The Original

MULTIVARIATE STATISTICAL ANALYSIS OF SOIL-WATER-IRRIGATION SYSTEMS: METHODS, APPLICATIONS AND RESEARCH GAPS

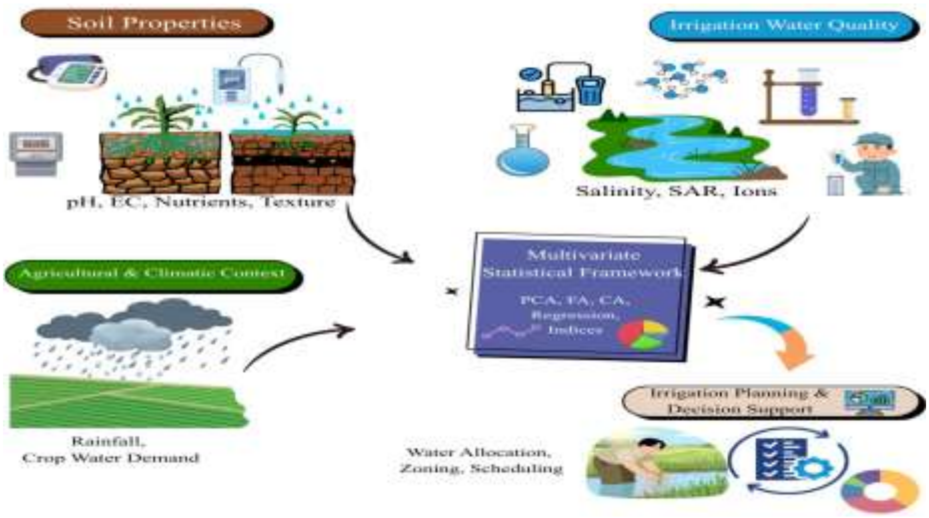
Kanika K[1], Gangai Selvi R[2*], Balaji Kannan[3], Angles S[4], Sridevi G[5]

[1]PG Scholar, Department of Physical Sciences and Information Technology, TNAU, Coimbatore, Tamil Nadu, India. E-mail: kkaagri2020@gmail.com
[2]Professor (Statistics), Department of Physical Sciences and Information Technology, TNAU, Coimbatore, Tamil Nadu, India. E-mail: gangastat@tnau.ac.in
[3]Professor (Soil &Water Conservation Engineering), Department of Physical Sciences and Information Technology, TNAU, Coimbatore, Tamil Nadu, India. E-mail: balajikannan@tnau.ac.in
[4]Professor (Agricultural Economics), Department of Agricultural Economics, TNAU, Coimbatore, Tamil Nadu, India. E-mail: angles.s@tnau.ac.in
[5]Assistant Professor (Soil Science and Agricultural Chemistry), Department of Soil Science and Agricultural Chemistry, TNAU, Coimbatore, Tamil Nadu, India. E-mail: sridevi@tnau.ac.in
*Corresponding author: gangastat@tnau.ac.in

ABSTRACT
Soil-water-irrigation systems are increasingly being affected by the deterioration of water quality, soil conditions and climatic changes. These complex datasets are highly interrelated. Traditional univariate analysis methods have limitations in effectively addressing the complex interrelationship between variables. Therefore, there is a need for the application of robust multivariate statistical analysis. This study critically reviews the application of multivariate statistical analysis in the soil-water-irrigation system, with a focus on irrigation planning and management. The study integrates the most commonly used data structures for the analysis of the variables in the agricultural irrigation system. The variables in the study include soil physicochemical characteristics, irrigation water quality variables and agro-climatic variables. The most commonly used multivariate statistical analysis methods, such as Principal Component Analysis, Factor Analysis, Cluster Analysis, regression models, Principal Component Regression, Partial Least Square Regression and composite indices, are critically discussed in terms of their advantages, limitations and applicability in irrigation planning and management. The study focuses on the most critical methodological aspects of multivariate statistical analysis, which have a significant influence on the outcome of the analysis. In addition, the review also reveals some significant gaps in the research in terms of methodological validation, integration of multivariate models into decision support systems and difficulties in cross-regional applications. Some emerging trends in data analysis and opportunities for combining multivariate analysis with other advanced data-driven techniques are also presented in this review, providing a framework for future research and irrigation management strategies.

