https://doi.org/10.5194/nhess-2023-50
Preprint. Discussion started: 12 April 2023


Natural Hazards and Earth System Sciences
Discussions

# Multivariate regression trees as an 'explainable machine learning' approach to exploring relationships between hydroclimatic characteristics and agricultural and hydrological drought severity

Ana Paez-Trujilo[1,2,3], Jeffer Cañon[3], Beatriz Hernandez[3], Gerald Corzo[1], Dimitri Solomatine[1,2,4]

[1]IHE Delft Institute for Water Education, P.O. Box 3015, 2601 DA Delft, the Netherlands
[2]Delft University of Technology, Water Resources Section, P.O. Box 5048, 2600 GA Delft, the Netherlands
[3]Fundacion Natura Colombia, P.O. 111311, Carrera 21 No. 39–43 Bogotá D.C., Colombia
[4]RAS Water Problems Institute, 119333, Moscow, Russia

*Correspondence to*: Ana M. Paez-Trujilo[1] (A.M.PaezTrujillo@tudelft.nl)

**Abstract**

The typical causes of droughts are lower precipitation and/or higher than normal evaporation in a region. The region's characteristics and anthropogenic interventions may enhance or alleviate these events. Evaluating the multiple factors that influence droughts is complex and requires innovative approaches. To address this complexity, this study employs a combination of modelling and machine learning tools to assess the relationship between hydroclimatic characteristics and the severity of agricultural and hydrological droughts. The Soil Water Assessment Tool is used for hydrological modelling. Model outputs, soil moisture and streamflow are used to calculate the drought indicators for the subsequent drought analysis. Other simulated hydroclimatic parameters are treated as hydroclimatic drivers of droughts. A machine learning technique, the multivariate regression tree approach, is then applied to identify the hydroclimatic characteristics that govern agricultural and hydrological drought severity. The case study is the Cesar River basin (Colombia).

Our research indicates that multiple parameters influence the Cesar River basin's exposure to agricultural and hydrological droughts. Accordingly, the basin can be divided into three distinct areas. First is the upper part of the river valley. Due to precipitation shortfalls and high potential evapotranspiration, this region is very susceptible to agricultural and hydrological droughts. The second area is the middle part of the river valley. This area is likewise very susceptible to agricultural and hydrological droughts; however, severe drought conditions are brought on by inadequate rainfall partitioning and an unbalanced water cycle that favours water loss through percolation and evapotranspiration. Third, the Zapatosa marsh and the Serrania del Perijá foothills present moderate exposure to agricultural and hydrological droughts. Mild drought conditions appear to be related to the capacity of the subbasins to retain water, which lowers evapotranspiration losses and promotes percolation. Our results show that the presented methodology, in combining hydrologic modelling and machine learning

https://doi.org/10.5194/nhess-2023-50
Preprint. Discussion started: 12 April 2023


Natural Hazards and Earth System Sciences Discussions

techniques, provides valuable information about an interplay between the hydroclimatic factors that influence drought severity in the Cesar River basin.

## 1 Introduction

Projections indicate that drought frequency, severity and duration are expected to increase globally in the twenty-first century (UNDRR, 2021). Upcoming soil moisture drought scenarios predict statically significant, large-scale drying, especially in scenarios with strong radiative forcing in Central America and tropical South America (Lu et al., 2019). A similar trend is predicted for hydrological drought severity. This is expected to increase by the end of the twenty-first century, with regional hotspots in central and western Europe and South America, where the frequency of hydrological drought may increase by more than twenty percent (Prudhomme et al., 2014). The intensification of drought characteristics (in combination with other factors) could force the migration of up to 216 million people by 2050 (The World Bank, 2021), increase wildfire risk and tree mortality, and negatively affect regional air quality, among other ecosystem impacts (Vicente-Serrano et al., 2020).

It is essential that we better understand drought drivers if we are to foster preparedness and resilience to projected drought events. Remarkable progress has been achieved in understanding the drought-generating process. Drought occurs due to climatic extremes, which may be enhanced or alleviated by region characteristics and anthropogenic influence (Hao et al., 2022; Seneviratne et al., 2012; Tijdeman et al., 2018). Typically, droughts are triggered by atmospheric circulation and weather systems that combine to cause lower precipitation and/or higher than normal evaporation in a region (Destouni & Verrot, 2014; Sheffield & Wood, 2011a). Reduced precipitation leads to a decrease in soil moisture, causing agricultural drought. When soil moisture depletion is high, it is restored in the wet season, thus reducing subsurface flow and groundwater recharge and giving rise to hydrological drought (Iglesias et al., 2018). Regional characteristics such as soil type, stratigraphy, elevation, slope, vegetation cover, drainage networks, water bodies and groundwater systems play a relevant role in response to the climate anomalies that affect drought propagation and contribute to different levels of agricultural and hydrological drought (Sheffield & Wood, 2011a; Zhang et al., 2022). Equally important, human interventions in the hydrological cycle (e.g. reservoirs, water diversion, deforestation, over-pumping groundwater, overgrazing, urbanization) can reduce water supplies, triggering a drought situation or exacerbating a climate-driven drought (Rangecroft et al., 2019; Wang et al., 2021).

Drought drivers and propagating processes have been assessed following different approaches. Process-based models are suitable for explaining the physical mechanisms and anthropogenic activities that influence drought propagation and duration (Apurv & Cai, 2020; Li et al., 2018; Margariti et al., 2019; Mastrotheodoros et al., 2020; Shah et al., 2021; Shah & Mishra, 2020). Statistical approaches and machine learning (data-driven) models can be used as complementary tools to identify the dominant climate or basin characteristics that control drought propagation (Ganguli P. et al., 2022; Huang et al., 2017; Peña-Gallardo et al., 2019; Saft et al., 2016; van Loon, 2015; Xu et al., 2019). While several approaches are available, there is still

https://doi.org/10.5194/nhess-2023-50
Preprint. Discussion started: 12 April 2023


a need for studies that evaluate the combined effect of multiple hydroclimatic variables on drought characteristics (Valiya Veettil & Mishra, 2020).

We have found two studies that employ machine learning to analyse the non-linear relationship between climate and basin processes and identify the key hydroclimatic characteristics (and the critical thresholds) that govern agricultural and hydrological drought severity. Previous work has focused on the dominant variables that influence hydrological droughts in the same climatic region. Specifically, Konapala et al. (2020) used a random forest algorithm to identify the climate and basin characteristics influencing different hydrological drought regimes in the continental United States. Meanwhile, Valiya Veettil et al. (2020) used a classification tree algorithm to identify the variables and associated thresholds that control hydrological droughts in the Savannah River basin (USA). According to the results, basin properties significantly influence hydrological droughts, followed by climate and morphological variables. The mentioned research efforts show the potential of machine learning techniques as tools for drought-related analysis and demonstrate the potential for developing and testing other methodological approaches to enhance such analyses.

The multiple factors that influence droughts are in complex interplay. To understand the relationships between the various factors impacting the severity of agricultural and hydrological droughts at the basin scale, this study employs a methodology that consists of two steps. The first is hydrological modelling. Soil Water Assessment Tool (SWAT model) outputs, soil moisture and streamflow are used to calculate drought indicators. Other simulated hydroclimatic parameters are used as hydroclimatic drivers of droughts. Second, a machine learning technique, namely the Multivariate Regression Tree (MVRT) approach, is applied to identify the hydroclimatic characteristics that govern agricultural and hydrological drought severity, as well as their degree of influence. This approach yields a set of non-linear models, each a piece-wise linear regression model (of zero order). One of the main reasons for choosing this particular approach is that it allows for the so-called “interpretable machine learning” (Molnar, 2022). The study is developed in the Cesar River basin (Colombia, South America).

## 2 Materials and methods

Figure 1 presents a flowchart of the methodology applied in this study.

### 2.1 Case study

Figure 2 presents the Cesar River basin’s location, topography and land use. The basin is located between 72º53’W 74º04’W and 10º52’00’N 7º41’00’’N latitude (Colombia, South America). It extends for an area of 22,312 km$^2$. The basin’s topography is defined in three distinct climatic regions (Universidad del Atlantico, 2014). In the north is La Sierra Nevada de Santa. This sector is characterised by steeply sloped mountains reaching up to 5,700 meters above sea level (masl). The temperature ranges from 3°C to 6°C, and the mean annual precipitation is 1,000 mm. In the east is La Serranía del Perijá. This mountainous area

https://doi.org/10.5194/nhess-2023-50
Preprint. Discussion started: 12 April 2023


Natural Hazards and Earth System Sciences
Discussions

is an extension of the eastern branch of the Andes range. In this sector, the altitude ranges from 1,000 to 2,000 masl. The average temperature is 24°C, and the average annual precipitation varies from 1,000 mm to 2,000 mm. Lastly, the valley of the Cesar River and the Zapatosa marsh are in the west and south of the basin, respectively. The valley is characterised by flat topography and a complex system of marshes formed by the Cesar River floodplains and its confluence with the Magdalena River. The average temperature is 28°C, and the mean annual precipitation is 1,500 mm. At the basin, the annual rainfall pattern presents a dry season from December to April, followed by a rainy season from April to May. In the intermediate period from June to July, precipitation decreases. The main rainfall events occur between August and November.

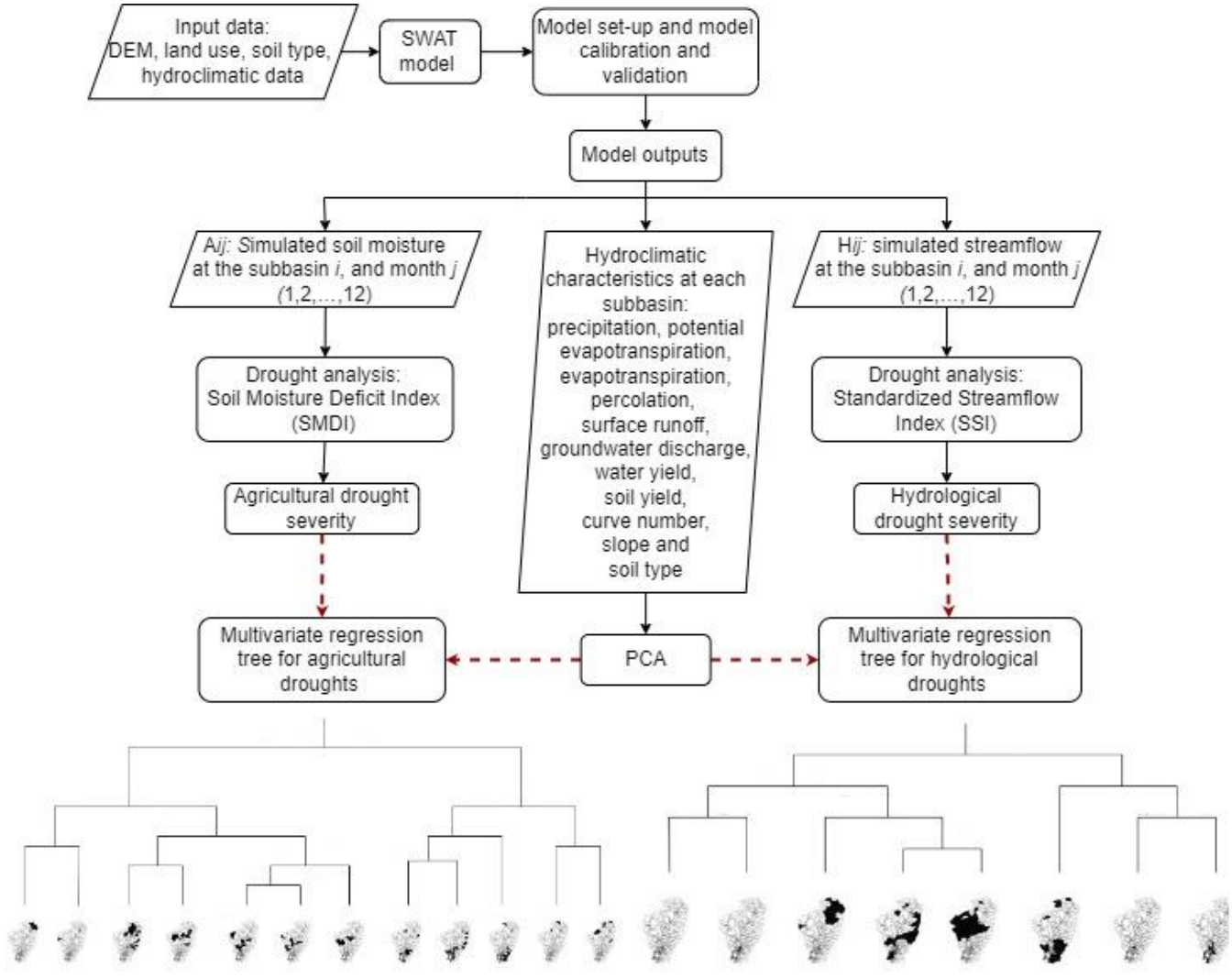
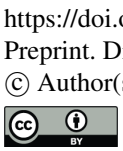



**Figure 1** Flow chart of the methodology

https://doi.org/10.5194/nhess-2023-50
Preprint. Discussion started: 12 April 2023

Natural Hazards and Earth System Sciences
Discussions

105 The predominant land use is pasture, followed by agriculture (Universidad del Atlantico, 2014). The primary land use in La Sierra Nevada foothills is pastures for cattle farming. In La Serranía del Perijá, the high altitude areas are covered by forests in very good condition; at the lower altitudes, the principal land use is agriculture, particularly subsistence crops. The Cesar River valley's soils are rich in nutrients, providing favourable conditions for agriculture. The riverbanks are covered by forest with low tree density.

