Title
Assessing land cover change impact on peak discharge using principal component analysis and random forest with remote sensing data assimilation

Journal
Environmental Monitoring and Assessment, 198(9)

ISSN
0167-6369

Authors
Salar-Khorasani, Seyed-Mahdi
Bozorg-Haddad, Omid
Loáiciga, Hugo A

Publication Date
2026-08-03

DOI
10.1007/s10661-026-15757-5

Peer reviewed

Environ Monit Assess (2026) 198:905
https://doi.org/10.1007/s10661-026-15757-5

RESEARCH

Assessing land cover change impact on peak discharge using principal component analysis and random forest with remote sensing data assimilation

Seyed-Mahdi Salar-Khorasani · Omid Bozorg-Haddad · Hugo A. Loáiciga

Received: 23 September 2025 / Accepted: 25 July 2026

Abstract The effect of land cover changes on annual peak discharge is evaluated using a novel method combining principal component analysis (PCA) and random forest regression (RF). This approach calculates original feature contributions by multiplying absolute PCA loadings with RF-derived component importances, enabling direct attribution of discharge variations to specific land cover types in each subbasin, which represents an integration unexplored in prior PCA–RF applications. The PCA–RF method is applied in the basin upstream of Golestan Dam, Iran. It uses multitemporal Landsat imagery for NDVI-based land cover classification, annual instantaneous peak discharge data from hydrometric stations, and elevation data for subbasin delineation over a limited period (post-2003, due to data constraints). Despite limitations like scarce hydrometric data and a short collection period, the method analyzes relationships between land cover classes and floods in subbasins. Remote sensing provides spatial vegetation patterns, organized via GIS; PCA reduces dimensionality and identifies key factors; RF models nonlinear relationships with high accuracy (R^2 0.78–0.90, per SHAP insights). Quantitatively, urban/barren lands dominated in Tangerah (26.3%) and Hajji Qushan (30.9%) and thin vegetation in Galikash (26.9%) and Tamar Gorgan (26.8%), with dense vegetation lowest in Hajji Qushan (2%). Findings confirm land cover changes significantly impact flood peaks, varying spatially: Urban/barren lands accelerate runoff, and vegetation mitigates it. The PCA–RF method captures these dynamics despite constraints and offers policymakers insights for flood risk reduction, vegetation restoration, urban runoff control, sustainable management, and adaptive planning. Furthermore, this work provides insights into the interactions between land cover changes and flood risks that assist policymakers and land managers in developing effective strategies for flood risk reduction, water resource management, and sustainable use practices.

Keywords Land cover changes · Flood · Principal component analysis (PCA) · Random forest regression · Remote sensing

S.-M. Salar-Khorasani
Department of Irrigation & Reclamation Engineering, Faculty of Agricultural Engineering & Technology, College of Agriculture & Natural Resources, University of Tehran, Karaj, Tehran, Iran
e-mail: Mahdi.Salar.Kh@ut.ac.ir

O. Bozorg-Haddad (✉) · H. A. Loáiciga
Department of Geography, University of California, Santa Barbara, CA 93106, USA
e-mail: OBHaddad@ut.ac.ir

H. A. Loáiciga
e-mail: hloaiciga@ucsb.edu

Introduction

Floods are destructive natural events that impact infrastructure, the environment, and the economy

Published online: 03 August 2026

around the world (Islam et al., 2021; Khosravi et al., 2019). Between 1995 and 2015, approximately 150,000 flood events worldwide resulted in approximately 157,000 deaths (Islam et al., 2021). Every year, floods affect nearly 250 million people and inflict costs of about 40 billion dollars (Nguyen, 2022). Different types of floods, such as in rivers, coastal areas, and urban areas, exhibit variable characteristics and damages (Islam et al., 2021). The impact of flooding is particularly severe in developing countries due to the effects of climate change and land cover change (Janizadeh et al., 2021).

There is an urgent need to investigate the factors affecting the occurrence and severity of floods due to the increasing number of flood events. These factors include changes in land cover and vegetation, topography, and weather patterns (Rautela et al., 2022; Sofi et al., 2021). Human activities, such as deforestation, urbanization, levee and dam construction, alter flood intensity (Boroughani et al., 2025; Junger et al., 2022; Pal et al., 2022; Roy et al., 2022; Salar-Khorasani & Singh, 2026). Advanced flood risk assessment methods are necessary for effective flood risk management and mitigation (Nguyen et al., 2022a). Remote sensing and analysis of land cover change detection, including the use of the normalized difference vegetation index (NDVI), are critical tools for monitoring these changes and assessing their impacts on flooding (Jacobberger-Jellison, 1994; Peters et al., 1993).

Land cover and vegetation changes significantly affect flood dynamics. For example, deforestation reduces the surface roughness and the infiltration capacity of the land, thereby increasing the flood risk (Nguyen et al., 2022b; Sugianto et al., 2022). Effective conservation and management of water and land resources are essential for sustainability and provide benefits such as soil erosion control, runoff utilization, increased groundwater recharge, and reduced land degradation (Bhattacharyya et al., 2015). Various catchment characteristics, including size, shape, drainage pattern, and slope, affect infiltration rates and peak flood discharges (Rautela et al., 2022; Sofi et al., 2021).

The NDVI index is commonly used to evaluate and monitor the response of vegetation to environmental changes such as drought (Jacobberger-Jellison, 1994; Peters et al., 1993). By analyzing satellite images before and after these changes, researchers identify changes at large spatial and spectral scales. The effects of climate change, changes in land use patterns, and natural and anthropogenic changes have reinforced the reliance on remote sensing data and digital image change detection techniques to document changes in ecological structure and function (Collins & Woodcock, 1996).

Principal component analysis (PCA) is a technique that can be applied to transform satellite images with uncorrelated groups to detect land cover changes easily (Byrne et al., 1980; Jensen & Christensen, 1986). The combination of PCA and the NDVI index is a powerful tool for detecting changes in vegetation due to factors such as economic development, population growth, urbanization, and industrialization (Caocao et al., 2008) that significantly enhance flood risk assessments (Al-Juaidi et al., 2018; Chen et al., 2020; Khoirunisa et al., 2021; Tehrany et al., 2014; Wang et al., 2021).

Remote sensing and machine learning techniques have become increasingly prominent in monitoring and identifying land use/land cover (LULC) changes and their impacts on hydrologic processes (Alarifi et al., 2022; Dewan et al., 2006). For example, Nigeria has experienced significant changes in LULC due to natural and human factors, which necessitate an in-depth consideration of the relationship of these changes with floods and the adoption of suitable management strategies (Ighile & Shirakawa, 2020; Meyer & Turner, 1992). The creation of thematic maps using the classification of satellite images is a common application of remote sensing that provides reliable information for land and disaster management (Yan et al., 2015).

Understanding the dynamics of LULC changes is imperative to achieve sustainability because these changes significantly affect ecosystems and hydrologic regimes (Bhaduri et al., 2000; Menzel & Blongewicz, 2000). Furthermore, deforestation, among other land use changes, significantly alters surface runoff, peak discharge, and flood hazards (Hammer, 1972; Klein, 1979; Schueler, 1994). For these reasons, the use of remote sensing technology combined with GIS is valuable in the evaluation of LULC changes and planning management strategies (Reid et al., 2000; Singh, 1989).

Machine learning algorithms, such as the random forest (RF) algorithm, have shown excellent performance in predicting floods and creating maps of flood-prone areas (Madhuri et al., 2021;

Roy et al., 2020; Tehrany et al., 2019). Compared to other machine learning models like support vector machines (SVMs) or single decision trees, RF offers several advantages in flood prediction contexts, including reduced overfitting through ensemble learning (averaging multiple decision trees), robustness to noise and imbalanced data, efficient handling of missing values, and the ability to provide feature importance measures for better interpretability (Breiman, 2001; Lee et al., 2017; Mosavi et al., 2018). These benefits make RF particularly suitable for capturing complex, nonlinear relationships in hydrological data with high accuracy and generalization, as demonstrated in flood susceptibility mapping, where RF often outperforms alternatives in metrics like AUC and sensitivity (Lee et al., 2017). Combining these algorithms with PCA provides a powerful tool for analyzing the impact of land cover changes on peak flow and flood risk.

Assessing land cover change impact on peak discharge is of utmost importance. This topic has been tackled by numerous authors but gaps remain in the assessment (Bozorg-Haddad & Mariño, 2007; Bozorg-Haddad et al., 2009a; Bozorg-Haddad et al., 2009b; Bozorg-Haddad & Mariño, 2011; Bozorg-Haddad et al., 2016; Fallah-Mehdipour et al., 2011; Karimi-Hosseini et al., 2011; Soltanjalili et al., 2011; Sabbaghpour et al., 2012; Fallah-Mehdipour et al., 2013; Shokri et al., 2013; Orouji et al., 2014; Beygi et al., 2014). To address these gaps, this study aims to evaluate and rank the effects of land cover changes on the annual instantaneous peak discharge by applying a novel framework that integrates principal component analysis (PCA), random forest (RF) regression, and remote sensing data assimilation. This approach is explicitly demonstrated through a case study in the basin upstream of the Golestan Dam, Iran. Despite regional data limitations, such as scarce hydrometric records and a short collection period (post-2003), this methodology leverages GIS and machine learning to effectively capture the complex, nonlinear relationships between land cover dynamics and hydrological responses. Ultimately, by providing a clearer understanding of these interactions, this study seeks to offer actionable insights for policymakers regarding flood risk reduction, vegetation restoration, and adaptive planning, while highlighting pathways for future research using expanded datasets.

Materials and methods

The "Materials and methods" section begins with an introduction of the study area, highlighting its location and the reasons for its selection. This is followed by a description of the datasets used, including their sources and time periods. Finally, the research methodology is presented in a structured manner, outlining the overall approach and analyses performed in the study.

Case study

The study area is located upstream of the Golestan Dam in the Golestan province of Iran. This region was selected due to its history of frequent and severe flood events, making it a critical case for studying the impacts of land cover changes on hydrological responses and flood risk management. Notably, Golestan province has experienced some of Iran's most devastating floods, including the catastrophic August 2001 flash flood—one of the worst in over two centuries—which claimed over 300 lives, submerged vast areas of forests and farmland, and caused extensive infrastructure damage in the Gorgan River basin. More recent events, such as the massive 2019 inundation, affected over 8500 km^2, displaced indigenous communities, filled the Golestan Dam to capacity, and highlighted vulnerabilities in water resource management amid climate variability and land use changes. Additional floods in 2014, 2016, and 2017 further underscore the area's susceptibility, with heavy rains leading to billions in damages. The basin's importance lies in its role as a key water resource hub for Golestan province, supporting agriculture, ecosystems, and downstream populations, while its diverse topography and climate exacerbate flood risks, thus necessitating advanced analytical methods for achieving sustainable mitigation strategies (Ardalan et al., 2009; Khosravi et al., 2019; Tourani & Çağlayan, 2021). This area is important in terms of water resources and flood management. The study area is located between latitudes 37.33° N and 37.42° N and longitudes 55.32° E and 55.90° E. The area of this watershed is 5151.70 km^2, which includes large parts of Golestan province. Important rivers, such as the Gorganrud and its branches, flow in this area, which direct the surface water and groundwater to the Golestan Dam. In terms of topography, this area has a significant elevation variation, with a difference of up to 2521 m between its lowest and highest points. This elevation difference contributes to the formation of complex water flows that are essential for hydrological management. The area, encompassing the Eastern Alborz Mountains, features high and semihumid plains with climates ranging from humid to semiarid. Vegetation varies from Hyrcanian forests in elevated regions to pastures and agricultural lands in the lower areas. As shown in Fig. 1, the region's ecological richness and diverse topography foster numerous streams and rivers that ultimately flow into the Golestan Dam.

Data

Shuttle Radar Topography Mission (SRTM) elevation data were acquired to provide a detailed topographical context (U.S. Geological Survey [USGS], 2024b). To depict this hydrological setting, satellite imagery from the Landsat archive (USGS Earth Explorer) (U.S. Geological Survey [USGS], 2024c) was used. The imagery was processed in QGIS (version 3.34) (QGIS.org, 2024), while subbasin delineation was carried out in HEC-HMS (version 4.11) (U.S. Army Corps of Engineers [USACE], 2024). The resulting basin and subbasin boundaries, together with the locations of hydrometric stations, were subsequently overlaid on the satellite image. Within this context, data on annual peak discharge across the Tangerah, Galikesh, Hajji Qushan, and Tamar Gorgan subbasins, as detailed in Table 1, are essential for understanding fluctuations in water flow. The annual instantaneous peak discharge data were obtained from Iran's Water Resources Management Company for the years 2004, 2009, 2013, 2016, and 2019 (Iranian Water Resources Management Company [IWRM], 2024). Furthermore, to monitor environmental changes over time,

Table 1 Annual instantaneous peak discharge (m^3/s) in the Tangerah, Galikesh, Hajji Qushan, and Tamar Gorgan subbasins

Year	Tangerah	Galikesh	Hajji Qushan	Tamar Gorgan
2004	36.5	15.1	185	127
2009	31	15.2	14	39
2013	26	53.2	50.6	39
2016	25	65	103	59
2019	151	310	130	230