KEYWORDS Multivariate; Soil Quality; Water Quality; Irrigation; PCA; Clustering

Graphical Abstract



INTRODUCTION

In the era of digitalization, the agriculture sector is taking up data-driven solutions to tackle challenges like water scarcity, soil degradation and efficient irrigation management. Current soil and irrigation water quality studies in agriculture rely on multi-parameter, multi-spatial and multi-temporal datasets of interrelated physicochemical variables. The datasets are frequently multi-dimensional and record soil properties as well as data on water properties used for irrigation, climatic information and other environmental parameters. Univariate or bivariate statistical analysis has been inadequate for analysing such datasets as they do not account for the relationships between multiple variables, which can result in incomplete or inaccurate conclusions (Shrestha and dimensionality reduction and understanding relationships between variables in complex environmental datasets include but are not limited to the aforementioned Principal Component Analysis (PCA), Factor Analysis (FA) and Cluster Analysis (CA) as well as regression-based techniques (Hallouz et al., 2022). These techniques allow for data to be represented in a compact and informative way that still conveys meaning and is easy to understand. In the past, multivariate statistical techniques have been proven to be useful for assessing soil and water quality as well as in the decision-making process for protecting the environment (Vega et al., 1998; Abdel-Fattah et al., 2020; de Andrade Costa et al., 2020).

Multivariate approaches for data interpretation and decision support are also gaining importance in recent methodological developments in agricultural and environmental research. These methods are useful for handling multicollinearity issues and natural variability found in agro-environmental datasets, as well as varying scales of measurements (Pullanagari and Cavalli, 2023; Ghoto et al., 2025). But most of the available literature still looks at soil properties and/or irrigation water quality in isolation and generally does not pay much attention to generalizing to regions and datasets. This is a major constraint on developing integrated insights that could help manage land and water resources at a regional level and promote sustainable water use (Minhas et al., 2020; Li and Ren, 2021).

The availability of secondary data sets from soil testing laboratories, irrigation monitoring programs, agricultural agencies and remote sensing platforms is growing and offers new opportunities for large-scale data-driven analysis. These data sources allow for researchers to study soil–water systems over a wide range of spatial and temporal scales, without depending on intensive field experiments alone. However, if such data sets are to be used effectively, they need a reliable statistical framework that can accommodate missing values, variability in scale, and complex interactions between the data (Helfenstein et al., 2024). Although multivariate techniques are increasing used, there is not an extensive review which integrates soil and water data analysis from a statistical point of view to manage irrigation. To fill this void, the present review combines and reviews literature in order to investigate the characteristics of data, the prevailing trends, and some research gaps in the use of multivariate statistical techniques for soil–water systems under agriculture conditions(Kazama, 2007; Sudhakaran et al., 2020).

2. METHODOLOGY

This study employs a systematic literature review and bibliometric analysis to examine research trends in water quality, groundwater, soil quality and agricultural studies using multivariate statistical techniques, particularly principal component analysis (PCA). The review was conducted following PRISMA 2020 guidelines using a Scopus-based dataset. Bibliometric analysis was performed using R (Bibliometrix/Biblioshiny), while VOSviewer was used for keyword co-occurrence and network visualization (Xiao *et al.,* 2025).

