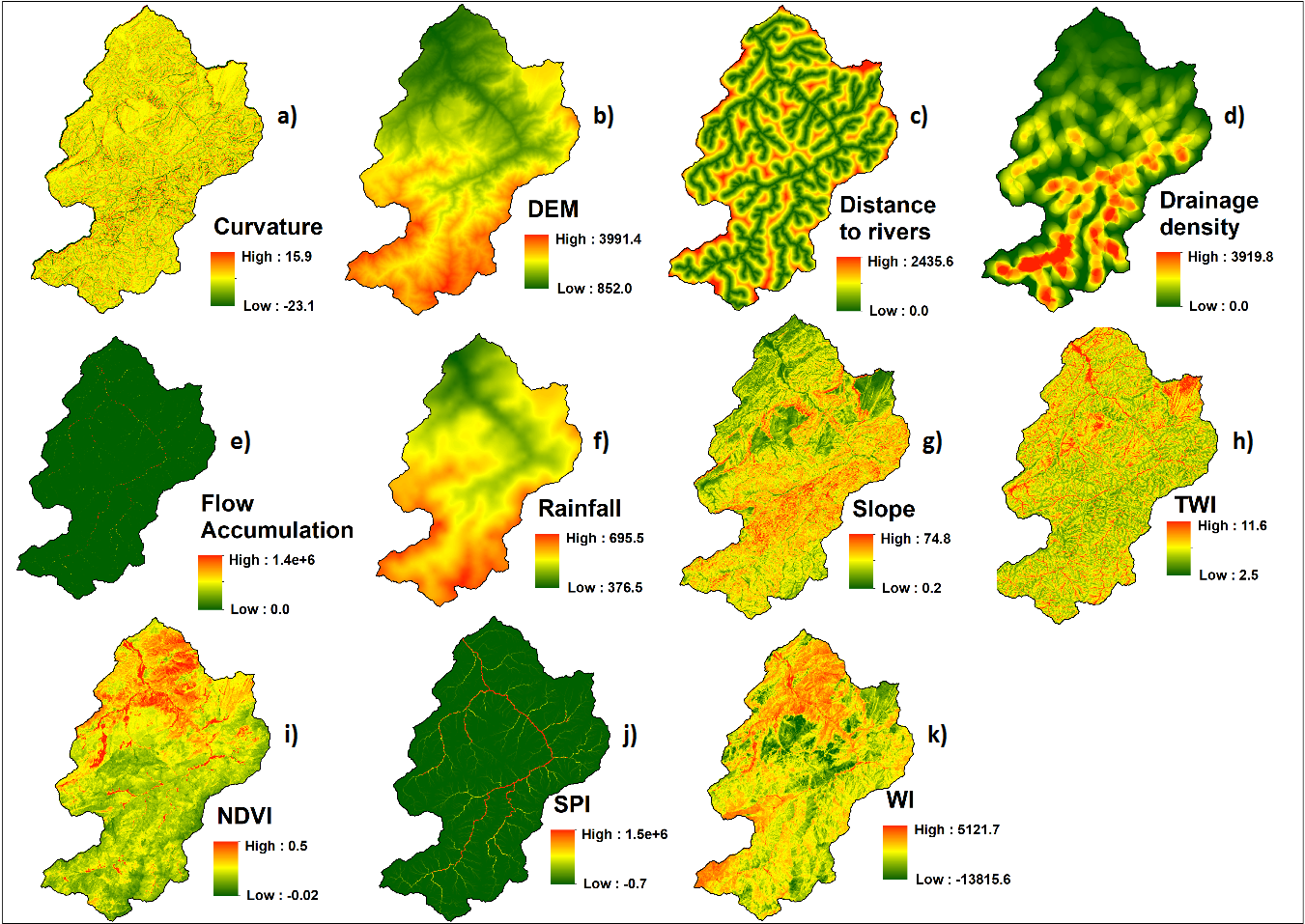
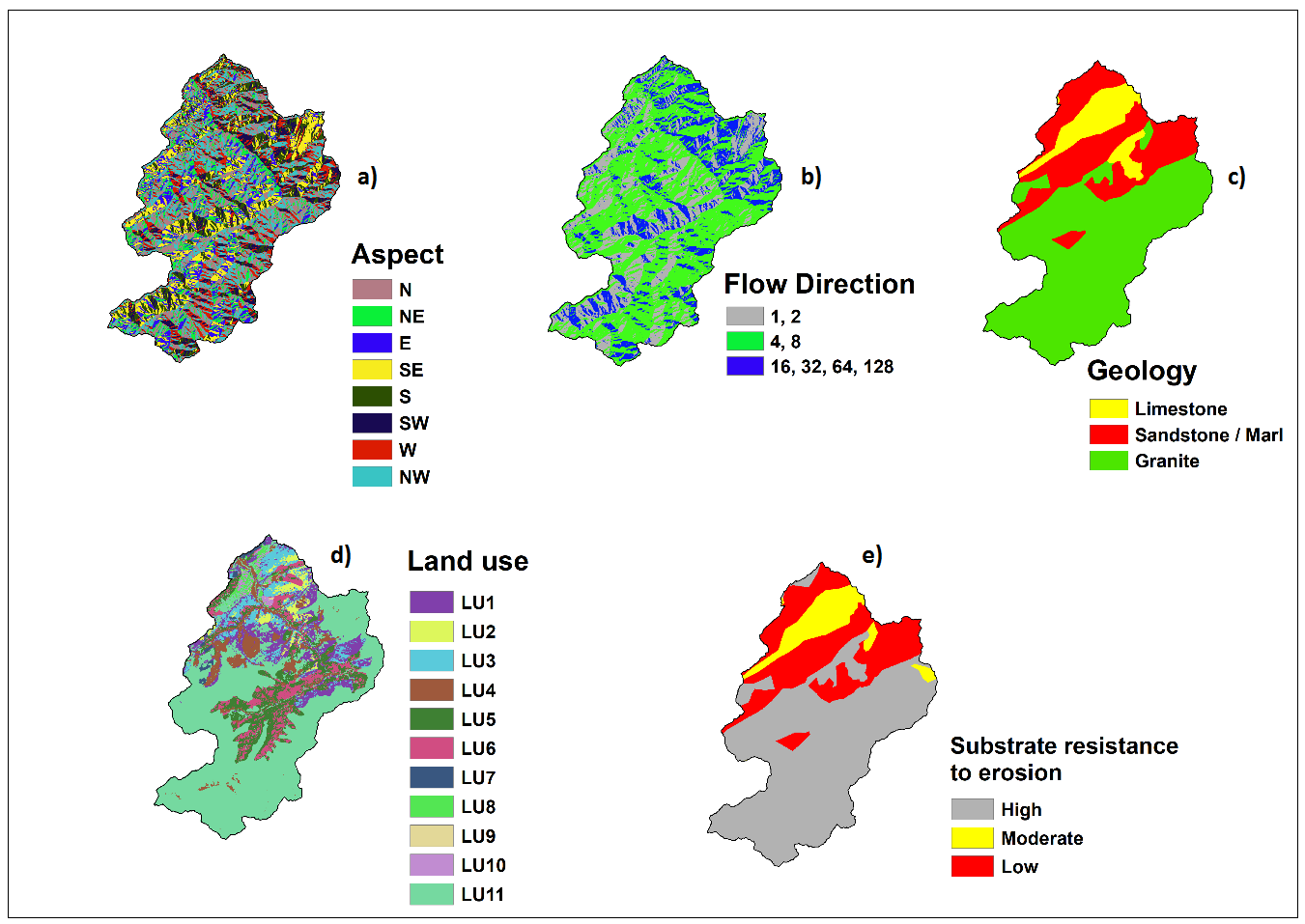
**Flood conditioning factors selected as inputs for modelling**