110

The Zapatosa marsh is recognised as one of the most important wetlands in the country, and considering the relevance of this ecosystem, it was declared a Ramsar site in 2018. Nevertheless, the region is threatened by the overexploitation of its forest resources and overfishing. In addition, climate change projections indicate that the basin's temperature may increase by 2.7°C, and precipitation may reduce by ten percent by 2071 (Universidad del Magdalena et al., 2017). Accordingly, multiple initiatives

115 are oriented to improve water management and create resilience to hydroclimatic extremes.

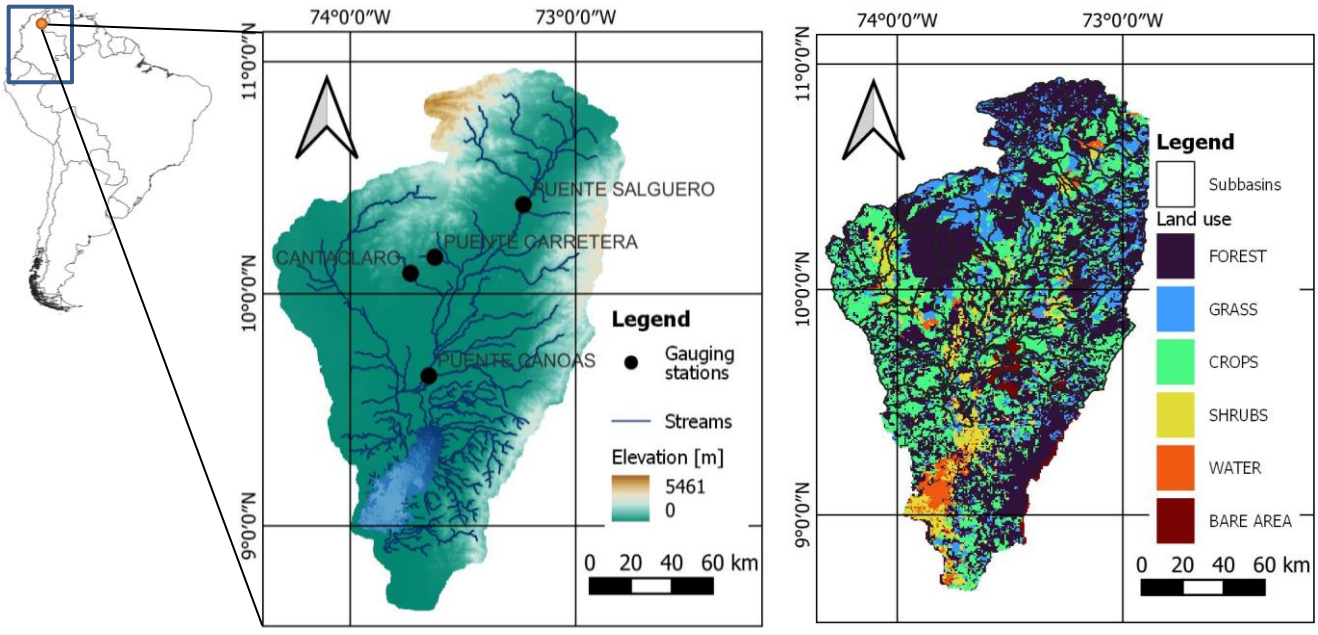



**Figure 2** Cesar River basin topography and land use

### 2.2 Hydrological modelling

A SWAT model with an ArcSWAT extension was used to develop the Cesar River basin model used in this research. The

120 Agricultural Research Service of the United States Department of Agriculture (ARS-USDA) developed the hydrological model. SWAT is a continuous-time, semi-distributed, process-based river watershed-scale model designed to simulate the quality and quantity of surface and groundwater and predict the environmental impacts of land use, land management and climate change (Neitsch et al., 2011). In SWAT, the basin area up to the outlet point is divided into several subbasins. Each

https://doi.org/10.5194/nhess-2023-50
Preprint. Discussion started: 12 April 2023


Natural Hazards and Earth System Sciences Discussions

subbasin is further split into multiple Hydrological Response Units (HRU), which are areas within the subbasin with common combinations of land cover, soil type and slope (Arnold et al., 2012).

### 2.2.1 Model setup

The model was built for the period from 1987 to 2018. The Cesar River basin was divided into 313 subbasins with a median area of 50 km$^2$. Four slope classes were set for the HRUs generation: flat (0–2%), gentle (2–10%), steep (10–35%) and considerably steep (>36%) (GEF et al., 2020, 2021). The following methods were used to model the principal hydrological processes: the soil conservation services-curve number (SCS-CN) was used to simulate surface runoff; potential evapotranspiration was estimated using the Hargreaves method; and water was routed through the channel network using the variable storage routing method. The details and sources of the SWAT model input data are presented in Table 1.

**Table 1.** SWAT model input data

| Data type | Details | Source |
|---|---|---|
| Digital elevation model | 25 × 25 m | Dataset ALOS PALSAR L1.0, Cartography 1:25000 Geographic Institute Agustín Codazzi (IGAC), Colombia |
| Soil map | 300 × 300 m | Soil profiles Project GEF Magdalena–Cauca VIVE, GEF, BID, Fundación Natura, Colombia |
| Land use map | 25 × 25 m | Land use map Geographic Institute Agustin Codazzi (IGAC), Colombia |
| Rainfall and temperature daily data (Gauging stations Figure 2) | Period 1985–2018 (34 years) | Institute of Hydrology, Meteorology and Environmental Studies (IDEAM), Colombia |
| Discharge data monthly data (Gauging stations Figure 2) | Period 1985–2018 (34 years) | Institute of Hydrology, Meteorology and Environmental Studies (IDEAM), Colombia |

### 2.2.2 Model calibration and validation

We used the SWAT-CUP software package with Sequential Uncertainty Fitting version 2 (SUFI-2) to evaluate the model's performance in simulating streamflow for automatic model calibration and validation. With SUFI-2, parameter sensitivities are determined by applying a multiple regression approach, which regresses the parameters sampled via a Latin hypercube technique against the objective function values (Abbaspour et al., 2018).

Based on expert judgment and the available literature (Arnold et al., 2012; Transactions of the ASABE (American Society of Agricultural and Biological Engineers), 2018), the following SWAT parameters were used in the calibration and validation process: ALPHA_BF (baseflow alpha factor), CH_K (effective hydraulic conductivity in main channel alluvium), CN2 (SCS runoff curve number), ESCO (soil evaporation compensation factor), GW_DELAY (groundwater delay), GWQMN (threshold

https://doi.org/10.5194/nhess-2023-50
Preprint. Discussion started: 12 April 2023


Natural Hazards and Earth System Sciences Discussions

depth of water in the shallow aquifer required for return flow to occur), RCHRG_DP (deep aquifer percolation fraction), REVAPMN (threshold depth of water in the shallow aquifer for percolation to the deep aquifer to occur) and SOL_AWC (available water capacity of the soil layer).

The model was calibrated using data from 1985 to 2002 and validated with data from 2003 to 2018 using the streamflow series from four stream gauges (Figure 2). The first two years were used as a warming-up period in both cases. Thus, performance indicators were calculated for 1987 to 2002 (calibration) and 2005 to 2018 (validation). The model was evaluated using the Nash-Sutcliffe Efficiency (NSE) and percent bias (PBIAS), represented by Eq. 1 and Eq. 2, respectively:

$$NSE = 1 - \frac{\sum_{i=1}^{N}(O_i - P_i)^2}{\sum_{i=1}^{N}(O_i - \bar{O})^2} \quad (1)$$

$$PBIAS = \frac{\sum_{i=1}^{N}(O_i - P_i) \times 100}{\sum_{i=1}^{N} O_i} \quad (2)$$

where $O_i$ is the observed data, $P_i$ the predicted data, $\bar{O}$ the mean of the observed data and $N$ the number of observations during the simulation period.

The NSE represents a dimensionless indicator ranging from -∞ to 1, with 1 representing a perfect match between the observed and simulated values (Moriasi et al., 2007). The PBIAS measures the average tendency of the simulated values to be larger or smaller than the observed values. A low PBIAS magnitude indicates accurate model simulation (Moriasi et al., 2007).

### 2.3 Agricultural and hydrological drought analysis

The present study used the soil moisture deficit index (SMDI) to analyse agricultural droughts (Narasimhan & Srinivasan, 2005). The input parameter used to calculate the SMDI was the simulated soil water in the soil profile at each subbasin. The computation procedure to determine the soil moisture deficit used the long-term soil moisture characteristics and the soil moisture conditions during the drought state evaluation period. The indicator was scaled between -4 to 4 to allow the spatial comparison of the drought index, regardless of climatic characteristics (Narasimhan & Srinivasan, 2005). Negative values of SMDI indicate dry periods, while positive values indicate wet periods (compared to the region's normal conditions). Per the SMDI, agricultural drought severity was divided into three categories: moderate drought (SMDI -2.0 to -2.99), severe drought (SMDI -3.0 to -3.99) and extreme drought (SMDI -4 or less). The following procedure was applied to compute the SMDI at each subbasin:

$$SD_{ij} = \frac{SW_{ij} - MSW_j}{MSW_j - minSW_j} \times 100, \qquad if\ SW_{ij} \le MSW_j \quad (3)$$

https://doi.org/10.5194/nhess-2023-50
Preprint. Discussion started: 12 April 2023


Natural Hazards and Earth System Sciences
Discussions

$$SD_{ij} = \frac{SW_{ij} - MSW_j}{maxSW_j - MSW_j} \times 100, \qquad if\ SW_{ij} > MSW_j \tag{4}$$

where $SD_j$ is the soil moisture deficit (%), $SW_j$ is the monthly soil water available in the soil profile (mm) and $MSW_j$, $maxSW_j$ and $minSW_j$ are long-term median, maximum and minimum soil water available in the soil profile (mm), respectively, ($i = 1987 - 2018\ and\ j = 1 - 12$).

175

The $SMDI_j$ of any given month was calculated using Eq. 5:

$$SMDI_j = 0.5 \times SMDI_{j-1} + \frac{SD_j}{50} \tag{5}$$

where $SMDI_{j-1}$ is the SMDI from the previous month.

180

SMDI was not calculated for the subbasins that correspond to the Zapatosa marsh. In these subbasins the predominant land cover is water. See Figure 5.

We used a standardised streamflow index (SSI) to represent hydrological droughts. The indicator was introduced by Modarres
185 (2007) and further investigated by Vicente-Serrano et al. (2011). The index is statically analogous to the commonly used standardised precipitation index (SPI) introduced by Mckee et al. (1993). SSI values mainly range from -2.0 (extremely dry) to 2.0 (extremely wet), and hydrological drought severity is divided into three categories: moderate drought (SSI -1.0 to -1.49), severe drought (SSI -1.5 to -1.99) and extreme drought (SSI -2.0 or less). The procedure to calculate SSI consists of converting streamflow values to standardised anomalies (i.e. z-scores). To this aim, the monthly simulated streamflow at each subbasin
190 in the present research was fitted to the gamma probability distribution function.

SMDI and SSI were calculated monthly using the simulated soil water and streamflow values at each subbasin. The drought events during the period of analysis were then identified. A drought (agricultural or hydrological) event was assumed to occur in the basin when at least thirty percent of the basin area was in a moderate drought category at the same time step (i.e. in this
195 study month). We opted to set a spatial extension minimum threshold because droughts typically extend regionally (Sheffield & Wood, 2011b). By setting the threshold, we avoided identifying a reduced number of subbasins experiencing water scarcity or short periods of water shortage as drought events.

https://doi.org/10.5194/nhess-2023-50
Preprint. Discussion started: 12 April 2023


### 2.4 Principal component analysis

A principal component analysis (PCA) is an unconstrained ordination technique that displays patterns in multivariate data. It aims to reduce the dimensionality of a data set, converting several potentially correlated variables into a new set of uncorrelated variables, known as principal components (PCs), that capture the variability in the original data (Jolliffe, 2002). PCs are ordered so that the first PCs retain most of the variation from the original variables.

