Asian Journal of Water, Environment and Pollution. Vol. 22, No. 5 (2025), pp. 231-244.
doi: 10.36922/AJWEP025250201

ORIGINAL RESEARCH ARTICLE

Multivariate analysis of evaporation drivers in Mbeya, Tanzania, using principal component analysis

Zacharia Katambara*

Department of Civil Engineering, College of Engineering and Technology, Mbeya University of Science and Technology, Mbeya, Mbeya Region, Tanzania

***Corresponding author:** Zacharia Katambara (zacharia.katambara@must.ac.tz)

Received: June 18, 2025; Revised: July 13, 2025; Accepted: July 17, 2025; Published online: August 8, 2025

Abstract: Evaporation is a vital process in the hydrological cycle, accounting for approximately 70% of water loss from the Earth's surface. In semi-arid and rapidly urbanizing regions, such as Mbeya, Tanzania, understanding the meteorological drivers of evaporation is critical for water resource management and agricultural planning. This study utilized principal component analysis (PCA) on a 10-year dataset comprising solar radiation, sunshine hours, minimum and maximum temperatures, and wind speed to identify key factors influencing evaporation. Descriptive statistics revealed significant non-normality in most variables, particularly radiation and wind speed. At the same time, correlation analysis showed a strong positive relationship between sunshine hours and radiation (r = 0.66) and a moderate negative correlation between radiation and minimum temperature (r = −0.30). PCA identified two principal components accounting for 66.61% of the total variance. Component 1 (38.06%) captured solar-driven variability, dominated by sunshine duration and radiation, whereas Component 2 (28.55%) reflected thermal influences, particularly maximum and minimum temperatures. Wind speed contributed minimally, suggesting a more localized or less consistent role in evaporation dynamics. These findings demonstrate the value of PCA in simplifying complex climatic datasets and improving the interpretation of evaporation processes. Solar radiation and sunshine hours emerged as the dominant drivers, with temperature as a secondary influence. The results emphasize the need to integrate surface-level variables, such as land use, vegetation cover, and soil moisture, in future studies to capture spatial heterogeneity and improve predictive accuracy, especially in data-scarce, climate-sensitive regions like Mbeya.

Keywords: Evaporation; Principal component analysis; Meteorological factors; Multivariate analysis; Dimensionality reduction

1. Introduction

Evaporation, a fundamental process in the hydrological cycle, is vital for water resource management, agriculture, and ecosystem sustainability, with approximately 70% of water loss from the Earth's surface attributed to evaporation.[1,2] Meteorological factors, including temperature, humidity, solar radiation, wind speed, and atmospheric pressure, are the primary drivers of evaporation.[3,4] Understanding the complex interplay of these variables is increasingly important, especially in the context of climate change and mounting water scarcity.[5] Recent advancements in data collection and analytical methods have enabled researchers to utilize multivariate analysis techniques, such as principal component analysis (PCA), to more accurately identify the key factors influencing evaporation and streamline the interpretation of complex datasets.[6] Traditional

univariate models, which analyze variables in isolation, often fail to account for the interdependence of meteorological factors, leading to oversimplified predictions.[7] In contrast, multivariate approaches such as PCA offer significant advantages by transforming correlated variables into uncorrelated principal components, each representing a distinct portion of variance within the dataset.[8,9] This dimensionality reduction allows for a more accurate and insightful analysis of the drivers of evaporation while preserving the integrity of the original data,[10] making it a powerful tool in environmental system studies.

PCA has been applied extensively in hydrological and meteorological research to identify key variables influencing evaporation.[11] It is particularly effective in analyzing large, complex datasets where multiple meteorological variables, such as temperature, humidity, wind speed, and solar radiation, interact simultaneously.[3,12] For example, Ullah *et al*.[13] applied PCA to assess the primary drivers of evaporation in South Asia, finding that solar radiation and air temperature were the most significant contributors to evaporation variability across the region. Similarly, Cao *et al*.[11] used PCA to evaluate the relationship between climate variability and evaporation rates in northern China, demonstrating that temperature and wind speed were the dominant factors in determining seasonal evaporation patterns. Their study highlighted that PCA simplifies the complexity of hydrological datasets, enabling a more straightforward interpretation of the critical variables that influence evaporation. Jafari and Dinpashoh[14] applied PCA in conjunction with multiple linear regression to model pan evaporation in Iran. Their results showed that the first three principal components accounted for more than 90% of the variance in the data, resulting in significant improvements in model performance, with R^2 values exceeding 0.74.

Cao *et al*.[11] used PCA alongside a radial basis function neural network to predict water surface evaporation. Their study demonstrated a 95.3% prediction accuracy, a marked improvement over traditional models such as the backpropagation network. In another study, Li *et al*.[15] used PCA to analyze evaporation factors in northern China, revealing that sunshine was the most significant factor affecting evaporation when temperature was held constant. These studies highlight the effectiveness of PCA in identifying key meteorological drivers of evaporation and enabling more accurate predictions. Therefore, this study utilized PCA on meteorological data in Mbeya, Tanzania, to identify key factors influencing evaporation.

2. Study area, materials, and methods

2.1. Description of the study area

Figure 1 illustrates the Mbeya Region, one of the fastest-growing urban centers in Africa, located in the southern highlands of Tanzania.[16,17] According to the 2012 national census, the Mbeya city population was 385,279, which increased significantly to 665,390 by 2015.[17,18] Spanning an area of approximately 253 km^2, the city serves as a strategic economic and administrative hub for Tanzania's southern zone, facilitating both domestic and cross-border trade with neighboring countries, including Zambia, Malawi, Mozambique, and the Democratic Republic of the Congo. The Mbeya Region boasts a diverse economy driven by agriculture, livestock, beekeeping, mining, and tourism.[18] It ranks third in national gross domestic product (GDP) contribution, after Dar es Salaam Region and Mwanza Region, accounting for approximately 7.44% of the country's total GDP. Agriculture, largely rainfed, remains the backbone of the regional economy, contributing about 40% of the GDP and employing nearly 80% of the population. The region is also known for its significant forestry products, including timber, fuelwood, and honey.