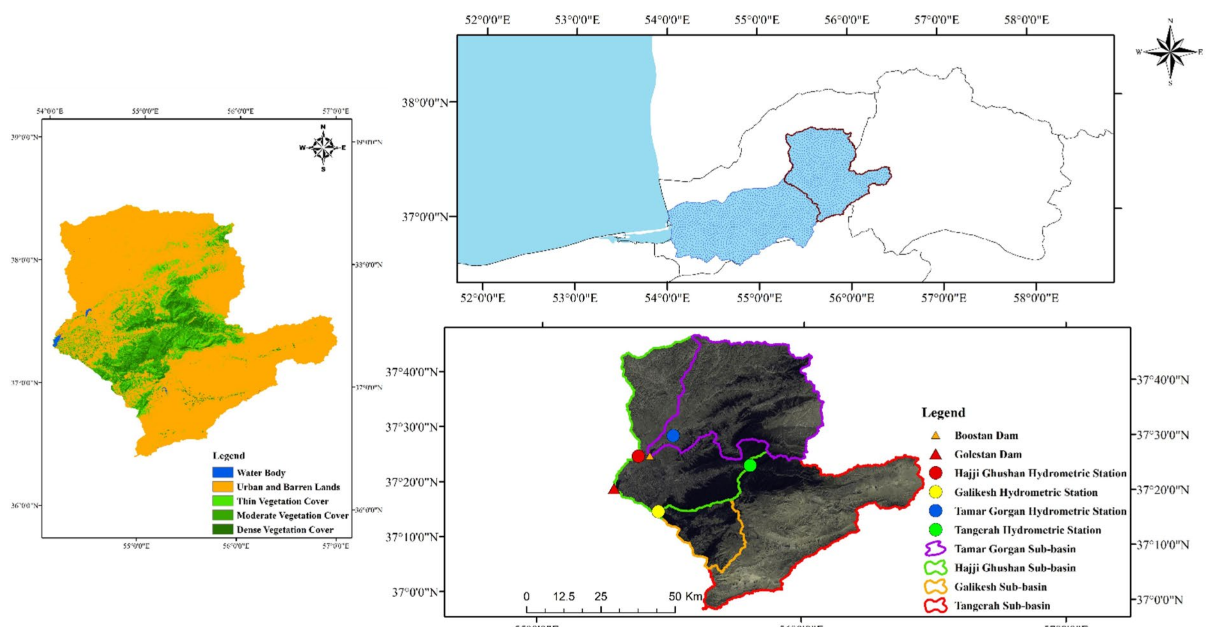


Fig. 1 Location of selected hydrometric stations, basin and subbasins upstream of Golestan Dam, and land use/land cover map of the study area

Landsat 5, 7, and 8 satellite images of collection 2 and level 2 with a resolution of 30 m, corresponding to October of the same years, were used. To ensure temporal consistency and minimize the influence of seasonal vegetation variability on NDVI values and land cover classification results, all images were intentionally selected from the same month (October) across the study period. These images, all projected in the UTM coordinate system, zone 40, level 2, provided critical spatial and temporal insights into land cover changes, as detailed in Table 2 (U.S. Geological Survey [USGS], 2024c). Hydrometric stations upstream of the Golestan Dam, whose details are outlined in Table 3, also play a fundamental role, as their data enable tracking of river dynamics and hydrological patterns. These are crucial elements for this research.

Methodology

The computational steps in the methodology section of this research begin with the selection of the flood risk watershed. The next steps include collecting pertinent datasets (annual instantaneous peak discharge, elevation data, and satellite images) and delineating the basin and subbasin using HEC-HMS (Hydrologic Engineering Center [HEC], 2024). Hydrometric stations are assigned based on HEC-HMS procedures. Following this image processing is performed, and land cover classification is made using the NDVI index. Principal component analysis (PCA) is applied next and the contribution of the main features is calculated. This paper's methodology ends by running a random forest algorithm regression to generate results, reporting R^2, MSE, and RMSE values, and calculating the total contributions of each feature or land cover class as shown in Fig. 2.

The annual instantaneous peak discharge, satellite imagery, and elevation date

Instantaneous peak discharge data were utilized alongside multitemporal Landsat satellite imagery and topographical data to analyze land feature changes over time. The satellite images, with suitable spatial resolution, were processed through steps including mosaicking and creating false-color composites to identify relevant land features. Additionally, elevation data were used to delineate subbasins and provide detailed topographical characterization, facilitating the examination of interactions between peak discharge trends and land features.

Basin and subbasin delineation and hydrometric station assignment

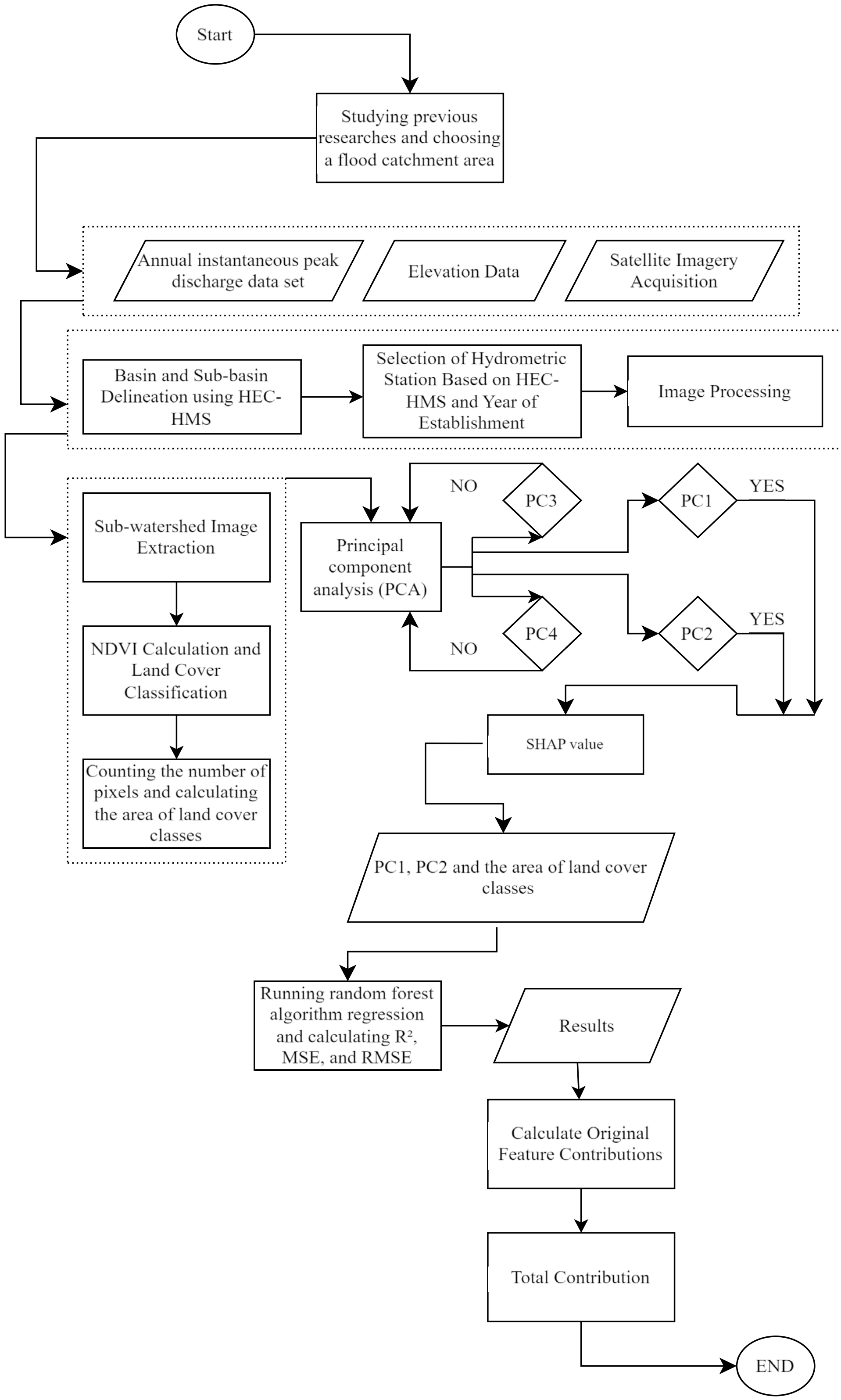
The catchment area and upstream subbasins were delineated using HEC-HMS software (U.S. Army Corps of Engineers [USACE], 2024). Subbasins were

Table 2 Specifications of Landsat satellite imagery data used in this study

Satellite	Sensor	Imaging date	Number of bands	Projection system	UTM zone	Level
LANDSAT 8	OLI/TIRS	2019-10–05	11	UTM	40	Level 2
LANDSAT 8	OLI/TIRS	2016-10–12	11	UTM	40	Level 2
LANDSAT 8	OLI/TIRS	2013-10–04	11	UTM	40	Level 2
LANDSAT 5	TM	2009-10–25	4	UTM	40	Level 2
LANDSAT 7	ETM +	2004-10–03	9	UTM	40	Level 2

Table 3 Specifications of hydrometric stations upstream of Golestan Dam (IWRM, 2024)

Row	Construction year	Elevation (m)	Latitude (°)	Longitude (°)	River name	Station name	Station code
1	1965	330	37.40	55.80	Dugh	Tangerah	12-001
2	1969	249	37.49	55.51	Gorganrud	Tamar Gorgan	12-005
3	1965	250	37.26	55.45	Oghan	Galikesh	12-007
4	1983	45	37.42	55.37	Gorganrud	Hajji Qushan	12–063


Start
Studying previous researches and choosing a flood catchment area
Annual instantaneous peak discharge data set
Elevation Data
Satellite Imagery Acquisition
Basin and Sub-basin Delineation using HEC-HMS
Selection of Hydrometric Station Based on HEC-HMS and Year of Establishment
Image Processing
Sub-watershed Image Extraction
NDVI Calculation and Land Cover Classification
Counting the number of pixels and calculating the area of land cover classes
Principal component analysis (PCA)
NO
PC3
PC1
YES
PC4
NO
PC2
YES
SHAP value
PC1, PC2 and the area of land cover classes
Running random forest algorithm regression and calculating R², MSE, and RMSE
Results
Calculate Original Feature Contributions
Total Contribution
END

◂**Fig. 2** Methodology flowchart for the watershed study, showing key steps including satellite image acquisition, NDVI calculation and land cover classification, PCA for feature reduction, random forest regression, and calculation of feature contributions

defined based on hydrometric data, considering factors such as the availability of flood records and the spatial arrangement of monitoring stations. Hydrometric stations were assigned to the subbasins according to their relevance to the study objectives. Detailed figures and tables illustrating these assignments will be included in the case study section.

Image processing

Image preprocessing is essential for minimizing errors and obtaining interpretable data. The effects of artifacts resulting from sensor products and atmospheric interferences were considered in this work to accurately recognize and separate relevant ground features based on reflectivity spectra. Landsat collection 2 level 2 images, which have undergone preapplied geometric and radiometric corrections by the USGS, were used to enhance the reliability of this work's analysis. Additionally, images with the lowest cloud cover were manually selected by comparing all available images within the same month to further minimize atmospheric disturbance. This procedure was performed using QGIS (QGIS.org, 2024), incorporating automated and manual preprocessing techniques through classification tools. Blue, green, red, NIR, SWIR1, and SWIR2 bands were utilized, while less significant bands were excluded. All images were corrected and displayed using the Transverse Mercator projection. Missing data, primarily from Landsat 7 ETM + images, were estimated using the Raster Interpolation tool, which interpolates missing values based on neighboring pixel information to ensure spatial continuity.

The Landsat Gap Fill tool is widely used to correct the SLC-off error of Landsat 7 ETM + images, which occurred after the scan line corrector failure in 2003, causing data gaps and black banding. In this study, the TIN-based interpolation method in QGIS was applied, interpolating missing pixels from adjacent valid data by forming triangles between surrounding pixels (U.S. Geological Survey [USGS], 2024a). Landsat 7 images from 2004 were corrected using this approach, after which the mosaicked images were clipped to the delineated subwatersheds for subsequent analysis.

NDVI calculation and land cover classification

The range of values of the NDVI index, which is between − 1 and 1, was calculated using single-date Landsat images acquired in October of each study year based on Eq. (1). For all subbasins, the land cover classes obtained using the NDVI value ranges were classified following Akbar et al. (2019). The urban land and barren land classes were combined to minimize potential classification errors, as listed in Table 4. Additionally, the water body class was removed because it was not consistently present across all subbasins and study years. Consequently, each subbasin was classified into four categories: urban and barren lands, thin vegetation, moderate vegetation, and dense vegetation.

$$NDVI = \frac{(NIR - \mathrm{Re}d)}{(NIR + \mathrm{Re}d)} \tag{1}$$

where *NIR* and *RED* denote the near-infrared and red bands of satellite images, respectively.

Statistical and machine learning analyses

This section describes the statistical and machine learning framework adopted to analyze the relationship between land cover changes and annual peak discharge. First, principal component analysis (PCA) is applied to reduce data dimensionality and extract dominant patterns of land cover variability. These components are then used as inputs to a random forest model to predict peak discharge, followed by a combined analysis of PCA loadings and random forest feature importance to quantify the contribution of individual land cover classes.

Principal component analysis (PCA) The Python programming language was implemented to perform the principal component analysis (PCA) (Microsoft, 2024). The PCA method was applied to reduce the number of land cover classes to two principal components by considering the maximum explained variance (explained variance was obtained from the eigenvalue calculations). This approach simplifies the

Table 4 NDVI range used for land cover classification

Class	NDVI range
Water bodies	−0.28 to 0.015
Urban and barren lands	0.015 to 0.18
Thin vegetation	0.18 to 0.27
Moderate vegetation	0.27 to 0.36
Dense vegetation	0.36 to 0.74

model by converting the input data into a set of main components and minimizes the risk of overfitting due to the reduced number of input data for each year.

PCA provides an alternative to subjective variable selection by objectively simplifying a large number of variables into a few uncorrelated factors that reveal most of the variability in the underlying data (Abdi & Williams, 2010). The PCA approach increases flexibility regarding the selection and number of variables, thereby allowing for a more robust and consistent set of variables. It also provides several potential advantages over aggregating spatially explicit and uncorrelated variables (Abson et al., 2012).

It should be noted that the principal components derived in this study are not intended to represent isolated physical land cover variables. Instead, they are latent constructs that summarize dominant patterns of land cover change over time, with each component being characterized by the relative contributions (loadings) of multiple land cover classes. The primary objective of applying PCA is to highlight years in which structural land cover changes were more pronounced, thereby improving the signal-to-noise ratio of the input data used for machine learning analysis.