PCA uses orthogonal linear transformation to identify a vector in the *N*-dimensional space to capture as much of the total variability in a set of N variables as possible (i.e. the first PC). A second vector or PC, orthogonal to the first, seeks to account for the remaining variability in the original variables. Each consecutive PC is linearly uncorrelated to the others and accounts for the remaining variability in the original variables. The eigenvalues associated with the vector for each PC define the PC's ranking, considering how much of the variability is captured by each PC. A detailed description of PCA can be found in Jolliffe (2002) and Legendre and Legendre (2012b).

Considering PCA is a suitable technique for exploring and reducing data dimensionality, we applied the technique to identify the most influential of the dataset of explanatory variables (i.e. potential drivers of droughts). Additionally, PCA results helped explain the key drivers of droughts obtained from the MVRT.

### 2.5 Multivariate regression tree approach for evaluating the relationships between hydroclimatic characteristics and droughts severity

MVRT is an extension of a regression tree (Breiman, 2001), but it differs in that it allows for multiple outputs (see De'ath, 2002). It allows the recursive split of a quantitative response variable (predictand, output) controlled by a set of numerical or categorical explanatory variables (predictors, input). An MVRT result is a tree whose terminal groups (leaves) of instances (input-output vectors) are comprised of subsets of instances selected to minimise the within-group sums of squares. Each successive split is given by a threshold value of the explanatory variables (Borcard et al., 2018). MVRT is applicable for dataset exploration, description and prediction (De'ath, 2002).

This study used MVRT as an explanatory approach to assess the relatedness between the hydroclimatic characteristics (explanatory variables) of the subbasins and the severity of agricultural and hydrological droughts (response variables). To compute the MVRT, we used R software; namely the package mvpart. Before the analysis, the sets of explanatory and response variables were transformed to compare the descriptors measured in different units and to modify the variables' weights. The matrix of explanatory variables was standardised to a mean of zero and a standard deviation of one. The matrix of response variables was standardised by the column maximum, then again by the row total (Wisconsin double standardisation). The analyses for agricultural and hydrological droughts were conducted separately; thus, two MVRTs were obtained.

https://doi.org/10.5194/nhess-2023-50
Preprint. Discussion started: 12 April 2023




230 **2.5.1 Datasets**

**Set of explanatory variables**

To select the set of explanatory variables, we used the outcomes of previous studies on governing drivers of droughts (Sheffield & Wood, 2011a; Zhang et al., 2022). Table 2 describes the eleven parameters selected as the potential drivers of droughts. The used value corresponded to the parameters' average at each subbasin. The averages were computed using the SWAT model
235 results obtained during the analysed period. We used the dominant category at each subbasin for the curve number and the two categorical parameters, the slope and the soil type. To validate the selection of the potential drivers of droughts, we applied the PCA technique to filter the most influential factors.

**Table 2.** Explanatory variables used in MVRT

| Hydroclimatic parameter | Abbreviation | Unit | Definition |
|---|---|---|---|
| Precipitation | PRECP | mm | Average precipitation at each subbasin |
| Potential evapotranspiration | PET | mm | Average potential evapotranspiration at each subbasin |
| Evapotranspiration | ET | mm | Average actual evapotranspiration at each subbasin |
| Percolation | PERC | mm | Average percolation past the root zone |
| Surface runoff | SURFQ | mm | Average surface contribution to the streamflow at each subbasin |
| Groundwater | GRWQ | mm | Average groundwater contribution to the streamflow at each subbasin |
| Water yield | WYLD | mm | Average amount of water that leaves the subbasin and contributes to the streamflow at each subbasin |
| Sediment yield | SYLD | metric tons/ha | Average sediment from the subbasin transported into the reach |
| Curve number | CN | – | Dominant curve number at each subbasin |
| Slope | SLP | – | Dominant slope at each subbasin |
| Soil type | STY | – | Dominant soil type at each subbasin |

240 **Set of response variables**

We used the drought analysis results as response variables. After identifying the drought events during the analysed period, we counted the months for each agricultural and hydrological drought category. The total number of months for each drought category at each subbasin were used as response variables. Table 3 describes these response variables.

https://doi.org/10.5194/nhess-2023-50
Preprint. Discussion started: 12 April 2023


Natural Hazards and Earth System Sciences Discussions

245 **Table 3.** Response variables used in MVRT

| Drought category | Abbreviation | Unit | Definition |
|---|---|---|---|
| Moderate agricultural/hydrological drought | MOD | month | Number of months in the moderate agricultural drought category during the drought events identified in the simulation period at each subbasin |
| Severe agricultural/hydrological drought | SEV | month | Number of months in the severe agricultural drought category during the drought events identified in the simulation period at each subbasin |
| Extreme agricultural/hydrological drought | EXT | month | Number of months in the extreme agricultural drought category during the drought events identified in the simulation period at each subbasin |

### 2.5.2 Building the MVRT

Building the MVRT consisted of two processes: (1) the constrained partitioning of the data, and (2) the cross-validation of the results. The mvpart package run both processes in parallel. The two procedures are briefly explained below, and a more detailed description can be found in Borcard et al. (2018).

250 **Constrained partitioning of the data**

The data partitioning consisted of three steps. First, all the possible partitions for the sites in the two groups were generated for each explanatory variable. Second, for each partition, it was calculated the resulting sum of within-group sums of squared distances to the group means for the response data (within-group SS, equivalent to standard deviation). Lastly, the partition into two groups to minimise the within-group SS and the threshold value/level of the explanatory variable was retained. These
255 steps were repeated within the two previously established subgroups until all the objects formed their own groups. For each tree that was computed, the relative error was calculated as the sum of the within-group SS of all leaves divided by the overall SS of the data. This procedure for MVRT is equivalent to the one originally proposed by Breiman (2001) for his regression tree technique.

**Cross-validation of the partitions and tree pruning**

260 A cross–validation procedure was used to prune the tree and identify the optimal tree size (Kuhn & Johnson, 2013; Legendre & Legendre, 2012b). The cross-validation procedure was performed automatically using mvpart. Per this procedure, the data was randomly divided into roughly equal-sized test groups. Each test group was held out in turn while the tree was fitted using the remaining groups. The distances between the centroids of the tree leaves and each object of the test group were then calculated. Finally, the objects of the test group were allocated to the closest leaf of the constructed tree. An overall relative
265 error statistic (relative cross-validation error, CVRE) was calculated for each group using all $n$ objects, per Eq. 6:

https://doi.org/10.5194/nhess-2023-50
Preprint. Discussion started: 12 April 2023


Natural Hazards and Earth System Sciences Discussions

$$CVRE_{(k)} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{p}\left(y_{ij(k)} - \hat{y}_{j(k)}\right)^2}{\sum_{i=1}^{n}\sum_{j=1}^{p}\left(y_{ij} - \bar{y}_j\right)^2} \quad 6$$

where $y_{ij(k)}$ is the value of variable $j$ for object $i$ belonging to test group $k$, $\hat{y}_{j(k)}$ is the value of that same variable at the centroid of the leaf closest to object $i$, and the denominator is the overall sum of the $Y$ data squares.

This cross-validation process was repeated several times for consecutive and independent divisions of the data into test groups. For each group, the mean and standard deviation of all CVRE were computed. The CVRE varied from 0 for perfect predictors to close to 1 for poor predictors (for large errors, CVRE may reach +∞). Among the mvpart function arguments, we used ten cross-validation groups (function argument, xval = 10) and 100 iterations (function argument xmult = 100). The tree was selected using interactive cross-validation (function argument xv = 'pick').

Two approaches were used to choose the size of the tree that retained the most descriptive partition. According to De'ath (2002), one should select the tree with the smallest CVRE. In contrast, Breiman (1984) recommends a tree with fewer splits, which have CVRE values within one standard error of the smallest CVRE value. In the present study, we followed the approach suggested by Breiman (1984). This tree offered the best combination of explanatory power and interpretability. Once the tree was built, the proportion of explained variance (EV) was calculated as $1 - RE_{tree}$ (tree relative error) (Cannon, 2012).

## 3 Results

### 3.1 SWAT model calibration and validation

Table 4 summarises the calibration and validation performance indicators for the SWAT model at each gauging station. The calibration and validation models simulated monthly stream flows with NSE values equal to or greater than 0.50 and relatively low PBIAS values (GEF et al., 2020, 2021). According to the performance ratings for calibrating and validating hydrological models, NSE and PBIAS values indicated that the model was appropriate for simulating streamflow (Moriasi et al., 2007). Figure 3 presents the model hydrographs at each gauging station for the calibration and validation periods. The locations of the stations can be found in Figure 2.

https://doi.org/10.5194/nhess-2023-50
Preprint. Discussion started: 12 April 2023




290 **Table 4.** SWAT model performance simulating streamflow

| Gauging station | Calibration | | Validation | |
|---|---|---|---|---|
| | NSE | PBIAS [%] | NSE | PBIAS [%] |
| Puente Salguero | 0.61 | 4.28 | 0.52 | -8.3 |
| Puente Carretera | 0.50 | -5.34 | 0.52 | 7.6 |
| Cantaclaro | 0.58 | -11.30 | 0.50 | -11.7 |
| Puente Canoas | 0.70 | -1.34 | 0.57 | 10.64 |

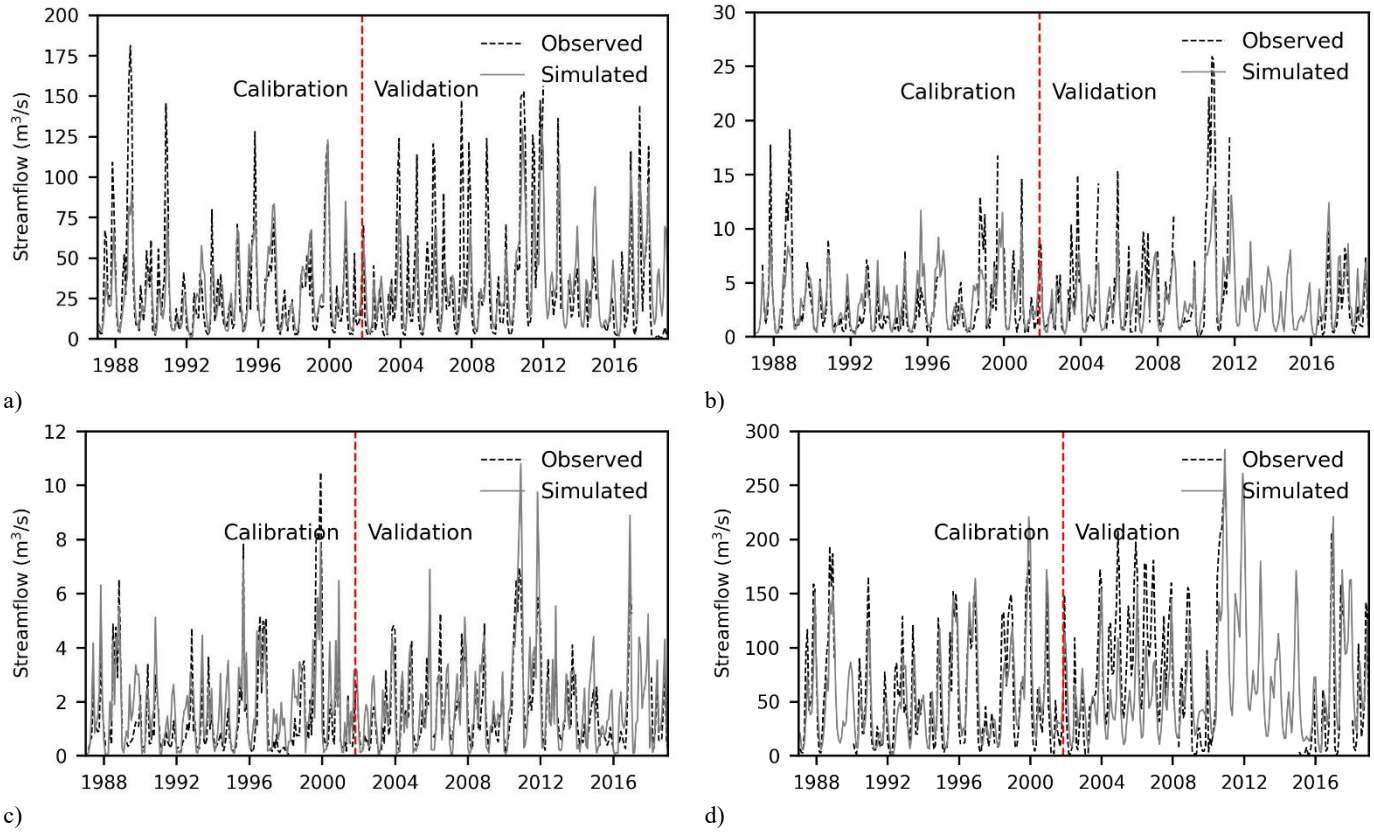



**Figure 3** Monthly calibration and validation for streamflow at: a) Puente Salguero, b) Puente Carretera, c) Cantaclaro and d) Puente Canoas.