However, the region's dependence on rainfed agriculture makes it particularly vulnerable to climate variability. In this context, understanding the dynamics of evaporation and its meteorological drivers has become increasingly important. Evaporation is a crucial component of the hydrological cycle, significantly impacting water availability, agricultural productivity, and ecological sustainability. Changes in key meteorological variables, such as temperature, humidity, solar radiation, wind speed, and atmospheric pressure, have been shown to significantly affect evaporation patterns and, consequently, water resource balance.[19]

Accurate analysis of these meteorological parameters is therefore a vital prerequisite for developing reliable and site-specific evaporation models, especially those based on the Penman or Penman-Monteith approaches. Their temporal and spatial variability must be thoroughly understood to enhance the accuracy and applicability of the model. In the context of Mbeya city and its surrounding areas, where climatic fluctuations are increasingly impacting water availability, such an analysis is foundational. It not only underpins the development of robust evaporation models but also informs sustainable water resource management, irrigation planning, and climate change adaptation. A comprehensive understanding of these variables will

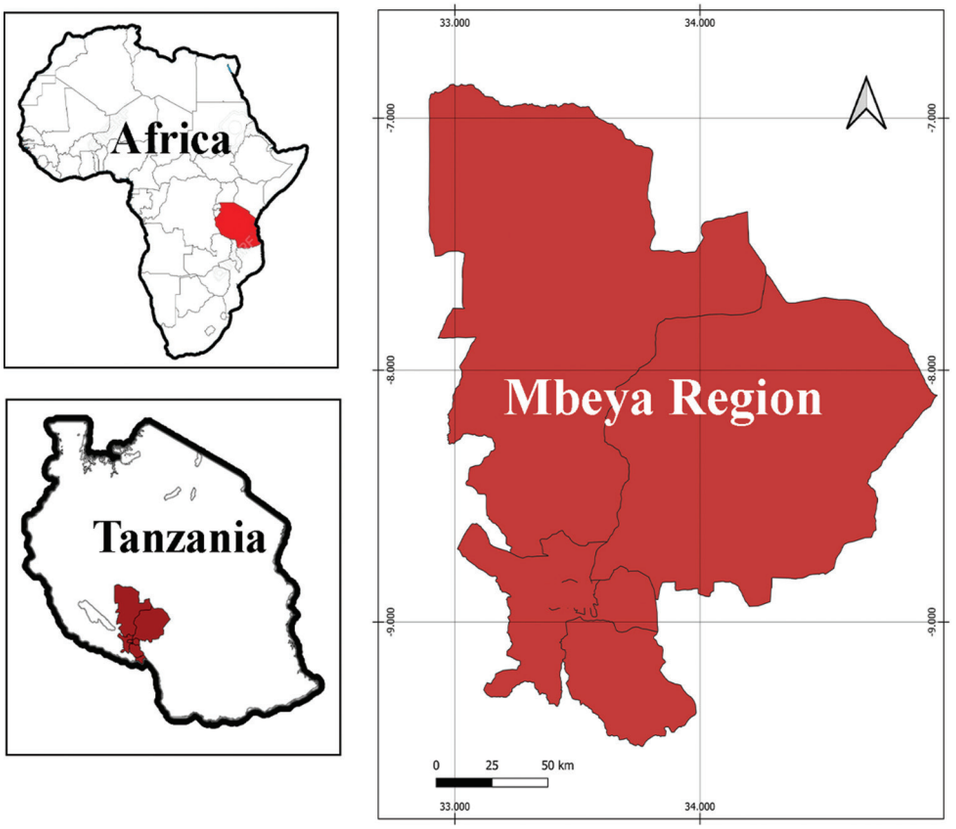

Figure 1. Location of the Mbeya Region in Tanzania, Africa

ultimately support evidence-based decision-making and increase the region's resilience to hydrological extremes.

2.2. Materials and methods

With the fact that PCA is a widely used technique for dimensionality reduction in large datasets, the following steps outline the PCA methodology used and are summarized in Figure 2:

(i) Data use: Meteorological data, including daily maximum and minimum temperatures, humidity, wind speed, and solar radiation, recorded over 10 years, were utilized. The data were averaged from stations within the region. The dataset was structured as a matrix where each row represents an observation, and each column represents a variable.

(ii) Data standardization: To ensure comparability across variables with differing scales and units, each variable x_i was standardized using the z-score transformation:

$$z_i = \frac{x_i - \mu_i}{\sigma_i} \quad \text{(I)}$$

Where μ_i is the mean and σ_i is the standard deviation of x_i. This process centers each variable around zero and scales it to unit variance, enabling meaningful comparison across diverse metrics. Although some variables exhibited non-normal distributions, z-score standardization was chosen over transformations such as the Box-Cox transformation for several reasons. First, the primary aim was not to normalize distributions but to standardize scales for comparability in downstream analyses, such as clustering or regression. Second, unlike the Box-Cox transformation, which requires assumptions about distributional form and can introduce nonlinearity, z-score standardization preserves the original variable relationships and is robust to mild deviations from normality. Thus, it was deemed more appropriate for the analytical goals of this study.

(iii) Calculation of the covariance matrix: The covariance matrix C is computed to assess the relationships between pairs of variables. For two variables x_i and x_j, the covariance is calculated as:

$$C(x_i, x_j) = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \mu_i)(x_{kj} - \mu_j) \quad \text{(II)}$$

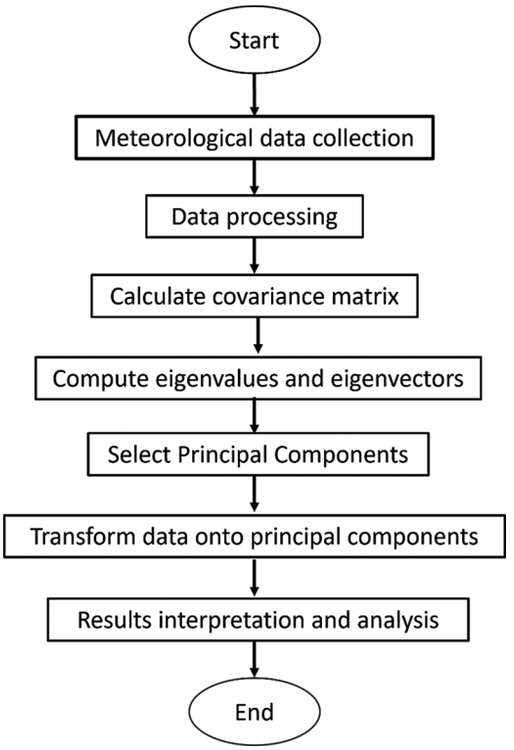

Figure 2. Flowchart of the principal component analysis

where n is the number of observations, and x_{ki} and x_{kj} are the values of variables i and j for the k-th observation.

(iv) Calculate the eigenvalue and eigenvector: Perform eigenvalue decomposition on the covariance matrix C. The eigenvalues λ_i represent the variance explained by each principal component, and the eigenvectors (principal components) v_i represent the direction of maximum variance in the dataset. The relationship is defined as:

$$Cv_i = \lambda_i v_i \quad \text{(III)}$$

(v) Selection of the principal component: Select the principal components corresponding to the largest eigenvalues. The variance explained by each component is proportional to its eigenvalue. The proportion of total variance explained by the k-th principal component is given by:

$$Explained\ variance\ ratio = \frac{\lambda_k}{\sum_{i=1}^{p}\lambda_i} \quad \text{(IV)}$$

where p is the total number of variables.

(vi) Component interpretation: The original data are projected onto the principal components by computing the score matrix Z as:

$$Z = X \cdot V \quad \text{(V)}$$

where X is the standardized data matrix, and V is the matrix of eigenvectors (also known as principal components). The resulting principal component scores can be analyzed to understand which variables contribute most to the variability in evaporation.

(vii) Results interpretation and analysis: The first few principal components, which explain the most variance, were used in regression models to identify the most influential meteorological factors driving the evaporation process.