2.1 Systematic review protocol

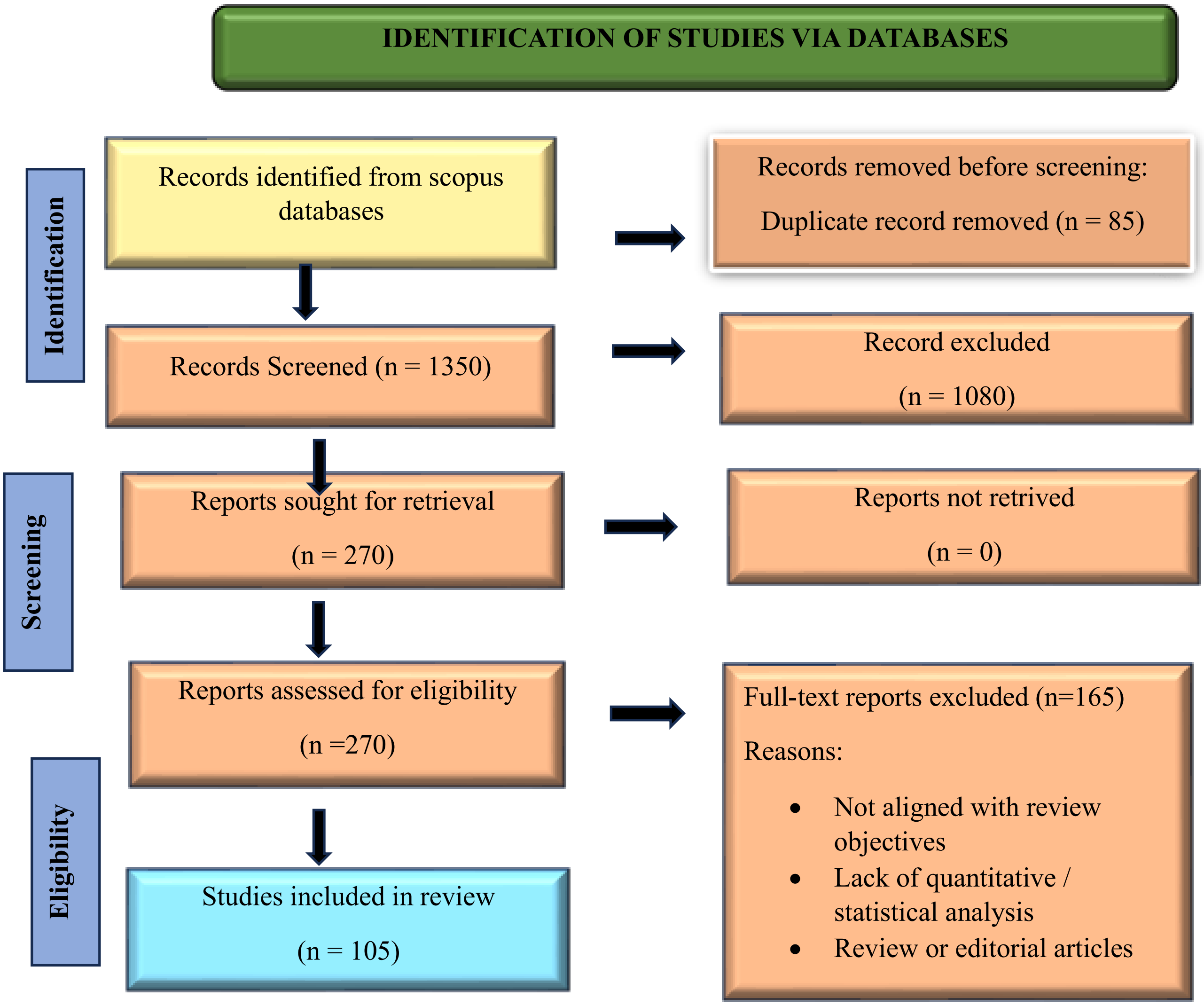
This review follows the PRISMA 2020 guidelines for the identification, screening, eligibility and inclusion of studies (Page *et al.,* 2020). The protocol defined the population as studies related to water and soil systems, the concept as water quality assessment using multivariate statistical approaches and the context as environmental and agricultural applications across all geographic regions. A combined systematic literature review and bibliometric approach was adopted, in which eligible Scopus records were first analysed bibliometrically and subsequently screened for qualitative synthesis.

2.2 Justification for Use of Scopus Database

The Scopus database was selected due to its broad coverage of peer-reviewed literature in agricultural and environmental sciences, enabling reliable systematic and bibliometric analysis. Using a single database ensures consistency in bibliometric indicators and analysis methods. However, it is acknowledged that some relevant publications indexed in other databases may not be captured, which may result in limited omission when assessing research trends and gaps.

2.3 Search strategy and data sources

The primary data source for this review was the Scopus database, from which relevant records were identified using an advanced TITLE-ABS-KEY search strategy. The search combined terms related to water quality, groundwater, soil quality, agriculture and multivariate statistical methods, using appropriate Boolean operators. The search was limited to studies published between 2020 and 2025, restricted to article document type, English language and relevant environmental and agricultural subject areas. To ensure thematic relevance, records unrelated to water-soil systems or statistical analysis were excluded during screening. The final dataset was used for bibliometric analysis and systematic synthesis.