### 3.2 Hydroclimatic drivers of droughts

Figure 4 *a* to *h* presents the average value of the numerical hydroclimatic drivers of droughts at each subbasin. The average
295 was calculated using the hydrological model's outputs during the simulation period (1987 to 2018). Figure 4 *i* to *k* show the dominant category of the parameters, the curve number slope and the soil type at each subbasin. The dataset of explanatory variables was created from the values presented in Figure 4.

https://doi.org/10.5194/nhess-2023-50
Preprint. Discussion started: 12 April 2023


Natural Hazards and Earth System Sciences
Discussions

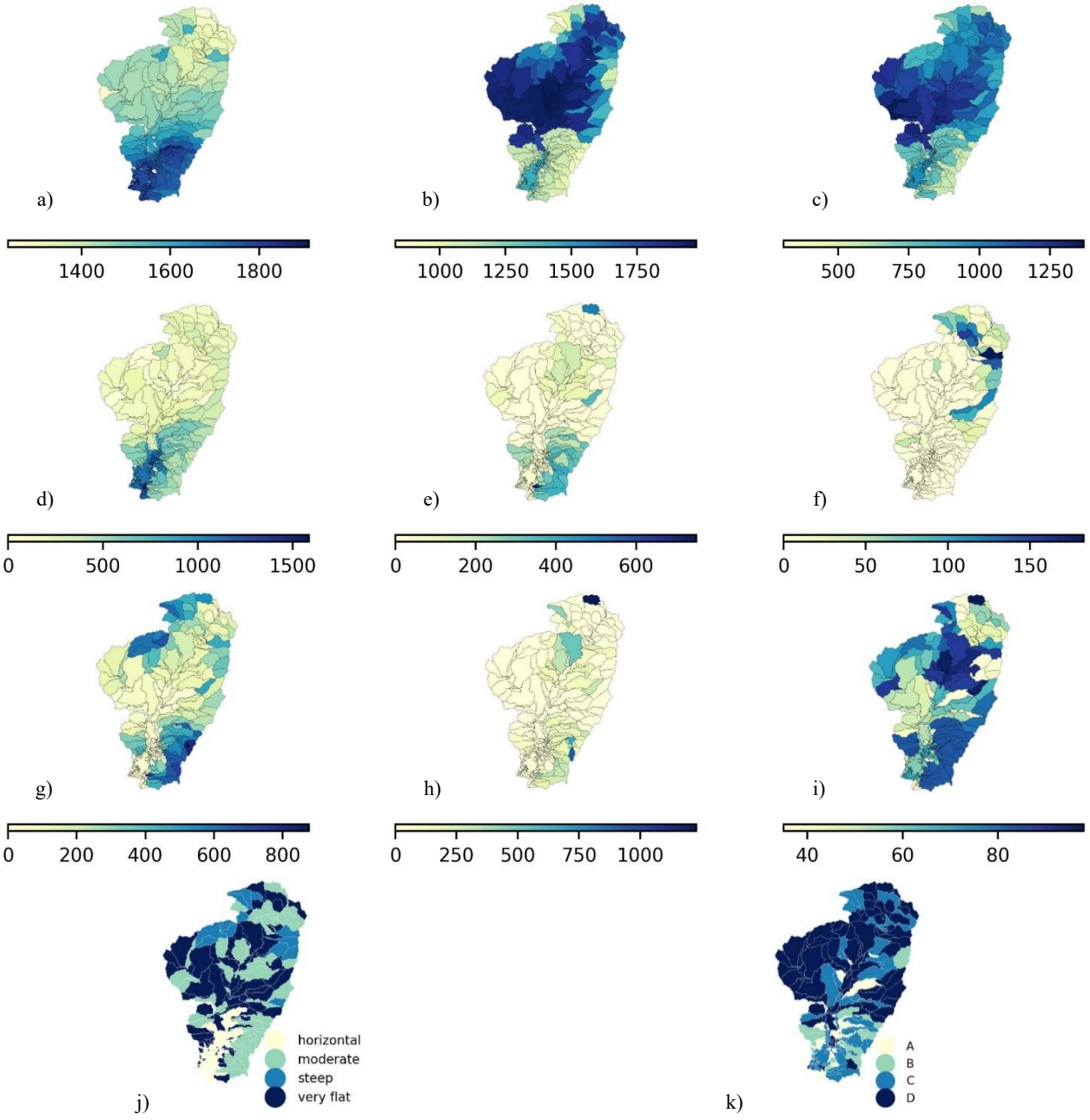



**Figure 4** Average value of hydroclimatic parameters during the simulation period at each subbasin: a) precipitation in mm, b) potential evapotranspiration in mm, c) actual evapotranspiration in mm, d) percolation in mm, e) surface runoff in mm, f) groundwater contribution to streamflow in mm, g) water yield in mm, h) sediment yield in metric tons/ha, i) curve number, j) slope and k) soil type.

https://doi.org/10.5194/nhess-2023-50
Preprint. Discussion started: 12 April 2023


Natural Hazards and Earth System Sciences
Discussions

### 3.3 Drought events during the simulation period and their duration

Following the methodology described in 2.3, we identified drought events in the analysis period. Table 5 shows the dates and durations of these events. The identified drought events in the simulation period were in good agreement with the chronology of drought events in Colombia described at the National Study of Water (Instituto de Hidrología, 2019).

305

**Table 5.** Agricultural and hydrological droughts during the period of analysis

| Event | Agricultural droughts | | Hydrological droughts | |
|---|---|---|---|---|
| | Date | Duration [months] | Date | Duration [months] |
| I | May 1991 – Jun 1992 | 13 | Apr 1991 – May 1992 | 14 |
| II | Jun 1997 – April 1998 | 11 | Apr 1997 – Feb 1998 | 11 |
| III | Jun 2001 – Aug 2001 | 3 | May 2001 – Jun 2001 | 2 |
| VI | Oct 2009 – Jan 2010 | 4 | Sep 2009 – Nov 2009 | 3 |
| V | Jun 2014 – Aug 2014 | 3 | Jun 2014 – Jul 2014 | 2 |
| VI | May 2015 – Jul 2016 | 15 | Apr 2015 – Apr 2016 | 13 |

After identifying the agricultural and hydrological drought events, it was possible to count the number of months for each drought category in each subbasin. Figure 5 presents the number of months for each agricultural drought category, and Figure 6 presents the number of months for each hydrological drought category. The results presented in Figures 5 and 6 are the response variables for the MVRT technique.

310

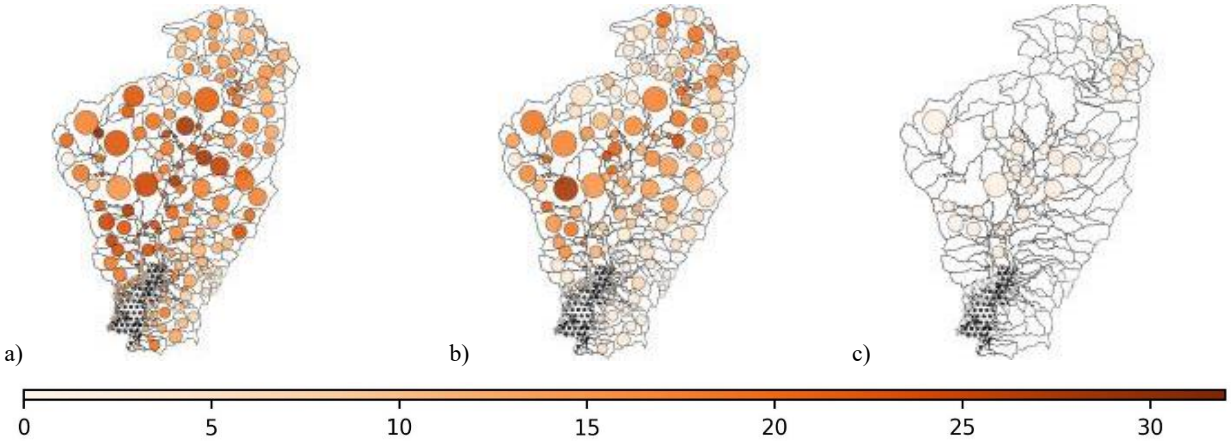



**Figure 5** Months counted in each agricultural droughts category: a) moderate, b) severe and c) extreme. Agricultural drought (SMDI) was not calculated in the wetland subbasins (i.e. hatched area).

https://doi.org/10.5194/nhess-2023-50
Preprint. Discussion started: 12 April 2023


Natural Hazards and Earth System Sciences
Discussions

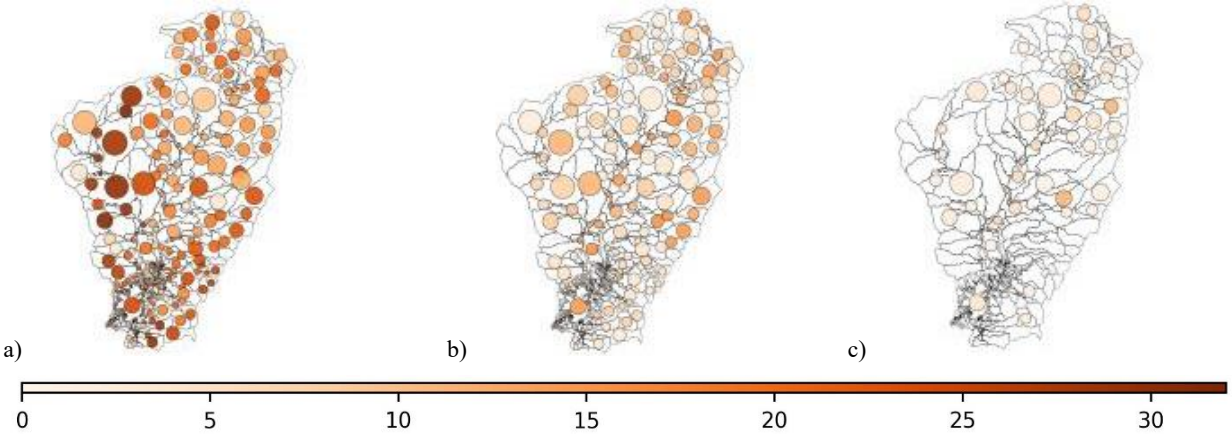



**Figure 6** Months counted in the hydrological droughts category: a) moderate, b) severe and c) extreme.

### 3.4 PCA

The PCA identified three PCs; together, they explain seventy-seven percent of the variation in the original eleven parameters included in the set of explanatory variables (i.e. parameters influencing droughts). Figure 7a presents the loading factors of the eleven hydroclimatic parameters for the first three PCs. The loading factors reveal that all the original variables heavily influenced at least one of the PCs retained. Accordingly, the eleven original parameters were retained as explanatory variables to compute the MVRT. The loading factors of the first two PCs are presented in a biplot (Figure 7b).

a)

| Parameter | PC1 | PC2 | PC3 |
|---|---|---|---|
| PRECP | -3.291 | 0.076 | -1.284 |
| PET | 3.303 | 1.285 | -1.702 |
| ET | 3.109 | 1.589 | -2.121 |
| PERC | -3.198 | 1.123 | 0.759 |
| SURFQ | -0.710 | -3.754 | -1.078 |
| GRWQ | 1.598 | -0.016 | 5.177 |
| WYLD | 0.261 | -3.838 | 1.563 |
| SYLD | 0.089 | -3.328 | -0.608 |
| CN | 2.492 | -1.938 | 2.405 |
| Slope | 2.536 | 0.013 | -1.150 |
| Soil type | 0.638 | -2.954 | -3.214 |
| Eigenvalue | 3.957 | 3.263 | 1.273 |
| Cumulative | 0.360 | 0.656 | 0.772 |

b)

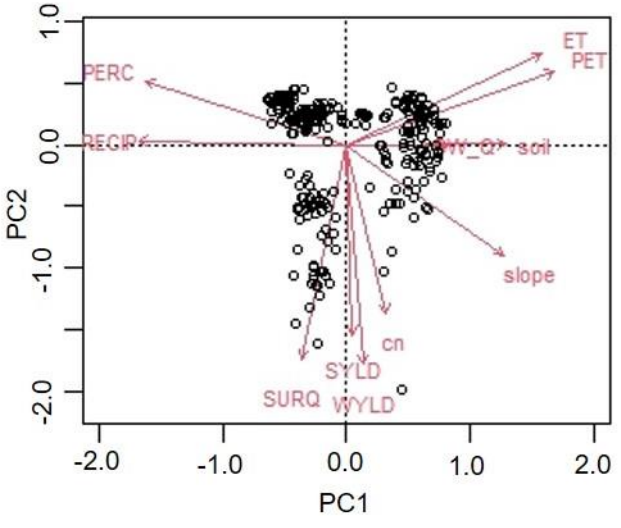



**Figure 7** Loading factors of the eleven hydroclimatic parameters for the first three PCs (a) and the biplot of the first two PCs (b).

https://doi.org/10.5194/nhess-2023-50
Preprint. Discussion started: 12 April 2023


Natural Hazards and Earth System Sciences Discussions

### 3.5 Multivariate regression tree

In this section, we describe the results of the MVRT technique applied to identify the governing drivers of agricultural and hydrological drought severity and the critical thresholds of these drivers.