Skewness and kurtosis can substantially impair the performance of classical PCA, which assumes that data are symmetrically distributed with moderate tail behavior. High skewness distorts component orientation by pulling variance toward extreme values, while elevated kurtosis increases sensitivity to outliers, potentially misrepresenting underlying structure. To address this, robust PCA methods, such as those developed by Hubert *et al*.,[20] have been introduced, specifically tailored to maintain performance on skewed or heavy-tailed data by minimizing the undue influence of outliers and asymmetric distributions. In addition, transformation techniques, such as those proposed by Yang *et al*.,[21] which use Mardia's skewness and kurtosis statistics, can help normalize the data before PCA, improving robustness and interpretability.

Despite these corrections, some variables demonstrated weak or localized correlations, suggesting that latent factors not captured in the global PCA structure may play a role. For example, terrain features, such as elevation gradients or slope orientation, can modulate environmental patterns at microscales. At the same time, urbanization effects, including surface sealing, heat island dynamics, or pollution pockets, can generate spatial heterogeneity not aligned with broader trends. These localized influences may dampen the explanatory power of PCA components, underscoring the need to incorporate spatial contextual information or consider geographically weighted PCA in future analyses.

3. Results and discussion

3.1. Visual analysis of meteorological drivers of evaporation

Figure 3A illustrates the average daily maximum and minimum temperatures, showing consistently high maximum temperatures throughout the year, with minor

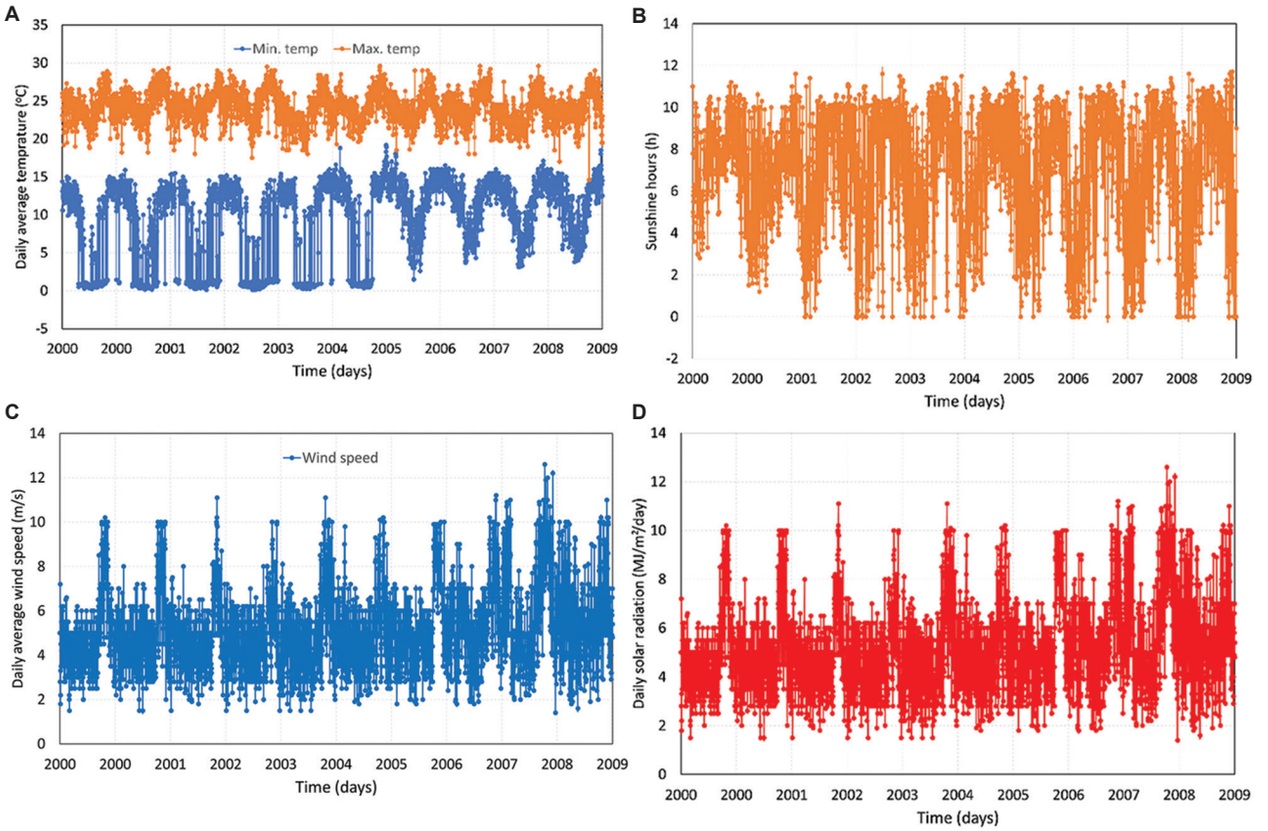

Figure 3. Meteorological variables. (A) Maximum temperature (max. temp.) and minimum temperature (min. temp.). (B) Sunshine hours. (C) Wind speed. (D) Solar radiation

seasonal variation. In contrast, minimum temperatures dropped significantly within the dry season (June to August). This seasonal cooling is influenced by continental climatic factors, where clear skies and low atmospheric humidity during the dry season promote nocturnal radiative cooling and increase diurnal temperature ranges.[18,22] Figure 3B shows the distribution of daily sunshine hours, which peaked during the dry season (May–October) and declined sharply during the rainy season (November–April). This pattern is consistent with semi-arid climates, where reduced cloud cover enhances solar exposure during the dry season.[23,24] In addition, Figure 3C depicts the seasonal trend in wind speed, which increased gradually during the dry season and peaked between August and October. In semi-arid regions such as Mbeya, stronger winds contribute significantly to evaporative demand by increasing the vapor pressure deficit and enhancing moisture loss from soil and vegetation surfaces.[6,24] Figure 3D displays the average daily solar radiation, which mirrors the pattern of sunshine hours, with higher radiation levels during the dry season and lower levels during the wet season. This trend reflects the reduced atmospheric moisture and cloud cover during the dry season, facilitating greater penetration of solar energy to the Earth's surface.[19]

3.2. Descriptive statistics and normality assessment of meteorological variables

Table 1 presents the descriptive statistics and normality test results for five meteorological variables that are considered key drivers of evaporation: radiation, wind speed, maximum temperature, sunshine hours, and minimum temperature. Most variables exhibited substantial deviations from normality. Radiation displayed a moderately high mean (17.60) with a standard deviation of 6.00, and a pronounced left skew (−1.48), suggesting a higher concentration of observations at the upper end of the scale. Although its kurtosis (1.94) indicated a relatively normal peak, the Shapiro-Wilk test decisively rejected the assumption of normality ($W = 0.84$, $p<0.001$). Wind speed, with a modest mean of 5.16 and a standard deviation of 2.00, was notably right-skewed (2.10) and leptokurtic (kurtosis = 17.36), indicating the presence of extremely high values; this was also supported by the Shapiro-Wilk test ($W = 0.89$, $p<0.001$). Meanwhile, maximum temperature was the

most normally distributed variable, with a high mean (24.16), low variability (standard deviation = 2.06), and near-zero skewness (−0.08) and moderate kurtosis (0.32). Yet, it still failed the Shapiro-Wilk test (W = 1.00, p<0.001), indicating subtle non-normality potentially due to large sample size. Sunshine hours exhibited moderate dispersion (standard deviation = 2.87), mild left skewness (−0.64), and a platykurtic profile (−0.51), also violating normality (W = 0.94, p<0.001). In addition, minimum temperature, with a mean of 9.37 and a standard deviation of 5.35, was modestly left-skewed (−0.67) and had a slightly flatter-than-normal distribution (kurtosis = −1.11), also deviating significantly from normality as indicated by the Shapiro-Wilk test (W = 0.84, p<0.001). Overall, wind speed and radiation showed the most pronounced departures from normality, necessitating caution when applying parametric statistical methods that assume a normal distribution.