The vulnerability index is constructed as a function of the principal components (PC) and their subsequent aggregation. PC aggregation refers to a method used to combine transformed, normalized, and weighted indices into a simpler metric, reducing the amount and complexity of information that must be used during the process of classifying areas into vulnerability classes (Nguyen et al., 2016). It is important to note that existing aggregation methods do not guarantee that these regions have similar characteristics in terms of the value variables of interest and, therefore, vulnerability. In addition, the choice of aggregation method affects the results and makes PC aggregation a subjective decision in the index construction process (Böhringer & Jochem, 2007).

Random forest analysis The random forest algorithm and the selected acceptable principal components were implemented to perform regression analysis between the area of the cover classes and the annual instantaneous peak discharge. The random forest (RF) algorithm (Breiman, 2001) solves classification and regression problems. The RF algorithm is a combination of a bagging algorithm and a random subspace method. RF works by constructing many decision trees (DT) following three main steps:

1. Using the bootstrap method to build a data subset similar in size to the original dataset.
2. Making the DT based on subsets.
3. Predicting the outcome by majority voting.

RF handles multidimensional data well, and its robustness is evident in its ability to accommodate missing data (Chen et al., 2020). Its performance depends on the number of decision trees (Ntree) and the number of variables used to build the tree (mtry). The model was trained to predict peak discharge rates, and feature importance was calculated based on the reduction in mean squared error (MSE).

RF is proficient in managing large datasets, capturing intricate associations, and mitigating overfitting concerns. Its application in assessing and forecasting flood-prone regions has gained popularity (Chen et al., 2020; Roy et al., 2020). RF can integrate diverse influencing factors, accommodate nonlinear associations, and yield precise and dependable outcomes. The model's thoroughness is evident in its comprehensive analysis of input variables, which facilitates the identification of pivotal factors contributing to flood susceptibility. The selection of the RF model for flood susceptibility modeling was based on its superior performance compared to other algorithms. RF produces accurate results with lower variances and biases and higher precision rates (Chen et al., 2020; Pourhashemi et al., 2025).

Calculation of feature contributions The contribution of major land cover features in predicting annual peak flow discharge was calculated by combining PCA loadings with feature importance obtained from the random forest model. This process involves the following computational steps.

Principal component analysis (PCA) loadings PCA is a crucial step in our method, as it effectively reduces the data's dimensionality while preserving as much variance as possible. This transformation involves finding the eigenvalues and eigenvectors of the covariance matrix S, key variables in the implementation of PCA.

A crucial step of PCA involves solving the eigenvalue problem for the covariance matrix of the dataset. This problem is represented by Eq. (2).

$$S\,a = \lambda\,a \tag{2}$$

where S, a, and λ denote respectively the sample covariance matrix, eigenvector, and the corresponding eigenvalue. The eigenvalues represent the variances of the principal components, and the eigenvectors represent the directions of these components (Jolliffe & Cadima, 2016).

Random forest feature importances This paper's methodology also leverages the power of the random forest model to determine feature importance. Each importance score i_k indicates the significance of the k-th principal component in predicting the target variable (annual peak discharge rates), facilitating the interpretation of the research analysis's results.

The importance score i_k is given by Eq. (3) for each principal component. k is computed by evaluating the error reduction (mean squared error) at each node in the decision trees where the principal component k is used to split the data. For each tree (t) in the random forest, the importance of the principal component k is calculated by summing the reductions in prediction error across all nodes that use k as a splitting feature.

$$i_k^{(t)} = \sum\nolimits_{j \in \, nodes \; using \; k} (reduction \; in \; error \; at \; node \; j) \tag{3}$$

where $i_k^{(t)}$ = the importance score of the k-th principal component in the t-th tree; j = index for the nodes in the tree that use the k-th principal component for splitting; nodes using k = the set of all nodes in the t-th tree that use the k-th principal component as a splitting feature; and (reduction in error at node j) = the amount by which prediction error decreases when the split at node j is made using the k-th principal component.

The overall importance score i_k given by Eq. (4) for the principal component k across the entire forest is obtained by averaging these scores across all m trees in the forest.

$$i_k = \frac{1}{m}\sum\nolimits_{t=1}^{m} i_k^{(t)} \tag{4}$$

where i_k = the aggregated importance score of the k-th principal component across the entire random forest and m = the total number of trees in the random forest.

Total contribution

For each original feature j, the contribution to the flow rate prediction is calculated by summing the products of the absolute values of the PCA loadings $\left|p_{kj}\right|$ and the corresponding feature importances i_k from the random forest model across the selected principal components, which are chosen based on their high variance, and, thus, their significant contribution to the prediction of the annual peak flow.

The formula for the contribution C_j of the j-th original feature is given by Eq. (5).

$$C_j = \sum\nolimits_k \left|p_{kj}\right| \times i_k \tag{5}$$

where C_j = contribution of the j-th original feature; $\left|p_{kj}\right|$ = absolute value of the PCA loading for the j-th original feature and the k-th principal component; i_k = importance of the k-th principal component from the random forest model; and k = index that ranges over the number of principal components.

The overall contribution of the data features is given by Eq. (6), which represents the sum of the contributions of all the principal components to the variance of the dependent variable (i.e., the annual peak flow).

$$C_{total} = \sum\nolimits_j C_j \tag{6}$$

Equation (6) ensures that every aspect of the data is considered in the analysis. These calculations provide a measure of how each land cover type (original feature) contributes to the prediction of discharge rates, therefore helping to identify the most influential land cover types in each subbasin. This methodology is grounded in the established principles of PCA and the interpretation of feature importance from random forest models (Abedi et al., 2021).

Results and discussion

The study provides an overview of factors affecting flow prediction across the Tangerah, Galikesh, Hajji Qushan, and Tamar Gorgan subbasins. PCA identifies key components for subsequent analyses, while the random forest model performance is evaluated using R^2, MSE, RMSE, and SHAP values. The contributions of land cover types are also discussed, offering insights into their influence on flow predictions. The following sections present detailed results for each analysis.

PCA results

Figure 3 shows graphs of the variance explained by each of the principal components (PCs), while Table 5 presents the PCA loadings for land cover classes across the subbasins. In each graph, the first principal component (PC1) records the highest variance, followed by the second principal component

Table 5 Principal component loadings and explained variance for land cover classes across four subbasins, representing the PCA component matrix

Subbasin	Class	PC1	PC2
Tangerah	Urban and barren lands	−0.50	−0.52
	Thin vegetation cover	−0.31	0.79
	Moderate vegetation cover	0.58	0.21
	Dense vegetation cover	0.57	−0.25
	Variance explained	69%	30%
Galikesh	Urban and barren lands	−0.48	−0.60
	Thin vegetation cover	−0.46	0.66
	Moderate vegetation cover	0.53	0.33
	Dense vegetation cover	0.53	−0.30
	Variance explained	81%	18%
Hajji Qushan	Urban and barren lands	−0.49	0.71
	Thin vegetation cover	−0.49	−0.70
	Moderate vegetation cover	0.51	−0.03
	Dense vegetation cover	0.50	0.04
	Variance explained	64%	31%
Tamar Gorgan	Urban and barren lands	0.58	−0.22
	Thin vegetation cover	−0.35	0.76
	Moderate vegetation cover	−0.57	−0.23
	Dense vegetation cover	−0.47	−0.56
	Variance explained	70%	28%

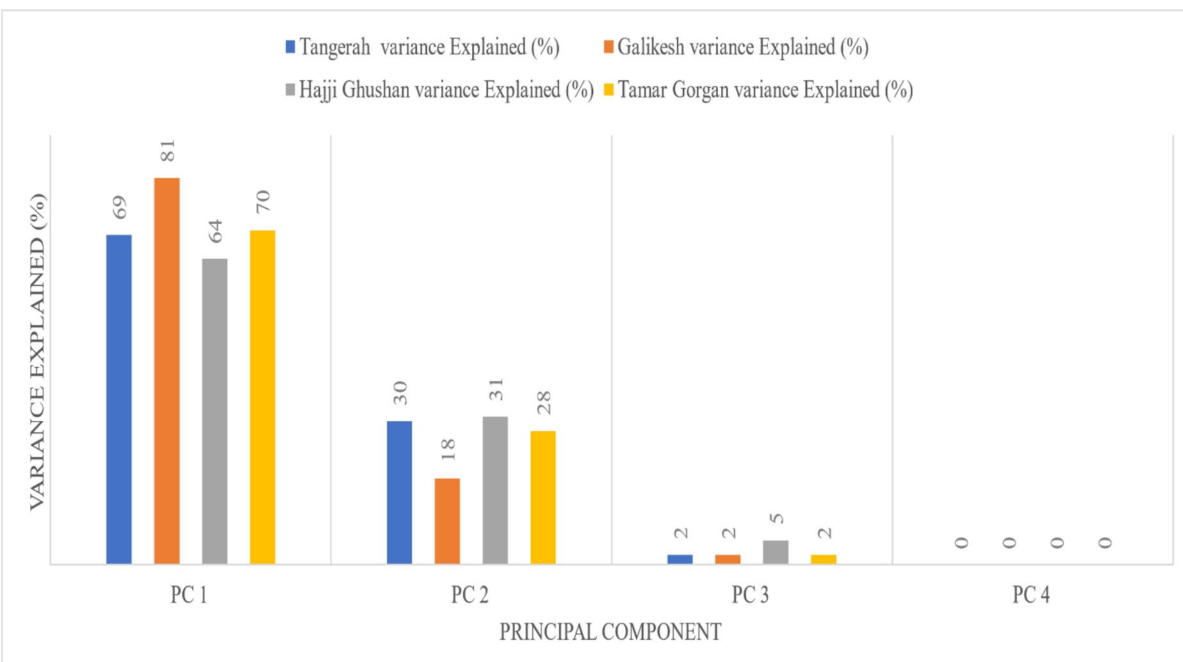


Fig. 3 The variance explained by principal components in the four subbasins: Tangerah, Galikesh, Hajji Qushan, and Tamar Gorgan

(PC2). The components PC3 and PC4 gradually explain less variance and become almost insignificant. The variance explained by PC1 is high in all subbasins, fluctuating between 64 and 81%, which demonstrates its dominant role in explaining the variance in land cover classes. In comparison, PC2 accounts for 18 to 31% of the variance, playing a supporting role in explaining the distribution between cover classes and adding depth to the analysis. Given this pattern, PC1 and PC2 were selected for further analysis, as they collectively explain most of the variance in the data. This choice reduces the dimensionality while preserving the most informative features of the dataset, ensuring the representativeness of the analysis.

The PCA loadings indicate that in the Tangerah subbasin, urban and barren lands have negative loadings for PC1 and PC2, while thin vegetation cover has positive loading for PC2, and moderate and dense vegetation show positive loadings for PC1. The Galikash subbasin exhibits a similar pattern: Urban and barren lands have negative loadings on both components, and thin vegetation is strongly positive with respect to PC2. The Hajji Qushan subbasin presents a different trend, such that urban and barren lands have positive loadings for PC2 and negative loadings for PC1, while thin vegetation shows a negative loading on both components. The Tamar Gorgan subbasin exhibits urban and barren lands that have positive loadings for PC1 and negative loadings for PC2, and thin vegetation has a positive loading for PC2.

The Tangerah subbasin

The survey of land cover changes in the Tangerah subbasin during 2004, 2009, 2013, 2016, and 2019 reveals significant temporal dynamics. Figures 4 and 8a and Table 6 illustrate these fluctuations, showing that urban and barren lands peaked at 1632.70 km^2 in 2009 before decreasing to 1499.21 km^2 in 2019.

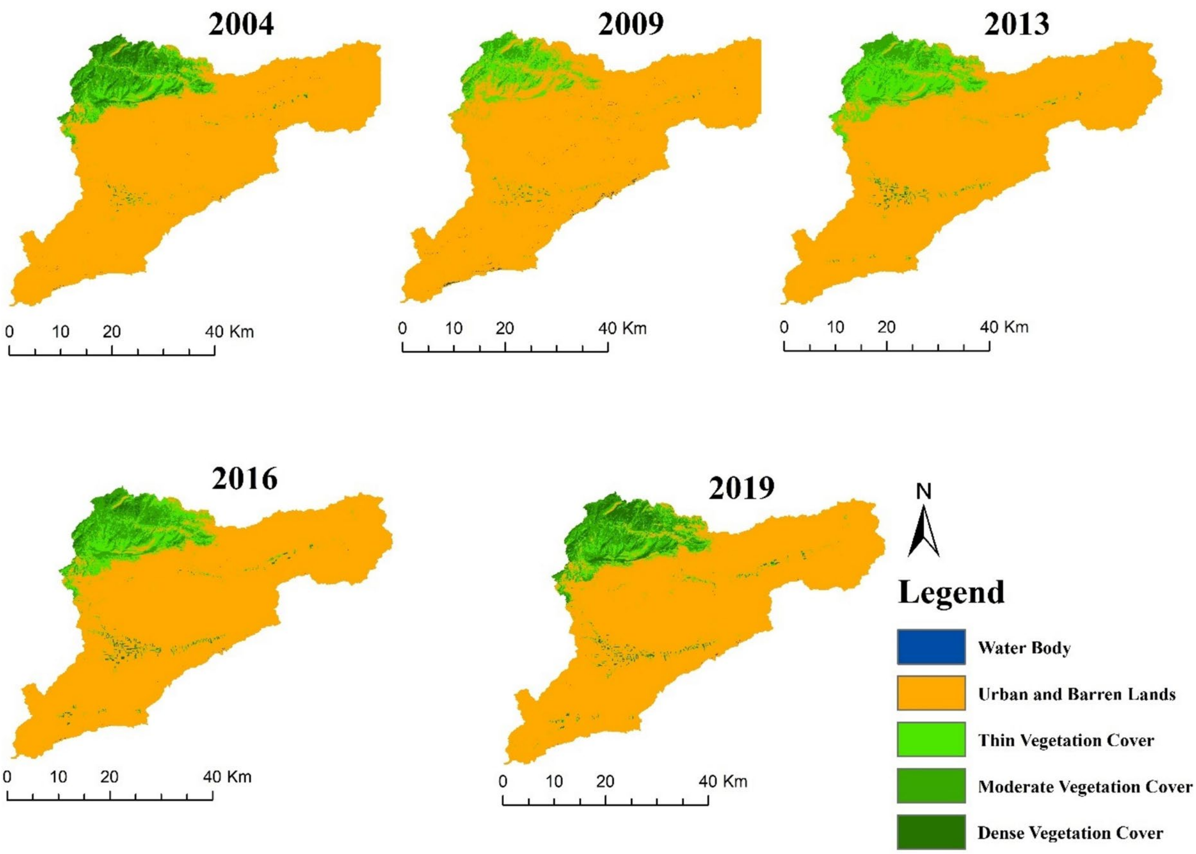


Fig. 4 Land cover class maps of the Tangerah subbasin for the years 2004, 2009, 2013, 2016, and 2019

Simultaneously, thin vegetation cover declined from 95.79 km^2 in 2004 to 95.46 km^2 in 2019, with a notable increase in 2013. Moderate vegetation generally decreased, while dense vegetation increased from 46.06 km^2 in 2004 to 61.80 km^2 in 2019, despite some fluctuations.

