EARTH SCIENCES
RESEARCH JOURNAL
Earth Sci. Res. J. Vol. 29, No. 4 (December, 2025): 379-386

UNIVERSIDAD
NACIONAL
DE COLOMBIA

SOILS

Performance of Random Forest in predicting soil loss based on values calculated by USLE

Arthur Pereira do Santos*, Liliane Moreira Nery, Letícia Tondato Arantes, Bruno Pereira Toniolo, Darllan Collins da Cunha e Silva, & Roberto Wagner Lourenço
São Paulo State University (UNESP), Institute of Science and Technology, Sorocaba, São Paulo, Brazil
*Corresponding author: arthur.p.santos@unesp.br

ABSTRACT

Soil erosion directly affects agricultural productivity and water resource quality, but estimating soil loss is complex and costly. This study proposes a machine learning (ML) approach to predict soil loss using selected factors from the Universal Soil Loss Equation (USLE) and the Normalized Difference Vegetation Index (NDVI). We applied the Random Forest (RF) algorithm to train and validate two models using different combinations of predictors: (1) NDVI, topographic factor (LS), and land cover/management factor (CP); and (2) NDVI, LS, and soil erodibility factor (K). These variables represent land use, conservation practices, and topographic conditions in the Sorocabuçu River Basin (SRB), part of Brazil's Atlantic Forest biome with high environmental and socioeconomic value. Soil loss was classified into three classes (in ton/ha): low (0–10.0), moderate (10.1–50.0), and high (≥50.1). A total of 3348 samples were randomly selected and proportionally distributed to reflect class representation across the study area. We used a 70/30 train-test split and standardized parameters (50 trees and four variables per node) to enable reproducibility. The model using NDVI, LS, and CP achieved 93.43% accuracy with a kappa index of 0.90. The performance was especially strong for the low-loss class, the most prevalent in the area. The second model using NDVI, LS, and K achieved 97.14% accuracy with a kappa index of 0.90, showing excellent results, particularly for the high-loss class, which poses the greatest environmental risk. These models prove effective in identifying areas at risk of severe erosion using fewer, more accessible parameters. The approach offers a scalable and practical tool for decision-makers, environmental managers, and public agencies to monitor and mitigate soil degradation, particularly in sensitive and ecologically important regions.

Keywords: NDVI; Machine Learning; Topographic factor (LS); Erosion processes.

Desempeño del algoritmo Random Forest en la predicción de la pérdida de suelo basada en valores calculados por la USLE

RESUMEN

La erosión del suelo afecta directamente la productividad agrícola y la calidad de los recursos hídricos; sin embargo, la estimación de la pérdida de suelo es un proceso complejo y costoso. Este estudio propone un enfoque de aprendizaje automático (Machine Learning (ML)) para predecir la pérdida de suelo utilizando factores seleccionados de la Ecuación Universal de Pérdida de Suelo (USLE) y el Índice de Vegetación de Diferencia Normalizada (NDVI). Se aplicó el algoritmo Random Forest (RF) para entrenar y validar dos modelos con diferentes combinaciones de variables predictoras: (1) NDVI, factor topográfico (LS) y factor de cobertura y manejo del suelo (CP); y (2) NDVI, LS y factor de erodabilidad del suelo (K). Estas variables representan el uso del suelo, las prácticas de conservación y las condiciones topográficas en la cuenca del río Sorocabuçu (SRB), ubicada en el bioma de la Mata Atlántica de Brasil, una región de alto valor ambiental y socioeconómico. La pérdida de suelo se clasificó en tres categorías (en t/ha): baja (0–10,0), moderada (10,1–50,0) y alta (≥50,1). Se seleccionaron aleatoriamente un total de 3348 muestras, distribuidas proporcionalmente para reflejar la representatividad de las clases en el área de estudio. Se utilizó una división de los datos del 70% para entrenamiento y 30% para validación, junto con parámetros estandarizados (50 árboles y cuatro variables por nodo) para garantizar la reproducibilidad del análisis. El modelo basado en NDVI, LS y CP alcanzó una precisión del 93,43% y un índice kappa de 0,90, con un desempeño destacado en la clase de baja pérdida de suelo, la más frecuente en el área. El segundo modelo, que utilizó NDVI, LS y K, obtuvo una precisión del 97,14% y un índice kappa de 0,90, mostrando resultados excelentes, especialmente en la clase de alta pérdida de suelo, que representa el mayor riesgo ambiental. Los resultados demuestran que ambos modelos son eficaces para identificar áreas con riesgo de erosión severa utilizando un conjunto reducido de parámetros más accesibles. Este enfoque constituye una herramienta práctica y escalable para la toma de decisiones por parte de gestores ambientales y organismos públicos, contribuyendo al monitoreo y la mitigación de la degradación del suelo, particularmente en regiones sensibles y de gran importancia ecológica.

Palabras clave: NDVI; Aprendizaje automático; Factor topográfico (LS); Procesos de erosión.

Record
Manuscript received: 30/06/2025
Accepted for publication: 15/12/2025

How to cite this item:
Pereira do Santos, A., Moreira Nery, L., Tondato Arantes, L., Pereira Toniolo, B., da Cunha e Silva, D. C., & Wagner Lourenço, R. (2025). Performance of Random Forest in predicting soil loss based on values calculated by USLE. *Earth Sciences Research Journal,* 29(4), 379-386. https://doi.org/10.15446/esrj.v29n4.121271

ISSN 1794-6190 e-ISSN 2339-3459
https://doi.org/10.15446/esrj.v29n4.121271

1. Introduction

The latest report by the Food and Agriculture Organization of the United Nations (FAO) warned that the planetary agricultural system, associated with the interconnected network of soil, land, and water, is at breaking point (FAO, 2023). Soil erosion is leveraging this discontinuity. It is responsible for removing 75 billion tons of soil by 2050 and discarding 10% of the world's agricultural production (FAO, 2021).

Due to the impacts of soil loss on food production, water quality, ecosystem services, and infrastructure, soil loss measurements and estimates have become increasingly important (Boardman and Poesen, 2006; Ganasri and Ramesh, 2016). However, its measurement is costly and localized, often making it necessary to estimate methods that can consider this loss using computational mathematical models.

Related studies show that the average global soil loss is between 12 and 15 tons per hectare per year, negatively impacting the environment and human life (Zhang et al., 2021; Tanyas et al., 2015). Therefore, the models that calculate its loss are fundamental for land use planning based on sustainable development principles.