3.3. Correlation matrix of the meteorological variables

Table 2 presents the correlation matrix for the five key meteorological variables: minimum temperature, sunshine hours, maximum temperature, wind speed, and radiation. The strongest observed relationship was between sunshine hours and radiation (r = 0.66), reflecting a significant positive correlation. This aligns with established climatological principles, as greater solar exposure typically results in higher radiation levels.[25] The minimum temperature exhibited a weak to moderate negative correlation with sunshine hours (r = −0.35) and radiation (r = −0.30), suggesting that cooler nights tend to coincide with shorter daylight periods and reduced solar radiation. These patterns may reflect seasonal or diurnal thermal dynamics, where reduced insolation during specific periods contributes to lower minimum temperatures.[26] A weak positive correlation existed between minimum and maximum temperatures (r = 0.21), implying a slight association between warmer days and warmer nights, though the link remains modest. In addition, maximum temperature exhibited generally weak correlations with other variables, with its strongest relationship to minimum temperature (r = 0.21), followed by sunshine hours (r = 0.12) and wind speed (r = 0.26). These weak associations suggest that maximum temperature may be influenced by more complex or independent atmospheric factors.[27] Wind speed showed the weakest inter-variable correlations, including negligible associations with radiation (r = 0.01) and sunshine hours (r = 0.03), indicating that wind may be shaped by localized or stochastic dynamics rather than broader solar or thermal influences.[28]

Table 1. Descriptive statistics of the meteorological variables

Descriptives	Mininum temperature	Sunshine hours	Maximum temperature	Wind speed	Radiation
Mean	9.37	6.95	24.16	5.16	17.60
Standard deviation	5.35	2.87	2.06	2.00	6.00
Minimum	0.09	0.00	14.30	1.40	0.39
Maximum	19.20	11.70	36.50	30.00	31.40
Skewness	−0.67	-0.64	−0.08	2.10	−1.48
Standard error skewness	0.04	0.04	0.04	0.04	0.04
Kurtosis	−1.11	−0.51	0.32	17.36	1.94
Standard error of kurtosis	0.08	0.08	0.08	0.08	0.08
Shapiro-Wilk (W)	0.84	0.94	1.00	0.89	0.84
Shapiro-Wilk (p)	<0.001	<0.001	<0.001	<0.001	<0.001

Table 2. Correlation matrix of the meteorological variables

Meteorological parameters	Minimum temperature	Sunshine hours	Maximum temperature	Wind speed	Radiation
Min. temperature	1				
Sunshine hours	−0.35	1			
Max. temperature	0.21	0.12	1		
Wind speed	0.21	0.03	0.26	1	
Radiation	−0.30	0.66	0.14	0.01	1

Given these varying degrees of correlation, it is essential to consider their implications for PCA. High correlations, such as those between sunshine hours and radiation, contribute to strong shared variance and can dominate the component structure, potentially overshadowing the unique contributions of individual variables. Conversely, weakly correlated variables, such as wind speed, may load onto separate components or contribute minimally to the explained variance. PCA mitigates multicollinearity by transforming correlated input variables into orthogonal components; however, understanding the interrelationships between these components is crucial for accurate interpretation. Recent studies, including that by Chan *et al*.,[29] emphasize the importance of addressing multicollinearity in predictive modeling using PCA, showing that while it can handle correlated inputs, interpretation must be guided by the underlying data structure. Furthermore, Yonaba *et al*.[4] have demonstrated how machine learning models, when applied to evapotranspiration estimation in the West African Sahel, can be complemented by variable importance tools, such as Shapley additive explanations, to clarify model behavior. Their results revealed that, although models such as random forest, support vector machines, and XGBoost effectively estimated reference evapotranspiration, they tended to overemphasize the role of wind speed. This finding resonates with the weak correlations observed in our analysis. These insights highlight the importance of integrating correlation analysis, PCA, and interpretability methods to ensure robust and meaningful outcomes in meteorological modeling and water resource applications.

3.4. Scatterplot matrix

Figure 4 shows a scatterplot matrix that reveals nuanced interrelationships among key climatic variables,

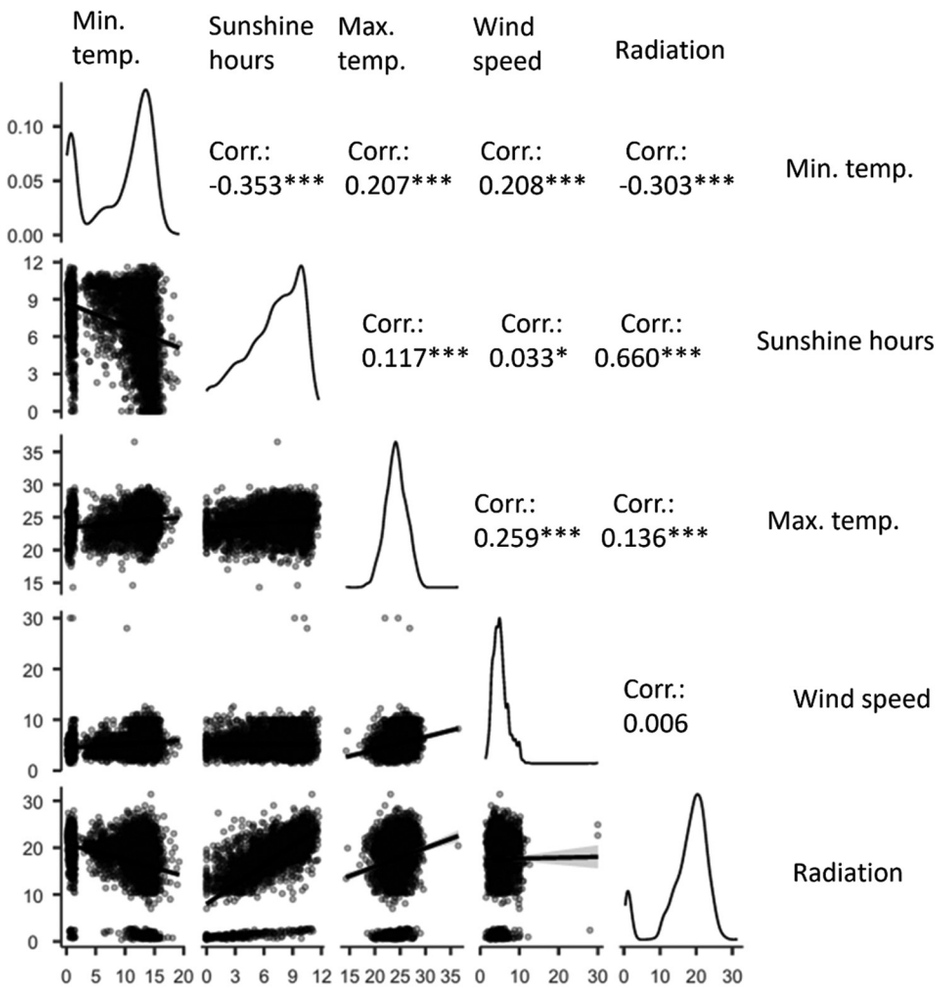

Figure 4. Scatterplot matrix on interrelationships among key climatic variables
Abbreviation: Temp.: Temperature

