Principal component analysis of factors influencing soil erosion on Da river basin, Vietnam

Le Van Thinh[1]

Abstract: This study applies Principal Component Analysis (PCA) to identify and quantify the dominant factors influencing soil erosion in the Da River Basin, Vietnam. Using the Revised Universal Soil Loss Equation (RUSLE) and long-term hydro meteorological, soil, and land use data, ten key parameters were analyzed. PCA reduced these interrelated variables to four principal components explaining 77.3% of total variance. The first component, accounting for 36.7%, highlights rainfall erosivity, precipitation, and sand content as primary drivers of erosion. The second and third components represent the effects of slope, soil erodibility, and silt content, while the fourth reflects the influence of land management and conservation practices. Regression analysis confirmed a strong positive association between rainfall–soil texture interactions and erosion intensity.
Keywords: Soil erosion, Da river basin, principal component analysis, erosion factors.

1. Introduction

Soil erosion is one of the most severe forms of land degradation, directly impacting agricultural productivity, ecological stability, and water quality across many watersheds worldwide. In Vietnam, the combination of steep mountainous terrain and intense seasonal rainfall has led to widespread and severe soil erosion, particularly in upstream river basins. Among these, the Da River Basin stands out as one of the country's largest and most strategically important regions, especially for national energy security. However, it is increasingly threatened by both natural and human-induced pressures, which heighten the risks of soil erosion and riverbank instability.

In recent years, land use within river basins has undergone significant transformation due to population growth and economic development. These changes have directly affected vegetation cover and altered the physical and chemical properties of the soil, thereby accelerating environmental degradation, sediment erosion, and sediment transport. Understanding these dynamics is crucial for developing effective soil conservation and land management strategies.

Given the complexity of land use changes and reservoir development—and their profound effects on upstream hydrodynamics and erosion processes—accurately predicting sediment yield has become essential. Such predictions are vital for informed decision-making in reservoir management, soil conservation, and land restoration planning. For instance, Rulli et al. (2012) observed increased erosion in deforested areas of the Sila Massif in Calabria, Italy, where simulated rainfall experiments in Liguria showed significantly higher erosion rates in deforested zones. Similarly, in Vietnam, Chinh et al. (2021) applied the SWAT model to demonstrate the significant influence of land use on sediment yield in northern reservoirs. These findings underscore the critical role of land management in mitigating sediment production and controlling erosion.

To better understand the drivers of erosion, Principal Component Analysis (PCA) has been employed to identify the key contributing factors. This analytical approach follows a logical progression, beginning with the selection of potential erosion-related variables and culminating in the identification of key indicators that best represent the biophysical and socio-economic characteristics of the study area. Multivariate statistical methods, particularly PCA, are instrumental in deciphering complex data matrices and providing a comprehensive assessment of the factors influencing soil erosion. PCA is especially valuable for uncovering latent components that exert the greatest influence on erosion processes (Damiba et al., 2024). This study focuses on evaluating the interactions and key factors governing soil erosion dynamics in the Da River Basin, Vietnam.

2. Study area and data sources

2.1. Characteristics of the study area

The Da River Basin (see Fig. 1) is located in a humid climatic region and is the largest tributary of the Red River. Its distinctive reddish-brown hue is due to the high concentration of iron-rich sediments in its waters. Spanning a total area of 52,900 km², the basin extends across three countries: 24,980 km² in China, 1,120 km² in Laos, and 26,900 km² in Vietnam. Rainfall distribution varies significantly across the basin, with the upstream region in China receiving an average annual rainfall of approximately 1,200 mm, while the Vietnamese section experiences higher levels, averaging around 1,900 mm annually.

[1]Faculty of Civil Engineering, Thuyloi University, Vietnam
Received 4th Nov. 2025
Accepted 23rd Dec. 2025
Publication date 31st Dec. 2025

Fig 1. The Da River Basin

2.2. Data collection and analysis

Data were collected from sixteen meteorological stations and two hydrological stations that continuously monitor streamflow and turbidity. Topographic information was derived from the Global Digital Elevation Model, which provides a spatial resolution of 90 metres. Land use patterns were analysed using maps from 2000 and 2010, sourced from GeoNetwork. Soil classification data were compiled from multiple references, including the Atlas of Vietnam (1999) for the Vietnamese portion and FAO (1995) for the Chinese section. Additionally, long-term streamflow and sediment data from the Hoa Binh station were analysed for the period spanning 1986 to 2020.