2.4 Inclusion and exclusion criteria

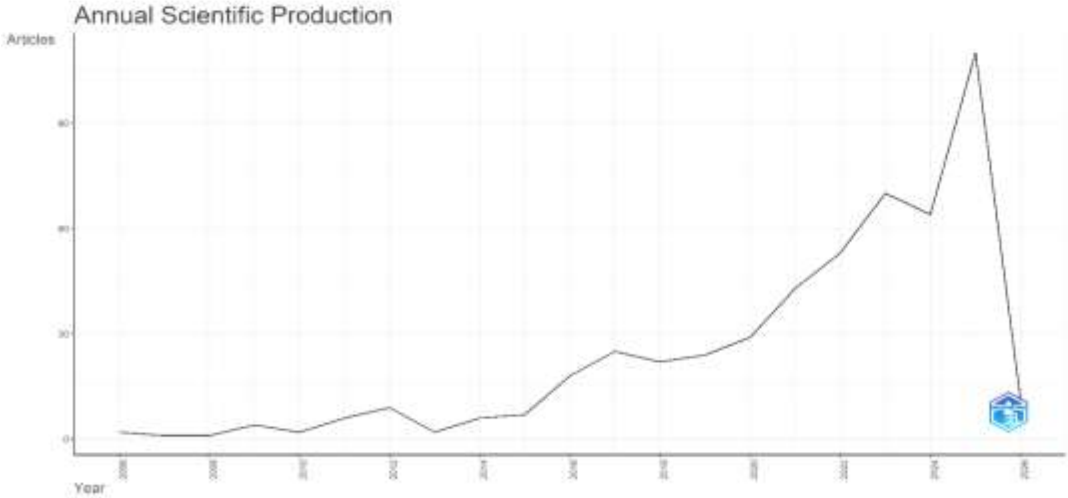
The inclusion criteria applied to the Scopus database were as follows: document type restricted to research articles, English language, final publication stage, all open-access categories and relevant environmental, agricultural and earth science subject areas. The search strategy targeted studies addressing water quality, groundwater, soil quality and agricultural systems, with a particular focus on the application of multivariate and statistical analysis methods. During the screening process, records were excluded if they were unrelated to water or soil quality assessment, lacked quantitative or statistical analysis, or focused solely on review articles, editorials, or conceptual discussions. Additional exclusions were applied to studies outside the scope of environmental and agricultural applications. Following title, abstract and full-text screening, only studies meeting the predefined eligibility criteria were retained for bibliometric analysis and qualitative synthesis.

2.5 Bibliometric analysis

Bibliometric analysis was conducted to examine the growth, structure and thematic evolution of research on water quality, groundwater, soil quality and agricultural applications using multivariate statistical approaches. Based on the Scopus-derived dataset, publication and keyword data were analysed using Biblioshiny (R) and VOS viewer to identify research trends, dominant themes and relationships among keywords through co-occurrence and thematic mapping.

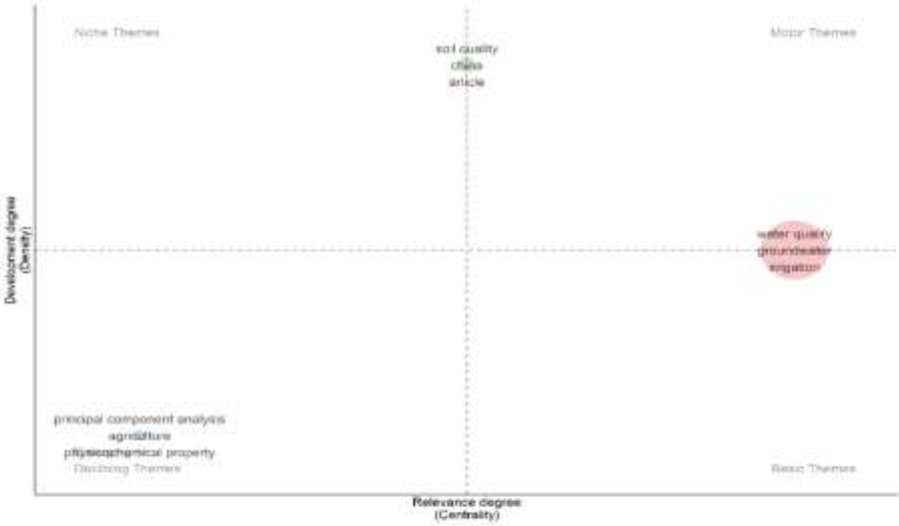
2.5.1 Annual scientific production

Figure 1. Illustrates the annual scientific production in the study area. The results show limited publication activity before 2012, followed by a steady increase after 2015 and a sharp growth from 2020 onwards. This trend indicates a rapidly expanding research interest in water and soil quality assessment, driven by increasing concerns over environmental sustainability and the wider adoption of multivariate statistical techniques.



