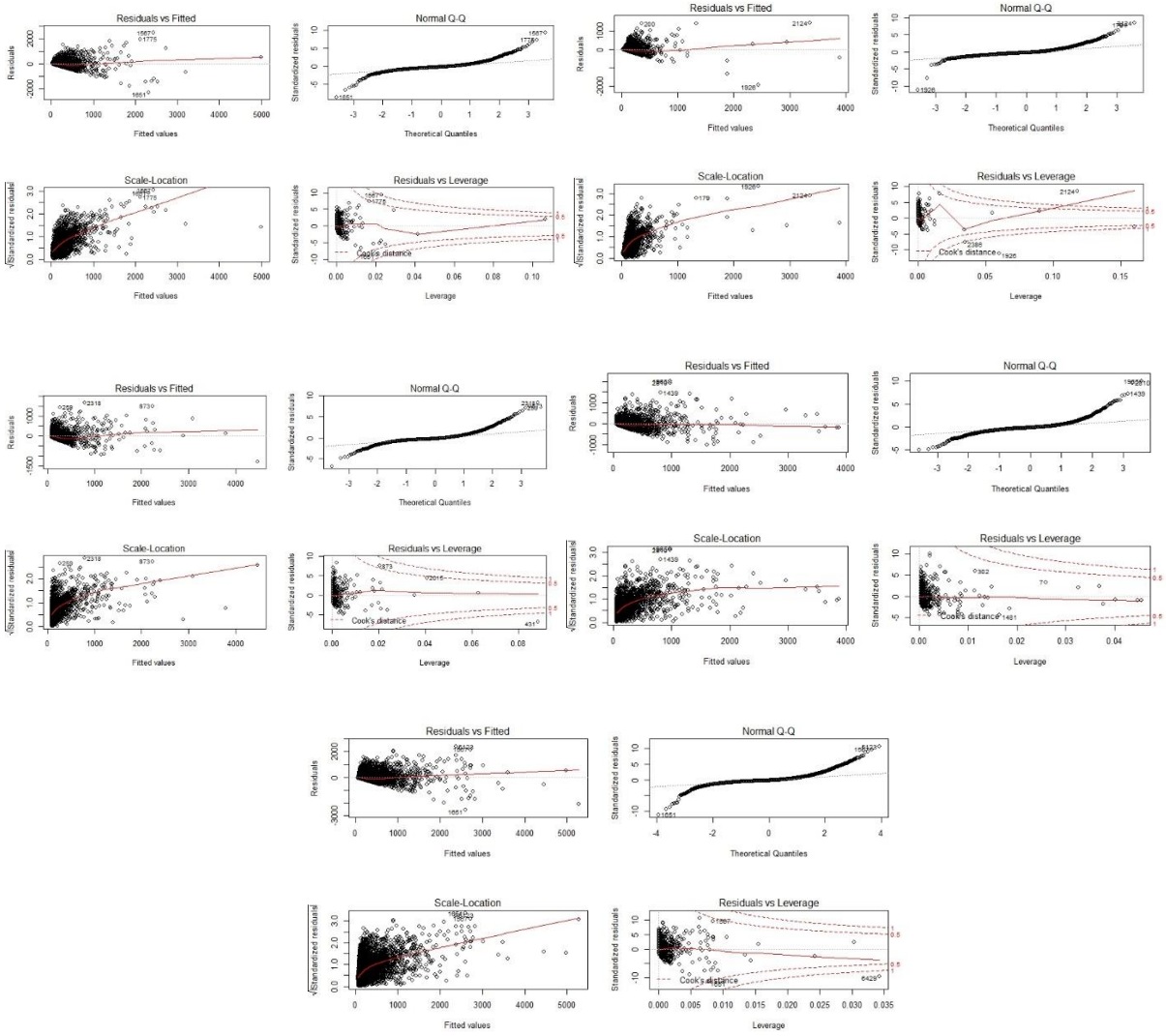

Figure 6. Regression diagnostics for Climate Zone A (top-left), B (top-right), C (bottom-left), D (bottom-right), and the whole South America regions.

Aside from the regression assumptions discussed above, Taylor diagrams were also generated to examine compare regression models regarding data standard deviation, correlation and Root Mean Square Error (RMSE). In Figure 7, the Taylor diagrams for all climatic zones produced

results that indicate the derived regression models have a strong ability to reproduce the spatial pattern of erosivity in the study area, with a standard deviation smaller than 1 to the reference point. Furthermore, the correlation coefficients for all the climate zones are above 60%. Finally, the RMSE differences are all within 1. The Taylor diagrams successfully validate these best-fit models using visual summary charts.