The Galikesh subbasin

The survey of land cover changes in the Galikesh subbasin from 2004 to 2019 reveals significant temporal dynamics. Figures 5 and 8b and Table 6 illustrate these fluctuations, showing that urban and barren lands peaked at 256.02 km^2 in 2009 before decreasing to 176.88 km^2 in 2019. Thin vegetation cover declined from 84.77 km^2 in 2004 to 70.35 km^2 in 2019, with variations in between. Moderate vegetation increased from 97.40 km^2 in 2004 to 105.94 km^2 in 2019, reaching its lowest level of 32.43 km^2 in 2009. Dense vegetation initially dropped from 20.46 km^2 in 2004 to 0.12 km^2 in 2013 before rising to 43.04 km^2 in 2019.

The Hajji Qushan subbasin

The survey of land cover changes in the Hajji Qushan subbasin from 2004 to 2019 reveals significant temporal dynamics. Figures 6 and 8c and Table 6 illustrate these variations, showing that urban and barren lands increased from 814.95 km^2 in 2004 to 908.69 km^2 in 2019, peaking at 970.40 km^2 in 2009. Thin vegetation initially rose from 170.43 km^2 in 2004 to 290.25 km^2 in 2009, fluctuated over the years, and finally declined to 139.47 km^2 in 2019. Moderate vegetation reached 95.87 km^2 in 2009 and further increased to 225.39 km^2 in 2019. Dense vegetation followed a decreasing trend, dropping from 106.21

Table 6 Area of land cover classes (km^2) in the Tangerah, Galikesh, Hajji Qushan, and Tamar Gorgan subbasins for the years 2004, 2009, 2013, 2016, and 2019

Year	Area of urban and barren lands (km^2)	Area of thin vegetation cover (km^2)	Area of moderate vegetation cover (km^2)	Area of dense vegetation cover (km^2)
Tangerah subbasin				
2004	1516.08	95.79	133.00	46.06
2009	1632.70	113.00	40.69	0.80
2013	1524.17	181.05	83.48	3.21
2016	1500.05	136.19	132.53	22.71
2019	1499.21	95.46	135.13	61.80
Galikesh subbasin				
2004	194.73	84.77	97.40	20.46
2009	256.02	108.45	32.43	0.32
2013	209.46	130.64	56.91	0.12
2016	165.84	104.38	108.31	17.99
2019	176.88	70.35	105.94	43.04
Hajji Qushan subbasin				
2004	814.95	170.43	276.55	106.21
2009	970.40	290.25	95.87	2.24
2013	943.52	238.31	177.62	4.77
2016	897.04	180.15	249.40	39.19
2019	908.69	139.47	225.39	90.78
Tamar Gorgan subbasin				
2004	1324.89	156.71	104.48	7.26
2009	1471.66	103.93	12.07	0.046
2013	1390.40	163.86	37.12	0.069
2016	1325.87	208.58	55.80	0.57
2019	1313.19	156.65	106.20	14.78

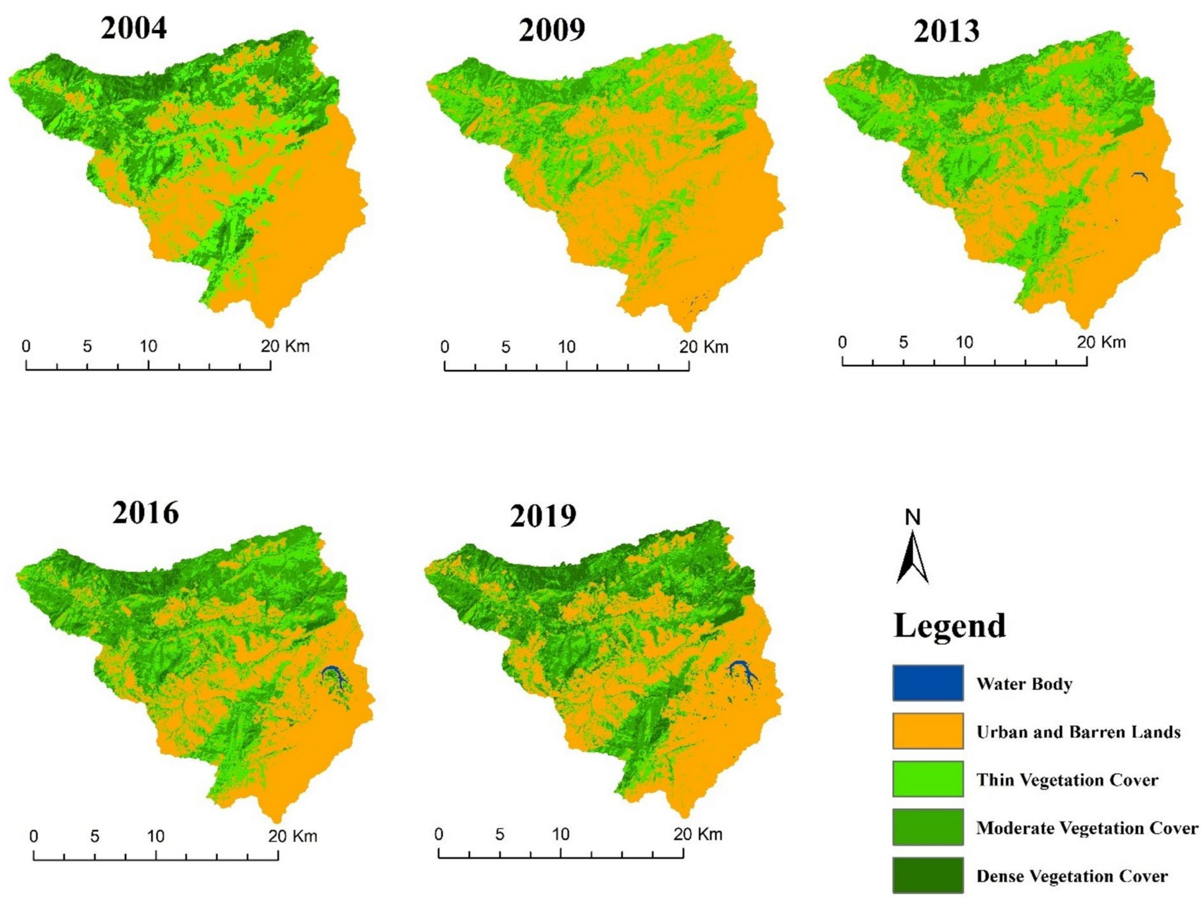


Fig. 5 Land cover class maps of the Galikesh subbasin for the years 2004, 2009, 2013, 2016, and 2019

km^2 in 2004 to 90.78 km^2 in 2019, with its lowest value recorded in 2009.

The Tamar Gorgan subbasin

The survey of land cover changes in the Tamar Gorgan subbasin from 2004 to 2019 highlights significant temporal dynamics. Figures 7 and 8d and Table 6 illustrate these variations. Urban and barren lands increased from 1324.89 km^2 in 2004 to 1331.19 km^2 in 2019, reaching a peak of 1471.66 km^2 in 2009. Thin vegetation fluctuated, starting at 156.71 km^2 in 2004, dropping to 103.93 km^2 in 2009, and then increasing again to 156.65 km^2 by 2019. Moderate vegetation followed a fluctuating pattern, rising from 104.48 km^2 in 2004 to 106.20 km^2 in 2019, with its lowest point recorded at 12.06 km^2 in 2009. Dense vegetation initially decreased from 7.26 km^2 in 2004 to 0.57 km^2 in 2016 but then increased to 14.78 km^2 by 2019.

The RF results

In the following sections, the performance of the random forest model is evaluated across the studied subbasins. Key metrics, including R^2, MSE, and RMSE, are presented, along with SHAP value analysis to interpret the contribution of principal components and land cover types to peak discharge predictions. Overall, the model demonstrates strong performance, explaining 78–90% of the variance in discharge, and highlights the key role of vegetation and impervious surfaces in controlling watershed hydrological responses.

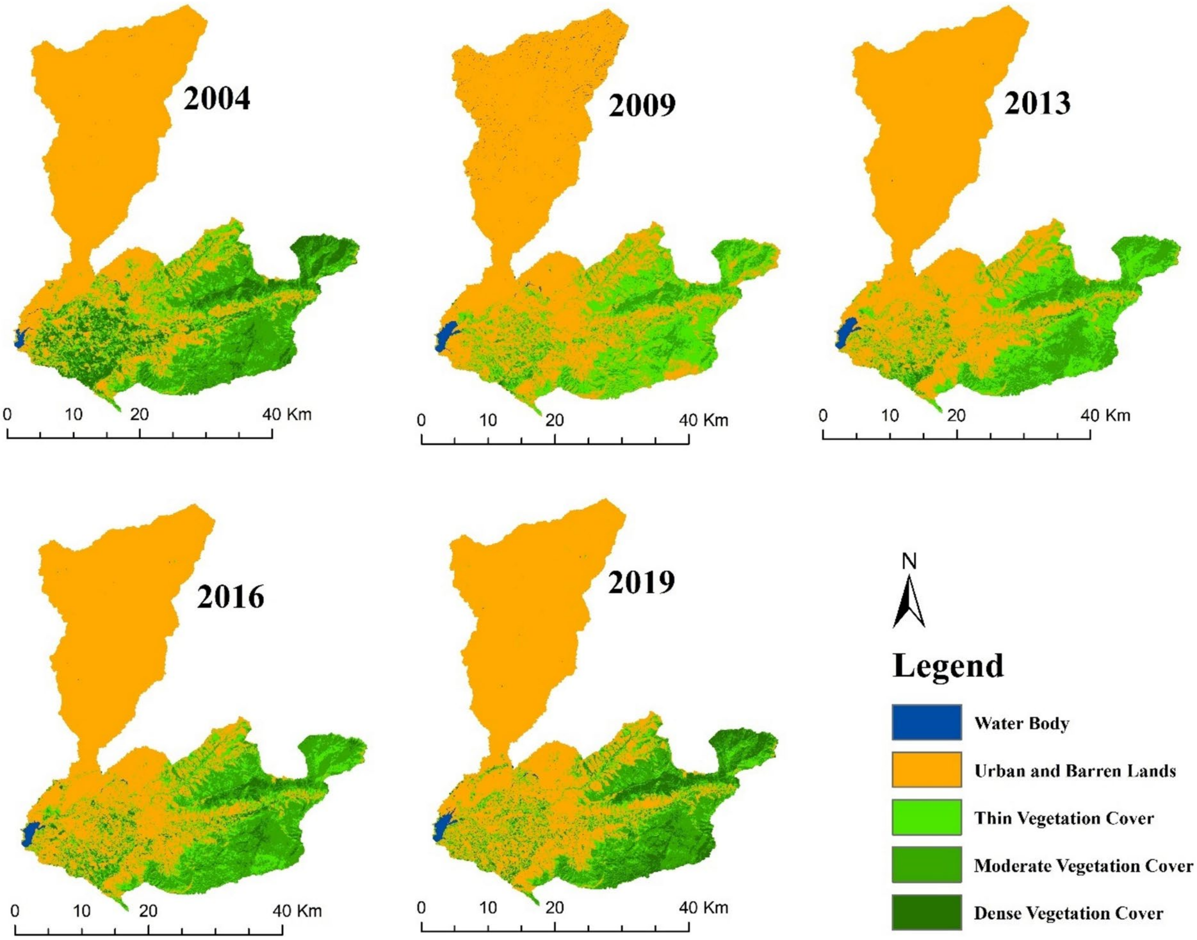


Fig. 6 Land cover class maps of the Hajji Qushan subbasin for the years 2004, 2009, 2013, 2016, and 2019

Model performance and SHAP value analysis

From a hydrological standpoint, the principal components derived from PCA capture distinct land cover patterns that affect watershed response. PC1, strongly associated with vegetation indices and forest cover, generally indicates higher infiltration, greater soil moisture storage, and more stable baseflow, whereas negative PC1 values often reflect reduced vegetation or increased impervious areas, leading to higher runoff. PC2, influenced by croplands, bare soil, and water bodies, reflects variability in rapid runoff generation and flow timing, with positive values indicating higher water retention potential. These interpretations help explain the direction and magnitude of SHAP values observed across subbasins.

The SHAP value plot in Fig. 9 displays the effect of the two main components PC1 and PC2 on the model. The top of the colored line representing the SHAP value is shown in red, while the bottom is shown in blue. The SHAP plot reveals the influencing factors and the direction (positive or negative) of each main component on the predicted discharge for the related subbasins.