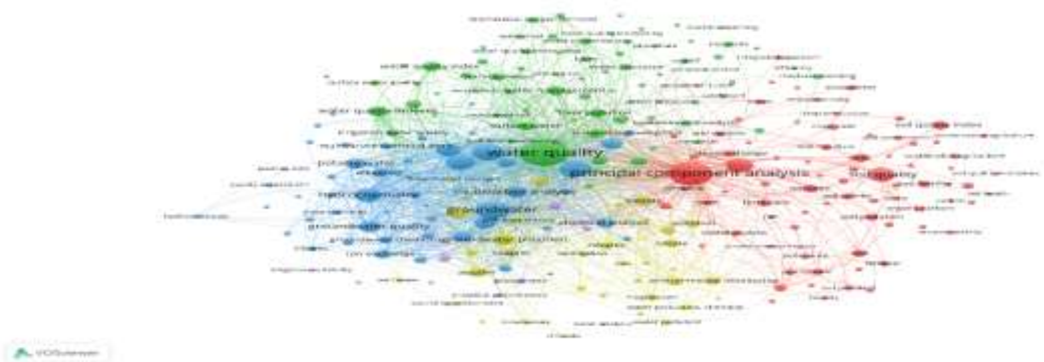
2.5.2 Thematic structure

Figure 2. Presents the thematic map based on keyword centrality and density. Motor themes, including *water quality*, *groundwater* and *irrigation*, represent well-developed and influential research areas. Basic themes highlight foundational topics that structure the field but require further methodological integration. Niche themes reflect specialized applications with high internal development; while emerging or declining themes suggest areas that are either gaining attention or becoming less central to current research.



2.5.3 Keyword co-occurrence network

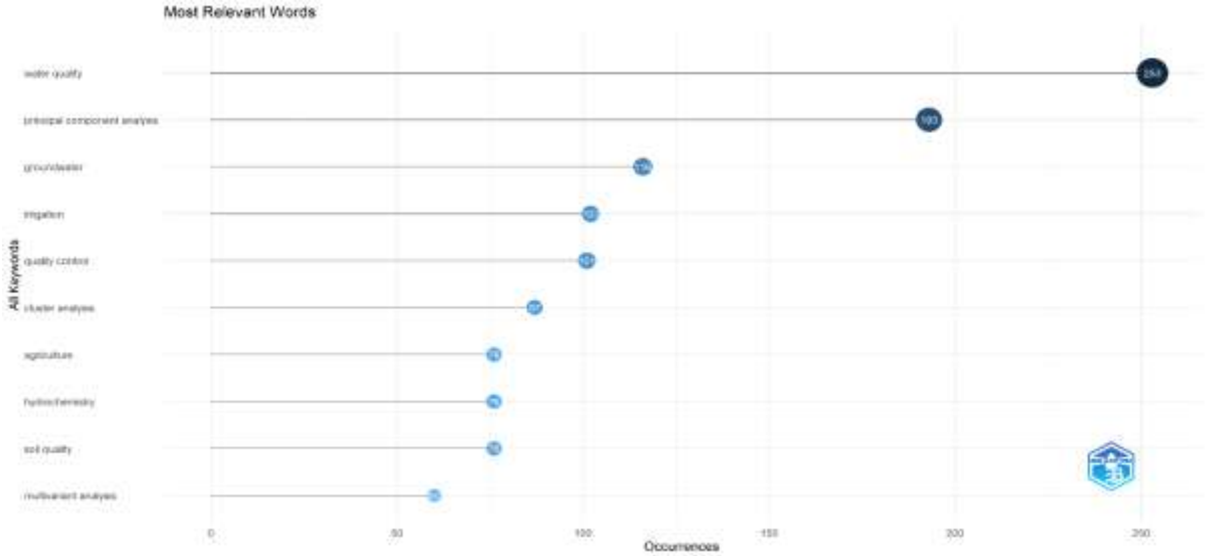
Figure 3. Shows the keyword co-occurrence network, revealing the intellectual structure of the field. Prominent keywords such as *water quality*, *groundwater*, *soil quality* and *principal component analysis* form central nodes, indicating their dominant role across studies. The network highlights distinct thematic clusters corresponding to



hadrochemical assessment, agricultural water management and multivariate analytical approaches, demonstrating both methodological concentration and thematic fragmentation within the literature.

2.5.4 Most relevant keywords

Figure 4. Displays the most frequently occurring keywords. *Water quality* emerges as the most dominant term, followed by *principal component analysis*, *groundwater*, *irrigation* and *soil quality*. This confirms the strong emphasis on statistical and multivariate techniques for assessing environmental quality in agricultural and water-resource studies.



3. Data Frameworks and Variable characteristics

3.1 Types of Data in Soil-Water-Agriculture Studies

The study of soil, water and agriculture involves multivariate data, which are collected using field measurements, laboratory analysis, secondary data and remote sensing techniques. The data are collected at different scales of space and time, but their integration is increasingly being used in irrigation planning and agricultural water management practices. From an analytical point of view, the use of multivariate data with several interrelated variables is more appropriate for multivariate statistical analysis, owing to their natural complexity and dimensionality (Shrestha and Kazama, 2007; Singh *et al.,* 2004). The major data sources used in such studies and their statistical relevance are summarized in Table 1.

Table 1. Data sources used in soil-water-agriculture studies.