Among these models, the surface water erosion model should be highlighted. It is considered the empirical model of the Universal Soil Loss Equation (USLE) (Sheikh et al., 2011; Panagos et al., 2015) proposed by Wischmeier e Smith (1978). It was designed to predict the average soil loss due to surface water erosion over the long term and in areas under specific soil cover and management conditions.

The USLE assumes that the amount of soil lost to water erosion is conditioned by rainfall erosivity (R), soil erodibility (K), topography (LS), soil cover or type of soil management (C), and the soil conservation practices employed (P) (Wischmeier and Smith, 1978).

Given the above context and the numerous obstacles involved in standardizing a model that covers all the peculiarities of soil loss, images from Remote Sensing (RS) and geoprocessing techniques are currently widely used to calculate the USLE (Pandey et al., 2007; Sheikh et al., 2011; Helmi, 2023). They provide new ways to obtain estimates of its factors (Silva et al., 2017) and new approaches, which are still very little tested based on Machine Learning (ML) techniques (Cheng et al., 2018; Nguyen et al., 2019). Thus, they seek to obtain more accurate estimates consistent with the nature of the erosion process.

Machine Learning (ML) techniques, such as the Random Forest (RF) algorithm (Breiman, 2001), have proven effective for extracting information directly from remote sensing data and topographic models without the need for empirical model calibration (Cheng et al., 2018). An example is the study by Santos et al. (2025), who applied RF to classify soil texture in the Sorocabuçu River Basin (SBR), achieving high accuracy in textural classification. Complementarily, Poletti et al. (2025) employed Fuzzy Logic to map land-use suitability in the same basin, demonstrating the usefulness of approximate-logic-based methods for integrating environmental variables and identifying land-use conflicts.

Therefore, this study aims to (1) evaluate the performance of the RF classifier in predicting soil loss in a river basin using only two of the parameters comprising the USLE and coupled with the Normalized Difference Vegetation Index (NDVI); (2) identify the USLE factor that most influenced the prediction model for the study area; and (3) analyze the metric performance of each class in the models generated.

Despite advances in modeling techniques applied to watershed studies, an important gap remains: the strong dependence on large data volumes and the inherent complexity of traditional soil-loss estimation models, such as the USLE. In many contexts, however, these data are scarce, difficult to obtain, or characterized by high uncertainty. In this scenario, the present study proposes an alternative method based on machine learning (ML), capable of estimating soil loss using a reduced set of USLE variables while preserving the representativeness of the basin's environmental conditions and increasing the applicability of the model in data-limited areas.

2. Materials and methods

2.1 Study Area

Part of the Atlantic Forest biome (IBGE, 2021), the SRB (Figure 1) covers 202.67 km². It is located in the municipality of Ibiúna, São Paulo (SP), occupying approximately 19% of its land area. It is also located within the Itupararanga Environmental Preservation Area (EPA), covering 22% of its spatial extension (Maia Júnior and Lourenço, 2020; Paula, 2025). Thus, it is of great environmental importance.

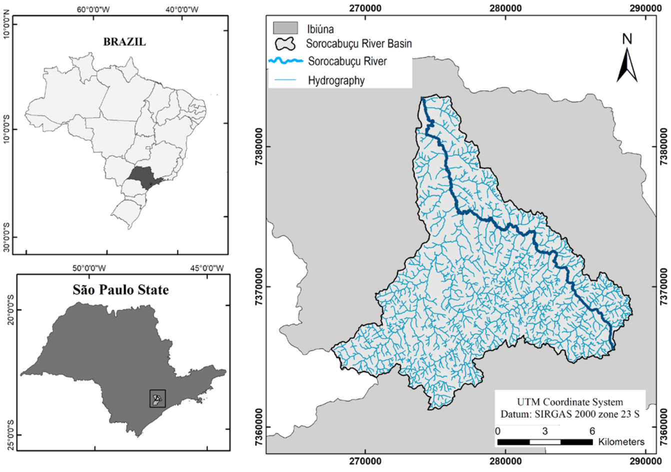

Figure 1. Geographic location of the Sorocabuçu River basin.
Source: The Authors (2024).

Regarding the climate, Dubreuil (2017) updated the Köppen (1948) classification of the area in which the SRB is located. It is of the Cwa type from 50 to 80%. It belongs to the subtropical climate of dry winter and hot summer, with average temperatures below 18º C and 47.5 mm per month of rainfall and averages above 22º C and 196 mm per month of rainfall, respectively.

Currently, the river basin has a low population density. However, there are small rural properties (Santos et al., 2025). Although some areas are suitable for temporary and permanent crops, this predominance is limited mainly by the risk of erosion, which is favored by its topography (Andreoti, 2012).

The SBR exhibits a set of pedological characteristics that increases its susceptibility to erosion. According to Rossi (2017), the area is predominantly composed of Latosols (89.26%), with smaller occurrences of Gleysols (10.48%) and Cambisols (0.26%). Red-Yellow Latosols, developed under intense weathering of Precambrian crystalline rocks such as granites and gneisses, are typically deep and well-drained but tend to show low natural fertility and reduced nutrient-retention capacity (EMBRAPA, 2018).

When combined with the steep slopes found in the region, these pedological attrna intensify the risk of soil loss, reinforcing the need for more robust and efficient approaches to assess erosionnability (Arantes 2024b; Toniolo et al., 2024; Costa et al., 2025).

3. Methodological procedures

USLE Data Preparation and Classification

We designed the USLE information plans based on the methodology proposed in Silva et al. (2017). First, we standardized all the data (R, K, LS, and CP) into the matrix format (raster). The CP factor considered the types of land use, management, and conservation practices employed. We interpreted the final result into three soil loss classes (Table 1).

Table 1. Soil loss classes according to the Universal Soil Loss Equation (USLE). Source: The Authors (2024).

Amount of Soil Loss (ton.ha^{-1})	Classification	Class
0.0 – 10.0	Low	1
10.1 – 50.0	Moderate	2
>= 50.1	High	3

Then, we used the LS, K, and CP factors in the RStudio software (RStudio Team, 2023) for validation purposes. We chose to work with two USLE factors in each prediction model added to the NDVI.

First, we carried out the modeling with the LS, CP, and NDVI factors (Model 1). After that, we used the LS, K, and NDVI factors (Model 2). We chose these parameters based on the agricultural and topographical characteristics of the study area.

Thus, it is worth noting that we disregarded the R factor in both models and used the LS factor in both because the river basin has a high degree of slope and ramp length.