The graphs in Fig. 10 show the performance of RF model in predicting discharge for the Tangerah, Galikesh, Hajji Qushan, and Tamar Gorgan subbasins, where blue points represent observed values and red points represent predictions, with vertical distances indicating prediction errors. While the R^2 values (0.78–0.90) indicate a generally high level of accuracy, larger differences between predicted and observed flows occur in extremely wet or dry years. In

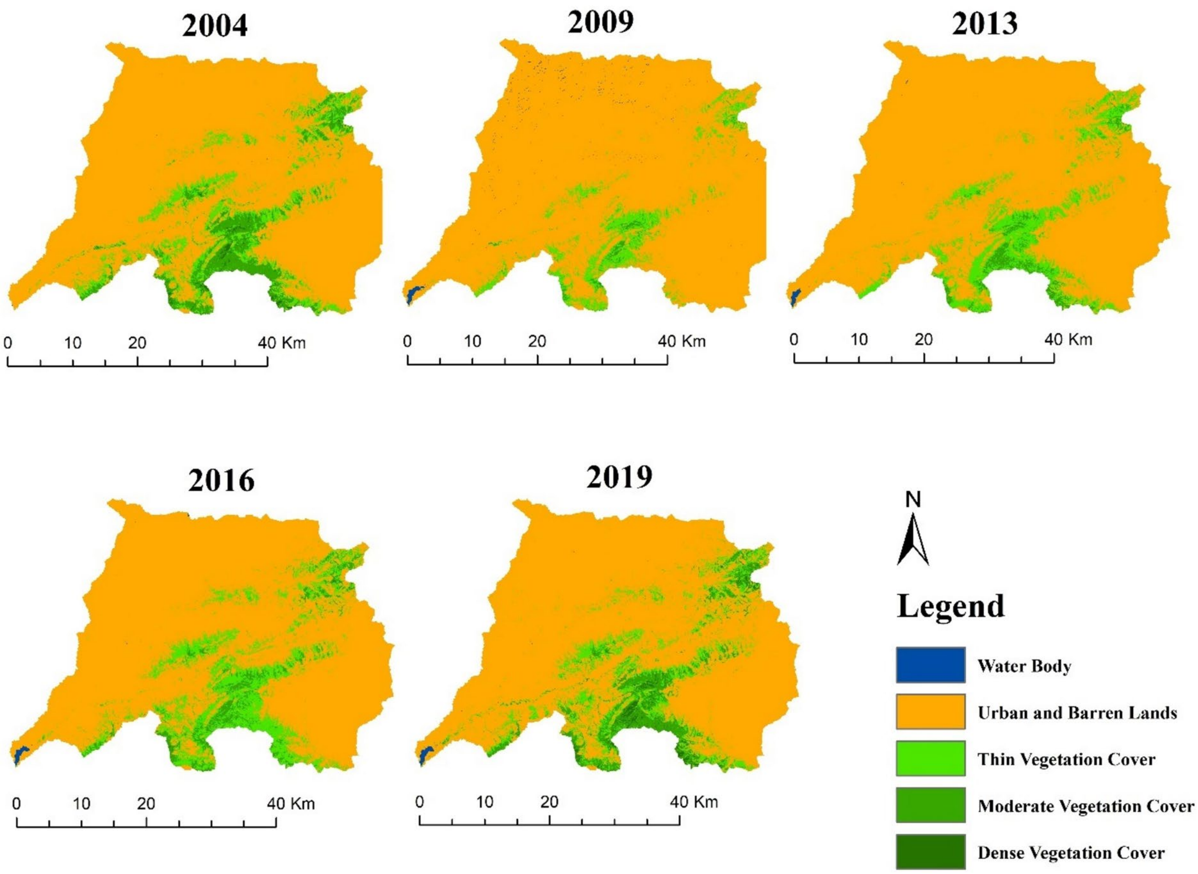


Fig. 7 Land cover class maps of the Tamar Gorgan subbasin for the years 2004, 2009, 2013, 2016, and 2019

most subbasins, the largest gap between observed and predicted values is seen in 2019, which is attributable to the occurrence of a major flood event in that year with a return period of several decades in the region. These discrepancies are likely related to the influence of hydrological variability and land cover–flow dynamics under extreme conditions, where factors such as intense rainfall, prolonged drought, vegetation loss, agricultural expansion, or changes in impervious surfaces can significantly alter runoff generation and peak discharge. When such climatic extremes coincide with notable land cover changes, the resulting flow responses may deviate from patterns represented in the training data, leading to a higher probability of prediction errors in peak discharge values.

The model performance metrics (R^2, MSE, RMSE) and the SHAP values for the random forest regression model across the four subbasins (Tangerah, Galikesh, Hajji Qushan, and Tamar Gorgan) are listed in Table 7, which underscores the influence of the principal components (PC1 and PC2) on the model output, and it also reveals the diverse contributions and impacts of different land cover types in each subbasin.

According to Table 7 in the section corresponding to the Tangerah subbasin, the random forest regression model reached an R^2 of 0.78. This value indicates that approximately 78% of the variance in the target variable is explained by the model. MSE equals 512.82. The RMSE equals 22.65, which provides an estimate of the standard deviation of the prediction errors, indicating relatively high accuracy of the model. The SHAP value diagram of the Tangerah subbasin shows that PC1 mainly has high values and has a positive effect on the model output. The PC2 shows a combination of positive and negative contributions and affects the model with high and low feature values.

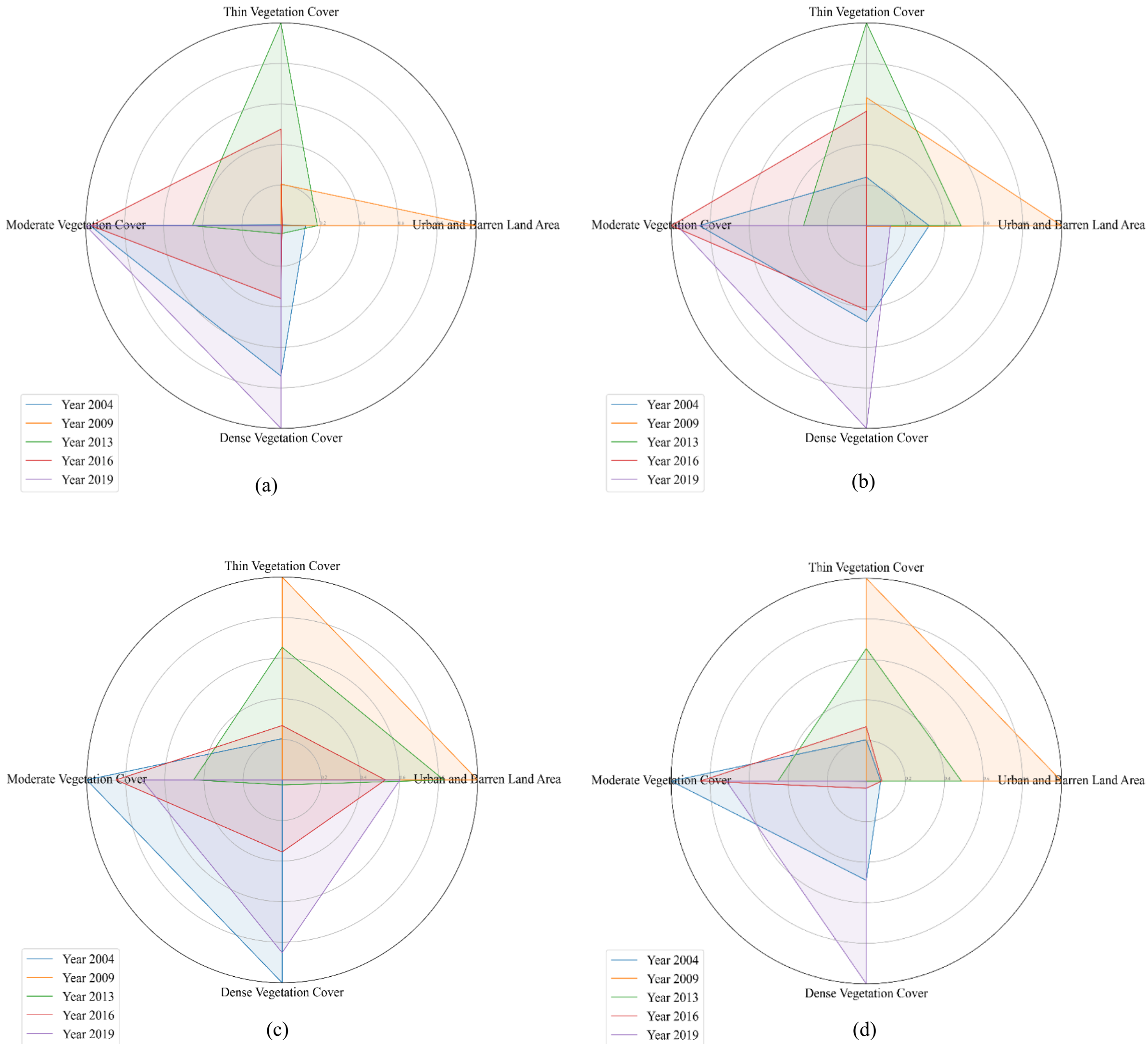


Fig. 8 Radar charts depicting the distribution of land cover types across different years in four subbasins. **a** Tangerah. **b** Galikesh. **c** Hajji Qushan. **d** Tamar Gorgan

Concerning the Galikesh subbasin, it is seen in Table 7 that the model achieves R^2 equal to 0.83 and this shows that about 83% of the variance of the target variable or discharge is explained by the model. The MSE in this subbasin is about 2067.99. This error is larger compared to other subbasins and the reasons for this can be the increase in variability in the data or imperfections in the model fit. The RMSE is 45.48. This larger RMSE indicates that the model predictions for the Galikesh subbasin are less accurate than in other subbasins. The SHAP value plot for the Galikesh subbasin shows that PC1 has high feature values that positively affect the model output. On the other hand, PC2 shows a wide range of effects, with some of the above characteristics having a positive effect and others having a negative effect.

According to Table 7, the random forest model reaches R^2 of about 0.87, which shows that about 87% of the variance and dispersion of the data is explained by the model in the Hajji Qushan subbasin. The MSE is relatively low and its value equals 461.02. This value is smaller compared to

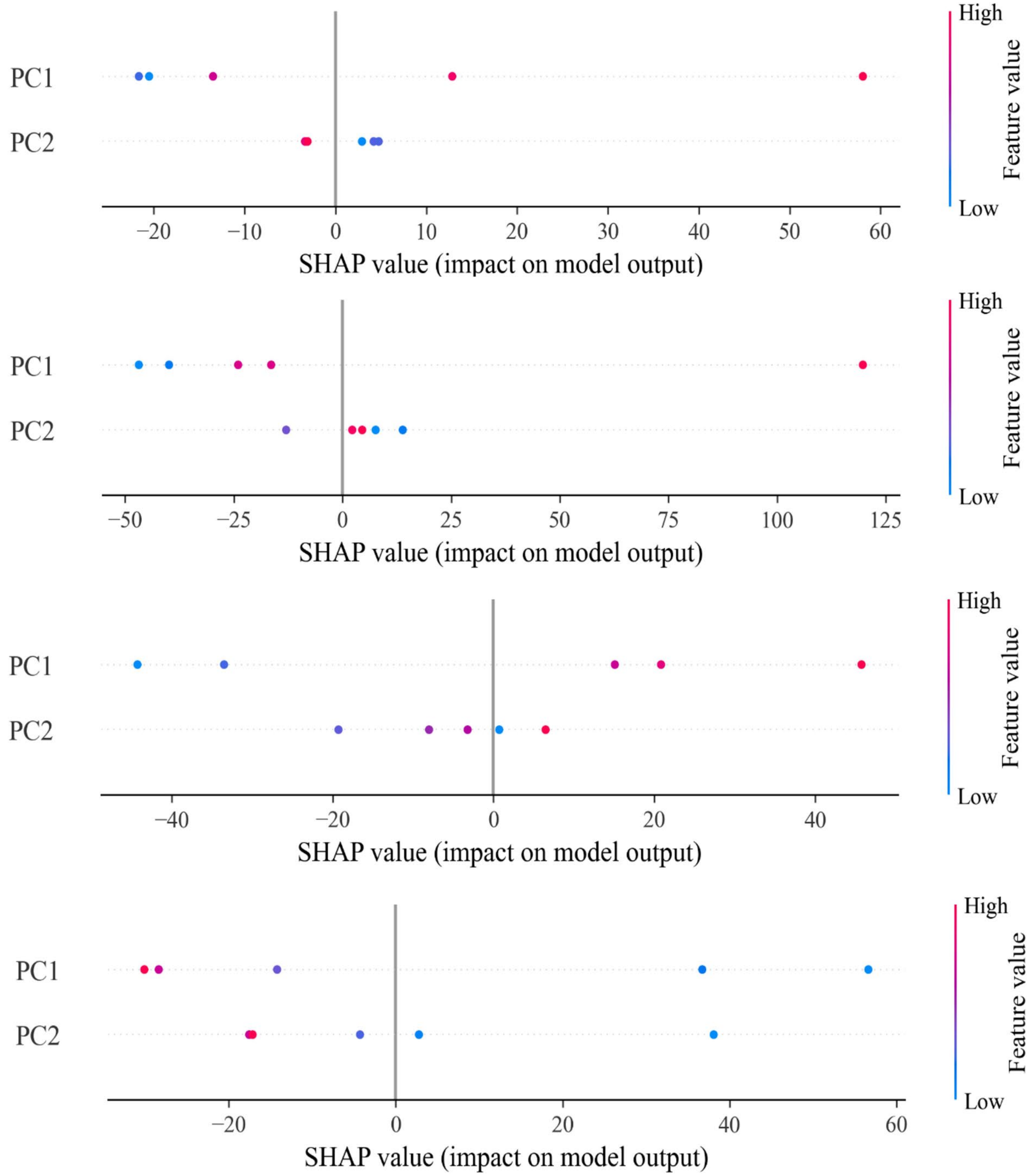


Fig. 9 SHAP value plots for the four subbasins: Tangerah (**a**), Galikesh (**b**), Hajji Qushan (**c**), and Tamar Gorgan (**d**) respectiy

other subbasins. The RMSE is 21.47, which indicates high accuracy with low standard deviation and more accuracy in forecasting. The graph of the SHAP value associated with the Hajji Qushan subbasin shows PC1 values mostly corresponding to negative SHAP values. This means a negative influence on the output of the model corresponding to low feature values. In contrast, PC2 exhibits mostly