3. Methods

3.1. Soil erosion model

a) RUSLE erosion calculation model

The RUSLE erosion calculation model, revised by Renard et al. (1997), is designed to estimate average soil erosion within a river basin using six parameters; the RUSLE model is represented as follows:

$$A = R \cdot K \cdot LS \cdot C \cdot P \quad (1)$$

where: Annual soil loss (A) indicates the average rate of soil erosion for a specific area over a certain period (tons/ha/year); The rainfall erosivity factor (*R,* MJ/ha·mm/h); The topographic factor (*LS)* ; The soil erodibility factor (*K,* tons·h/MJ·mm); The cover management factor (*C);* The support practice factor (*P).*

b) Sediment transport in the river

Based on the study by Ranzi et al., (2012), the sediment transport capacity of the river was determined. Simultaneously, assuming sediment transport follows the linear reservoir law:

$$q_{s,i,k} = q_{s,i-1,k} e^{-T/\theta_k} + S_{k,i}(1 - e^{-T/\theta_k}) \text{ (tons/day)} \quad (2)$$

where: $\mathbf{q_{s,i,k}}$ **(ton/day)** – suspended sediment discharge at each measurement station; **T** – measurement duration (day); $\boldsymbol{\theta_k}$ – travel time of the basin, depending on basin length and flow velocity (day); $\mathbf{S_{k,i}}$ – accumulated monthly sediment production (tons/day), calculated using the RUSLE model. Volume of soil eroded and transported over one month:

$$V_{s,i,k} = q_{s,i-1,k}\theta_k(1 - e^{-T/\theta_k}) + S_{k,i}T + S_{k,i}\theta_k(e^{-T/\theta_k} - 1) \text{ (tons)} \quad (3)$$

The rainfall erosivity factor (R) was derived from 16 rainfall stations within the Da River Basin. The soil erodibility factor (K), which depends on soil structure, was calculated based on soil classification maps and validated against selected sites with field survey data. The topographic factor (LS) was computed using the Digital Elevation Model (DEM). The cover management factor (C) was determined from land cover maps, while the support practice factor (P) was identified based on erosion control measures implemented in the basin, with specific practices corresponding to crop types. The model was calibrated and validated by optimizing the C and P parameters to best fit the basin conditions, ensuring the highest Nash–Sutcliffe Efficiency (NSE) coefficient. The estimation of the model parameters is described in detail in Ranzi et al. (2012), with additional information can be found in Le Van et al. (2018).

3.2. Steps to perform principal component analysis

Step 1: Standardize the Data

Each variable is centred (subtracting the mean) and scaled (dividing by the standard deviation) so that all variables have zero mean and unit variance:

$$Z = \frac{C_{ij} - C_j}{S_j} \quad (1)$$

where Z is the standardized matrix, C_{ij} is the value for observation i and variable j, C_j is the mean and S_j is the standard deviation.

Step 2: Compute the Covariance Matrix

The covariance matrix is calculated from the standardised data to capture relationships between variables:

$$C = \frac{X'X}{n-1} \quad (2)$$

where C is the covariance matrix, X is the standardised data matrix, X' is its transpose, and $n$is the number of observations.

Step 3: Calculate Eigenvalues and Eigenvectors

Eigenvalues (λ') and eigenvectors of the covariance matrix $C$are obtained by solving the characteristic equation:

$$|(C - \lambda' I)| = 0 \quad (3)$$

where I is the identity matrix of the same size as C. For each eigenvalue λ', the corresponding eigenvector $v$is found by solving:

$$|(C - \lambda' I) \times v| = 0 \quad (4)$$

Eigenvalues are ranked in descending order ($\lambda'_1 > \lambda'_2$), with corresponding eigenvectors defining PC1 (v_1) and PC2 (v_2).

Step 4: Construct the Principal Components

PC loading matrix is obtained by multiplying the square roots of eigenvalues of $C$with the characteristic values of the correlation matrix P

$$R = PD^{\frac{1}{2}} \quad (5)$$

3.3. Use of R platform for principal component analysis (PCA)

Principal Component Analysis (PCA) was conducted within the R environment, primarily utilizing the packages *stats*, *tidyverse*, *FactoMineR*, and *factoextra*. PCA was performed on the correlation matrix to eliminate the influence of differing measurement units across variables. To enhance interpretability, an orthogonal varimax rotation was applied to the factor loading matrix.