We disregarded the R factor based on the results of Arantes (2024b) and Silva et al. (2017). In the first study, erosivity values were obtained from the erosivity database of the State of São Paulo (SP-Erosividade), which compiles historical series from the National Water Resources Information System. The authors reported an average value of 8,636 MJ.mm/ha.h.year and a range of only 240 MJ.mm/ha.h.year for the watershed study, evidencing low spatial variability. Silva et al. (2017) also observed a practically homogeneous behaviour of erosivity for the adjacent watershed these authors used nearby meteorological stations without any being located within the low number of meteorological stations as a limitation for calculating erosivity in this study area.

In the Brazilian context, the low density of rain gauge stations in relation to the country's territorial extension is recognized, which compromises the spatial representation of high-resolution dependent variables, such as dependentessivity. Under these conditions, the spatial interpolation of erosivity becomes limited, and traditional methods —such as Thiessen, isohyets, weighted averages, or geostatistical tools— may incorporate uncertainties when applied in areas with few available rain gauge stations.

Considering that there is not a sufficient number of rain gauge stations in the basin or its surroundings to calculate erosivity with a spatial distribution (Arantes et al., 2024b; Silva et al., 2017), both indicating low spatial variability of erosivity with low amplitude of values, it was decided to disregard the R factor in the models, assuming it as constant, since it would have little influence on the model proposed in this study.

Xiao et al. (2021) and Ghosal and Das Bhattacharya (2020) also highlighted the importance of the LS parameter. They obtained results indicating that the volume and rate of cumulative runoff increase as the slope length becomes steeper. Furthermore, both studies pointed out that increasing the slope also increases the runoff speed, contributing to soil erosion.

Kashiwar et al. (2022) and Kulimushi et al. (2021) also pointed out that the LS factor and soil erosion have a direct relationship. In other words, runoff erosion tends to be high when the slope's length and steepness are high, and vice versa. Finally, they concluded that the LS factor in a river basin impacts soil loss.

NDVI calculation, sampling and data integration

We calculated the NDVI shown by Rouse et al. (1973) using the Qgis software, version 3.30.1(Qgis, 2023). Thus, we used bands four (B4) and three (B3) of the CBERS satellite from a scene dated August 29, 2020. We chose the satellite and its date to maintain the spatial resolution standard of the study conducted by Arantes (2024b), which we used as the basis for calculating the USLE factors here.

We calibrated the model by proportionally sampling the pixels according to each class in Table 1. Therefore, we randomly predicted 3348 samples at the centroid of each pixel: a) 1036 samples for the low class; b) 1275 samples for the moderate class; and c) 1037 samples for the high class.

Then, we reclassified the soil loss classes from text to numerical format: low = 1; moderate = 2; and high = 3. This step was necessary due to the characteristics of the class format required by RF. In order to do so, we used the r.reclass tool in the Qgis software.

We merged the matrix files containing the USLE factors and the NDVI in the RStudio software. This procedure was necessary to obtain a data matrix in which each column comprised a predictive variable (USLE and NDVI factors) and a dependent variable (soil loss classes).

It is worth noting that when stacking matrix files, the data values are not modified, as the program's native 'stack' function transforms a list of data into a single column, which is then used in mathematical models. Finally, for the model's training (calibration), we decided to use the standard 70/30 ratio. In other words, 70% of the samples are for training and 30% for testing.

Model training, evaluation and exporting

In order to program the model, two parameters had to be set: a) the number of trees (ntree) and b) the number of features in each division (ntry). Studies claim that satisfactory results can be achieved with the standardized values of the (Zhang and Roy, 2017; Noi and Kapas, 2018) algorithm.

Therefore, Breiman (2001) ratifies that using more trees than required may be unnecessary, not detrimental to the model, but rather to computational performance. Thus, the author recommends an ntree value of up to 100. This study included different tests using the Out Of Bag (OOB) error to assess the best number of trees for the model since this parameter uses the ratio between the wrong classifications and the total number of elements sampled, resulting in a generalization estimate (OOB) (Breiman, 2001).

It is worth noting that the variation for fewer or more trees and nodes and the variation in training and test samples did not show significant differences. However, we decided to work with 50 ntress and 4 ntrys because, given the variability of river basins' chemical and biophysical parameters, future studies can use this quantity and obtain satisfactory results.

Figure 2 shows an example of training and validation behavior based on the classes (Table 1) used here.

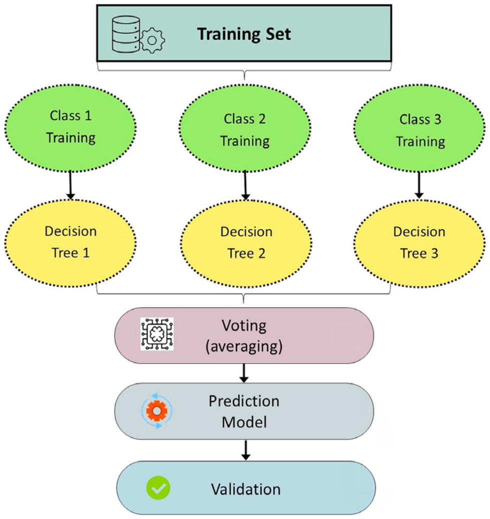

Figure 2. Application of RF in USLE prediction. Source: The Authors (2024).

Moreover, we evaluated the relative importance of the different input factors in the classification. This step is important for verifying how each predictive variable influences the generated classification model (Ghimire et al., 2010). Thus, we used the Gini criterion (Equation 1), obtained from the difference between the Gini$_{\text{index}}$ before and after splitting the node (Equation 2) (Filho, 2014).

$$\text{Gini}_{\text{index}}\ (\text{node}) = 1 - \sum_{i=1}^{c} p\ \frac{i}{\text{nó}} \tag{1}$$

Where: p(i/node) is the proportion of class i in the node.

$$\mathrm{Gini} = \mathrm{Gini}_{\mathrm{index}}(\mathrm{father}) - \sum_{j=1}^{n}\left[\frac{N\left(v_j\right)}{N}\,\mathrm{Gini}_{\mathrm{index}}\left(v_j\right)\right] \tag{2}$$

Where: n refers to the number of child nodes, N is the total number of observations of the parent node, and N(v_j) is the number of observations associated with the child node v_j.

In order to check the algorithm's performance, we cross-checked the information calculated from the USLE with that predicted by the classifier. We used the global accuracy metric (Accuracy), which is mathematically defined as the proportion of the number of predictions made correctly by the algorithm for the total data set (Equation 3), and the kappa index, calculated from the (Weiss and Zhang, 2003) confusion matrix.