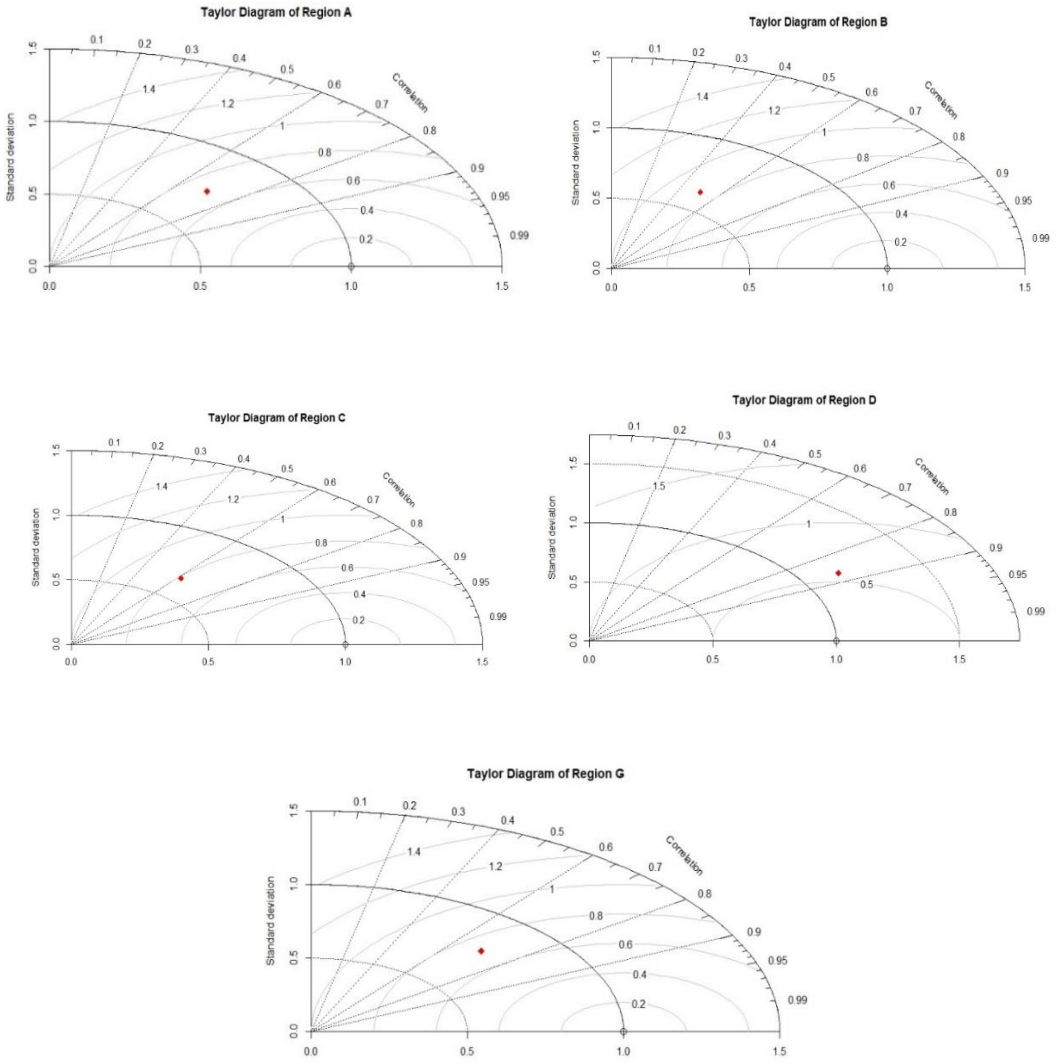

Figure 7. Taylor diagrams for Climate Zone A (top-left), B (top-right), C (bottom-left), D (bottom-right), and the whole South America regions.

Correlation between estimated erosivity and ENSO indicators

Pearson's correlation was used to perform the correlation between estimate erosivity and ENSO indicators. Figure 8 shows a series of scatterplots showing the monthly data distribution and cross-correlation coefficient of SOI, SST, MEI, and estimated rainfall erosivity data in four climate zones (A, B, C, and D) from January to December through the period between 1980 and 2017. M, H, and O represent MEI, SST and SOI indicators, respectively. A, B, C, D represent four classified climate zones within South America. The results show that MEI and SST have a strong negative relationship with SOI indicators. When it comes to the erosivity correlation with these three ENSO indicators, most correlation coefficients between region A's erosivity data and these ENSO indicators remain relatively low, with values smaller than 0.4. The only abnormal month is July when the correlation coefficients reach 0.54 between region A's erosivity data and SOI index. For other regions, the abnormal month is January for regions B, C, and D.