#### 3.5.1 Drivers of agricultural drought

Figure 8 presents the tree generated by R software, the number of subbasins clustered at each terminal group (variable "*n*"), and the spatial distribution of these subbasins. The tree consists of five levels of split and twelve leaves. The minimum value of the cross-validation error (CVRE = 0.46) was used to select the tree size. The relative error of the MVRT was 0.19, and the EV was 0.81. Figure 9 presents the tree's numerical output: namely, the number of months for each drought category. The scattering of the outputs in each leaf allows us to identify the subbasins most exposed to agricultural droughts.

The MVRT indicated that evapotranspiration was a strong driver of agricultural droughts; it appeared three times at different tree levels in the splitting rules. The subbasins were split at the first level according to ET (924 mm). At the second level of split, precipitation (1,318 mm) was used for the left branch of the tree and percolation (271 mm) for the right branch. Then, the left branch was recursively split as follows: at the third level, according to potential evapotranspiration (1,888 mm) and evapotranspiration (1,191 mm); at the fourth level, according to evapotranspiration (1,064 mm) and percolation (111 mm); and at the fifth level, according to potential evapotranspiration (1,679 mm) and sediment yield (101 mm). The left branch account for seven out of the tree's twelve leaves. Regarding the right branch, it was divided according to evapotranspiration (729 mm) and the curve number (67) at the third level and according to the water yield (352 mm) at the last level. In the following, we describe agricultural drought MVRT terminal groups.

Leaves *a* and *b*: Leaf *a* clusters seven subbasins in the north part of the basin. In this area, evapotranspiration and potential evapotranspiration were above the basin average, while precipitation was below average. Figure 9a shows that these subbasins experienced the highest number of months in extreme agricultural drought, with a median condition of fifteen months in severe agricultural drought. Leaf *b* clusters two subbasins in the western part of the basin. This leaf contains no instance of severe agricultural drought. The median of months in the moderate and severe agricultural drought categories is ten months, one of the highest among the terminal groups (Figure 9b).

Leaves *c* and *d* cluster twenty-four and nineteen subbasins, respectively. In leaf *c*, evapotranspiration was above 1,064 mm, a value above the basin average. These subbasins experienced no extreme agricultural droughts, and the median of months in the severe category was below ten, representing the lowest of the left branch of the MVRT (Figure 9c). Leaf *d* shows the highest median of months in the severe drought category (Figure 9d).

https://doi.org/10.5194/nhess-2023-50
Preprint. Discussion started: 12 April 2023


Natural Hazards and Earth System Sciences
Discussions

Leaves *e*, *f* and *g* cluster twenty-four, six and twelve subbasins, respectively, located in the river valley. In these subbasins, evapotranspiration was higher than 1,991 mm. For leaf *f*, percolation was considerably above the basin average (111 mm), and sediment yield was notably above the median (101 metric tons/ha). Figure 9f, g and h show that the median of months in the moderate drought category was above twenty months; the severe category was above ten months; and the three leaves exhibited months in the extreme drought category.

Leaves *h*, *i* and *j* cluster twenty-six, fifty-two and fifty-six subbasins, respectively. These subbasins are mainly located in the wetland surroundings. Evapotranspiration at terminal groups *i* and *j* was higher than 729 mm. At terminal group *i*, the water yield was higher than the median (352 mm), while at terminal group *j*, it was lower. At leaf *k*, evapotranspiration was lower than 729 mm. The agricultural drought situation in these subbasins was mild. The median of months in the moderate drought category was slightly higher than ten, and the median for months in the severe and extreme categories was the lowest for the area of study (Figure 9h, i, j).

Leaves *k* and *l* cluster two and six subbasins, respectively. In these subbasins, percolation was lower than 271 mm. In leaf *l*, the curve number was lower than sixty-seven, while in leaf *l* it was higher. These subbasins experienced intermediate drought exposure. Figure 9k and l demonstrate that there no months were registered in the extreme drought category.

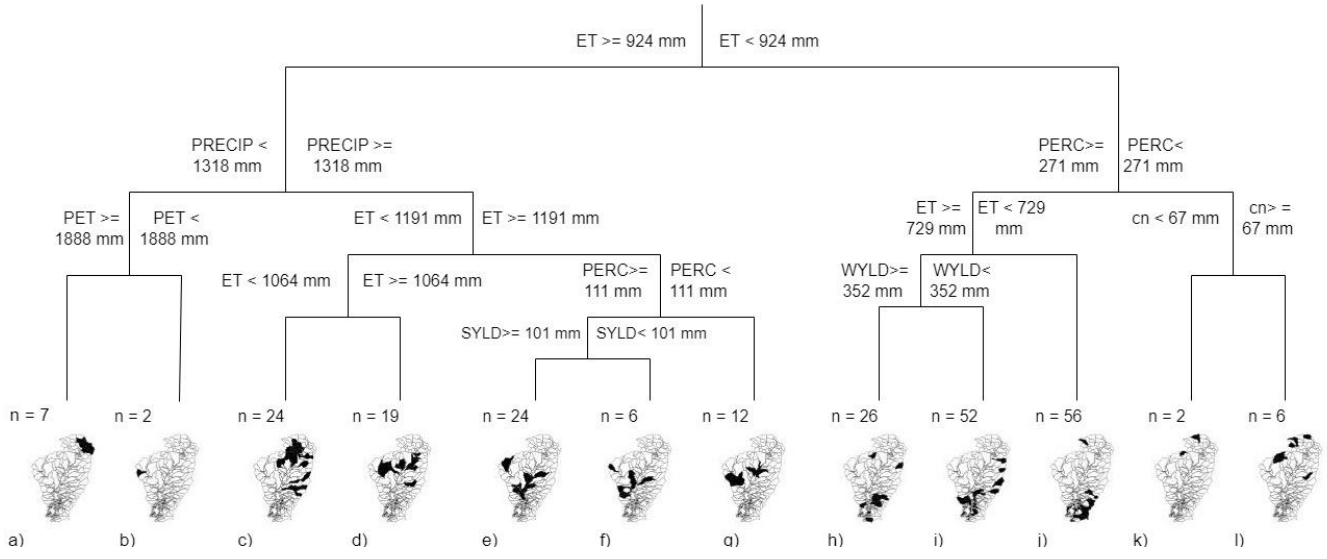



**Figure 8** MVRT of hydroclimatic drivers of agricultural droughts at the Cesar River basin, and spatial distribution of the subbasins clustered at each leaf.

https://doi.org/10.5194/nhess-2023-50
Preprint. Discussion started: 12 April 2023

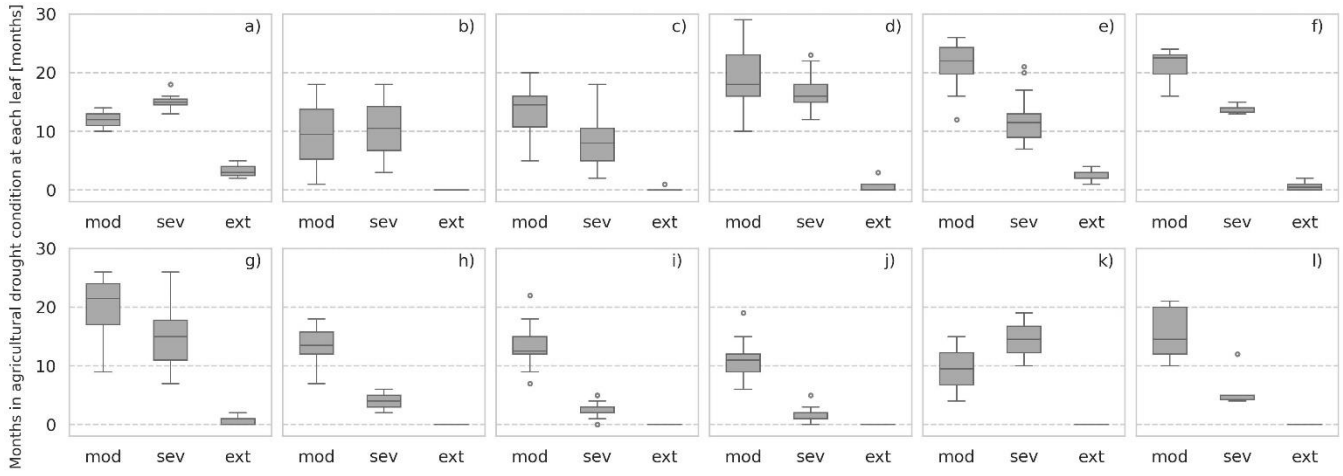




**Figure 9** Number of months in agricultural drought categories (moderate, severe, extreme) at each leaf.

### 3.5.2 Drivers of hydrological drought

Figure 10 presents the hydrological drought MVRT, the number of subbasins clustered at each terminal group (variable "*n*") and the spatial distribution of these subbasins. The tree consists of four levels of split and eight leaves. The minimum value of the cross-validation error (CVRE = 0.67) was used to select the tree size. The relative error of the MVRT was 0.52, and the EV was 0.48. Figure 11 presents the tree's numerical output: namely, the number of months for each drought category. This information allowed us to identify the clusters of subbasins most exposed to hydrological droughts.

The MVRT demonstrated that precipitation was a primary driver of hydrological drought; it appeared two times at different levels of split. The subbasins were separated at the first split level according to precipitation (1362 mm). At the second split level, precipitation (1398 mm) was used as the left branch of the tree, and water yield was used as the right branch (29 mm). The left branch was then further divided according to percolation at the third level and according to curve number at the fourth level. Also at the third level, the right branch was split according to evapotranspiration (833 mm) and surface runoff (0.5 mm). The MVRT terminal groups were then examined in detail.

Leaf *a* clusters twenty-eight subbasins located in the upper basin (Figure 10a). Here, precipitation was below the basin average (1,398 mm). Figure 11a shows that the subbasins in this terminal group were considerably exposed to severe and extreme hydrological drought.

Leaves *b*, *c* and *d* cluster thirty-seven, thirteen and twenty-nine subbasins, respectively. In these subbasins, precipitation was higher than 1,398 mm. In leaf *b*, percolation was higher than 153 mm, and the curve number was higher than 51. These results suggest that the subbasins in leaf *b* experienced considerable exposure to hydrological drought (Figure 11b). Concerning leaf

https://doi.org/10.5194/nhess-2023-50
Preprint. Discussion started: 12 April 2023


Natural Hazards and Earth System Sciences
Discussions

*d*, percolation was considerably lower than the median. Figure 11d indicates that in this terminal group, the subbasins experienced fewer months in the severe and extreme drought categories compared to leaf *a*. However, they also experienced one of the highest medians of months at moderate drought.

In leaves *e* (n = 72) and *f* (n = 18), precipitation was higher than 1,632 mm and water yield lower than 29 mm. Both terminal groups described moderate exposure to hydrological drought. Notably, at leaf *e*, the median of months in the severe and extreme drought categories was below five, while the median of months in the moderate drought category was twenty (Figure 11e). The hydrological drought exposure of the subbasins clustered at leaf *f* was also mild. In these subbasins, evapotranspiration was higher than 833 mm. These subbasins presented the lowest median of months at all drought categories. Notably, the Zapatosa marsh and upstream subbasins were clustered in this terminal group (Figure 10f).

Leaves *g* and *h* cluster seventy-one and thirty-seven subbasins, respectively. The water yield (29 mm) was considerably below the basin average at these subbasins. The MVRT indicated that surface runoff was higher than 0.5 mm. Our analysis showed that the subbasins grouped at leaf *g* presented the lowest exposure to hydrological drought. The median months at all categories were the lowest in the basin (Figure 11g). All the subbasins clustered in these leaves were upstream of the Zapatosa marsh. In leaf *h*, the surface runoff was lower than 0.5 mm. In these subbasins, the medians of months in the severe and extreme categories were relatively low, while the median of months in the moderate category was eighteen (Figure 11h).