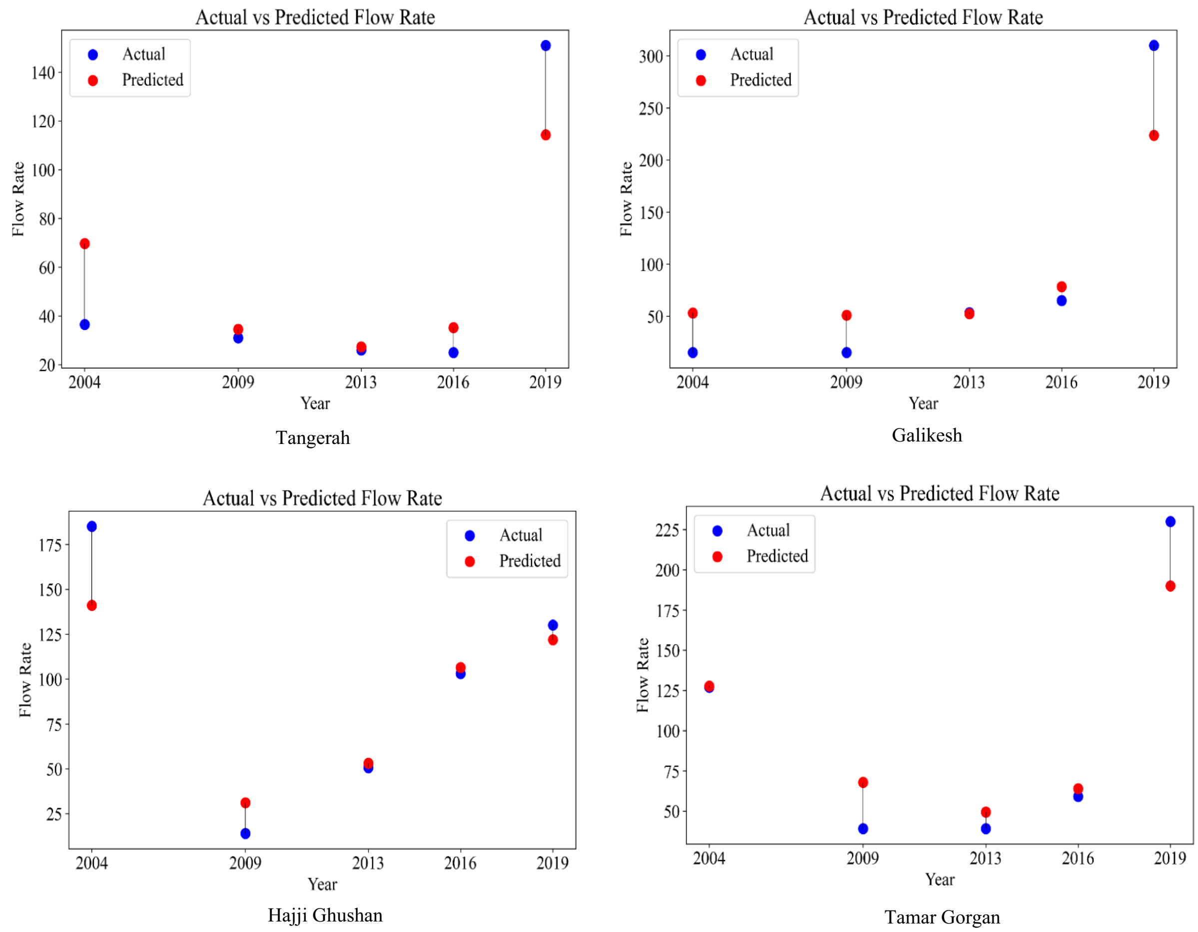


Fig. 10 Actual vs. predicted flow rates (m^3/s) for subbasins

Table 7 The model performance metrics and key insights from the SHAP value plots for each subbasin

Subbasin	R^2	MSE (m^3/s)	RMSE (m^3/s)	Key SHAP insights
Tangerah	0.784	512.82	22.65	PC1 has high positive feature values, while PC2 has mixed positive and negative contributions
Galikesh	0.832	2067.99	45.48	PC1 has high positive impacts, while PC2 shows wide-ranging impacts with high feature values contributing both positively and negatively
Hajji Qushan	0.871	461.02	21.47	PC1 generally has negative SHAP values, and PC2 shows predominantly negative impacts with some positive contributions
Tamar Gorgan	0.904	513.20	22.65	PC1 has negative SHAP values, indicating negative influences, while PC2 has positive SHAP values for higher feature values

negative effects, although it contributes positively at high feature values.

Concerning the Tamar Gorgan subbasin, the results listed in Table 7 establish that the model shows the best performance with an R^2 value of 0.9. This value represents the explanation of approximately 90% of the variance of the target variable by the model. The MSE equals 513.20 and the RMSE equals 22.65. These values indicate the good performance of the model. The SHAP value plot for Tamar

Gorgan shows that PC1 generally has negative SHAP values, which indicates a negative effect on the model output. Conversely, PC2 has positive SHAP values for higher feature values, indicating a positive contribution to model output.

In summary, the random forest regression model exhibits strong performance in all subbasins. In the Tamar Gorgan subbasin is achieved the highest value of R^2, which constitutes the best-fitting model. From these results, it is clear from the comprehensive and accurate SHAP values of the positive influence of the principal components PC1 and PC2 on the forecasting models, with those influences varying among the subbasins. This paper's analysis ensures that all contributing factors are considered and reveals the different characteristics and overall performance of the model for each subbasin, which provides valuable insights about hydrological modeling and land cover analysis.

Land cover distribution and feature contributions

The contributions of dense vegetation, moderate vegetation, thin vegetation, and urban/barren lands to annual instantaneous peak discharge prediction were calculated for the Tangerah, Galikash, Hajji Qushan, and Tamar Gorgan subbasins using Fig. 11 and Table 8. Results show that urban/barren lands had the greatest effect in Tangerah (26.3%, total share 0.502) and Hajji Qushan (30.9%, total share 0.566), while thin vegetation dominated in Galikash (26.9%, total share 0.527) and Tamar Gorgan (26.8%, total share 0.507). The Tangerah subbasin exhibited the most balanced influence across all classes (22–26%), whereas Hajji Qushan showed the least uniformity, with dense vegetation contributing only 2%. The higher impact of urban/barren lands in Tangerah and Hajji Qushan reflects the role of impervious surfaces or sparse vegetation in reducing infiltration and accelerating flood response, while the dominance of thin vegetation in Galikash and Tamar Gorgan indicates heightened sensitivity to small-scale vegetation changes. The variations observed among subbasins indicate that the effects of vegetation cover on peak discharge are not uniform. These findings emphasize the important role of vegetation cover in controlling runoff and reducing peak discharge, which can help prioritize flood mitigation efforts focused on vegetation management. These results are essential for understanding how vegetation cover influences the annual instantaneous peak discharge in each subbasin.

Discussion

Results from this study indicate that land cover changes significantly influence annual instantaneous peak discharge in the subbasins upstream of Golestan Dam, Iran, with urban/barren lands exerting the strongest impact in Tangerah (26.3%) and Hajji Qushan (30.9%), due to their impervious surfaces that reduce infiltration and accelerate runoff during rainfall events, thereby amplifying flood peaks in line with established hydrological principles; in contrast, thin vegetation predominates in Galikash (26.9%) and Tamar Gorgan (26.8%), highlighting increased sensitivity to minor vegetative shifts that diminish water retention and promote rapid surface flow, whereas dense vegetation provides a mitigating effect, notably minimal in Hajji Qushan (only 2%), through enhanced interception, evapotranspiration, and soil permeability that attenuate discharge variations. The balanced distribution across land cover classes in Tangerah (22–26%) reflects a heterogeneous landscape that moderates hydrological responses via interactions among different covers, while the uneven pattern in Hajji Qushan reveals vulnerability to dominant barren areas, underscoring that subbasin-specific factors such as topography and soil characteristics modulate these impacts, as effectively captured by the novel PCA–RF integration despite data limitations. These results align with previous research on land cover dynamics and flood responses; for instance, in the Nyando River Basin, Kenya, historical land cover changes from 1973 to 2000, including deforestation, led to a 16% increase in peak discharges and a 10% rise in runoff volumes, with greater effects in upstream deforested subcatchments, paralleling the observed dominance of urban/barren areas in amplifying peaks, although the PCA–RF approach enables more precise attribution to specific cover types compared to the process-based models employed in that analysis. Similarly, evaluations in East African watersheds, such as Kenya's River Njoro, demonstrate that land cover alterations substantially affect peak discharges and runoff, reinforcing vegetation's role in flood mitigation but differing in scope owing to the emphasis on subbasin variability and machine learning integration rather than

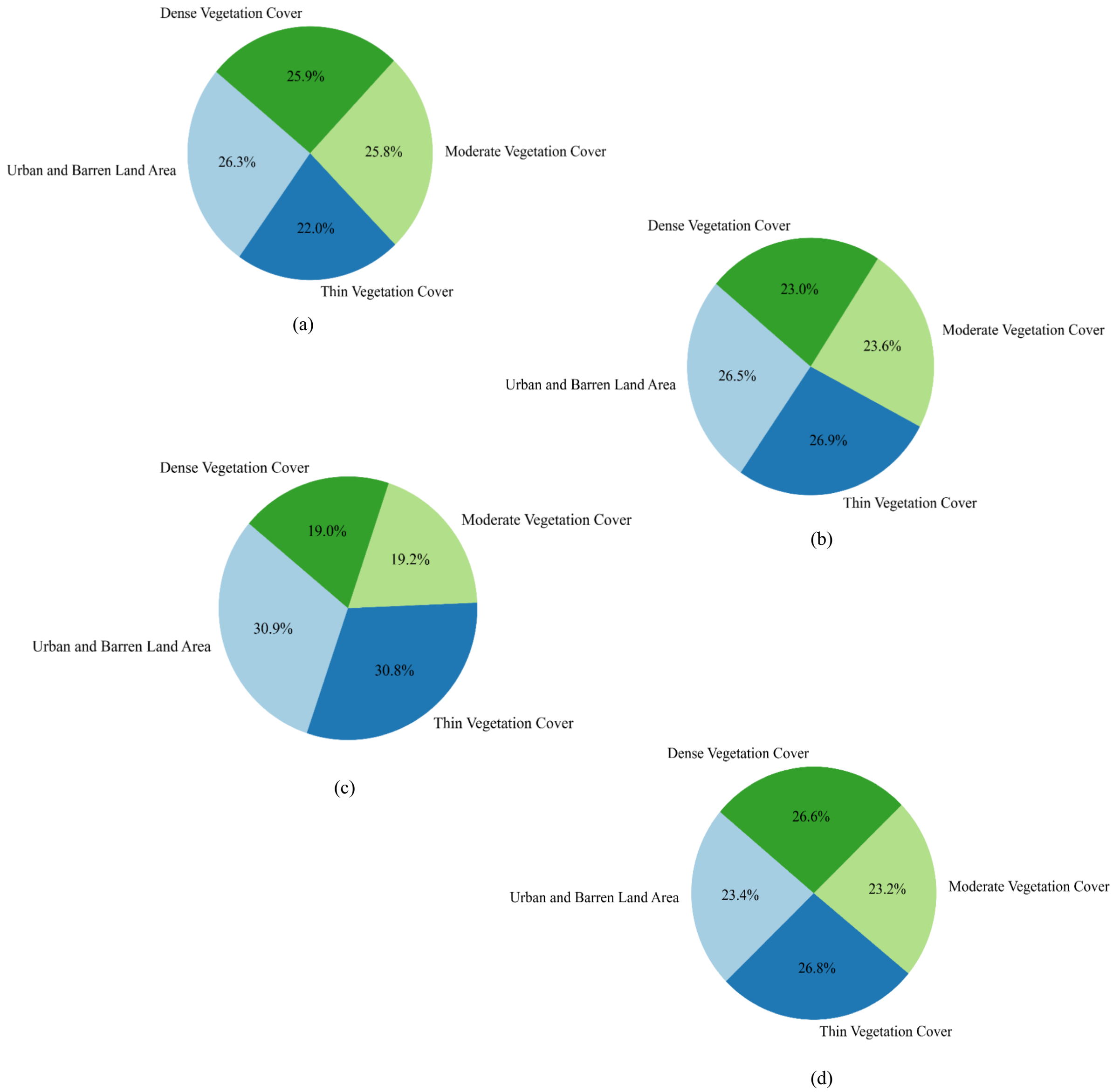


Fig. 11 Contributions of land cover types to discharge prediction for subbasins. **a** Tangerah. **b** Galikesh. **c** Hajji Qushan. **d** Tamar Gorgan

solely remote sensing-based hydrologic modeling. Conversely, investigations using the SWAT model in Iranian watersheds, such as the Maroon River, report elevated runoff over decades attributable to vegetation loss and land use transitions, consistent with the highlighted protective function of vegetation but varying in quantification—without specific percentage increases for peak flows—likely due to differences in model parameterization and the inclusion of climate variables not addressed in this data-constrained analysis. Scientifically, these outcomes clarify biophysical interactions wherein impervious and sparse covers serve as runoff amplifiers by shortening response times and elevating coefficients, while denser vegetation acts as a natural regulator, consistent with ecosystem service frameworks that emphasize vegetative contributions to hydrological stability; practically, they suggest targeted management strategies