But how about the relationship during the typical months in El Niño and La Niña years? To examine the difference of erosivity-ENSO relationship between normal years and ENSO years, the erosivity records and ENSO values were extracted based on El Niño and La Niña years. Figure 9 and Figure 10 show the resulting scatterplots of erosivity and ENSO for specific El Niño and La Niña years. In Figure 9, A1, B1, C1, and D1 represent the statistical distribution of monthly average erosivity estimates during 1982, 1983, 1991, 1992, 1997, 1998, 2015 and 2016 in four different regions across the South America continent. In Figure 10, A2, B2, C2, and D2 represent the statistical distribution of monthly average erosivity estimates during 1988, 1989, 1999, 2000, 2007, 2008, 2010 and 2011 in four different regions across the South America continent (J, 2015). Additionally, a boxplot was also created based on these two groups of sub-datasets to visualize the

rainfall erosivity values during different ENSO years, which helps to determine if the ENSO exerts any influence on rainfall erosivity.

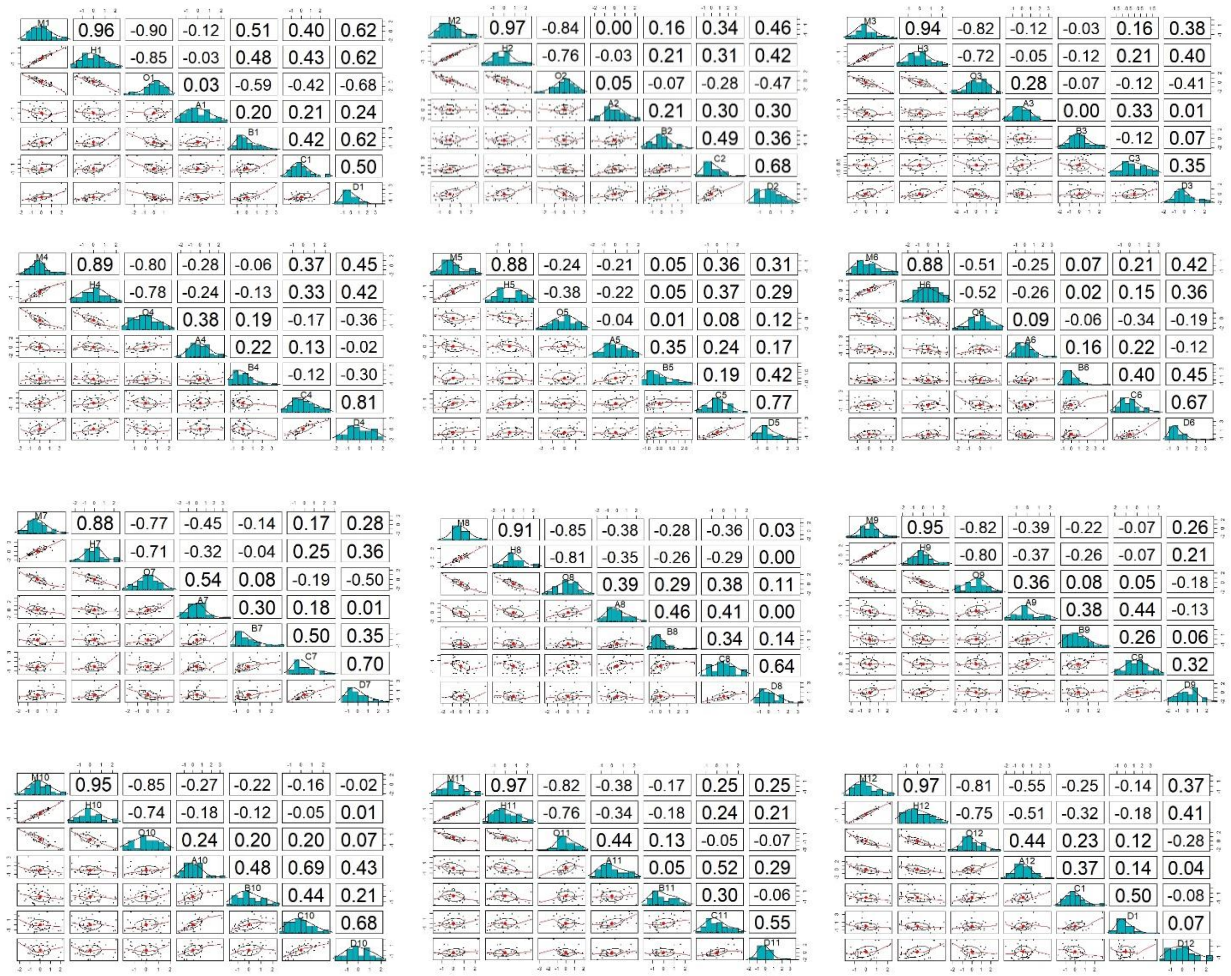

Figure 8. Scatterplots of El Niño indicators and estimated rainfall erosivity by month. M, H, and O represent MEI, SST and SOI indicators, respectively. A, B, C, D represent four classified climate zones within South America.

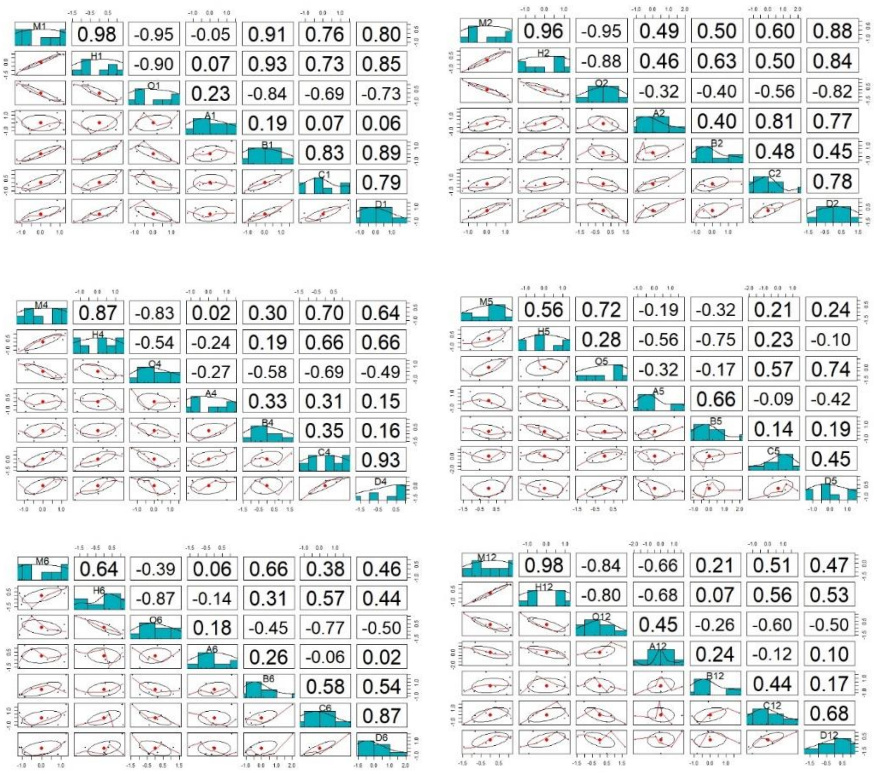

Figure 9. Scatterplots of El Niño indicators and estimated rainfall erosivity during El Niño years.

Figure 10. Scatterplots of El Niño indicators and estimated rainfall erosivity during La Niña years.