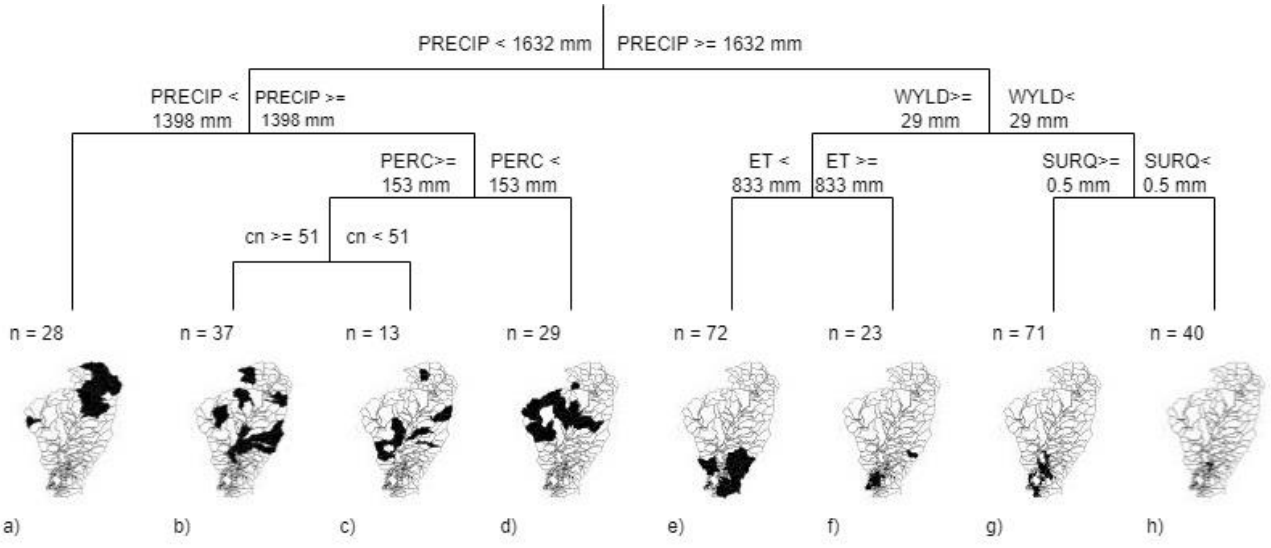



**Figure** 10 MVRT of hydroclimatic drivers of hydrological drought at the Cesar River basin, and spatial distribution of the subbasins clustered at each leaf.

https://doi.org/10.5194/nhess-2023-50
Preprint. Discussion started: 12 April 2023

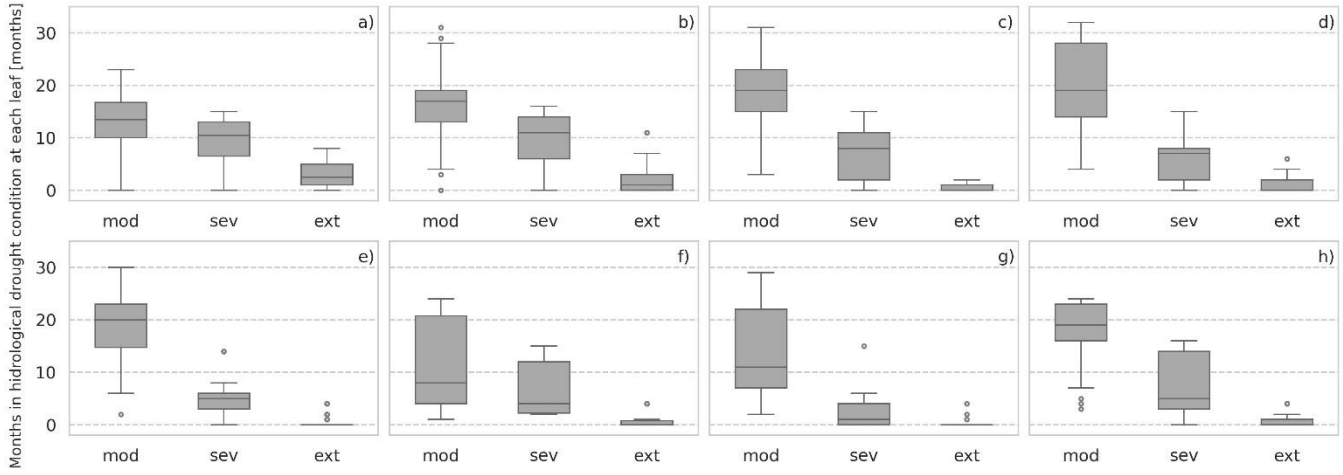




**Figure 11** Months in agricultural drought categories (moderate, severe, extreme) at each leaf.

## 4. Discussion

### 4.1 Hydroclimatic drivers of agricultural drought

The left branch of the MVRT clusters the subbasins most exposed to agricultural drought (Figure 9a, e, f, g). Conversely, the right branch of the MVRT clusters the subbasins with moderate exposure to agricultural drought. The subbasins in leaves *h*, *i* and *j* predominately experienced months in the moderate drought category (Figure 9i, j, k).

Interestingly, the levels of exposure to agricultural drought shown in leaves *a*, *e*, *f* and *g* were comparable but governed by different parameters. This suggests that the interactions between climate and different land surface processes can lead to similar exposure to agricultural drought. For instance, leaf *a* presented the highest median of months for severe and extreme agricultural drought (Figure 9a). The drought drivers in this terminal group, namely precipitation and potential evapotranspiration, indicate that agricultural drought results from an imbalance between the soil moisture supply (i.e. precipitation relatively close to the minimum value at the basin) and soil moisture demand (i.e. moderately high potential evapotranspiration). This finding aligns well with studies demonstrating that potential evapotranspiration considerably enhances the severity of agricultural droughts in water-limited areas (Ding et al., 2021; Manning et al., 2018; Teuling et al., 2013). According to such studies, potential evapotranspiration influence on agricultural drought severity may be explained by the significant increase in net radiation during droughts, as the lack of rainfall usually concurs with decreased cloud cover.

In contrast, the MVRT outcomes suggest that a lack of precipitation was not a primary driver of agricultural drought in the subbasins clustered at leaf *e*. This leaf grouped the subbasins that experienced the most severe agricultural drought in the analysis period. The median of months in the moderate drought category was above twenty; the severe category was above ten; and the extreme drought category was six (Figure 9f). The observed evapotranspiration and percolation thresholds might

https://doi.org/10.5194/nhess-2023-50
Preprint. Discussion started: 12 April 2023


Natural Hazards and Earth System Sciences Discussions

indicate poor precipitation partitioning and a disturbed water regime that favours water lost by percolation and evapotranspiration. Furthermore, the sediment yield threshold (notably above the median) may be linked to poor soil structure, thus compromising soil water retention capacity and enhancing drought exposure.

The results from leaf *f* show that a lower sediment yield slightly reduces the severity of extreme droughts (Figure 9f), as compared to the results from leaf *e*. This is in good agreement with the findings by Masroor et al. (2022), Trnka et al. (2016) and Santra et al. (2020), who found that soil erosion enhances the characteristics of agricultural drought. Further, our results are consistent with previous studies that indicate the incidence of droughts is not only caused by extreme weather events but also by the inefficient soil–water management associated with land and soil degradation (Cornelis et al., 2019; Wildemeersch et al., 2015).

The right branch of the tree provides valuable information on the hydroclimatic parameters that reduce exposure to agricultural droughts. Moderate drought exposure levels for leaves *h*, *i* and *j* were associated with relatively low evapotranspiration thresholds; accordingly, it may be asserted that evapotranspiration controlling measures (e.g. surface cover, crop rotation, agroforestry, intercropping) are relevant interventions for building resistance to agricultural drought. At terminal groups *i* and *j*, water yield was found to influence the severity of agricultural drought. Notably, the subbasins at leaf *i* were slightly more resistant to drought (Figure 9i); this indicates that measures aimed at increasing the subbasins' water storage capacity (e.g. rainwater and floodwater harvesting techniques) are suitable interventions to reduce the severity of agricultural drought.

Some of the subbasins grouped at leaf *i* showed high exposure to hydrological drought (Figure 11b and c). Contrasting exposure to agricultural and hydrological droughts revealed that the water retention capacity in these subbasins reduced the severity of agricultural drought events but limited the contribution of surface runoff, lateral flow and groundwater to the streamflow, thus exacerbating the water deficit and hydrological drought severity. Therefore, drought management interventions require the prior assessment of the potential effects on both types of droughts.

### 4.2 Hydroclimatic drivers of hydrological droughts

The subbasins clustered on the left branch of the tree were the most exposed to hydrological drought (Figure 11a, b, c, d). The MVRT indicated that the severe exposure of leaf *a* to hydrological drought was driven by limited precipitation. Leaf *a* presented the highest median for months in the severe and extreme hydrological categories. The analysis results confirmed that precipitation deficits caused the severe drought (agricultural and hydrological) conditions in the upper part of the basin.

Conversely, the MVRT also showed that in terminal groups *b*, *c* and *d*, hydrological drought severity was linked to the inefficient partition of precipitation. Selected drivers of hydrological droughts are widely recognised as predominant drivers of hydrological droughts (Iglesias et al., 2018; Stoelzle et al., 2014; van Lanen et al., 2013; van Loon, 2015). The difference

https://doi.org/10.5194/nhess-2023-50
Preprint. Discussion started: 12 April 2023


Natural Hazards and Earth System Sciences Discussions

observed between the precipitation and percolation thresholds suggests that a large part of rainwater was lost either by evapotranspiration or surface runoff (or other water abstractions, e.g., human consumption, agriculture). Low percolation values limited the groundwater contribution to the streamflow, enhancing the streamflow deficit during drought periods.

Interestingly, the curve number was selected as a driver of hydrological drought for leaves *b* and *c* (Figure 10b and c). The subbasins in leaf *b* showed higher curve numbers than those in leaf *c* and higher exposure to hydrological drought. High curve number values are commonly the result of anthropogenic changes in land cover, which modifies evapotranspiration and the division of precipitation into evapotranspiration and streamflow. The present selection of the curve number at the third level of split is consistent with previous studies, which established that hydroclimatic parameters and human activities influence hydrological droughts; however, the influence of both drivers is uneven. Reports indicate that hydroclimatic parameters are more influential (Jehanzaib et al., 2020; Saidi et al., 2018).

The right branch of the MVRT grouped subbasins with moderate and intermediate exposure to hydrological drought. The hydroclimatic parameters and the thresholds used to define leaves *e* and *f* (precipitation, water yield and evapotranspiration) demonstrate that in these subbasins, precipitation values compensated for the water abstraction by evapotranspiration. When we compare the severity of the hydrological droughts observed in leaves *e* and *f*, we find that lower values of evapotranspiration reduce exposure to severe and extreme hydrological drought but increase the incidence of moderate hydrological drought.

The subbasins in terminal group g experienced the lowest median number of months for all hydrological drought categories (Figure 11g). The observed low water yield values indicate that the subbasins in this leaf had significant water retention capacity, which reduced their exposure to hydrological drought. This, together with the subbasins' proximity to the marsh (which acted as a natural control), the low slope in the area (which reduced streamflow velocity) and the presence of water bodies (which collected and stored runoff during the rainy season) may have enhanced the water retention capacity in these areas. The observed moderate exposure of these subbasins fits the results of earlier analyses, which found that wetlands exert significant impacts on the alleviation of hydrological drought severity when direct evaporation from the water body does not significantly reduce water storage (Wu et al., 2023). Thus, the present findings indicate that the water storage capacity of the Zapatosa marsh can compensate for the increased evaporation that occurs during drought events, thereby alleviating hydrological drought severity upstream.

The hydrological drought conditions in the subbasins clustered at leaf *h* were mild, despite water yield values below 29 mm. Negligible values of surface runoff increased the exposure of these subbasins to hydrological drought as compared to leaf *g*. The observed drought drivers and thresholds indicate that rainfall retention in the soil profile and percolation for groundwater recharge limited the amount of water that reached the streamflow. This resulted in a water deficit that raised the chances of hydrological drought.

https://doi.org/10.5194/nhess-2023-50
Preprint. Discussion started: 12 April 2023


### 4.3 Accuracy and validation of the MVRT-based analysis

The high EV (0.81) value reflects the good explanatory power of the tree built for agricultural drought. This confirms that the selected explanatory variables significantly influence the severity of agricultural drought.