Table 8 Principal component analysis (PCA) contributions of different land cover types to flow rate prediction in four subbasins

Subbasin	Urban and barren	Thin vegetation cover	Moderate vegetation cover	Dense vegetation cover	Total contribution (urban and barren)	Total contribution (thin vegetation)	Total contribution (moderate vegetation)	Total contribution (dense vegetation)
Tangerah	PC1: 0.379, PC2: 0.123	PC1: 0.234, PC2: 0.186	PC1: 0.443, PC2: 0.050	PC1: 0.436, PC2: 0.058	0.502	0.420	0.493	0.494
Galikesh	PC1: 0.320, PC2: 0.201	PC1: 0.305, PC2: 0.221	PC1: 0.353, PC2: 0.109	PC1: 0.353, PC2: 0.099	0.520	0.527	0.462	0.452
Hajji Qushan	PC1: 0.328, PC2: 0.237	PC1: 0.329, PC2: 0.235	PC1: 0.341, PC2: 0.011	PC1: 0.334, PC2: 0.013	0.566	0.564	0.351	0.347
Tamar Gorgan	PC1: 0.359, PC2: 0.084	PC1: 0.216, PC2: 0.291	PC1: 0.352, PC2: 0.088	PC1: 0.290, PC2: 0.213	0.443	0.507	0.440	0.503

for flood-prone regions, including prioritization of vegetation restoration in thin-cover-dominated subbasins like Galikash and Tamar Gorgan to enhance resilience and deployment of urban controls such as green infrastructure in Tangerah and Hajji Qushan to mitigate peak amplification, thereby supporting improved flood forecasting, sustainable land use planning, and adaptive water resource policies in the context of environmental changes. For future directions, research could integrate extended time series data, higher-resolution imagery, and supplementary variables like rainfall intensity or soil properties into the PCA–RF framework to bolster model robustness and applicability across diverse basins.

Study limitations

This study was conducted under the assumption that hydrometric station data provide the most reliable source for assessing flood response in the study area. Land cover was considered the sole factor affecting flood variations to rapidly determine the influence of different land cover classes on flood generation, while other hydrological parameters, such as soil infiltration capacity changes, were not incorporated. It should be noted that errors in hydrometric flood data, which are common in Iran, may be more pronounced in certain subbasins such as the Galikesh subbasin, contributing to model inaccuracies beyond our control. This approach aims to assist decision-makers in identifying which land cover class changes have the most significant impact on flooding in their respective regions.

Another limitation of this study is related to the limited availability of direct rainfall data suitable for consistent integration into the analysis. Because long-term and spatially homogeneous precipitation records were not available across the entire study area and because one of the primary objectives of the study was to investigate the relationship between land cover changes and flood response without explicitly modeling the rainfall–runoff process, the hydrological assessment was primarily based on observed flood discharge data from hydrometric stations. Within this framework, the analysis was designed to emphasize the relative influence of land cover changes on flood behavior under broadly comparable hydrological conditions, rather than to isolate rainfall-driven variability. Although this approach may introduce some uncertainty in peak discharge estimation, particularly during extreme events that are sensitive to rainfall intensity, duration, and spatial variability, the results remain robust for comparative evaluation of land cover effects. Accordingly, the estimated peak discharges should be interpreted as relative indicators of flood response, with appropriate consideration of the potential influence of unaccounted rainfall variability.

Furthermore, the reliance on NDVI-based classification introduces seasonal dependencies, as vegetation indices vary throughout the year. All satellite images were selected from the same month (October) in different years to minimize seasonal variations. Nevertheless, the NDVI remains sensitive to climate fluctuations, which may have influenced vegetation growth in specific years, such as 2009. In addition, possible shifts in the onset of seasons—for example, earlier leaf senescence occurring during summer instead of autumn—could lead to significant NDVI differences between years, even within the same month of observation. Additionally, differences in sensor characteristics may have contributed to observed variations, as the 2009 image was captured by Landsat 5, which has different spectral responses compared to newer Landsat missions. The use of Landsat 7 for that year was not feasible due to uncorrectable image distortions in the study area.

Potential classification errors due to satellite image resolution and processing algorithms may also introduce uncertainties in land cover identification. Additionally, the absence of field data for validating certain findings represents a limitation that, if addressed in future studies, could further enhance model accuracy.

Another limitation is related to the short and non-contiguous nature of the available time series data. The relatively limited temporal coverage and gaps between observation years restrict the ability to fully capture continuous land cover dynamics and hydrological responses. This discontinuity reduces the robustness of trend analysis and may obscure intermediate changes, which should be considered when interpreting the results.

Despite these limitations, the study provides a structured framework for analyzing the relationship between land cover changes and flood generation, offering valuable insights for flood management and mitigation strategies. It is recommended that future studies consider incorporating additional variables alongside vegetation cover to improve model accuracy and better capture the complex dynamics influencing flood variations.

Concluding remarks

This study evaluated the important and effective role of changes in land cover classes in affecting the annual instantaneous peak discharge in the basin and subbasins upstream of Golestan Dam, Iran. A novel approach that integrates PCA and RF regression was applied in this work to analyze and predict flood response in several interconnected subbasins despite limitations in data quantity and a limited time frame. The analysis revealed that urban and barren lands were the dominant contributors to peak discharge in some subbasins, while thin vegetation had the strongest influence in others. Dense vegetation generally played a mitigating role by reducing flood peaks. The effects were not uniform, with some subbasins showing a more balanced influence of different land cover classes and others being strongly dominated by a single type. These findings confirm that (1) land cover changes have a significant and measurable effect on flood peaks, (2) the nature and strength of this effect vary from one subbasin to another, and (3) the integrated PCA–RF approach is effective in capturing these spatially variable relationships despite data limitations.

The integration of remote sensing data, GIS, and advanced machine learning techniques was instrumental in obtaining this paper's results. Moreover, our findings emphasize the importance of monitoring LULC changes for sustainable resource management and disaster mitigation. From a methodological perspective, future research can benefit from refining the PCA–RF integration by incorporating longer time series, higher-resolution imagery, and complementary hydrological variables (e.g., soil infiltration capacity, rainfall intensity) to enhance model robustness and transferability.

For practical application, the insights from this study can guide flood risk management in other basins by identifying the most sensitive land cover types and prioritizing vegetation restoration or urban runoff control measures according to local conditions. Such adaptation of the framework will allow policymakers to target interventions where changes in LULC have the highest hydrological impact.

Furthermore, this methodology can be extended by integrating additional datasets such as climate projections, socioeconomic land use scenarios, or finer-scale DEM data, enabling the prediction of flood responses under future environmental and developmental changes. This will help build adaptive strategies for basin-wide flood mitigation and water resource planning.

This research provides a framework and insight for understanding the complex interactions between land cover dynamics and hydrological responses. Finally, it provides valuable insights for policymakers and land managers with the aim of developing strategies for flexibility and adaptation in the face of environmental changes.

Acknowledgements The authors are indebted to the Iranian Water Resources Management Company for providing data to conduct this research. The authors thank Iran's National Science Foundation (INSF) for its support.

Author contributions Seyed-Mahdi Salar-Khorasani; Software, Formal analysis, Writing—Original Draft Omid Bozorg-Haddad; Conceptualization, Supervision, Project administration Hugo A. Loáiciga; Validation, Writing—Review & Editing.

Funding The authors thank Iran's National Science Foundation (INSF) for its support for this research.

Data availability The data that support the findings of this study are available from the corresponding author upon reasonable request.

Code availability The codes that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Ethics approval All authors have read, understood, and complied as applicable with the statement on "ethical responsibilities of authors" as found in the instructions for authors.

Consent to participate All authors consent to participate.

Consent for publication All authors consent to publish.

Competing interests The authors declare no competing interests.

Clinical trial number Not applicable.

References

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics, 2*(4), 433–459. https://doi.org/10.1002/wics.101

Abedi, R., Costache, R., Shafizadeh-Moghadam, H., & Pham, Q. B. (2021). Flash-flood susceptibility mapping based on XGBoost, random forest and boosted regression trees. *Geocarto International, 37*(19), 5479–5496. https://doi.org/10.1080/10106049.2021.1920636

Abson, D. J., Dougill, A. J., & Stringer, L. C. (2012). Using principal component analysis for information-rich socio-ecological vulnerability mapping in Southern Africa. *Applied Geography, 35*(1–2), 515–524. https://doi.org/10.1016/j.apgeog.2012.08.004

Akbar, T. A., Hassan, Q. K., Ishaq, S., Batool, M., Butt, H. J., & Jabbar, H. (2019). Investigative spatial distribution and modelling of existing and future urban land changes and its impact on urbanization and economy. *Remote Sensing, 11*(2), Article 105. https://doi.org/10.3390/rs11020105

Alarifi, S. S., Abdelkareem, M., Abdalla, F., & Alotaibi, M. (2022). Flash flood hazard mapping using remote sensing and GIS techniques in southwestern Saudi Arabia. *Sustainability, 14*(21), Article 14145. https://doi.org/10.3390/su142114145

Al-Juaidi, A. E., Nassar, A. M., & Al-Juaidi, O. E. (2018). Evaluation of flood susceptibility mapping using logistic regression and GIS conditioning factors. *Arabian Journal of Geosciences, 11*(24), 765. https://doi.org/10.1007/s12517-018-4095-0

Ardalan, A., Holakouie Naieni, K., Kabir, M. J., et al. (2009). Evaluation of Golestan province's early warning system for flash floods, Iran, 2006–7. *International Journal of Biometeorology, 53*, 247–254. https://doi.org/10.1007/s00484-009-0210-y

Beygi, S., Bozorg-Haddad, O., Fallah-Mehdipour, E., & Mariño, M. A. (2014). Bargaining models for optimal design of water distribution networks. *Journal of Water Resources Planning and Management, 140*(1), 92–99. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000324

Bhaduri, B., Harbor, J. O. N., Engel, B., & Grove, M. (2000). Assessing watershed-scale, long-term hydrologic impacts of land-use change using a GIS-NPS model. *Environmental Management, 26*(6), 643–658. https://doi.org/10.1007/s002670010122

Bhattacharyya, R., Ghosh, B. N., Mishra, P. K., Mandal, B., Rao, C. S., Sarkar, D., Das, K., Anil, K. S., Lalitha, M., Hati, K. M., & Franzluebbers, A. J. (2015). Soil degradation in India: Challenges and potential solutions. *Sustainability, 7*(4), 3528–3570. https://doi.org/10.3390/su7043528

Böhringer, C., & Jochem, P. E. (2007). Measuring the immeasurable—A survey of sustainability indices. *Ecological Economics, 63*(1), 1–8. https://doi.org/10.1016/j.ecolecon.2007.03.008

Boroughani, M., Zandi, R., Pourhashemi, S., Gholami, H., & Kaskaoutis, D. G. (2025). Linking sand/dust storms hotspots and land use over Iran. *Atmospheric Pollution Research, 16*(2), Article p.102380. https://doi.org/10.1016/j.apr.2024.102380

Bozorg-Haddad, O., Afshar, A., & Mariño, M. A. (2009b). Optimization of non-convex water resource problems by honey-bee mating optimization (HBMO) algorithm. *Engineering Computations, 26*(3), 267–280. https://doi.org/10.1108/02644400910943617

Bozorg-Haddad, O., Janbaz, M., & Loáiciga, H. A. (2016). Application of the gravity search algorithm to multi-reservoir operation optimization. *Advances in Water Resources, 98*, 173–185. https://doi.org/10.1016/j.advwatres.2016.11.001

Bozorg-Haddad, O., & Marino, M. A. (2007). Dynamic penalty function as a strategy in solving water resources combinatorial optimization problems with honey-bee mating optimization (HBMO) algorithm. *Journal of Hydroinformatics, 9*(3), 233–250. https://doi.org/10.2166/hydro.2007.025

Bozorg-Haddad, O., & Marino, M. A. (2011). Optimum operation of wells in coastal aquifers. *In Proceedings of the Institution of Civil Engineers-Water Management, 164*(3), 135–146. https://doi.org/10.1680/wama.1000037. Thomas Telford Ltd.

Bozorg-Haddad, O., Moradi-Jalal, M., Mirmomeni, M., Kholghi, M. K., & Mariño, M. A. (2009a). Optimal cultivation rules in multi-crop irrigation areas. *Irrigation and Drainage: THe Journal of the International Commission on Irrigation and Drainage, 58*(1), 38–49. https://doi.org/10.1002/ird.381

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Byrne, G. F., Crapper, P. F., & Mayo, K. K. (1980). Monitoring land-cover change by principal component analysis of multitemporal Landsat data. *Remote Sensing of Environment, 10*(3), 175–184. https://doi.org/10.1016/0034-4257(80)90021-8

Caocao, C., Gaodi, X., Lin, Z., & Yunfa, L. (2008). Analysis on Jinghe watershed vegetation dynamics and evaluation on its relation with precipitation. *Acta Ecologica Sinica, 28*(3), 925–938. https://doi.org/10.1016/S1872-2032(08)60032-3

Chen, W., Li, Y., Xue, W., Shahabi, H., Li, S., Hong, H., ... & Ahmad, B. B. (2020). Modeling flood susceptibility using data-driven approaches of naïve bayes tree, alternating decision tree, and random forest methods. *Science of The Total Environment,* 701, 134979. https://doi.org/10.1016/j.scitotenv.2019.134979

Collins, J. B., & Woodcock, C. E. (1996). An assessment of several linear change detection techniques for mapping forest mortality using multitemporal Landsat TM data. *Remote Sensing of Environment, 56*(1), 66–77. https://doi.org/10.1016/0034-4257(95)00233-2

Dewan, A. M., Kumamoto, T., & Nishigaki, M. (2006). Flood hazard delineation in Greater Dhaka, Bangladesh using an integrated GIS and remote sensing approach. *Geocarto International, 21*(2), 33–38. https://doi.org/10.1080/10106040608542381

Fallah-Mehdipour, E., Bozorg-Haddad, O., Beygi, S., & Mariño, M. A. (2011). Effect of utility function curvature of Young's bargaining method on the design of WDNs. *Water Resources Management, 25*(9), 2197–2218. https://doi.org/10.1007/s11269-011-9802-5

Fallah-Mehdipour, E., Bozorg-Haddad, O., & Mariño, M. A. (2013). Extraction of multicrop planning rules in a reservoir system: Application of evolutionary algorithms. *Journal of Irrigation and Drainage Engineering, 139*(6), 490–498. https://doi.org/10.1061/(ASCE)IR.1943-4774.0000572

Hammer, T. R. (1972). Stream channel enlargement due to urbanization. *Water Resources Research, 8*(6), 1530–1540. https://doi.org/10.1029/WR008i006p01530

Hydrologic Engineering Center (HEC). HEC-HMS user's manual. *MUD history*, https://www.hec.usace.army.mil/confluence/hmsdocs/hmsum/latest (May 1, 2024).