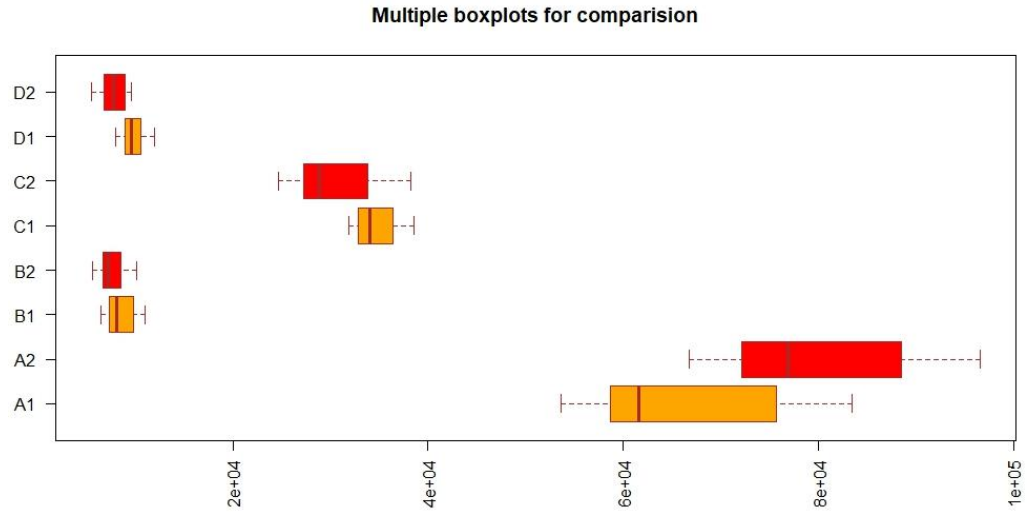

Figure 11. Boxplots of rainfall erosivity estimates during El Niño and La Niña years.

After comparing the scatterplots and boxplot generated from normal years and ENSO years, the findings are obvious. In El Niño years, the high-correlated erosivity-ENSO pairs are found frequently in January, February, April, May, June, and December. In La Niña years, the high-correlated erosivity-ENSO pairs are found frequently in January, March, April, May, July, and August. For both El Niño and La Niña years, the counts of abnormal months are all bigger than that of normal years, proving that ENSO events did influence rainfall erosivity a lot in the past climate period. Besides, in the boxplot showing in Figure 11, it is evident to see that the monthly rainfall erosivity is stronger during El Niño years while weaker during La Niña years in region B, C, and D. The reason why the result for region A is different from others might be that the rainfall amounts in tropical climate areas are continuous and abundant, which may not be substantially impacted by heavy precipitation events during El Niño and La Niña years. Also, the erosivity values are much bigger in region A compared to other regions due to its unique climate environment. In this case, any abnormal precipitation value may influence the overall calculation. Further study will improve the outlier remove and data clean process to deal with potential data uncertainty in region A.

Conclusion

This paper presented multiple regression models to estimate rainfall erosivity in different regions based on climate classification across the South America continent and examined the relationship between rainfall erosivity and ENSO index at different spatial and temporal scales. With smaller AIC and bigger R^2, best-fit models were selected and applied to predict rainfall erosivity. The estimated rainfall erosivity data was then validated and used to examine the potential relationship with ENSO indicators. The resulting relationship between rainfall erosivity and ENSO indicators is obvious, with correlation coefficients greater than 0.6, which indicates that rainfall erosivity has a strong correlation to the ENSO indicators. Also, abnormal rainfall erosivity data was observed during El Niño and La Niña years. The results show that rainfall erosivity tends to be higher during El Niño events, while lower during La Niña years.

Reference

1. Mello, C. R., Norton, L. D., Curi, N., Yanagi, S. N. M., & Silva, A. M. (2011). El-Niño southern oscillation and rainfall erosivity in the headwater region of the Grande River Basin, Southeast Brazil. *Hydrology and Earth System Sciences Discussions*, *8*(6), 10707-10738.

2. Nearing, M. A., Yin, S. Q., Borrelli, P., & Polyakov, V. O. (2017). Rainfall erosivity: A historical review. *Catena*, *157*, 357-362.

3. D'Odorico, P., Yoo, J. C., & Over, T. M. (2001). An assessment of ENSO-induced patterns of rainfall erosivity in the southwestern United States. Journal of Climate, 14(21), 4230-4242.

4. Lee, M. H., & Lin, H. H. (2015). Evaluation of annual rainfall erosivity index based on daily, monthly, and annual precipitation data of rainfall station network in Southern Taiwan. *International Journal of Distributed Sensor Networks*, *11*(6), 214708.

5. Wischmeier, W. H., & Smith, D. D. (1978). *Predicting rainfall erosion losses: a guide to conservation planning* (No. 537). Department of Agriculture, Science, and Education Administration.

6. Psd. (n.d.). PSD : Climate Indices: Monthly Atmospheric and Ocean Time Series. Retrieved from https://www.esrl.noaa.gov/psd/data/climateindices/list/.

7. Köppen, W., & Geiger, R. (1954). Klima der Erde (Climate of the earth) Wall Map. *Gotha: Klett-Perthes*.

8. Earthdata. (2019, December 10). Retrieved from https://earthdata.nasa.gov/.

9. Renard, K. G., & Freimund, J. R. (1994). Using monthly precipitation data to estimate the R-factor in the revised USLE. Journal of hydrology, 157(1-4), 287-306.

10. J. (2015). El Niño and La Niña years and intensities. *Golden Gate Weather Services (5 Sep 2013). Article Back to top*.