4. Results

4.1. Results of RUSLE model simulation

Simulation results using observed data for the period 1986–2005 (calibration phase; see Fig. 2) revealed that the Nash–Sutcliffe Efficiency (NSE) coefficient at Hoa Binh was 0.72. For the validation period of 2006–2020, the NSE coefficient remained at 0.72 for Hoa Binh. Potential soil erosion map of the Da River Basin (see Fig. 3).

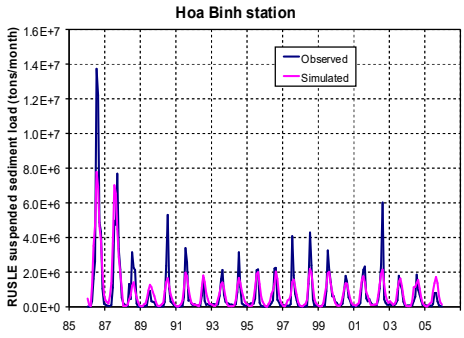


Fig 2. Simulation of monthly sediment yield at Hoa Binh for the period 1986-2005

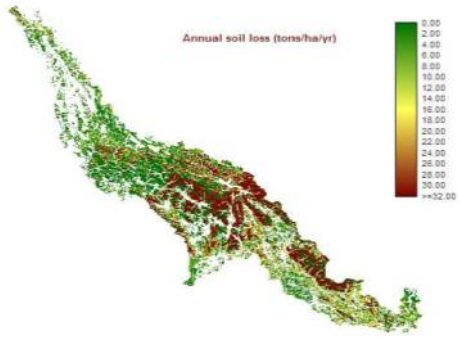
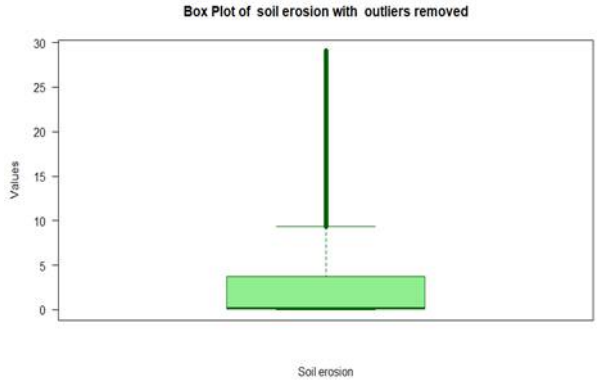
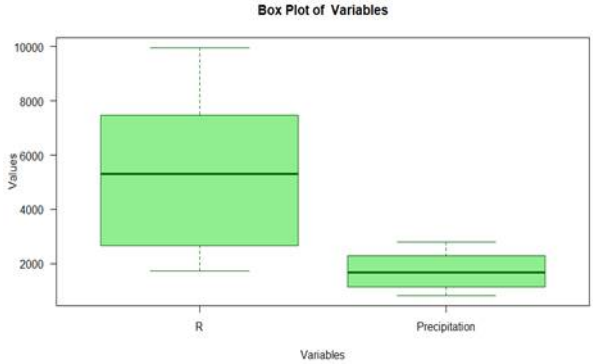


Fig 3. Average soil loss (tons ha^{-1} yr^{-1}) in the Da River Basin

4.2. Data distribution and variable correlation

Ten parameters were chosen based on two primary considerations: (i) These variables represent the core components of the RUSLE model and have been widely employed in previous erosion studies; (ii) They comprehensively capture the three groups of factors governing soil erosion, including: climate (R, precipitation); topography–soil (slope%, LS, K, silt%, sand%, clay%); and land management–use (C, P).