Ighile, E. H., & Shirakawa, H. (2020). A study on the effects of land use change on flooding risks in Nigeria. *Geographia Technica*. https://doi.org/10.21163/gt_2020.151.08

Iranian Water Resources Management Company (IWRM). Annual peak flow data for hydrometric stations. *MUD history*, https://www.wrm.ir (May 1, 2024).

Islam, A. R. M. T., Talukdar, S., Mahato, S., Kundu, S., Eibek, K. U., Pham, Q. B., Kuriqi, A., & Linh, N. T. T. (2021). Flood susceptibility modelling using advanced ensemble machine learning models. *Geoscience Frontiers, 12*(3), Article 101075. https://doi.org/10.1016/j.gsf.2020.09.006

Jacobberger-Jellison, P. A. (1994). Detection of post-drought environmental conditions in the Tombouctou region. *International Journal of Remote Sensing, 15*(16), 3183–3197. https://doi.org/10.1080/01431169408954321

Janizadeh, S., Pal, S. C., Saha, A., Chowdhuri, I., Ahmadi, K., Mirzaei, S., ... & Tiefenbacher, J. P. (2021). Mapping the spatial and temporal variability of flood hazard affected by climate and land-use changes in the future. *Journal of Environmental Management,* 298, 113551. https://doi.org/10.1016/j.jenvman.2021.113551

Jensen, J. R., & Christensen, E. J. (1986). Solid and hazardous waste disposal site selection using digital geographic information system techniques. *Science of the Total Environment, 56*, 265–276. https://doi.org/10.1016/0048-9697(86)90331-1

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society a: Mathematical, Physical and Engineering Sciences, 374*(2065), Article 20150202. https://doi.org/10.1098/rsta.2015.0202

Junger, L., Hohensinner, S., Schroll, K., Wagner, K., & Seher, W. (2022). Land use in flood-prone areas and its significance for flood risk management—A case study of Alpine regions in Austria. *Land, 11*(3), Article 392. https://doi.org/10.3390/land11030392

Karimi-Hosseini, A., Bozorg-Haddad, O., & Marino, M. A. (2011). Site selection of raingauges using entropy methodologies. *In Proceedings of the Institution of Civil Engineers-Water Management, 164*(7), 321–333. https://doi.org/10.1680/wama.2011.164.7.321. Thomas Telford Ltd.

Khoirunisa, N., Ku, C. Y., & Liu, C. Y. (2021). A GIS-based artificial neural network model for flood susceptibility assessment. *International Journal of Environmental Research and Public Health, 18*(3), 1072. https://doi.org/10.3390/ijerph18031072

Khosravi, K., Shahabi, H., Pham, B. T., Adamowski, J., Shirzadi, A., Pradhan, B., ... & Prakash, I. (2019). A comparative assessment of flood susceptibility modeling using multi-criteria decision-making analysis and machine learning methods. *Journal of Hydrology,* 573, 311–323. https://doi.org/10.1016/j.jhydrol.2019.03.073

Klein, R. D. (1979). Urbanization and stream quality impairment 1. *JAWRA Journal of the American Water Resources Association, 15*(4), 948–963. https://doi.org/10.1111/j.1752-1688.1979.tb01074.x

Lee, S., Kim, J. C., Jung, H. S., Lee, M. J., & Lee, S. (2017). Spatial prediction of flood susceptibility using random-forest and boosted-tree models in Seoul metropolitan city, Korea. *Geomatics, Natural Hazards and Risk, 8*(2), 1185–1203. https://doi.org/10.1080/19475705.2017.1308971

Madhuri, R., Sistla, S., & Srinivasa Raju, K. (2021). Application of machine learning algorithms for flood susceptibility assessment and risk management. *Journal of Water and Climate Change, 12*(6), 2608–2623. https://doi.org/10.2166/wcc.2021.051

Menzel, L., & Blongewicz, M. (2000). LADEMO–a user supported model for the development of land use scenarios.

In European Conference on Advances in Flood Research, Proceedings. PIK-Report, 65, 43–51.

Meyer, W. B., & Turner, B. L. (1992). Human population growth and global land-use/cover change. *Annual Review of Ecology and Systematics, 23*, 39–61.

Microsoft. (2024). Visual Studio Code. *Microsoft Corporation*, https://code.visualstudio.com

Mosavi, A., Ozturk, P., & Chau, K. W. (2018). Flood prediction using machine learning models: Literature review. *Water, 10*(11), Article 1536. https://doi.org/10.3390/w10111536

Nguyen, H. D. (2022). GIS-based hybrid machine learning for flood susceptibility prediction in the Nhat Le–Kien Giang watershed, Vietnam. *Earth Science Informatics, 15*(4), 2369–2386. https://doi.org/10.1007/s12145-022-00825-4

Nguyen, H. D., Dang, D. K., Nguyen, Q. H., Bui, Q. T., & Petrisor, A. I. (2022a). Evaluating the effects of climate and land use change on the future flood susceptibility in the central region of Vietnam by integrating land change modeler, machine learning methods. *Geocarto International, 37*(26), 12810–12845. https://doi.org/10.1080/10106049.2022.2071477

Nguyen, H. D., Quang-Thanh, B., Nguyen, Q. H., Nguyen, T. G., Pham, L. T., Nguyen, X. L., … Petrisor, A. I. (2022b). A novel hybrid approach to flood susceptibility assessment based on machine learning and land use change. Case study: A river watershed in Vietnam. *Hydrological Sciences Journal,* 67(7), 1065–1083. https://doi.org/10.1080/02626667.2022.2060108

Nguyen, T. T., Bonetti, J., Rogers, K., & Woodroffe, C. D. (2016). Indicator-based assessment of climate-change impacts on coasts: A review of concepts, methodological approaches and vulnerability indices. *Ocean & Coastal Management, 123*, 18–43. https://doi.org/10.1016/j.ocecoaman.2015.11.022

Orouji, H., Bozorg-Haddad, O., Fallah-Mehdipour, E., & Mariño, M. A. (2014). Extraction of decision alternatives in project management: Application of hybrid PSO-SFLA. *Journal of Management in Engineering, 30*(1), 50–59. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000186

Pal, S. C., Chowdhuri, I., Das, B., Chakrabortty, R., Roy, P., Saha, A., & Shit, M. (2022). Threats of climate change and land use patterns enhance the susceptibility of future floods in India. *Journal of Environmental Management, 305*, Article 114317. https://doi.org/10.1016/j.jenvman.2021.114317

Peters, A. J., Reed, B. C., Eve, M. D., & Havstad, K. M. (1993). Satellite assessment of drought impact on native plant communities of southeastern New Mexico, USA. *Journal of Arid Environments, 24*(3), 305–319. https://doi.org/10.1006/jare.1993.1027

Pourhashemi, S., Asadi, M. A. Z., & Boroughani, M. (2025). Multi-hazard susceptibility mapping in the Salt Lake watershed. *Environmental Challenges, 18*, Article 101079. https://doi.org/10.1016/j.envc.2024.101079

QGIS.org. (2024). *QGIS Geographic Information System* (Version 3.34). Open Source Geospatial Foundation Project, https://www.qgis.org

Rautela, K. S., Kumar, M., Khajuria, V., & Alam, M. A. (2022). Comparative geomorphometric approach to understand the hydrological behaviour and identification of the erosion prone areas of a coastal watershed using RS and GIS tools. *Discover Water, 2*(1), Article 1. https://doi.org/10.1007/s43832-021-00009-z

Reid, R. S., Kruska, R. L., Muthui, N., Taye, A., Wotton, S., Wilson, C. J., & Mulatu, W. (2000). Land-use and land-cover dynamics in response to changes in climatic, biological and socio-political forces: The case of southwestern Ethiopia. *Landscape Ecology, 15*(4), 339–355. https://doi.org/10.1023/A:1008177712995

Roy, P., Pal, S. C., Chakrabortty, R., Chowdhuri, I., Malik, S., & Das, B. (2020). Threats of climate and land use change on future flood susceptibility. *Journal of Cleaner Production, 272*, Article 122757. https://doi.org/10.1016/j.jclepro.2020.122757

Roy, P. S., Ramachandran, R. M., Paul, O., Thakur, P. K., Ravan, S., Behera, M. D., ... & Kanawade, V. P. (2022). Anthropogenic land use and land cover changes—A review on its environmental consequences and climate change. *Journal of the Indian Society of Remote Sensing,* 50(8), 1615–1640. https://doi.org/10.1007/s12524-022-01569-w

Sabbaghpour, S., Naghashzadehgan, M., Javaherdeh, K., & Bozorg-Haddad, O. (2012). HBMO algorithm for calibrating water distribution network of Langarud city. *Water Science and Technology, 65*(9), 1564–1569. https://doi.org/10.2166/wst.2012.045

Salar-Khorasani, S.-M., & Singh, V. P. (2026). The role of remote sensing in water resources management under climate change. In K. Behzadian, A. Ferdowsi, & B. Zolghadr-Asli (Eds.), *Planning and management for sustainable water resources and infrastructure under climate change* (pp. 235–250). Elsevier. https://doi.org/10.1016/B978-0-443-33687-4.00013-7

Schueler, T. (1994). The importance of imperviousness. *Watershed Protection Techniques, 1*(3), 100–101.

Shokri, A., Bozorg-Haddad, O., & Mariño, M. A. (2013). Algorithm for increasing the speed of evolutionary optimization and its accuracy in multi-objective problems. *Water Resources Management, 27*(7), 2231–2249. https://doi.org/10.1007/s11269-013-0285-4

Singh, A. (1989). Review article digital change detection techniques using remotely-sensed data. *International Journal of Remote Sensing, 10*(6), 989–1003. https://doi.org/10.1080/01431168908903939

Sofi, M. S., Rautela, K. S., Bhat, S. U., Rashid, I., & Kuniyal, J. C. (2021). Application of geomorphometric approach for the estimation of hydro-sedimentological flows and cation weathering rate: Towards understanding the sustainable land use policy for the Sindh Basin, Kashmir Himalaya. *Water, Air, & Soil Pollution, 232*(7), Article 280. https://doi.org/10.1007/s11270-021-05217-w

Soltanjalili, M., Bozorg-Haddad, O., & Mariño, M. A. (2011). Effect of breakage level one in design of water distribution networks. *Water Resources Management, 25*(1), 311–337. https://doi.org/10.1007/s11269-010-9701-1

Sugianto, S., Deli, A., Miswar, E., Rusdi, M., & Irham, M. (2022). The effect of land use and land cover changes on flood occurrence in Teunom Watershed, Aceh Jaya. *Land, 11*(8), Article 1271. https://doi.org/10.3390/land11081271

Tehrany, M. S., Kumar, L., & Shabani, F. (2019). A novel GIS-based ensemble technique for flood susceptibility mapping using evidential belief function and support vector machine: Brisbane, Australia. *PeerJ, 7*, Article e7653. https://doi.org/10.7717/peerj.7653

Tehrany, M. S., Pradhan, B., & Jebur, M. N. (2014). Flood susceptibility mapping using a novel ensemble weights-of-evidence and support vector machine models in GIS. *Journal of Hydrology, 512*, 332–343. https://doi.org/10.1016/j.jhydrol.2014.03.008

Tourani, M., & Çağlayan, A. (2021). *Example of the biggest flood disaster in Iranian history: Golestan province (NE Iran)*. Geoscience for Society.

U.S. Army Corps of Engineers (USACE). (2024). HEC-HMS Hydrologic Modeling System (Version 4.11). *Hydrologic Engineering Center*, https://www.hec.usace.army.mil/software/hec-hms/

U.S. Geological Survey (USGS). (2024a). Landsat 7 SLC-off gap-filled products phase two methodology. *MUD history*, https://www.usgs.gov/media/files/landsat-7-slc-gap-filled-products-phase-two-methodology

U.S. Geological Survey (USGS). (2024b). Shuttle Radar Topography Mission (SRTM) elevation data. *MUD history*, https://earthexplorer.usgs.gov

U.S. Geological Survey (USGS). (2024c). Landsat Collection 2 and Level 2 satellite images. *MUD history*, https://earthexplorer.usgs.gov

Wang, H., Wang, H., Wu, Z., & Zhou, Y. (2021). Using multi-factor analysis to predict urban flood depth based on Naive Bayes. *Water, 13*(4), 432. https://doi.org/10.3390/w13040432

Yan, E., Wang, G., Lin, H., Xia, C., & Sun, H. (2015). Phenology-based classification of vegetation cover types in Northeast China using MODIS NDVI and EVI time series. *International Journal of Remote Sensing, 36*(2), 489–512. https://doi.org/10.1080/01431161.2014.999167