It is worth noting that all the methodological and statistical procedures used here adopted a 95% confidence level. Regarding the performance measurement of the models generated, we analyzed Sensitivity, Specificity, Pos. Pred. Value (PPV), and Neg. Pred. Value (NPV) of each class.

In this case, Sensitivity (Equation 4) is explained by the percentage of (positive) hits predicted in the real cases in each soil loss class. Meanwhile, Specificity (Equation 5) represents the percentage of that same real case being predicted as negative in the other classes (Glaros and Kline 1988; Van Stralen et al. 2009). On the other hand, PPV (Equation 6) indicates the percentage of cases correctly identified as true. Meanwhile, NPV (Equation 7) indicates the number of cases correctly identified as negative.

We must emphasize that we used the Balanced Accuracy metric (Equation 8) in addition to Accuracy because, in the presence of unbalanced data, the Accuracy metric may not provide adequate information regarding the classifier's performance and could be misleading.

Thus, if the classifier performs equally well in the three soil loss classes analyzed, this metric will be similar to conventional Accuracy. However, suppose this measure is high simply because the classifier does a good job of classifying the majority class and does not do a good job of predicting the minority class in an unbalanced data set. In that case, Balanced Accuracy will reflect it with a low result.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{4}$$

$$Specificity = \frac{TN}{TN + FP} \tag{5}$$

$$PPV = \frac{TP}{TP + FP} \tag{6}$$

$$NPV = \frac{TN}{TN + FN} \tag{7}$$

$$Balanced\ accuracy = \frac{Sensitivity\ +\ Specificity}{2} \tag{8}$$

Where: TP = True Positives; TN = True Negatives; FP = False Positives; FN = False Negatives.

Finally, we exported the generated classifications, containing the classification models using NDVI, LS, and CP and NDVI, LS, and K so that the layout could include the standard USLE and the generated model. We carried out this procedure using the Qgis software.

3. Results

When using the LS and CP factors as USLE predictive variables, coupled with the NDVI index, the classifier achieved an accuracy of 93.43% and a kappa index of 0.9009. The OOB error estimate was 5.12%, and Table 2 shows the confusion matrix for this evaluation.

Table 3 shows the statistical characteristics of each class after model classification. Figure 3 shows the USLE soil loss classification and the model generated using the analyzed factors.

Table 2. Confusion matrix of the training and the soil loss map predicted by the RF, using the LS, CP and NDVI.
Source: The Authors (2024).

Training				
Class	**Low**	**Moderate**	**High**	**Error**
Low	699	26	0	0.03586207
Moderate	20	835	37	0.06390135
High	0	37	689	0.05096419
OOB Error Rate				5.12%
Prediction				
Class	**Low**	**Moderate**	**High**	**Error**
Low	297	9	0	0.02941176
Moderate	14	355	24	0.12213740
High	0	19	287	0.06209150
Average				0.07121355

Table 3. Statistics of the classes.
Source: The Authors (2024).

	Class		
Statistics	**Low**	**Moderate**	**High**
Sensitivity	0.9550	0.9269	0.9228
Specificity	0.9870	0.9389	0.9726
Pos Pred Value	0.9706	0.9033	0.9379
Neg Pred Value	0.9800	0.9542	0.9657
Balanced Accuracy	0.9710	0.9329	0.9477

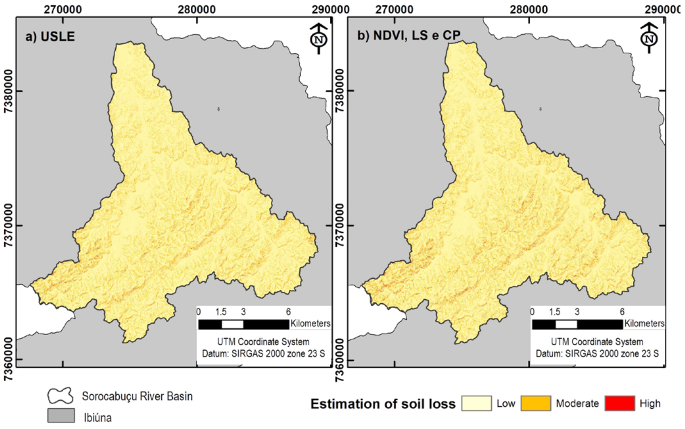

Figure 3. Soil loss map – USLE – (a) and the model generated by the RF using the LS and CP factors and the NDVI index (b).
Source: The Authors (2024).

When using the LS and K factors as USLE predictive variables, coupled with the NDVI index, the classifier achieved an accuracy of 97.14% and a kappa index of 0.9015. The OOB error estimate was 6.79%, and Table 4 shows the confusion matrix for this evaluation.

Table 5 shows the statistical characteristics of each class after model classification. Figure 4 shows the USLE soil loss classification and the model generated using the analyzed factors.

Table 4. Confusion matrix of the training and the soil loss map predicted by the RF, using LS, K and NDVI. Source: The Authors (2024).

Training				
Class	**Low**	**Moderate**	**High**	**Error**
Low	657	66	2	0.09379310
Moderate	24	841	27	0.05717489
High	0	40	686	0.05509642
OOB Error Rate				6.79%
Prediction				
Class	**Low**	**Moderate**	**High**	**Error**
Low	281	8	0	0.02768166
Moderate	30	361	10	0.09975062
High	0	14	301	0.04651162
Average				0.0579813

Table 5 – Statistics of the classes. Source: The Authors (2024).

	Class		
Statistics	**Low**	**Statistics**	**Low**
Sensitivity	0.9035	0.9426	0.9678
Specificity	0.9885	0.9357	0.9798
Pos Pred Value	0.9723	0.9002	0.9556
Neg Pred Value	0.9581	0.9636	0.9855
Balanced Accuracy	0.9460	0.9391	0.9738

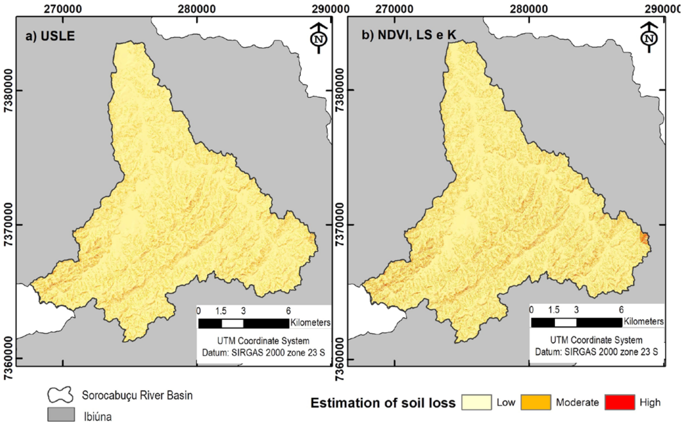

Figure 4. Soil loss map – USLE – (a) and the model generated by the RF using the LS and K factors and the NDVI index (b). Source: The Authors (2024).