including minimum temperature, sunshine hours, maximum temperature, wind speed, and solar radiation. A notably strong positive correlation was observed between sunshine hours and solar radiation (r = 0.660, p<0.001), confirming the intuitive link between daylight duration and solar energy availability. This relationship reflects the essential role of sunshine in determining surface radiative flux, as longer sunny periods increase insolation and energy input to the system.[30] In contrast, minimum temperature exhibited a moderate but significant inverse correlation with sunshine hours (r = −0.353, p<0.001) and solar radiation (r = −0.303, p<0.001), suggesting that clearer skies at night promote enhanced radiative cooling and lower minimum temperatures—an effect well-documented in diurnal temperature regulation literature.[31]

Furthermore, maximum temperature was positively associated with sunshine hours (r = 0.117), wind speed (r = 0.259), and radiation (r = 0.136), all statistically significant at p<0.001. These findings support the thermodynamic understanding that higher solar radiation and airflow contribute to elevated daytime temperatures by enhancing heat flux and convective mixing.[32] Wind speed, however, showed negligible correlation with solar radiation (r = 0.006), implying limited influence on radiative transfer processes. This is consistent with studies indicating that cloud cover and atmospheric moisture dominate radiation variability, while wind affects thermal distribution more than radiation magnitude.[24]

Interestingly, the weak yet significant correlation between sunshine hours and wind speed (r = 0.033, p<0.05) may indicate that clearer, drier days tend to coincide with breezier conditions, though the relationship is marginal. Overall, the matrix highlights the central role of sunshine duration and radiation in modulating other climatic variables, especially temperatures, while wind plays a secondary role. These results carry implications for agroclimatic planning, solar energy forecasting, and hydrological modeling in similar climatic zones.

3.5. Suitability of the dataset for PCA

3.5.1. Bartlett's test of sphericity

To assess the suitability of the dataset for PCA, Bartlett's test of sphericity was conducted. This test assesses whether the observed correlation matrix deviates significantly from an identity matrix, assuming that the variables are uncorrelated. The results yielded a Chi-square statistic of 3,353.04 with 10 degrees of freedom and a p<0.001, indicating a highly significant result (Table 3). This confirmed the presence of substantial correlations among the variables and rejected the null hypothesis of independence. Therefore, the dataset possessed adequate shared variance, justifying the use of PCA for dimensionality reduction and latent structure detection.[9]

3.5.2. Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy

The KMO test further supported the dataset's appropriateness for PCA by evaluating the proportion of variance that might be attributed to underlying common factors, as indicated in Table 4. The overall KMO value of 0.59 fell within the "mediocre" range (0.5–0.7), which is considered acceptable for PCA applications in exploratory environmental data analysis.[21] Inspection of individual measures of sampling adequacy (MSA) scores indicated that all variables exceeded the minimum threshold of 0.5, with minimum temperature showing the highest MSA (0.66), followed by wind speed (0.58), radiation (0.58), sunshine hours (0.57), and maximum

Table 3. The Bartlett's test of sphericity

Chi-square statistic (χ^2)	Degree of freedom	p-value
3,353.04	10	<0.001

Table 4. Kaiser-Meyer-Olkin measure of sampling adequacy (MSA)

Description	MSA
Overall	0.59
Minimum temperature	0.66
Sunshine hours	0.57
Maximum temperature	0.55
Wind speed	0.58
Radiation	0.58

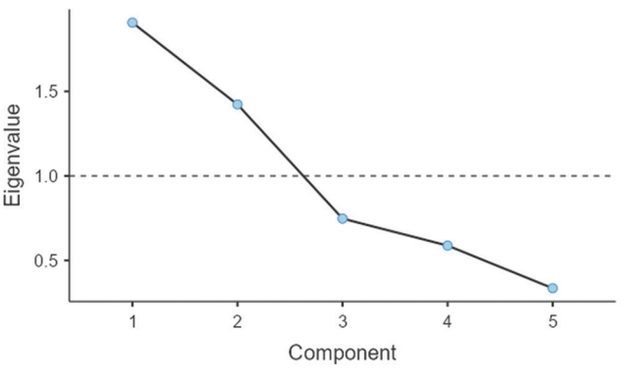

Figure 5. Scree plot of the principal components and eigenvalues

temperature (0.55). These values suggest that each variable shares sufficient common variance with others to justify its inclusion in PCA, with no variables falling below the critical cutoff.

3.6. Eigenvalues and component retention

The PCA was conducted to reduce dimensionality and identify latent patterns within the meteorological dataset. Figure 5 presents the scree plot, whereas Tables 5 and 6 summarize the initial eigenvalues, variance explained, and rotated component loadings. The scree plot illustrates a sharp decline from the first to the second component, followed by a more gradual slope, forming a distinct "elbow" at the transition from the first to the second component. This visual inflection point supported the application of the elbow method, indicating that the first two components captured the most meaningful variance in the dataset. Consistent with Kaiser's criterion, which recommends retaining components with eigenvalues >1.0, only the first two components met this threshold. These were Component 1 with an eigenvalue of 1.90 and Component 2 with 1.43, jointly explaining 66.61% of the total variance.[33]

Component loadings following varimax rotation further clarified the interpretation of these components. Component 1 exhibited strong loadings for radiation (0.87) and sunshine hours (0.89), indicating a latent dimension associated with solar energy or exposure. Meanwhile, Component 2 exhibited high loadings for minimum temperature (0.53) and maximum temperature (0.77), suggesting a thermal gradient or temperature-related variation. Wind speed demonstrated the highest uniqueness value (0.46), implying that its variance is mainly independent of the shared components and may reflect localized meteorological influences.

The cumulative variance explained by Components 1 and 2–38.06% and 28.55%, respectively-totals 66.61%, which falls within the acceptable range (60–70%) for dimensionality reduction in environmental data studies.[7] This level of explanatory power supported reducing the dataset to two principal dimensions, effectively simplifying the structure while preserving most of the relevant information. Overall, the retained components offered meaningful insight into the dominant factors driving variation in the dataset, capturing both solar-driven and temperature-driven environmental processes.