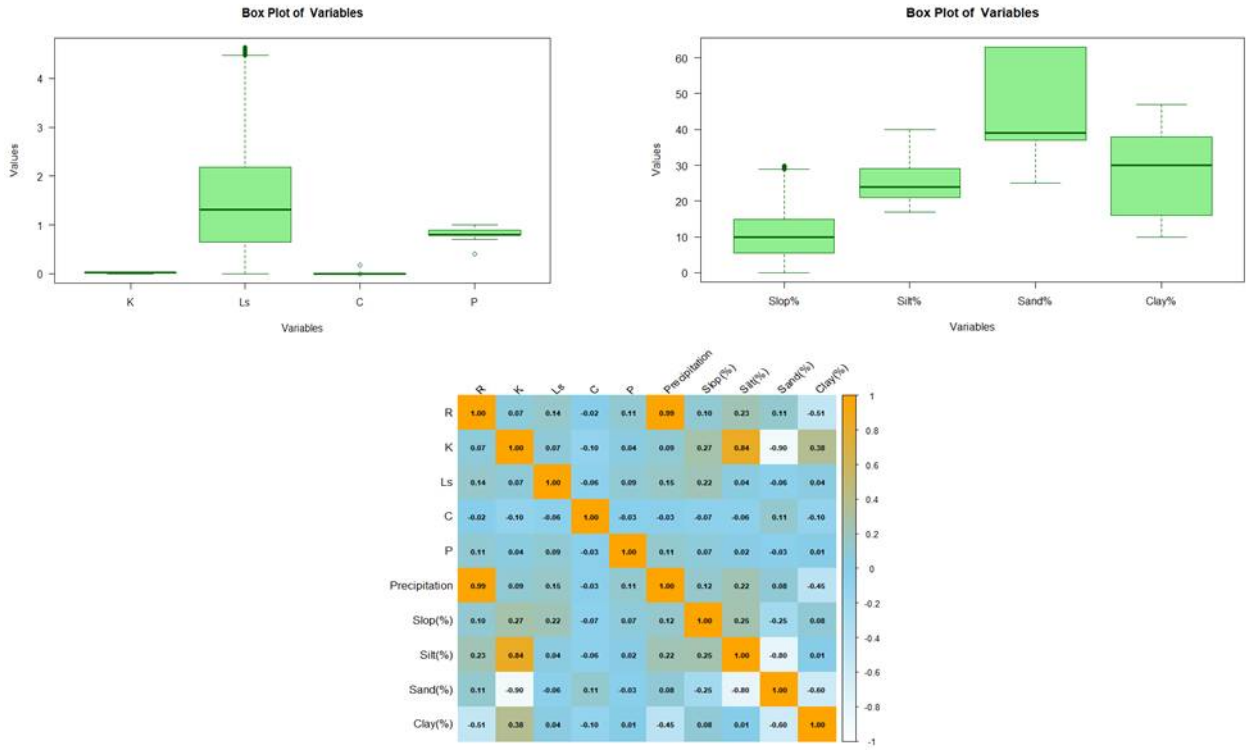


Fig 4. Variable distribution correlation matrix

Figure 4 shows the correlation matrix and the distribution of variables. Variables with differing units were grouped separately to improve the clarity and interpretability of the box plots. Although some variables show high correlations, all were retained for PCA, which effectively manages multicollinearity and preserves the comprehensive information needed for analysis.

4.3. Principal component analysis results

Table 1 presents the eigenvalues, variance explained, and cumulative percentages for the principal components. Principal components PC1 through PC4, each with eigenvalues above 1 (Kadam et al., 2019), collectively account for 77.3% of the total variance. Table 2 details the factor loadings for each of the five principal components. The first principal component (PC1) explains 36.7% of the total variance and shows positive correlations with sand proportion, precipitation, and the rainfall erosivity (R). The second principal component (PC2) explains 17.8% of the variance and is strongly associated with precipitation, rainfall erosivity (R), land slope (LS), K, and silt factor. The third principal component (PC3) explains 12.5% and correlates with silt and the cover-management (C). The fourth principal component (PC4) shows a positive correlation with the support practice (P).

Figure5 presents the spatial clustering of mesh blocks according to their erosion-related characteristics. Each point in the plot represents a mesh block, and the proximity between points indicates similarity in morphometric and environmental attributes. Mesh blocks that are located close together share comparable features, such as slope, soil composition, and rainfall exposure. This visualisation helps identify patterns and groupings that may correspond to areas of similar erosion risk, providing a spatially explicit understanding of soil erosion dynamics across the study area.

Figure 6 complements the clustering analysis by illustrating how the original variables contribute to the principal components derived from the dataset. Each arrow represents a variable, with its direction indicating alignment with the principal components and its length reflecting the strength of its contribution-longer arrows denote greater influence. The angle between arrows signifies correlation: small angles indicate strong positive correlation, right angles suggest no correlation, and opposing directions imply negative correlation.