Conversely, the explanatory power of the tree built for hydrological drought is not very high (EV = 0.48). This may be related to the inaccurate representation of groundwater contribution to the streamflow. Streams depend significantly on groundwater during droughts to maintain flow; nevertheless, groundwater contribution to the streamflow was not included as a key drought driver in this study, although it was in the list of explanatory variables. It is possible that the model's simplifications for the simulation of groundwater flow and storage did not adequately represent the system (Molina-Navarro et al., 2019). The lack of information about a relevant factor influencing hydrological drought may have compromised the MVRT's accuracy. Unexplained variability may also have been due to factors that influence hydrological drought (e.g. abstractions such as water for irrigation, industry or human consumption) but were not considered in the dataset of explanatory variables.

It is worth comparing the results obtained using the MVRT with the analysis carried out using PCA. The MVRT results are compatible with the PCA. Eleven out of nine potential drivers of drought were selected to build the MVRT. This is in good agreement with PCA results, which revealed that all eleven original variables heavily influenced at least one of the PCs retained. The parameters selected at the first levels of split were the same parameters that considerably influenced the PCs (PRECP, PET, ET, PERC). Other factors used at subsequent levels of split showed relatively lower loading factors. The main difference between the PCA and MVR results was the effect of slope and soil type on drought severity. Although both parameters showed significant loading factors in the PCA analysis, neither was used in the computation of the MVRT.

Interestingly, the PCA correlation biplot clusters presented similarities with the clusters created using the MVRT (Figure 7b). For instance, the cluster on the top left of the biplot corresponds to subbasins with high precipitation values and percolation. Overall, these subbasins coincided with subbasins clustered at the right side of the MVRT (i.e. leaves with moderate exposure to droughts). Points on the top right of the biplot represent the subbasins with high evapotranspiration. These subbasins were clustered at the left side of the MVRT, where terminal groups exhibited high exposure to droughts triggered mainly by evapotranspiration. The remaining subbasins are clustered at the lower part of the biplot. This clustering corresponds with the subbasins that presented high exposure to droughts influenced by different hydroclimatic parameters.

## 5. Conclusion

This study applied the MVRT technique, which served not only as a predictive machine learning model but also as an explanatory approach (in the line of 'explanatory AI') to assess the relationship between a subbasin's hydroclimatic characteristics (i.e. explanatory variables) and the severity of agricultural and hydrological drought (i.e. response variables).

https://doi.org/10.5194/nhess-2023-50
Preprint. Discussion started: 12 April 2023


The results show that the employed machine learning technique successfully identified the primary drivers of drought severity and their critical thresholds. The MVRT also provided valuable information on which parameters contribute to the reduction of agricultural and hydrological drought severity. These MVRT outcomes are easy to interpret and visualise.

The present findings reveal that the Cesar River basin's exposure to hydrological and hydrological drought can be divided into three different areas. In the upper part of the river valley, exposure to agricultural and hydrological drought is high and linked to precipitation deficits and high potential evapotranspiration. In the middle part of the river valley, exposure to agricultural and hydrological drought is also high. However, severe drought conditions are triggered by poor partitioning of precipitation and a disturbed water regime that exacerbates water lost by percolation and evapotranspiration. These findings confirm that the combined effect of different hydroclimatic parameters and human interventions can drive comparable exposure to agricultural and hydrological drought. Third, the Zapatosa marsh and the Serrania del Perijá foothills present moderate exposure to agricultural and hydrological drought. Mild drought conditions seem connected to appropriate subbasins' retention capacity, which reduces water loss by evapotranspiration and favours percolation. Our results provide valuable information on the drought-generating process in the Cesar River basin. It can also be concluded that the MVRT (and other machine learning techniques that generate 'explainable AI' models based on progressive tree-like data partitioning and simplified models in leaves) is a relevant tool for defining drought management strategies. The tool helps researchers prioritise the areas most vulnerable to droughts and design strategies and interventions for the disturbed hydroclimatic parameters.

The study's limitations include its simplified approach to modelling a complex phenomenon using SWAT software (e.g. representing the groundwater components that impact hydrological drought conditions) and using only a single ML technique to build explainable models. Further extensions of this research may address these limitations. For example, candidate ML techniques could include M5 model trees (rather than regression trees), which have shown their effectiveness in solving water-related problems (see Solomatine & Dulal, 2004; Solomatine & Xue, 2004). These result in linear models in tree leaves, rather than constants like in MVRT.

The issue of combining human and artificial intelligence (and knowledge of physics with machine learning) is currently a point of great interest (see Jiang et al., 2020, on 'physics-aware deep learning models' and Moreido et al., 2021, on the role of experts in constraining machine-learning hydrological models). This study can be seen as one that contributes to developing and testing tools to better incorporate 'explanatory' ML techniques into existing modelling and management practices.


*Data Availability*. The data is available on request.

*Authors contributions*. All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Ana Paez-Trujillo. Jeffer Cañon developed the hydrological model. The first draft of the manuscript was written by Ana Paez-Trujillo, and all authors commented on previous versions of the manuscript. Ana Paez-Trujillo, Jeffer Cañon, Beatriz Hernandez, Gerald Corzo and Dimitri Solomatine read and approved the final manuscript.

*Competing interests.* The contact author has declared that neither she nor her co-authors have any competing interest.

*Acknowledgements.* The authors would like to express their gratitude to the Natura Foundation and the project GEF Magdalena–Cauca VIVE for proving the hydrological model of the Cesar River basin.

*Financial support.* This study was financially supported by the Ministry of Education Colombia, Programa Colombia Cientifica, Grant No. 3597287. The authors have no relevant financial interest to disclose.

**References**

Abbaspour, K. C., Vaghefi, S. A., & Srinivasan, R. (2018). A Guideline for Successful Calibration and Uncertainty Analysis for Soil and Water Assessment: A Review of Papers from the 2016 International SWAT Conference. *Water*, *10*(6). https://doi.org/10.3390/w10010006

Apurv, T., & Cai, X. (2020). Drought Propagation in Contiguous U.S. Watersheds: A Process-Based Understanding of the Role of Climate and Watershed Properties. *Water Resources Research*, *56, e2020WR027755*. https://doi.org/10.1029/2020WR027755

Arnold, J. G., Moriasi, D. N., Gassman, P. W., Abbaspour, K. C., White, M. J., Srinivasan, R., Santhi, C., Harmel, R. D., van Griensven, A., Liew, M. W. van, Kannan, N., Jha, M. K., Harmel, D., Member, A., Liew, M. W. van, & Arnold, J.-F. G. (2012). SWAT: MODEL USE, CALIBRATION, AND VALIDATION. *Transactions of the ASABE*, *55*(4), 1491–1508. http://swatmodel.tamu.edu

Borcard, D., Gillet, F., & Legendre, P. (2018). Cluster analysis. In *Numerical Ecology with R. Use R!* Springer, Cham. https://doi.org/10.1007/978-3-319-71404-2_4

Breiman, L. (1984). *Classification and Regression Trees* (1 st ed.). Routledge. https://doi.org/doi.org/10.1201/9781315139470

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

Cannon, A. J. (2012). Köppen versus the computer: Comparing Köppen-Geiger and multivariate regression tree climate classifications in terms of climate homogeneity. *Hydrology and Earth System Sciences*, *16*(1), 217–229. https://doi.org/10.5194/HESS-16-217-2012

https://doi.org/10.5194/nhess-2023-50
Preprint. Discussion started: 12 April 2023


Cornelis, W., Waweru, G., & Araya, T. (2019). Building Resilience Against Drought and Floods: The Soil-Water Management Perspective. In R. Lal & R. Francaviglia (Eds.), *Sustainable Agriculture Reviews 29. Sustainable Agriculture Reviews, vol 29*. Springer, Cham. https://doi.org/https://doi.org/10.1007/978-3-030-26265-5_6

De'ath, G. (2002). MULTIVARIATE REGRESSION TREES: A NEW TECHNIQUE FOR MODELING SPECIES-ENVIRONMENT RELATIONSHIPS. *Ecology*, *83*(4), 1105–1117.

Destouni, G., & Verrot, L. (2014). Screening long-term variability and change of soil moisture in a changing climate. *Journal of Hydrology*, *516*(1), 131–139. https://doi.org/10.1016/J.JHYDROL.2014.01.059

Ding, Y., Gong, X., Xing, Z., Cai, H., Zhou, Z., Zhang, D., Sun, P., & Shi, H. (2021). Attribution of meteorological, hydrological and agricultural drought propagation in different climatic regions of China. *Agricultural Water Management*, *255*. https://doi.org/10.1016/J.AGWAT.2021.106996

Ganguli P., Singh B., Reddy N.N, & et al. (2022). Climate-catchment-soil control on hydrological droughts in peninsular India. *Sci Rep*, *12 8014*. https://doi.org/doi.org/10.1038/s41598-022-11293-7

GEF, BID, & Fundación Natura. (2020). *Proyecto manejo sostenible y conservacion de la biodiversidad en la cuenca del Río Magdalena. Modelo hidrológico refinado 1 en la cuenca del Río Cesar*. https://drive.google.com/file/d/1X5-iuqRPAeCjIDE-pktc0mC9cV-qKOnM/view

GEF, BID, & Fundación Natura. (2021). *Proyecto manejo sostenible y conservacion de la biodiversidad en la cuenca del Río Magdalena. Modelo hidrológico refinado 2 en la cuenca del Río Cesar*. https://drive.google.com/file/d/1sECdhG_SOYICKpjhILBtGP3fqc6y1XTk/view

Hao, Z., Hao, F., Xia, Y., Feng, S., Sun, C., Zhang, X., Fu, Y., Hao, Y., Zhang, Y., & Meng, Y. (2022). Compound droughts and hot extremes: Characteristics, drivers, changes, and impacts. *Earth-Science Reviews*, 104241. https://doi.org/10.1016/J.EARSCIREV.2022.104241

Huang, S., Li, P., Huang, Q., Leng, G., Hou, B., & Ma, L. (2017). The propagation from meteorological to hydrological drought and its potential influence factors. *Journal of Hydrology*, *547*, 184–195. https://doi.org/10.1016/J.JHYDROL.2017.01.041

Iglesias, A., Assimacopoulos, D., & Van, L. H. A. J. (Eds. ). (2018). *Drought: science and policy*. John Wiley & Sons, Incorporated. https://doi.org/10.1002/9781119017073.ch1

Instituto de hidrología, meteorología y estudios ambientales (IDEAM). (2019). *Estudio Nacional del Agua 2018*.

Jehanzaib, M., Shah, S. A., Yoo, J., & Kim, T. W. (2020). Investigating the impacts of climate change and human activities on hydrological drought using non-stationary approaches. *Journal of Hydrology*, *588 (Article 125052)*. https://doi.org/10.1016/J.JHYDROL.2020.125052

Jiang, S., Zheng, Y., & Solomatine, D. (2020). Improving AI System Awareness of Geoscience Knowledge: Symbiotic Integration of Physical Approaches and Deep Learning. *Geophysical Research Letters*, *46, e2020GL088229*. https://doi.org/10.1029/2020GL088229

https://doi.org/10.5194/nhess-2023-50
Preprint. Discussion started: 12 April 2023


Jolliffe, I. . T. (2002). *Principal Component Analysis (Second, Ser. Springer series in statistics)*. Springer. https://doi.org/10.1007/B98835

Konapala, G., & Mishra, A. (2020). Quantifying Climate and Catchment Control on Hydrological Drought in the Continental United States. *Water Resources Research*, *56*, e2018WR024620. https://doi.org/10.1029/2018WR024620

Kuhn, M., & Johnson, K. (2013). Over-Fitting and Model Tuning. In *Applied Predictive Modeling*. Springer, New York, NY. https://doi.org/10.1007/978-1-4614-6849-3_4

Legendre, P., & Legendre, L. (2012a). Ordination in reduced space. In *Developments in Environmental Modelling* (Vol. 24, Issue C, pp. 425–520). https://doi.org/10.1016/B978-0-444-53868-0.50009-5

Legendre, P., & Legendre, L. (2012b). Cluster analysis. In *Developments in Environmental Modelling* (Vol. 24, pp. 337–424). https://doi.org/10.1016/B978-0-444-53868-0.50008-3

Li, J., Guo, Y., Wang, Y., Lu, S., & Chen, X. (2018). Drought propagation patterns under naturalized condition using daily hydrometeorological data. *Advances in Meteorology*, *2018*, Article ID 2469156, 14 pages. https://doi.org/10.1155/2018/2469156

Lu, J., Carbone, G. J., & Grego, J. M. (2019). Uncertainty and hotspots in 21st century projections of agricultural drought from CMIP5 models. *Sci Rep*, *9 4922*. https://doi.org/10.1038/s41598-019-41196-z

Manning, C., Widmann, M., Bevacqua, E., van Loon, A. F., Maraun, D., & Vrac, M. (2018). Soil Moisture Drought in Europe: A Compound Event of Precipitation and Potential Evapotranspiration on Multiple Time Scales. *Journal of Hydrometeorology*, *19*(8), 1255–1271. https://doi.org/10.1175/JHM-D-18-0017.1

Margariti, J., Rangecroft, S., Parry, S., Wendt, D. E., & Van Loon, A. F. (2019). Anthropogenic activities alter drought termination. *Elementa: Science of the Anthropocene 1*, *7 27*. https://doi.org/10.1525/elementa.365

Masroor, M., Sajjad, H., Rehman, S., Singh, R., Hibjur Rahaman, M., Sahana, M., Ahmed, R., & Avtar, R. (2022). Analysing the relationship between drought and soil erosion using vegetation health index and RUSLE models in Godavari middle sub-basin, India. *Geoscience Frontiers*, *13*(2), 101312. https://doi.org/10.1016/J.GSF.2021.101312

Mastrotheodoros, T., Pappas, C., Molnar, P., Burlando, P., Manoli, G., Parajka, J., Rigon, R., Szeles, B., Bottazzi, M., Hadjidoukas, P., & Fatichi, S. (2020). More green and less blue water in the Alps during warmer summers. *Nature Climate Change 2020 10:2*, *10*(2), 155–161. https://doi.org/10.1038/s41558-019-0676-5

Mckee, T. B., Doesken, N. J., & Kleist, J. (1993). The relationship of drought frequency and duration to time scales. *8th Conference on Applied Climatology, Anaheim, 17-22 January 1993*, 179–184.