Data type	Examples	Statistical relevance
Primary field data	Soil moisture, pH, EC	High-resolution multivariate inputs
Laboratory data	Nutrient concentrations	Continuous predictors
Secondary datasets	District irrigation statistics	Panel and longitudinal analysis
Remote sensing data	NDVI, ET	Spatial and time-series covariates

3.1.1 Challenges in Data Quality and Pre-processing

While the soil-water-agriculture data sets provide considerable scope for analysis, these are often hindered by problems of data quality and consistency. The primary field measurements, including sensor-based observations, are often marred by problems of soil moisture monitoring system calibration, sensor deterioration, incorrect sensor installation and environmental interference, especially in long-term monitoring and irrigation scheduling applications(Anil *et al.,* 2022; Ebstu *et al.,* 2025; Rasheed *et al.,* 2022).

There are also uncertainties in the soil quality and irrigation water quality data obtained from laboratory measurements, owing to the detection limits, precision and variability of the sample protocols, as well as the background soil conditions. The differences in analytical protocols, risk assessment models and approaches between different laboratories also add to the heterogeneity of the data, making the integration of different data sources difficult.

While the remote sensing data sets provide considerable scope for analysis, certain challenges are associated with the use of these data, including the spatial resolution of the satellite data, atmospheric interference, cloud interference and the indirect estimation of soil and crop variables, often resulting in uncertainties and missing observations during the integration of satellite-based data with other data sources, especially in agro-environmental applications (Abidi *et al.,* 2024; Atzberger, 2013).If data quality issues are not addressed, they can impact the accuracy of multivariate statistical outputs, leading to unstable clusters, biased component loadings, or incorrect regression relationships. In this regard, data pre-processing is considered a key process in soil-water-agriculture data structures. General pre-processing operations involve detecting outliers, handling missing values, normalization and data transformation to enable comparison between heterogeneous scales of measurement. Sophisticated pre-processing operations have been proven to significantly enhance the accuracy of outcomes from multivariate statistical analysis on complex agricultural data (García *et al.,* 2015; Heru *et al.,* 2021; Nugroho *et al.,* 2021).

3.1.2 Data Integration Strategies for Heterogeneous Datasets

Recent research on the relationship between soil, water and agriculture has begun to place greater emphasis on data integration techniques involving heterogeneous datasets. These datasets include information on soils, irrigation water quality, climatic variables and remotely sensed data. The heterogeneous nature of such datasets complicates direct statistical integration due to differences in spatial resolution, temporal frequency and data formats. The spatial integration process involves integrating field-based point data with gridded remotely sensed data. The temporal integration process involves resampling or aggregating data sets with disparate temporal frequencies. If not properly handled, it can negatively impact the ability of multivariate models to explain phenomena (Hagage *et al.,* 2025).

At the data framework level, an effective data integration strategy aims to create an integrated dataset with sufficient coherence to be used in multivariate analysis. This approach has been shown to increase the ability of multivariate models to describe phenomena involving soil-water interactions. The approach also increases the relevance of multivariate models to irrigation planning and management (Lobell and Burke, 2010) In addition to the issues involved in integrating heterogeneous datasets, the nature of the data also presents certain issues to be considered before applying multivariate statistical techniques.

3.2 Variable Characteristics and Measurement Scales

Variables used in the soil-water datasets are mostly continuous in nature. However, they are measured on heterogeneous scales and units. These characteristics necessitate preprocessing techniques such as normalization and transformations before applying multivariate techniques. Moreover, high inter-correlations among variables are observed, which often lead to multicollinearity. This limits the effectiveness of univariate techniques; therefore, dimension-reducing techniques are used in multivariate techniques (Abdel-Fattah *et al.,* 2020; Chaudhary *et al.,* 2024; Gautam *et al.,* 2024). The measurement characteristics and typical preprocessing methods of commonly used variables are presented in Table 2.

Table 2. Measurement characteristics of soil and irrigation water variables.

Variable	Measurement scale	Unit	Typical preprocessing
Soil pH	Interval	–	None
Electrical conductivity	Ratio	dS m^{-1}	Log transformation
Organic carbon	Ratio	%	Square-root
SAR	Ratio	Dimensionless	None

3.2.1. Incorporating Spatio-temporal Dependencies in Multivariate Analysis

Soil and irrigation water datasets manifest significant spatio-temporal dependencies, which are a function of environmental, management and climatic factors. Spatial dependency is a function of observations at different locations having similar characteristics with regard to soil properties and water quality, while temporal dependency is a function of factors related to irrigation cycles, hydrological cycles, crop growth stages and climatic trends. Significant water quality studies have highlighted that spatio-temporal variability affects multivariate relationships considerably and must be incorporated while analyzing water quality data (Helena *et al.,* 2000; Shrestha and Kazama, 2007; Singh *et al.,* 2004).