|  |  |
| --- | --- |
|  | **Description** |
| Numerical variables | **Curvature** is the rate of change of the slope gradient in a specific direction and is considered an important factor influencing flooding. A positive curvature value implies that the slope is convex upward, while a negative value indicates concavity upward. Zero represents a flat surface. Curvature in the study area varied between -23.1 and 15.9. **Elevation** is considered one of the most influential factors affecting flooding. Low-lying areas of a watershed have a higher flooding potential than higher elevation areas by virtue of water flowing from the higher elevation to the lower elevation. The DEM representation of the study area was obtained from the ASTER GDEM website. **Distance to the river** through its direct control of the flood event is a critical conditioning factor, helping to identify areas prone to flooding. The greater the distance, the lower the likelihood of flooding. **Drainage density** is the ratio of the total length of channels in the watershed to the total area of the watershed. It has a direct relationship with flooding, as a higher probability of flooding is directly related to a higher drainage density through its role in surface runoff. **Rainfall** is directly correlated with river flows, and heavy rainfall in a short period can result in flash floods, especially in semi-arid regions. For this study, rainfall data was obtained from the WorldClim website (<https://www.worldclim.org/>), with the mean annual rainfall in the study area ranging from 376.5 to 695.5mm. **TWI** describes the spatial distribution of attributes such as soil moisture, water table depth, and soil wetness. High values represent areas that favor water accumulation and runoff. As a result, the resulting saturated land can lead to flooding. **Vegetation** plays a defensive role against flooding, as it slows down runoff. Indeed, plant roots directly reduce substrate erodibility by increasing the stability of soil aggregates, which is further promoted by increased root density. To assess vegetation cover and, thus, its impact on flooding in the study area, **NDVI** was selected. It ranged between -0.02 and 0.5. **SPI** refers to the rate of flow with erosive power of flowing water in a watershed. It is inversely proportional to the frequency of flood occurrence. Thus, low SPI values indicate a higher probability of flood occurrence while high values indicate a lower probability of flooding. |
| Categorical variables | **Aspect** corresponds to the horizontal direction of a slope. It conditions hydrological processes, soil moisture regimes, local climatic conditions, etc. Aspect has been shown to indirectly affect channel flooding, therefore playing an important role in flood assessment. **Land use** plays a key role in influencing flooding. Thus, a thorough understanding of land use and land cover is critical to studies of natural hazards. Indeed, areas with vegetation are less prone to flooding due to their role in limiting runoff. This is particularly true for forested areas compared to areas dominated by agriculture. Conversely, urban areas are generally composed of impervious surfaces and lack vegetative cover, making them more vulnerable to flooding. **Geology** and **substrate** are critical factors in flood risk assessment, as permeable rocks allow water to pass through pores and cracks, while impermeable rocks do not. Thus, the latter contributes to increased flood risk due to increased surface runoff. The study area was dominated by granite rocks, while being highly resistant to erosion in most of the southern portion. |