The initial eigenvalues provided insight into how variances were distributed across principal components. Component 1 exhibited the highest eigenvalue (1.90), accounting for 38.06% of the total variance, followed by Component 2 (eigenvalue = 1.43, 28.55%). Together, these three components explain 66.61% of the total variance, which is approximately 70% benchmark often considered sufficient for reliable dimensionality reduction in environmental modelling.[7]

Table 5. Sum of square (SS) loadings of the principal components

Component	SS loadings	Percentage of variance	Cumulative percentage
1	1.90	38.06	38.06
2	1.43	28.55	66.61

Table 6. Summary of component statistics

Variable	Component 1	Component 2	Uniqueness
Min. temperature	−0.58	0.53	-
Sunshine hours	0.89	-	0.21
Max. temperature	-	0.77	0.38
Wind speed	-	0.73	0.46
Radiation	0.87	-	0.23

3.7. Factor score coefficients for the first two principal components

Figure 6 and Table 6 present the factor score coefficients for the first two principal components, illustrating how the five meteorological variables–minimum temperature, sunshine hours, maximum temperature, wind speed, and radiation–contributed to the underlying dimensions extracted through PCA. Component 1 was dominated by strong positive coefficients for sunshine hours (0.89) and radiation (0.87), suggesting that this

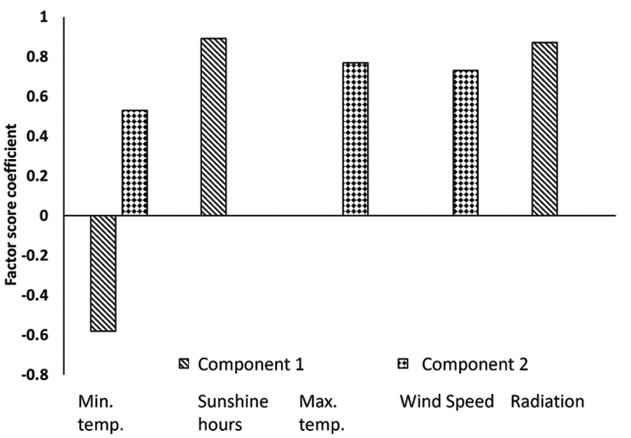

Figure 6. Factor score coefficients for the two principal components
Abbreviation: Temp.: Temperature

component primarily captures variability related to solar energy exposure. The high magnitude of these coefficients indicates that this dimension reflects a solar-driven environmental gradient. Interestingly, minimum temperature displayed a negative coefficient (−0.58) on this component, implying an inverse relationship; lower minimum temperatures tend to occur during periods of reduced solar exposure, such as during colder, cloudier seasons. This negative association aligns with climatological patterns where low nocturnal temperatures coincide with limited daytime radiation.

Component 2, in contrast, was predominantly shaped by maximum temperature (0.77) and minimum temperature (0.53), indicating its role in capturing the thermal variability within the dataset. Wind speed also contributed moderately (0.73), suggesting its role in modulating or reflecting thermal conditions, potentially through processes such as convective heat transfer. The dual loading of minimum temperature–negatively on Component 1 and positively on Component 2–reveals its complex behaviour as both inversely related to solar input and directly associated with broader thermal dynamics. The separation of solar and thermal variables into distinct components is consistent with previous findings in environmental multivariate analyses, where PCA often reveals interpretable dimensions aligned with natural climatic processes.[7,9] This apparent dimensionality enhances the interpretability of the PCA results and provides a solid foundation for subsequent modeling or clustering analyses.

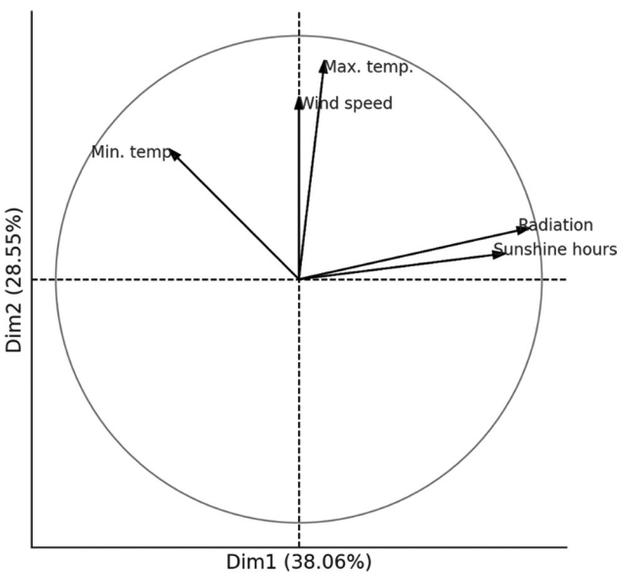

Figure 7. Principal component analysis biplot
Abbreviation: Temp.: Temperature

3.8. PCA biplot

Figure 7 shows the PCA biplot, illustrating the dimensional contributions of the five key climatic variables–minimum temperature, maximum temperature, sunshine hours, wind speed, and solar radiation–projected onto the first two principal components (Dim1 and Dim2), which explain 38.06% and 28.55% of the total variance, respectively. Together, these two dimensions captured 66.61% of the overall variability in the dataset, indicating a reasonably efficient dimensionality reduction.

The variables solar radiation and sunshine hours were positioned closely and strongly aligned along Dim1, indicating a strong positive correlation between them. This alignment suggests that solar exposure is the dominant axis of climatic variation, consistent with prior findings showing sunshine duration as a key driver of solar energy availability and daily temperature fluctuation.[30] Their long vector lengths confirmed high contributions to the total variance, indicating that these variables were well-represented in the PCA space and exhibited minimal redundancy. On the other hand, minimum temperature pointed in the opposite direction to solar radiation and sunshine hours, implying a negative relationship. This potentially reflects radiative cooling at night, which is enhanced during clear-sky conditions, leading to low nighttime temperatures when sunshine duration is high.[31] This divergence is key in distinguishing between daytime heating and nocturnal cooling dynamics. Furthermore, maximum temperature and wind speed were projected moderately along Dim2, showing some shared variance but lower explanatory power compared to radiation-related variables. Wind speed's vector was shorter, indicating it contributes less to total variability, potentially because it is influenced more by synoptic conditions than local radiative or thermal effects.[24] Maximum temperature aligned partially with both sunshine hours and wind speed, consistent with its role as an integrator of radiative and advective processes.[32]

The PCA biplot provides compelling evidence of the dominant role played by solar exposure, specifically radiation and sunshine hours, in driving climatic variability in Mbeya. It also reveals the clear differentiation between daytime warming (maximum temperature) and nighttime cooling (minimum temperature) influences, underscoring the thermal dichotomy inherent in local climatic processes. These findings support the utility of PCA as a robust statistical tool for disentangling complex inter-variable relationships within climate datasets, enabling practical

simplification for use in meteorological modeling and environmental forecasting.

When compared with similar regional-scale studies,[11] which analyzed climate-driven productivity trends across China from 1980 to 2018, the contrast is notable. While Cao *et al*.[11] found that precipitation and temperature were the most influential variables shaping ecosystem productivity across various Chinese biomes, solar radiation played a relatively secondary role. This divergence highlights the regional uniqueness of Mbeya, where solar exposure, not moisture or broad thermal trends, emerges as the primary climatic driver of evaporation. Such discrepancies underscore the importance of localized analysis, as climatic influence hierarchies can vary significantly based on geography, topography, and land use intensity.