Spatio-temporal dependency, if ignored, leads to biased results while performing multivariate analysis on water quality data. For example, cluster analysis will identify clusters that are a function of spatial proximity, while principal component analysis will identify major temporal trends, which are not necessarily a function of process relationships. This effect is well documented with regard to surface water and groundwater quality, where temporal and spatial changes manifest a strong effect on multivariate relationships (Helena *et al.,* 2000; Singh *et al.,* 2004).

Recent reviews on methodological advancements have, in fact, emphasized the need to address spatial structures in an explicit manner in the statistical modeling frameworks that are applied to water quality or agro-environmental problems. Spatial statistical methods have been found to help in effectively distinguishing between real environmental gradients and spatial artifacts, making it easier to interpret multivariate results in situations where there is a high level of spatial autocorrelation in the data (Mainali *et al.,* 2019). In addition, addressing spatial autocorrelation in water quality modeling has been found to improve the accuracy of predictions, especially in heterogeneous regions (Miralha and Kim, 2018).

Although there is no explicit use of spatio-temporal models in this research, it is essential to consider spatial and temporal dependence in preparing, stratifying, or interpreting data. In fact, spatio-temporal dependence is found to enhance the robustness of multivariate inference, making it more relevant to irrigation planning or agricultural water management, where decisions are, in fact, influenced by spatial variability or temporal dynamics (Shrestha and Kazama, 2007).

3.3 Spatial Structure of Soil-Water-Agriculture Data

It has been observed that soil data, as well as irrigation water data, exhibit considerable spatial variability, which can be attributed to the variation in processes that control soil formation, irrigation practices, environmental conditions, etc. It has also been noted that data of this nature tend to be collected at different scales, i.e., from a plot scale up to a block/district scale. It has also been observed that with the increase in scale, the complexity of the variation in soil/water properties also increases, which makes the statistical analysis more complex.

To address the issue of spatial heterogeneity, statistical methods, along with multivariate techniques, have also been used for the classification of regions, derivation of homogeneous zones and the establishment of patterns in soil/water data. It has also been observed that digital soil mapping, along with spatial statistical approaches, has shown considerable potential for the derivation of the spatial variability of the data at different scales, which can be used for regionalization purposes (Asma and Şener, 2024). The major spatial scales and their corresponding statistical applications are summarized in Table 3.

Table 3. Spatial scales of soil-water datasets and associated statistical applications.

Spatial scale	Description	Statistical application
Field level	Plot-wise observations	Regression, local PCA
Block level	Aggregated administrative units	Cluster analysis
District/region	Large-area datasets	Zoning, classification models

3.4 Temporal Structure and Data Challenges in Soil-Water Studies

Temporal patterns of variability in the soil and irrigation water data are driven by seasonal changes in the climate, irrigation patterns, crop growth patterns and long-term changes in the environment. In practice, the resulting data often manifests as non-uniformly spaced observations, missing data, non-normal distributions and strong correlations between variables, making the integration of multi-year data challenging and limiting the validity of inferences obtained from single-parameter analyses.

Multivariate statistical approaches represent an attractive means of managing temporal patterns of variability, summarizing dominant patterns of temporal behaviour, managing correlations and reducing the effect of noise variables and seasonality (Sayel *et al.,* 2026). Previous research has clearly demonstrated the effectiveness of multivariate statistical approaches in successfully differentiating between seasonality and long-term temporal behaviour in water quality data, thereby improving interpretability and decision-making in water quality environments (Vega *et al.,* 1998). The common temporal data challenges and corresponding statistical solutions are summarized in Table 4.

Table 4. Common temporal data challenges and corresponding statistical solutions.

Data challenge	Description	Statistical solution
Missing values	Irregular monitoring	Imputation methods
Non-normality	Skewed distributions	Data transformation
Multicollinearity	Correlated predictors	PCA, FA
Unequal time steps	Seasonal/annual mismatch	Standardization

4. Multivariate Statistical Techniques Used in Agricultural Irrigation Research

4.1 Principal Component Analysis and Factor Analysis

Agricultural irrigation datasets generally contain numerous interrelated variables describing soil characteristics, irrigation water quality parameters and climatic conditions (Dormann *et al.,* 2013; Shrestha and Kazama, 2007). The presence of many correlated variables often complicates data interpretation and statistical modeling. To address this challenge, multivariate statistical techniques such as Principal Component Analysis (PCA) and Factor Analysis (FA) are widely applied in irrigation-related studies.