**Predisposing factors used as numerical variables in floods susceptibility modeling**



**Predisposing factors used as categorical variables in floods susceptibility modeling**



*LU1 = Bare areas/built-up, LU2 = moderately dense Quercus rotundifolia stands, LU3 = Agriculture, LU4 = Open Tetraclinis articulata stands, LU5 = Open Quercus rotundifolia stands, LU6 = Open Juniperus (phoenicea, oxycedrus, thurifera) stands, LU67 = Reforestation, LU8 = Dense Tetraclinis articulata stands, LU9 = Dense Quercus rotundifolia stands, LU10 = moderately dense Juniperus (phoenicea, oxycedrus, thurifera) stands, LU11 = moderately dense Tetraclinis articulata stands,*

**Machine learning algorithms used for modelling flood susceptibility**

|  |  |
| --- | --- |
| **Name** | **Description** |
| Random Forest (RF) | RF is relatively intuitive, fast to train and produces generalizable results. Its only drawback is that it can be a bit of a black box, giving results that are not very interpretable. RF consists of a set of independent decision trees. Each tree has a fragmented view of the problem due to a double random draw. This consists of a random draw with replacement on the observations in a process called "tree bagging" and a random draw on the variables in a process referred to as "feature sampling". This algorithm provides a useful tool for classification and regression problems, and operates on multiple decision trees randomly trained on different subsets. |
| Extreme Gradient Boost (XGB) | XGB is based on aggregates trees where a new tree is trained according to the error of the preceding tree during each iteration. Thus, even if each tree has a low predictive power, the decision rule built by aggregating the result of each tree is reliable. Sparse and dense matrices are used as input, while the output can be an integer for classification problems and real number in the case of regression. A linear regression is defined by default in the prediction. |
| Artificial Neural Network (ANN) | A neural network is composed of several layers. The output layer is composed of an affine transformation and takes as input the learned representation of an intermediate layer called the hidden layer, which itself transforms an input layer. The input layer corresponds either to a normalization of the data or to an identity. Thus, the data can be considered as pre-normalized. This layered architecture gives the neural network a higher capacity than a model with a flatter architecture, for the same number of parameters. ANNs are unique in their ability to extract sophisticated patterns from ambiguously complex data thus making it possible to accomplish tasks that would otherwise have been impossible without assistance or with existing computational approaches. |
| K-Nearest Neighbors (KNN) | KNN is widely popular of due to its robustness and feasibility. It requires the choice of two parameters: (i) the distance measure and (ii) the number K of neighbors to be considered in the ranking. KNN uses the Euclidean, Manhattan and Minkowski distance functions to solve problems, with the Euclidean distance being the most commonly used. With the normalization of the training data, the accuracy of the KNN model can be greatly improved since the algorithm is distance-based for classification purposes. |