3.9. Broader environmental influences on evaporation beyond meteorological drivers in urban climates

While the current study has successfully identified primary meteorological drivers of evaporation using PCA, it is imperative to recognize that evaporation is a multifactorial process influenced by a broader array of environmental determinants beyond atmospheric variables. Chief among these are surface characteristics and land use and land cover (LULC) dynamics, which significantly modulate the surface energy balance and, consequently, evaporation rates. Parameters such as surface albedo, aerodynamic roughness, soil moisture content, and vegetative cover play crucial roles in mediating how meteorological inputs, such as solar radiation and temperature, are converted into latent heat fluxes.

Urban environments, particularly in rapidly developing regions such as Mbeya, exhibit distinct surface energy dynamics due to expanding impervious surfaces, fragmented green spaces, and altered thermal properties. Impervious materials, such as asphalt, concrete, and roofing tiles, increase surface albedo and thermal storage, intensifying daytime heating while simultaneously limiting water retention and soil moisture availability. This decouples incoming solar energy from latent heat processes, reducing evaporation despite high solar radiation levels.[34,35] These altered dynamics also suppress transpiration by reducing vegetation density and interception storage, weakening the correlation between meteorological variables and observed evaporation patterns.

Recent studies have underscored the critical role of land surface heterogeneity in shaping urban evapotranspiration. For example, Chen *et al*.[35] found that urbanization led to significant decreases in evapotranspiration and its vegetative and soil components, while increasing sensible heat flux and Bowen ratio. These findings support our PCA results showing localized and weak correlations in Mbeya's urban areas, potentially due to unaccounted spatial variation in surface properties. Integrating LULC data into modeling frameworks, especially through hybrid methods, such as PCA-RF or PCA-artificial neural network (ANN), can enhance the representation of nonlinear interactions and feedbacks between land cover and atmospheric drivers.[11,14]

For stakeholders, especially in agriculture and urban water management, actionable strategies are essential. One promising solution is solar-based irrigation scheduling, which leverages real-time solar radiation data to optimize irrigation frequency and volume in response to crop evapotranspiration needs. This is particularly relevant for Mbeya, which experiences strong diurnal radiation cycles. In urban settings, expanding green infrastructure, such as vegetative corridors, permeable pavements, and rooftop gardens, can mitigate the effects of impervious surfaces, restore soil moisture feedbacks, and enhance microclimatic conditions. These integrative, evidence-based strategies can guide adaptive planning in Mbeya under intensifying urbanization and climate change.

4. Conclusion

This study highlights solar radiation and sunshine hours as the dominant meteorological determinants of evaporation in Mbeya, Tanzania, with thermal parameters, particularly minimum and maximum temperatures, playing secondary yet significant roles. Through the application of PCA, the study effectively reduced dimensionality while retaining essential variance, thereby enhancing the interpretability of complex atmospheric interactions. Component 1, representing radiative processes, accounted for 38.06% of the total variance and featured strong loadings from sunshine hours (0.89) and solar radiation (0.87). Component 2, associated primarily with thermal variability, explained an additional 28.55%, with maximum temperature (loading = 0.77) as its leading contributor. Together, these two components explained 66.61% of the total variance, underscoring the dominant influence of solar and thermal dynamics on evaporation. Wind speed, in contrast, demonstrated weak intervariable correlations and moderate loadings, suggesting its role may be more

localized or influenced by stochastic factors rather than broader climatic trends. These findings are consistent with existing literature that emphasizes the significance of radiative inputs in evaporation modeling. However, the exclusion of critical surface-related variables, such as land use, vegetation cover, albedo, and soil moisture contents, limits the comprehensiveness of the model. Such variables are known to mediate how meteorological inputs translate into actual evaporation rates and should be integrated into future studies to provide a holistic understanding of evapotranspiration processes.

To enhance the accuracy and applicability of future evaporation modeling in urban climates such as Mbeya, some recommendations are provided here. Researchers should incorporate both meteorological and surface-based environmental variables, particularly those related to land use and surface characteristics. Factors such as vegetation cover, impervious surfaces, soil moisture content, surface albedo, and vegetation indices significantly influence the surface energy balance and modulate how atmospheric drivers translate into actual evaporation rates. Furthermore, integrating high-resolution LULC data would enable more spatially explicit modeling, improve prediction accuracy, and reflect the heterogeneity of urban landscapes. Moreover, combining PCA with advanced hybrid modeling techniques, such as ANN, RF, or SVM, would leverage the strengths of dimensionality reduction and nonlinear pattern recognition, thereby enhancing model robustness. Seasonal disaggregation of meteorological data is also recommended, as it would provide granular insights into temporal variability and enable the formulation of season-specific water management strategies. In addition, expanding the meteorological dataset to include variables such as relative humidity, vapor pressure deficit, and cloud cover would offer a more comprehensive understanding of the atmospheric drivers of evaporation. Attention should also be given to the statistical distribution of input variables; significant skewness should be addressed through transformation techniques, such as the Box-Cox method, to improve model performance. Finally, conducting similar analyses across diverse climatic zones would improve the generalizability of findings and inform climate-resilient urban planning, irrigation management, and green infrastructure development.

Acknowledgments

The author gratefully acknowledges the moral support received from the Mbeya University of Science and Technology through the Department of Civil Engineering.

Funding

None.

Conflict of interest

The author declares no competing interests.

Author contributions

This is a single-authored article.

Availability of data

Data is available from the corresponding author upon reasonable request.

Further disclosure

The paper has been deposited in the preprint server Research Square (doi: 10.21203/rs.3.rs-5336289/v1).

References

1. Moges S, Katambara Z, Bashar K. Decision support system for estimation of potential evapo-transpiration in Pangani Basin. *Phys Chem Earth Parts A/B/C*. 2003;28(20-27):927-934. doi: 10.1016/j.pce.2003.08.038
2. Fan J, Zheng J, Wu L, Zang F. Estimation of daily maize transpiration using support vector machines, extreme gradient boosting, artificial and deep neural networks models. *Agric Water Manag*. 2021;245:106547. doi: 10.1016/j.agwat.2020.106547
3. Han Y, Calabrese S, Du H, Yin J. Evaluating biases in Penman and Penman-Monteith evapotranspiration rates at different timescales. *J Hydrol*. 2024;638:131534. doi: 10.1016/j.jhydrol.2024.131534
4. Yonaba R, Tazen F, Cissé M, *et al*. Trends, sensitivity and estimation of daily reference evapotranspiration ET0 using limited climate data: Regional focus on Burkina Faso in the West African Sahel. *Theor Appl Climatol*. 2023;153(1):947-974. doi: 10.1007/s00704-023-04507-z
5. Masson-Delmotte V, Zhai P, Pirani A, *et al*. *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Vol. 2. Cambridge: Cambridge University Press; 2021. p. 2391. doi: 10.1017/9781009157896
6. Tsujimoto K, Masumoto T, Mitsuno T. Seasonal changes in radiation and evaporation implied from the diurnal distribution of rainfall in the Lower Mekong. *Hydrol*