The LS factor was the most important parameter for the generated model in both cases analyzed. It showed a value of 1200 Mean Decrease Gini. Meanwhile, the CP and K factors showed an average value of 200, and the NDVI showed a value of 190.

Discussion

Based on the performance metrics presented by each generated model, the RF algorithm generally performed satisfactorily in predicting the USLE.

The significance of the LS factor for generating the model in both cases indicates that the SRB strongly depends on the slope and ramp length regarding the drag of soil particles, as shown by Arantes (2024b). The author(s) found that for the same area, the highest values of this parameter occurred predominantly in the southern portion of the river basin, which is characterized by having the highest slope values. Comparable results were also reported by Santos et al. (2025), who applied the RF algorithm in the same basin (SRB) to classify soil texture and similarly identified a strong influence of topographic characteristics on predictive patterns.

Silva et al. (2017) found similar results, in which the topographic condition was the most important factor in calculating soil loss in a river basin close to the area analyzed here and was used as a key variable for extrapolating soil loss values.

These results point to the need for proper soil management in the study area. This type of situation, related to water erosion processes, reduces the productivity of the land since it removes the fertile soil that contains most of the essential nutrients for plants and microorganisms. Furthermore, soil depletion can occur rapidly over a short period of months if use practices that do not respect the characteristics of its constitution are employed. As a resu lt, there is a current or potential loss of productivity and a severe environmental and socio-economic problem (Pacheco et al., 2018; Yang et al., 2020). A recent socioeconomic-scale study in the SRB itself (Arantes et al., 2024a) also points to inequalities and vulnerabilities that worsen in areas subject to environmental degradation, reinforcing the need for more efficient predictive tools.

It is already known that the intensification of erosion, dragging, and transportation of soil particles from higher areas to lower areas degrades the soil (Bertoni and Lombardi Neto, 1999; Lepsch, 2010) and water resources. In this context, the results point to a significant susceptibility to sedimentation and deterioration in the quality of water bodies since the erosion of soil sheets and furrows, coupled with the dragging of particles, also contributes to it.

Furthermore, it is advantageous for the key variables in soil loss models to be morphometric since they can be easily calculated from a Digital Elevation Model (DEM), for example, or from simple spatial analysis techniques. Both techniques can be analyzed remotely, obtaining quick and accurate results within a short period without the need to make decisions using time-consuming techniques.

On the other hand, although the R factor is a fundamental component of the USLE, its reliable spatial estimation requires a network of meteorological stations, historical series, and data capable of adequately representing the spatial distribution of rainfall. In the case of the study area, the availability and distribution of rain gauge stations are not sufficient to generate an erosivity surface with adequate precision.

In addition, studies carried out in areas with similar characteristics (Silva et al., 2017; Arantes, 2024b) indicate that the spatial variability of erosivity is very low at these scales, resulting in practically constant values. Thus, considering both the limitations of rainfall data and the reduced expected variability of the R factor, it was decided not to include it as a predictor variable in the modeling, in order to avoid introducing additional uncertainties into the model.

In addition, while using conventional mathematical models involves a complex approach and requires high computational power, the models generated by ML can achieve satisfactory results based on parameters that are easy to obtain, as shown by the statistical analysis (Table 3 and Table 5).

From this perspective, it is worth noting that soil erosion rates are higher than the production of this natural resource, and this situation is driven by predatory agricultural practices (Amundson et al., 2015). In South America, this loss can reach approximately 10 tons ha^{-1}.year^{-1}, while in Brazil losses can reach 4.9 10 tons ha^{-1}.year^{-1} (Yang et al., 2003).

Therefore, given the increase in agroforestry areas with inadequate management and their respective environmental problems, there is a growing need to estimate the spatial distribution of areas with erosion potential in a faster and simpler way. However, as Cheng et al. (2018) pointed out, there is a lack of appropriate approaches to this type of issue. However, we believe that the methodology presented here can fill this gap.

Regarding the training and modeling statistics, the most significant errors occurred between the moderate and high soil loss classes. However, the error values were insignificant compared to the performance of the accuracy, kappa index, and metrics for each class, which indicates the relevance of the results presented. Furthermore, when analyzed together, the performance metrics for each soil loss class indicate an excellent performance of the models generated.

Regarding Sensitivity and Specificity, we found values above 90%, with emphasis on the low class of the model containing NDVI, LS, and CP (95% and 98%, respectively) and the high class of the model generated by NDVI, LS, and K (96% and 97%, respectively).

Regarding the PPV and NPV of the predicted values, the two generated models showed excellent performance for the low class. It is worth noting that this class was the most representative in the study area, indicating the algorithm's prediction efficiency. We can also point out that the high class, an important parameter to be analyzed in soil loss, also achieved excellent performance in these metrics.

Sensitivity indicates how the model works. Therefore, values of this parameter combined with Specificity close to one corroborate and increase the results' reliability according to correlated studies (Tarek et al., 2023; Kulimushi et al., 2023). Thus, we can infer that the generated models are around 90% efficient.

Given the results presented by the Balanced Accuracy of each class, the high values showed excellent performance regarding the generated models, indicating that this type of prediction was able to strongly analyze whether the soil loss class would be low, moderate, or high.

It is worth highlighting the performance of the low soil loss class based on the model generated using LS, CP, and NDVI and of the high class when the model comprised the LS, K, and NDVI factors. It should also be noted that the moderate class showed constant values in both training sessions.

Finally, it is also worth noting that the results presented here proved to be excellent compared to conventional models and with similar results to other Artificial Intelligence (AI) techniques, such as the use of Artificial Neural Networks (ANN), according to He et al. (2024). The author(s) concluded that the most critical advantage of using ANNs is that they can estimate soil erosion with high precision and speed. However, they are limiting regarding the data's accuracy because they require assigning weights to each parameter.