**Accuracy and kappa of models based on the test data**

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm | | Accuracy | Kappa |
| ANN | ANN | 0.986 | 0.972 |
| ANN.SE | 0.741 | 0.481 |
| ANN.TR | 0.991 | 0.981 |
| ANN.TR.SE | 0.893 | 0.786 |
| KNN | KNN | 0.927 | 0.854 |
| KNN.SE | 0.857 | 0.714 |
| KNN.TR | 0.991 | 0.981 |
| KNN.TR.SE | 0.893 | 0.786 |
| RF | RF | 0.969 | 0.938 |
| RF.SE | 0.846 | 0.693 |
| RF.TR | 0.964 | 0.929 |
| RF.TR.SE | 0.891 | 0.783 |
| XGB | XGB | 0.983 | 0.966 |
| XGB.SE | 0.974 | 0.947 |
| XGB.TR | 0.978 | 0.957 |
| XGB.TR.SE | 0.919 | 0.839 |

**Sensitivity and specificity of models based on the test data**

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm | | Sensitivity | Specificity |
| ANN | ANN  ANN.SE  ANN.TR  ANN.TR.SE | 0.972  0.960  0.981  0.941 | 1.000  0.522  1.000  0.845 |
| KNN | KNN  KNN.SE  KNN.TR  KNN.TR.SE | 0.898  0.966  0.907  0.941 | 0.957  0.748  0.950  0.879 |
| RF | RF  RF.SE  RF.TR  RF.TR.SE | 0.997  0.981  0.997  0.969 | 0.944  0.714  0.932  0.811 |
| XGB | XGB  XGB.SE  XGB.TR  XGB.TR.SE | 0.978  0.957  0.969  0.922 | 0.988  0.991  0.988  0.916 |

**Hyperparameters of the different prediction models and the optimal parameters revealed during parameter tuning**

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm | | Hyperparameter | Optimal value |
| ANN | ANN | size decay | size = 5 decay = 0.1 |
| ANN.SE | size = 5 decay = 0 |
| ANN.TR | size = 5 decay = 0.1 |
| ANN.TR.SE | size = 5 decay = 0 |
| KNN | KNN | k | 5 |
| KNN.SE |
| KNN.TR |
| KNN.TR.SE |
| RF | RF | mtry | 23 |
| RF.SE | 2 |
| RF.TR | 34 |
| RF.TR.SE | 2 |
| XGB | XGB | nrounds max\_depth eta  gamma colsample\_bytree  min\_child\_weight subsample | nrounds = 715 max\_depth = 5 eta = 0.165 gamma = 5.592 colsample\_bytree = 0.455 min\_child\_weight = 4 subsample = 0.419 |
| XGB.SE |
| XGB.TR |
| XGB.TR.SE | nrounds = 697 max\_depth = 9 eta = 0.574 gamma = 4.905 colsample\_bytree = 0.405 min\_child\_weight = 11 subsample = 0.791 |

**Built-up areas based on flood susceptibility class in the watershed**

|  |  |
| --- | --- |
| Flood Susceptibility Class | Built-up Area (%) |
| Very low | 81.07 |
| Low | 2.80 |
| Moderate | 0.45 |
| High | 0.76 |
| Very high | 14.91 |

**Models results similarities and dissimilarities**

|  |  |  |
| --- | --- | --- |
| **Models** | **Classes** | **Area (%)** |
| All | Dissimilar results across all models | 46.10 |
| Very low risk (Similar results across all models) | 53.79 |
| Very high risk (Similar results across all models) | 0.11 |
| KNN | Dissimilar models results | 43.64 |
| Very low risk (Similar models results) | 55.16 |
| Very high risk (Similar models results) | 1.20 |
| ANN | Dissimilar models results | 20.84 |
| Very low risk (Similar models results) | 78.73 |
| Very high risk (Similar models results) | 0.43 |
| RF | Dissimilar models results | 3.97 |
| Very low risk (Similar models results) | 95.79 |
| Very high risk (Similar models results) | 0.24 |
| XGB | Dissimilar models results | 4.00 |
| Very low risk (Similar models results) | 95.42 |
| Very high risk (Similar models results) | 0.58 |