This figure highlights the dominant role of rainfall and soil texture-particularly sand content-in shaping soil erosion dynamics. These variables strongly influence the first principal component (Dim1), as shown by the alignment and length of their arrows. Dim2 is more influenced by clay and silt content, slope, and terrain-related factors such as the LS and K factors. In contrast, management-related variables like the C and P factors show shorter arrows and lighter colours, indicating a lower contribution to the principal components. The colour gradient, ranging from light to dark blue, reinforces the relative strength of each variable's influence, with darker arrows (e.g., precipitation and sand) standing out as key drivers of erosion. A regression analysis was performed with soil erosion as the dependent variable, and the results are summarized in Table 3.

Table 1. Eigenvalues, variance explained, and cumulative percentages of all components

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Eigenvalue	3.7	1.8	1.3	1.03	0.9	0.7	0.4	0.2	0.009	0.0003
Variability (%)	36.7	17.8	12.5	10.3	9.2	6.7	4.4	2.3	0.09	0.003
Cumulative (%)	36.7	54.5	67	77.3	86.4	93.1	97.5	99.8	99.9	100

Table 1. Factor loadings

Variable	PC1	PC2	PC3	PC4
R	**0.35**	**0.52**	0.17	0.02
K	-0.42	**0.30**	0.18	0.01
Ls	-0.04	**0.32**	-0.60	0.07
C	0.04	-0.17	**0.20**	0.60
P	-0.02	0.08	-0.14	**0.80**
Precipitation	**0.33**	**0.54**	0.17	0.01
Slope (%)	-0.12	0.22	-0.61	-0.04
Silt (%)	-0.31	**0.37**	**0.34**	0.01
Sand (%)	**0.50**	-0.14	-0.10	-0.01
Clay (%)	-0.48	0.00	-0.02	0.01

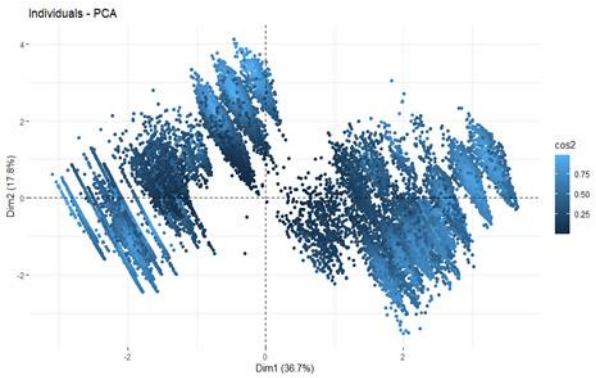


Fig 5. Clustering of mesh blocks based on erosion-related features

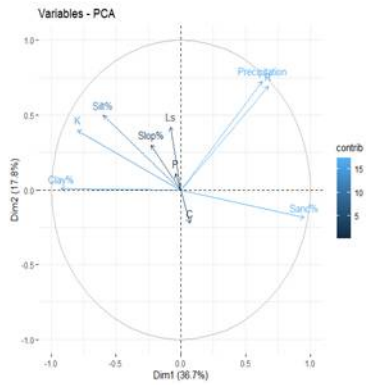


Fig 6. Contribution of variables to principal components

4.4. Discussion

PCA and regression analyses reveal the multidimensional nature of erosion in the Da River Basin. Table 3 summarises the results. PC1 is the most influential component, explaining 36.7% of the variance and showing the strongest positive regression coefficient (+0.63). It is mainly linked to high rainfall and sandy soils, indicating greater susceptibility to erosion in such areas.

PC2 accounts for 17.8% of the variance and has a moderate positive regression coefficient (+0.27). It reflects the combined effects of terrain and soil structure-particularly precipitation, R factor, slope, K, and silt-which moderately increase erosion risk, especially in complex topographies.

PC3 accounts for 12.5% of the variance, with positive loadings on silt and the C factor, but strong negative associations with slope and LS. Its negative coefficient (–0.46) suggests more stable or managed landscapes, likely due to flatter terrain or effective vegetation cover. PC4 explains 10.3% of the variance, dominated by positive loadings on P and C factors and a positive coefficient (+0.41), indicating that conservation practices exist but may be less effective because of poor implementation or site-specific limitations.

Table 2. Summary Regression Results

Component	Explained variance (%)	Regression coefficients	Key Loadings	Interpretation
PC1	36.7	+0.63	Sand%, Precipitation, R factor	Captures most data structure, strongly linked to erosion.
PC2	17.8	+0.27	Precipitation, R, LS, K, Silt	The second highest variance, moderately positively associated with erosion.