Given this limitation and the results presented here, we believe that the RF algorithm provides more efficient results for such modeling. However, while the two generated models help estimate soil loss, they are not designed to estimate gully erosion, which involves hydrological complexes and is a more advanced level of soil erosion. Thus, it is a limitation of the technique presented. Even so, the results presented can be used to plan and implement soil conservation measures. They help identify areas prone to erosion in locations with similar socio-environmental and economic characteristics and environmental parameters.

If conservation measures are applied in the areas of critical soil loss identified by the model, they will directly contribute to the United Nations Sustainable Development Goals (SDGs), especially SDG 2 (Zero Hunger and Sustainable Agriculture) and SDG 15 (Life on Land). Monitoring and mitigating erosion promote more fertile soils, increase agricultural productivity sustainably, and protect terrestrial ecosystems, reducing the risks of environmental degradation and associated economic losses (Simonetti et al., 2022; Rizzo et al., 2024; Nery et al., 2024). Furthermore, by preserving soil quality and water resources, the model helps reduce socioeconomic inequalities in rural communities, strengthening environmental and social resilience in the face of intensive agricultural practices (Nery et al., 2025).

Finally, it is worth mentioning that degraded areas need recovery projects, mainly to restore organic matter in the soil, reducing compaction, erosion potential, and increasing organic carbon levels, in addition to requiring inter-institutional action in conjunction with rural producers to promote more sustainable agricultural systems (Nery et al., 2023). Moreover, ensuring that agricultural systems are resilient, act in a conservationist manner, and ensure the conservation, recovery, and sustainable use of terrestrial ecosystems is an integral part of these SDGs (UN, 2023).

5. Conclusion

The RF classifier effectively predicted the soil loss classes calculated by the USLE, as shown by the excellent Accuracy and kappa values. Thus, we believe that by using this soil loss model with field-measured values as a reference, we can expect agreement that is at least equivalent to the estimates made by the USLE.

The predominance of the LS factor in the model indicates that the river basin strongly depends on slope and ramp length. This parameter should always be considered in small and medium-sized river basins. However, given its homogeneity, the R factor can be considered constant in these cases.

Analyzed together, the performance metrics also showed excellent values, which is directly reflected in the overall values of the generated models (Accuracy and kappa). Thus, no model or class performed much better than the others. Therefore, when considering the Balanced Accuracy values for each class, we understand that depending on the study area, the LS, CP, NDVI and LS, CP, and K parameters can be used to estimate local soil loss.

Finally, future studies evaluating the accuracy of machine learning approaches regarding data measured in the field will be important to consolidate the effectiveness of these procedures in estimating soil loss. However, the generated model can be a decision-making aid for agricultural and environmental managers and public authorities. Management and inspection agencies can use it to identify areas prone to erosion and apply the necessary conservation measures.

Furthermore, the results confirm the potential of ML techniques as tools to support environmental management, allowing for more precise identification of critical areas susceptible to erosion. This information strengthens territorial planning by supporting conservation practices, agricultural management strategies, and the definition of priority areas for recovery. Such applications broaden the model's usefulness for public policies aimed at soil and water resource conservation, as well as opening opportunities for future studies that integrate new data and improve the prediction of erosive processes.

References

Arantes, L. T., Santos, A. P., Silva, C. V., Nery, L. M., Toledo, M. V. L., Simonetti, V. C., Silva, D. C. C., & Lourenço, R. W. (2024a). Socioeconomic spatial analysis through fuzzy system as a tool for territorial planning applied to watersheds. *International Journal of River Basin Management,* 1–17. https://doi.org/10.1080/15715124.2024.2387579

Arantes, L. T., Santos, A. P., Silva, D. C. C., & Lourenço, R. W. (2024b). Indicador de vulnerabilidade ao carreamento de sedimentos integrado ao SIG e SR. *Geo UERJ,* (45). https://doi.org/10.12957/geouerj.2024.74164.

Amundson, R., Berhe, A. A., Hopmans, J. W., Olson, C., Sztein, A. E., & Sparks, D. L. (2015). Soil and human security in the 21st century. *Science*, 348(6235), 1261071. https://doi.org/10.1126/science.1261071

Andreoti, C. E. (2012). *Avaliação da eficiência de um sistema agroflorestal na recuperação de um solo degradado por pastoreio*. Dissertação (mestrado – Programa de Pós-graduação em Geografia Física) – Faculdade de

Filosofia, Letras e Ciências Humanas – São Paulo, Brasil. https://doi.org/10.11606/D.8.2012.tde-09012013-121619

Bertoni, J., & Lombardi Neto, F. (1999). *Conservação do solo*. 4. ed. São Paulo, SP: Ícone.

Boardman, J., & Poesen, J. (2006). *Soil erosion in Europe*. Ed. John Wiley & Sons Ltd. West Sussex. 855 p. https://doi.org/10.1002/0470859202.ch36

Breiman, L. (2001). Random Forests. *Journal Machine Learning,* 45, 5-32. https://doi.org/10.1023/A:1010933404324

Cheng, Z., Lu, D., Li, G., Huang, J., Sinha, N., Zhi, J., & Li, S. (2018). A Random Forest-Based Approach to Map Soil Erosion Risk Distribution in Hickory Plantations in Western Zhejiang Province, China. *Remote Sensing,* 10, 1-20. https://doi.org/10.3390/rs10121899

Costa, R. V. F., Leite, M. G. P., Leao, L. P., Nalini Junior, H. A., Silva, D. C. C., & Valente, T. M. F. (2025). Hydrogeochemistry of surface waters in the Iron Quadrangle, Brazil: High-Resolution Mapping of Potentially Toxic Elements in the Velhas and Paraopeba River Basins. *Water*, 17, 2446. https://doi.org/10.3390/w17162446

Dubreuil, V., Fante, K. P., Planchon, O., & Sant'Anna Neto, J. L. (2017). Les types de climats annuels au Brésil: une application de la classification de Köppen de 1961 a 2015. *EchoGéo,* 41, 1-27. https://doi.org/10.4000/echogeo.15017

Empresa Brasileira de Pesquisa Agropecuária [EMBRAPA]. (2018). *Solo e relevo: Influência no uso da terra*. https://www.embrapa.br/agencia-de-informacao-tecnologica/cultivos/eucalipto/pre-producao/escolha-da-area/solo-e-relevo Access: 02 nov. 2025.

FAO – Food and Agriculture Organization of the United Nations. (2023). *The State of Food and Agriculture 2023. Revealing the true cost of food to transform agrifood systems*. From: https://www.fao.org/documents/card/en/c/cc7724en. Access: 09 ago. 2023.