Process Int J. 2008;22(9):1257-1266
doi: 10.1002/hyp.69357
7. Haddad K, Rahman A, Stedinger JR. Regional flood frequency analysis using Bayesian generalized least squares: A comparison between quantile and parameter regression techniques. *Hydrol Process*. 2012;26(7):1008-1021.
doi: 10.1002/hyp.8189
8. Wold S, Ruhe A, Wold H, *et al*. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM J Sci Stat Comput*. 1984;5(3):735-743.
doi: 10.1137/0905052
9. Jolliffe IT, Cadima J. Principal component analysis: A review and recent developments. *Philos Trans A Math Phys Eng Sci*. 2016;374(2065):20150202.
doi: 10.1098/rsta.2015.0202
10. Ma J, Zhou L, Foltz GR, *et al*. Hydrological cycle changes under global warming and their effects on multiscale climate variability. *Ann N Y Acad Sci*. 2020;1472(1):21-48.
doi: 10.1111/nyas.14335
11. Cao W, Zhang SJ, Lu ZL, *et al*. Prediction for Water Surface Evaporation Based on PCA and RBF Neural Network. In: *International Conference on Information Computing and Applications*, Springer; 2011. p. 351-358.
doi: 10.1007/978-3-642-25255-6_45
12. James G, Witten D, Hastie T, Tibshirani R, Taylor J. *Linear Regression. An Introduction to Statistical Learning: With Applications in Python*. Berlin: Springer; 2023. p. 69-134.
13. Ullah I, Ma X, Ren G, *et al*. Recent changes in drought events over South Asia and their possible linkages with climatic and dynamic factors. *Remote Sens*. 2022;14(13):3219.
doi: 10.3390/rs14133219
14. Jafari M, Dinpashoh Y. Derivation of regression models for pan evaporation estimation. *Environ Resour Res*. 2019;7(1):29-42.
15. Tianxiao L, Qiang F, Shuqin X, *et al*. Application of Principal Component Analysis in Evaluating Influence Factors of Evaporation in Northern Cold Area. In: *2009 Fifth International Conference on Natural Computation: IEEE*; 2009. p. 514-518.
doi: 10.1109/icnc.2009.721
16. Dupre K. Trends and gaps in place-making in the context of urban development and tourism: 25 years of literature review. *J Place Manage Dev*. 2019;12(1):102-120.
doi: 10.1108/JPMD-07-2017-0072
17. Gwaleba MJ. Urban growth in Tanzania: Exploring challenges, opportunities and management. *Int J Soc Sci Stud*. 2018;6:47.
doi: 10.11114/ijsss.v6i12.3783
18. Tanzania URo. *The United Republic of Tanzania Agricultural Sector Development Programme (ASDP) Support through Basket Fund Government Programme Document*; 2017.
19. Liang S, Fang H, Chen M. Atmospheric correction of Landsat ETM+ land surface imagery. I. Methods. *IEEE Trans Geosci Remote Sens*. 2002;39(11):2490-2498.
doi: 10.1109/36.964986
20. Hubert M, Rousseeuw P, Verdonck T. Robust PCA for skewed data and its outlier map. *Comput Stat Data Anal*. 2009;53(6):2264-2274.
doi: 10.1016/j.csda.2008.05.027
21. Yang J, Ye M, Tang Z, *et al*. Using cluster analysis for understanding spatial and temporal patterns and controlling factors of groundwater geochemistry in a regional aquifer. *J Hydrol*. 2020;583:124594.
doi: 10.1016/j.jhydrol.2020.124594
22. New M, Lister D, Hulme M, *et al*. A high-resolution data set of surface climate over global land areas. *Clim Res*. 2002;21(1):1-25.
doi: 10.3354/cr021001
23. Nicholson SE. Climate and climatic variability of rainfall over eastern Africa. *Rev Geophys*. 2017;55(3):590-635.
doi: 10.1002/2016RG000544
24. Allen RG, Pereira LS, Raes D, Smith M. *Crop Evapotranspiration-Guidelines for Computing Crop Water Requirements-FAO Irrigation and Drainage Paper no 56*. Vol. 300. Rome: FAO; 1998. p. D05109.
25. Antuña-Sánchez JC, Estevan R, Román R, *et al*. Solar radiation climatology in Camagüey, Cuba (1981-2016). *Remote Sens*. 2021;13(2):169.
doi: 10.3390/rs13020169
26. Noel DD, Justin KGA, Alphonse AK, *et al*. Normality assessment of several quantitative data transformation procedures. *Biostat Biometr Open Access J*. 2021;10(3):51-65.
doi: 10.19080/BBOAJ.2021.10.5557786
27. Przeździecki K, Zawadzki JJ, Urbaniak M, *et al*. Using temporal variability of land surface temperature and normalized vegetation index to estimate soil moisture condition on forest areas by means of remote sensing. *Ecol Indic*. 2023;148:110088.
doi: 10.1016/j.ecolind.2023.110088
28. Chang TP, Liu FJ, Ko HH, *et al*. Oscillation characteristic study of wind speed, global solar radiation and air temperature using wavelet analysis. *Appl Energy*. 2017;190:650-657.
doi: 10.1016/j.apenergy.2016.12.149
29. Chan JY, Leow SMH, Bea KT, *et al*. Mitigating the multicollinearity problem and its machine learning approach: A review. *Mathematics*. 2022;10(8):1283.
doi: 10.3390/math10081283
30. Stanhill G, Cohen S. Global dimming: A review of the evidence for a widespread and significant reduction in global radiation with discussion of its probable causes and possible agricultural consequences. *Agric Forest Meteorol*. 2001;107(4):255-278.
doi: 10.1016/S0168-1923(00)00241-0

31. Matuszko D, Węglarczyk S. Effect of cloudiness on long-term variability in air temperature in Krakow. *Int J Climatol*. 2014;34(1):145-154. doi: 10.1002/joc.3672
32. Sarrat C, Lemonsu A, Masson V, Guedalia D. Impact of urban heat island on regional atmospheric pollution. *Atmos Environ*. 2006;40(10):1743-1758. doi: 10.1016/j.atmosenv.2005.11.037
33. Rojas-Valverde D, Pino-Ortega J, Gómez-Carmona CD, *et al*. A systematic review of methods and criteria standard proposal for the use of principal component analysis in team's sports science. *Int J Environ Res Public Health*. 2020;17(23):8712. doi: 10.3390/ijerph17238712
34. Barnes KB, Morgan J, Roberge M. *Impervious Surfaces and the Quality of Natural and Built Environments*. Baltimore: Department of Geography and Environmental Planning, Towson University; 2001.
35. Chen H, Huang JJ, Dash SS, McBean E, Wei Y, Li H. Assessing the impact of urbanization on urban evapotranspiration and its components using a novel four-source energy balance model. *Agric For Meteorol*. 2022;316:108853. doi: 10.1016/j.agrformet.2022.108853