Modarres, R. (2007). Streamflow drought time series forecasting. *Stochastic Environmental Research and Risk Assessment*, *21*(3), 223–233. https://doi.org/10.1007/S00477-006-0058-1/FIGURES/14

Molina-Navarro, E., Bailey, R. T., Andersen, H. E., Thodsen, H., Nielsen, A., Park, S., Jensen, J. S., Jensen, J. B., & Trolle, D. (2019). Comparison of abstraction scenarios simulated by SWAT and SWAT-MODFLOW. *Hydrological Sciences Journal*, *64*(4), 434–454. https://doi.org/10.1080/02626667.2019.1590583

https://doi.org/10.5194/nhess-2023-50
Preprint. Discussion started: 12 April 2023


Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd ed.)*. https://christophm.github.io/interpretable-ml-book/

Moreido, V., Gartsman, B., Solomatine, D. P., & Suchilina, Z. (2021). How Well Can Machine Learning Models Perform without Hydrologists? Application of Rational Feature Selection to Improve Hydrological Forecasting. *Water 2021, Vol. 13, Page 1696*, *13*(12), 1696. https://doi.org/10.3390/W13121696

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., & Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, *50*(3), 885–900.

Narasimhan, B., & Srinivasan, R. (2005). Development and evaluation of Soil Moisture Deficit Index (SMDI) and Evapotranspiration Deficit Index (ETDI) for agricultural drought monitoring. *Agricultural and Forest Meteorology*, *133*(1–4), 69–88. https://doi.org/10.1016/j.agrformet.2005.07.012

Peña-Gallardo, M., Vicente-Serrano, S. M., Hannaford, J., Lorenzo-Lacruz, J., Svoboda, M., Domínguez-Castro, F., Maneta, M., Tomas-Burguera, M., & Kenawy, A. el. (2019). Complex influences of meteorological drought time-scales on hydrological droughts in natural basins of the contiguous Unites States. *Journal of Hydrology*, *568*, 611–625. https://doi.org/10.1016/J.JHYDROL.2018.11.026

Prudhomme, C., Giuntoli, I., Robinson, E. L., Clark, D. B., Arnell, N. W., Dankers, R., Fekete, B. M., Franssen, W., Gerten, D., Gosling, S. N., Hagemann, S., Hannah, D. M., Kim, H., Masaki, Y., Satoh, Y., Stacke, T., Wada, Y., & Wisser, D. (2014). Hydrological droughts in the 21st century, hotspots and uncertainties from a global multimodel ensemble experiment. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(9), 3262–3267. https://doi.org/10.1073/PNAS.1222473110/SUPPL_FILE/PNAS.201222473SI.PDF

Rangecroft, S., Van Loon, A. F., Maureira, H., Verbist, K., & Hannah, D. M. (2019). An observation-based method to quantify the human influence on hydrological drought: upstream–downstream comparison. *Hydrological Sciences Journal*, *64*(3), 276–287. https://doi.org/10.1080/02626667.2019.1581365

Saft, M., Peel, M. C., Western, A. W., & Zhang, L. (2016). Predicting shifts in rainfall-runoff partitioning during multiyear drought: Roles of dry period and catchment characteristics. *Water Resources Research*, *52*(12), 9290–9305. https://doi.org/10.1002/2016WR019525

Saidi, H., Dresti, C., Manca, D., & Ciampittiello, M. (2018). Quantifying impacts of climate variability and human activities on the streamflow of an Alpine river. *Environmental Earth Sciences*, *77*(19), 1–16. https://doi.org/10.1007/S12665-018-7870-Z/TABLES/5

Santra, A., & Santra Mitra, S. (2020). Space-Time Drought Dynamics and Soil Erosion in Puruliya District of West Bengal, India: A Conceptual Design. *Journal of the Indian Society of Remote Sensing*, *48*(8), 1191–1205. https://doi.org/10.1007/S12524-020-01147-Y/TABLES/5

Seneviratne, S. I., Nicholls, N., Easterling, D., Goodess, C. M., Kanae, S., Kossin, Y., Luo, Y., Marengo, J., McInnes, K., Rahimi, M., Reichstein, M., Sorteberg, A., Vera, C., & Zhang, X. (2012). Changes in Climate Extremes and their Impacts

https://doi.org/10.5194/nhess-2023-50
Preprint. Discussion started: 12 April 2023


on the Natural Physical Environment. In C. B. Field, V. Barros, T. F. Stocker, D. Qin, D. J. Dokken, K. L. Ebi, M. D. Mastrandrea, Mach K.J., G.-K. Plattner, S. K. Allen, M. Tignor, & Midgley P.M (Eds.), *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation* (pp. 109–230). Cambridge University Press, UK, and New York, NY, USA.

Shah, D., & Mishra, V. (2020). Drought Onset and Termination in India. *Journal of Geophysical Research: Atmospheres*, *125*(15), e2020JD032871. https://doi.org/10.1029/2020JD032871

Shah, D., Shah, H. L., Dave, H. M., & Mishra, V. (2021). Contrasting influence of human activities on agricultural and hydrological droughts in India. *Science of the Total Environment*, *774*. https://doi.org/10.1016/J.SCITOTENV.2021.144959

Sheffield, J., & Wood, E. F. (2011a). The science of drought. In *Drought: Past Problems and Future Scenarios* (pp. 18–42). Taylor & Francis Group.

Sheffield, J., & Wood, E. F. (2011b). What is drought. In *Drought: Past Problems and Future Scenarios* (pp. 9–15). Taylor & Francis Group.

Solomatine, D. P., & Xue, Y. (2004). M5 Model Trees and Neural Networks: Application to Flood Forecasting in the Upper Reach of the Huai River in China. *Journal of Hydrologic Engineering*, *9*(6), 491–501. https://doi.org/10.1061/(asce)1084-0699(2004)9:6(491)

Stoelzle, M., Stahl, K., Morhard, A., & Weiler, M. (2014). Streamflow sensitivity to drought scenarios in catchments with different geology. *Geophysical Research Letters*, *41*(17), 6174–6183. https://doi.org/10.1002/2014GL061344

Teuling, A. J., van Loon, A. F., Seneviratne, S. I., Lehner, I., Aubinet, M., Heinesch, B., Bernhofer, C., Grünwald, T., Prasse, H., & Spank, U. (2013). Evapotranspiration amplifies European summer drought. *Geophysical Research Letters*, *40*(10), 2071–2075. https://doi.org/10.1002/GRL.50495

The World Bank. (2021). *Groundswell Part 2: Acting on Internal Climate Migration*. www.worldbank.org

Tijdeman, E., Barker, L. J., Svoboda, M. D., & Stahl, K. (2018). Natural and Human Influences on the Link Between Meteorological and Hydrological Drought Indices for a Large Set of Catchments in the Contiguous United States. *Water Resources Research*, *54*(9), 6005–6023. https://doi.org/10.1029/2017WR022412

Transactions of the ASABE (American Society of Agricultural and Biological Engineers). (2018). *Guidelines for Calibrating, Validating, and Evaluating Hydrologic and Water Quality (H/WQ) Models*. *61(4): 1393-1401*. https://doi.org/doi: 10.13031/trans.12806

Trnka, M., Semerádová, D., Novotný, I., Dumbrovský, M., Drbal, K., Pavlík, F., Vopravil, J., Štěpánková, P., Vizina, A., Balek, J., Hlavinka, P., Bartošová, L., & Žalud, Z. (2016). Assessing the combined hazards of drought, soil erosion and local flooding on agricultural land: A Czech case study. *Climate Research*, *70*(2–3), 231–249. https://doi.org/10.3354/cr01421

Universidad del Atlantico. (2014). *Plan de ordenamiento del recurso hidrico del Rio Cesar Formulacion Final*.

https://doi.org/10.5194/nhess-2023-50
Preprint. Discussion started: 12 April 2023


Universidad del Magdalena, CORPAMAG, & CORPOCESAR. (2017). *Documento sintesis para la declaratoria del complejo cenagso de la Zapatosa como area protegida*.

Valiya Veettil, A., & Mishra, A. k. (2020). Multiscale hydrological drought analysis: Role of climate, catchment and morphological variables and associated thresholds. *Journal of Hydrology*, *582*, 124533. https://doi.org/10.1016/J.JHYDROL.2019.124533

Van Lanen, H. A. J., Wanders, N., Tallaksen, L. M., & Van Loon, A. F. (2013). Hydrological drought across the world: Impact of climate and physical catchment structure. *Hydrology and Earth System Sciences*, *17*(5), 1715–1732. https://doi.org/10.5194/HESS-17-1715-2013

van Loon, A. F. (2015). Hydrological drought explained. *Wiley Interdisciplinary Reviews: Water*, *2*(4), 359–392. https://doi.org/10.1002/WAT2.1085

Vicente-Serrano, S. M., López-Moreno, J. I., Beguería, S., Lorenzo-Lacruz, J., Azorin-Molina, C., & Morán-Tejeda, E. (2011). Accurate Computation of a Streamflow Drought Index. *Journal of Hydrologic Engineering*, *17*(2), 318–332. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000433

Vicente-Serrano, S. M., Quiring, S. M., Peña-Gallardo, M., Yuan, S., & Domínguez-Castro, F. (2020). A review of environmental droughts: Increased risk under global warming? *Earth-Science Reviews*, *201*. https://doi.org/10.1016/J.EARSCIREV.2019.102953

Wang, M., Jiang, S., Ren, L., Xu, C. Y., Menzel, L., Yuan, F., Xu, Q., Liu, Y., & Yang, X. (2021). Separating the effects of climate change and human activities on drought propagation via a natural and human-impacted catchment comparison method. *Journal of Hydrology*, *603*. https://doi.org/10.1016/J.JHYDROL.2021.126913

Wildemeersch, J. C. J., Garba, M., Sabiou, M., Fatondji, D., & Cornelis, W. M. (2015). Agricultural drought trends and mitigation in Tillaberí, Niger. *Soil Science and Plant Nutrition*, *61*(3), 414–425. https://doi.org/10.1080/00380768.2014.999642

Wu, Y., Sun, J., Blanchette, M., Rousseau, A. N., Xu, Y. J., Hu, B., & Zhang, G. (2023). Wetland mitigation functions on hydrological droughts: From drought characteristics to propagation of meteorological droughts to hydrological droughts. *Journal of Hydrology*, *617*, 128971. https://doi.org/10.1016/J.JHYDROL.2022.128971

Xu, Y., Zhang, X., Wang, X., Hao, Z., Singh, V. P., & Hao, F. (2019). Propagation from meteorological drought to hydrological drought under the impact of human activities: A case study in northern China. *Journal of Hydrology*, *579*, 124147. https://doi.org/10.1016/J.JHYDROL.2019.124147

Zhang, X., Hao, Z., Singh, V. P., Zhang, Y., Feng, S., Xu, Y., & Hao, F. (2022). Drought propagation under global warming: Characteristics, approaches, processes, and controlling factors. *Science of The Total Environment*, *838*, 156021. https://doi.org/10.1016/J.SCITOTENV.2022.156021