FAO – Food and Agriculture Organization of the United Nations. (2021). The state of the world land and water resources for food and agriculture 2021. Disponível em: https://www.fao.org/3/cb7654en/online/cb7654en.html. Access: 09 ago. 2023.

Filho, J. P. (2014). *Capacidade preditiva de Modelos Credit Scoring em inferência dos rejeitados*. Dissertação (Mestrado em Estatística) – Centro de Ciências Exatas e de Tecnologia, Universidade federal de São Carlos, São Carlos, 95p.

Ganasri, B. P., & Ramesh, H. (2016). Assessment of soil erosion by RUSLE model using remote sensing and GIS-A case study of Nethravathi Basin. *Geoscience Frontiers*, 7(6), 953-961. https://doi.org/10.1016/j.gsf.2015.10.007

Ghimire, B., Rogan, J., & Miller, J. (2010). Contextual land-cover classification: incorporating spatial dependence in land-cover classification models using random forests and the Getis statistic. *Remote Sensing Letters,* 1(1), 45-54. https://doi.org/10.1080/01431160903252327

Ghosal, K., & Das Bhattacharya, S. (2020). A review of RUSLE model. *Journal of the Indian Society of Remote Sensing*, 48, 689-707. https://doi.org/10.1007/s12524-019-01097-0

Glaros, A. G., & Kline, R. B. (1998). Understanding the accuracy of tests with cutting scores: The sensitivity, specificity, and predictive value model. *Journal of clinical psychology,* 44(6), 1013-1023. https://doi.org/10.1002/1097-4679(198811)44:6%3C1013::aid jclp2270440627%3E3.0.co;2-z

He, Q., Zhao, H., Feng, Y., Wang, Z., Ning, Z., & Luo, T. (2024). Edge computing-oriented smart agricultural supply chain mechanism with auction and fuzzy neural networks. *Journal of Cloud Computing: Advances, Systems and Applications,* 13(1). https://doi.org/10.1186/s13677-024-00626-8

Helmi, A. M. (2023). Quantifying catchments sediment release in arid regions using GIS-based Universal soil loss equation (USLE). *Ain Shams Engineering Journal,* 14(8), 102038. https://doi.org/10.1016/j.asej.2022.102038

IBGE – Instituto Brasileiro de Geografia e Estatística. (2021). *Banco de Dados de Informações Ambientais.* From: https://bdiaweb.ibge.gov.br/. Acess 09 jun. 2023.

Kashiwar, S. R., Kundu, M. C., & Dongarwar, U. R. (2022). Soil erosion estimation of Bhandara region of Maharashtra, India, by integrated use of RUSLE, remote sensing, and GIS. *Natural Hazards,* 110(2), 937–959. https://doi.org/10.1007/s11069-021-04974-5

Köpen, W. (1948). *Climatologia*. Buenos Aires: Gráfica Panamericana. 478p.

Kulimushi, L. C., Choudhari, P., Mubalama, L. K., & Banswe, G. T. (2021). GIS and remote sensing-based assessment of soil erosion risk using RUSLE model in South-Kivu province, eastern, Democratic Republic of Congo. *Geomatics Natural Hazards and Risk,* 12(1), 961–987. https://doi.org/10.1080/19475705.2021.1906759

Kulimushi, L. C., Bashagaluke, J. B., Prasad, P., Heri-Kazi, A. B., Kushwaha, N. L., Masroor, M., Choudhari, P., Elbeltagi, A., Sajjad, H. & Mohammed, S. (2023). Soil erosion susceptibility mapping using ensemble machine learning models: A case study of upper Congo river sub-basin. *Catena*, 222, 106858. https://doi.org/10.1016/j.catena.2022.106858

Lepsch, I. F. (2010). *Formação e conservação dos solos*. 2. ed. São Paulo, SP: Oficina de Textos.

Maia Júnior, L. P. M. & Lourenço, R. W. (2020). Impactos das mudanças no uso e cobertura da terra sobre a variabilidade do albedo na Bacia Hidrográfica do Rio Sorocabuçu (Ibiúna-SP). *Revista Brasileira de Climatologia,* 27, 443-462. https://doi.org/10.5380/abclima.v27i0.72761

Nery, L. M., Sabonaro, D. Z. & Silva, D. C. C. (2023). A multicriteria analysis for decision making. *Environment, Development and Sustainability,* 25, 1-19. https://doi.org/10.1007/s10668-023-03261-6

Nery, L. M., Gomes, G., Nicomedes, N. P., Sabonaro, D. Z., & Silva, D. C. C. (2024). Análise socioambiental de sistemas de integração: quais seus benefícios, desafios e oportunidades? RISUS. *Journal on Innovation and Sustainability*, 15, 177–192. https://doi.org/10.23925/2179-3565.2024v15i2p177-192

Nery, L. M., Toniolo, B. P., Santos, A. P., Martins, A. C. G. & Silva, D. C. C (2025). Challenge of political integration in the territorial management of a protected area based on the analysis of land use and land cover change. *Journal of Environmental Studies and Sciences, 15*, 845–860. https://doi.org/10.1007/s13412-024-00990-6

Nguyen, K. A., Chen, W., Lin, B. S., Seeboonruang, U. & Thomas, K. (2019). Predicting Sheet and Rill Erosion of Shihmen Reservoir Watershed in Taiwan Using Machine Learning. *Sustainability*, 11, 1-18. https://doi.org/10.3390/su11133615

Noi, P. T. & Kappas, M. (2017). Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 imagery. *Sensors*, 18(1), 18. https://doi.org/10.3390/s18010018

Pacheco, F. A. L., Fernandes, L. F. S., Valle Júnior, R. F., Valera, C. A. & Pissarra, T. C. T. (2018). Land degradation: Multiple environmental consequences and routes to neutrality. *Current Opinion in Environmental Science & Health,* 5, 78-86. https://doi.org/10.1016/j.coesh.2018.07.002

Panagos, P., Borrelli, P., Poesen, J., Ballabio, C., Lugato, E., Meusburger, K., Montanarella, L. & Alewell, C. (2015). The new assessment of soil loss by water erosion in Europe. *Environmental Science & Policy,* (54), 438-447. https://doi.org/10.1016/j.envsci.2015.08.012

Pandey, A., Chowdary, V. M. & Mal, B. C. (2007). Identification of critical erosion prone areas in the small agricultural watershed using USLE, GIS and remote sensing. *Water Resources Management,* 21(4), 729-746. https://doi.org/10.1007/s11269-006-9061-z

Paula, A. L., Pereira dos Santos, A., Belfort Poletti, F., & Lourenço, R. W. (2025). Adjustment of the conservation practices factor calculation in estimating soil loss. *Ra'e Ga: O Espaço Geográfico em Análise, 63*(1), 125–151. https://doi.org/10.5380/raega.v63i1.100335

QGIS. (2023). *QGIS Geographic Information System*. QGIS Association. From: http://www.qgis.org. Acess: 09 jun. 2023.