Component	Explained variance (%)	Regression coefficients	Key Loadings	Interpretation
PC3	12.5	–0.46	Silt, C factor	Moderately linked to erosion, with a negative effect.
PC4	10.3	0.41	P factor	Moderately linked to erosion, with a positive effect.

Areas dominated by PC1-high rainfall, strong rainfall energy, and substantial sand content-pose the greatest erosion risk and should be prioritised for measures such as improved surface cover, vegetation restoration, and agroforestry. Conversely, zones influenced by terrain and soil structure (PC2–PC3), especially steep slopes with high clay or silt content, require interventions like contour farming, terracing, and slope stabilisation to curb runoff and soil loss.

This aligns with numerous studies in tropical mountainous watersheds, where intense rainfall combined with coarse soil texture accelerates soil detachment and transport. Kadam et al. (2019), examining western river basins in India, found that rainfall- and soil-related components often explain over 30–40% of PCA variance. Similarly, Damiba et al. (2024) in Kenya emphasised the dominant influence of climatic–edaphic factors on erosion dynamics in hilly terrains. However, compared with several Southeast Asian studies, this research shows notable differences. For instance, Wuttichaikitcharoen et al. (2014) applied PCA to estimate suspended sediment yield in ungauged basins of Northern Thailand, identifying five key factors: basin slope, hierarchical anomaly density, main channel length, forest area, and dry-season rainfall.

5. Conclusion

This study applied Principal Component Analysis (PCA) to identify the dominant factors influencing soil erosion in the Da River Basin, Vietnam. Key variables included slope, slope length, vegetation cover, soil type, rainfall, and soil texture (sand, silt, and clay content). The analysis revealed that rainfall and sand content are the most significant contributors to erosion risk, followed by terrain characteristics and soil structure.

While vegetation cover and conservation practices can play a crucial role in reducing erosion, the findings suggest that their effectiveness varies across the basin. In some areas, insufficient or poorly targeted interventions have limited their impact, allowing erosion to persist. These insights reinforce the need for site-specific, data-driven management strategies that consider both biophysical and socio-economic conditions to effectively mitigate erosion and support sustainable land use planning.

Acknowledgments

This research was financially supported by the IRD, the ACROSS initiative, and Thuyloi University, Vietnam, under grant number 27112024/IRD-TLU.

References

Chinh T.M et al. (2021), "*The study proposes the application of the universal soil loss equation (USLE) in predicting erosion caused by agricultural production activities in the mountainous areas of Northern Vietnam"*. Journal of Science and Technology of Irrigation and Environment 76, 39-45.

Damiba W.A.F et al. (2024), "*Soil quality index (SQI) for evaluating the sustainability status of Kakia-Esamburmbur catchment under three different land use types in Narok County, Kenya"*. Heliyon 10, e25611.

Kadam A.K et al. (2019), "*Identification of erosion-prone areas using modified morphometric prioritization method and sediment production rate: a remote sensing and GIS approach"*. Geomatics, Natural Hazards and Risk 10, 986–1006.

Le Van T, Ranzi R, Rulli M.C (2018), *"Modeling Soil Erosion and Sediment Load for Red River Basin (Vietnam): Impact of Land Use Change and Reservoirs Operation"*. IEEE Xplore, pp 1-6.

Meshram S.G, Sharma S.K (2017), "*Prioritization of watershed through morphometric parameters: a PCA-based approach"*. Appl Water Sci 7(3), 1505–1519.

Ranzi R, Le T.H, Rulli M.C (2012), "*A RUSLE approach to model suspended sediment load in the Lo river (Vietnam): Effects of reservoirs and land use changes"*. Journal of Hydrology 422–423, 17–29.

Renard K.G et al. (1997), *"Predicting Soil Erosion by Water: A Guide to Conservation Planning with the Revised Universal Soil Loss Equation (RUSLE)"*. Agriculture Handbook No.703, USDA, Washington DC.

Rulli M.C, Offeddu L, Santini M (2012), "*Modeling post-fire water erosion mitigation strategies"*. Hydrol. Earth Syst. Sci. 17, 2323–2337.

Wuttichaikitcharoen P, Babel M (2014), "*Principal Component and Multiple Regression Analyses for the Estimation of Suspended Sediment Yield in Ungauged Basins of Northern Thailand"*. Water 6, 2412–2435.