Rizzo, F. A., Santos, A. P., & Silva, D. C. C. (2024). Técnicas de geoprocessamento aplicadas para análise temporal do microclima na bacia hidrográfica do córrego do Pequiá, Maranhão. *Boletim Goiano de Geografia,* 44, e78032. https://doi.org/10.5216/bgg.v44i1.78032

Rossi, M. (2017). *Mapa pedológico do Estado de São Paulo: Revisado e ampliado.* Instituto Florestal. https://www.infraestruturameioambiente.sp.gov.br/institutoflorestal.

Rouse, J. J. R., Haas, R. H., Schell, J. A. & Deering, D. W. (1973). *Monitoring the vernal advancement and retrogradation (green wave effect) of natural vegetation*. Remote Sensing Center Texas A&M University College Station, Texas. 93p. From: https://core.ac.uk/download/pdf/42887948.pdf. Acess: 09 jun. 2023.

RStudio Team (2023). *RStudio: Integrated Development Environment for R.* RStudio, PBC, Boston. From: http://www.rstudio.com/. 09 jun. 2023.

Santos, A. P., Silva Junior, A. X., Nery, L. M., Gomes, G., Toniolo, B. P., da Cunha e Silva, D. C., & Lourenço, R. W. (2025). Random forest algorithm applied to model soil textural classification in a river basin. *Environmental Monitoring and Assessment*, 197, 330. https://doi.org/10.1007/s10661-025-13786-0.

Sheikh, A. H., Palria, S. & Alam, A. (2011). Integration of GIS and Universal Soil Loss Equation (USLE) for soil loss estimation in a Himalayan watershed. *Recent Research in Science and Technology,* 3(3), p. 51-57. https://www.researchgate.net/publication/286921198_INTEGRATION_OF_GIS_AND_UNIVERSAL_SOIL_LOSS_EQUATION_USLE_FOR_SOIL_LOSS_ESTIMATION_IN_A_HIMALAYAN_WATERSHED

Silva, D. C. C., Albuquerque Filho, J. L., Sales, J. C. A. & Lourenço, R. W. (2017). Identificação de áreas com perda de solo acima do tolerável usando NDVI para o cálculo do fator C da USLE. *Ra'e Ga,* 42, 72-85. http://dx.doi.org/10.5380/raega.v42i0.45524

Simonetti, V. C., Silva, D. C. C., & Rosa, A. H. (2022). Correlação espacial compartimentada dos padrões de drenagem com características morfométricas da bacia hidrográfica do rio Pirajibu-Mirim. *Revista Brasileira de Geomorfologia*, 23, 1134–1154. https://doi.org/10.20502/rbg.v23i1.2037

Tanyas, H., Kolat Ç. & Süzen, M. L. (2015). A new approach to estimate cover-management factor of RUSLE and validation of RUSLE model in the watershed of Kartalkaya Dam. *Journal of Hydrology*, 528, 583-598. https://doi.org/10.1016/j.jhydrol.2015.06.048

Tarek, Z., Elshewey, A. M., Shohieb, S. M., Elhady, A. M., El-Attar, N. E., Elseouf, S., Shams, M. Y. (2023). Soil Erosion Status Prediction Using a Novel Random Forest Model Optimized by Random Search Method. *Sustainability,* 15(9), 7114. https://doi.org/10.3390/su15097114

Toniolo, B. P., Nery, L. M., & Silva, D. C. C. (2024). Modelagem espacial para identificação de áreas potenciais à geração de poluição difusa na Bacia Hidrográfica do Rio Cotia - SP. URBE. Revista Brasileira de Gestão Urbana, 16, e20220207. https://doi.org/10.1590/2175-3369.016.e20220207.

UN – United Nations. The 17 goals. (2023). From: https://sdgs.un.org/goals. Acess: 30 nov. 2023.

Van Stralen, K. J., Stel, V. S., Reitsma, J. B., Dekker, F. W., Zoccali, C. & Jager, K. J. (2009). Diagnostic methods I: sensitivity, specificity, and other measures of accuracy. *Kidney international,* 75(12), 1257-1263. https://doi.org/10.1038/ki.2009.92

Weiss, S. M. & Zhang, T. (2003). *Performance analysis and evaluation. In: The handbook of Data Mining*. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, 14, 425 – 440.

Wischmeier, W. H. & Smith, D. D. (1978). *Predicting rainfall erosion losses – A guide to conservation planning*. Washington, USDA, 1978. 58p. (USDA AH-537). file:///C:/Users/simio/Downloads/USLE.pdf

Xiao, Y. G. B., Lu, Y., Zhang, R., Zhang, D., Zhen, X., Chen, S., Wu, H., Wei, C., Yang, L. & Zhang, Y. (2021). Spatial–temporal evolution patterns of soil erosion in the Yellow River Basin from 1990 to 2015: impacts of natural factors and land use change. *Geomat Nat Hazard Risk*, 12(1), 103–122. https://doi.org/10.1080/19475705.2020.1861112

Yang, D., Kanae, S., Oki, T., Koike, T. & Musiake, K. (2003). Global potential soil erosion with reference to land use and climate changes. *Hydrological Processes*, 17, 2913-2918. https://doi.org/10.1002/hyp.1441

Yang, T., Siddique, K. H. M. & Liu, K. (2020). Cropping systems in agriculture and their impact on soil health - A review. *Global Ecology and Conservation,* 23, e01118. https://doi.org/10.1016/j.gecco.2020.e01118

Zhang, H. K. & Roy, D. P. (2017). Using the 500 m MODIS land cover product to derive a consistent continental scale 30 m Landsat land cover classification. *Remote Sensing of Environment,* 197, 15-34. https://doi.org/10.1016/j.rse.2017.05.024

Zhang, X., Song, J., Wang, Y., Deng, W. & Liu, Y. (2021). Effects of land use on slope runoff and soil loss in the Loess Plateau of China: A meta-analysis. *Science of The Total Environment*, 755(1), 142418. https://doi.org/10.1016/j.scitotenv.2020.142418