Principal Component Analysis is a dimension-reduction technique that transforms a large set of correlated variables into a smaller number of uncorrelated principal components while retaining most of the variability in the dataset (Helena *et al.,* 2000; Singh *et al.,* 2004). These components help identify dominant environmental variables influencing irrigation water quality and soil conditions.

Factor Analysis extends this approach by extracting latent factors that explain the relationships among observed variables. These factors often represent underlying hydro-chemical processes such as salinity dynamics, nutrient loading, or anthropogenic influences within irrigation systems (Kannel *et al.,* 2007; Liu *et al.,* 2003). Consequently, PCA and FA provide a useful foundation for further multivariate analyses such as clustering, regression modeling and composite index development.

4.1.1 Comparative Application and Interpretation of PCA versus FA

Although PCA and FA are frequently applied together in soil-water studies, their objectives differ. PCA is primarily a variance-based technique used to summarize variability in datasets and address multicollinearity without assuming an underlying latent structure (Tejashvini *et al.,* 2024; Wiafe *et al.,* 2025). In contrast, FA assumes that correlations among variables arise from a smaller number of latent factors representing physical or chemical processes.

Therefore, PCA is often used for exploratory analysis and as a preprocessing step prior to clustering or regression modeling. FA, however, is more suitable when the goal is to interpret the underlying environmental processes affecting irrigation water quality (Abdi and Williams, 2010; Bro and Smilde, 2014).

4.1.2 Advanced Dimension Reduction for Non-linear and High-Dimensional Data

Traditional dimension-reduction methods such as PCA and FA are linear techniques and may not fully capture complex non-linear relationships that often exist in high-dimensional datasets derived from hyperspectral imagery or dense environmental sensor networks. As a result, several advanced techniques have been explored in recent studies.

Independent Component Analysis (ICA) has been used to separate mixed hydro-chemical signals and improve source identification in complex environmental systems. Additionally, non-linear dimension-reduction techniques such as t-distributed Stochastic Neighbour Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) enable visualization of complex high-dimensional datasets and reveal structures that may not be detectable using linear methods (Coifman and Lafon, 2006; McInnes *et al.,* 2018).

Although these methods are primarily exploratory, they provide valuable support for identifying patterns in soil-water datasets. However, these approaches also have limitations; for example, ICA relies on assumptions of statistical independence among sources, whereas t-SNE and UMAP are mainly visualization tools rather than predictive models.

4.2 Cluster Analysis: Principles and Methodological Selection

Cluster Analysis (CA) is an important multivariate statistical technique used in soil-water-agriculture studies to classify samples or spatial units based on similarities in multiple variables. Unlike PCA or FA, which focus on dimensionality reduction, CA groups observations into homogeneous clusters. This classification helps identify hydro-chemical patterns and supports region-specific irrigation planning for surface and groundwater resources (Cloutier *et al.,* 2008; Simeonov *et al.,* 2003; Ali et al., 2022).

CA is also useful for evaluating spatial variability in water chemistry and distinguishing areas influenced by natural geochemical processes from those affected by anthropogenic activities such as agriculture or urbanization (Yidana *et al.,* 2008). In many studies, PCA or FA component scores are used as input variables for CA to enhance classification accuracy and delineate irrigation suitability zones with similar soil-water characteristics.

The effectiveness of clustering largely depends on the selected clustering framework. Two main clustering approaches are commonly used as summarized in Table 5.

Table 5. Comparison of major clustering techniques used in soil and agricultural data analysis.

Technique	**Best Use Case**	**Key Characteristics**
Hierarchical Clustering	Exploratory zoning where cluster number is unknown	Uses linkage methods such as Ward's method and produces dendrograms illustrating relationships among clusters (Murtagh and Contreras, 2012).
Non-Hierarchical Clustering (K-means)	Large datasets with predefined cluster numbers	Computationally efficient but sensitive to initial cluster centres and therefore requires validation (Xu and Wunsch, 2005).

4.3 Advanced Regression-Based and Multivariate Modeling

Regression-based models are widely used in agricultural statistics to quantify relationships between dependent variables, such as crop yield or water-use efficiency and soil or irrigation water quality parameters (Anil *et al.,* 2022). However, datasets containing highly correlated predictors often lead to multicollinearity, which reduces the reliability and stability of traditional regression models.

To overcome this limitation, multivariate approaches such as Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR) are commonly applied. These techniques transform correlated predictors into orthogonal components, thereby improving model robustness and predictive performance (Abdi, 2010; Abdi and Williams, 2010; Wold *et al.,* 2001).

Model validation is essential for ensuring reliable predictions and preventing overfitting. Common evaluation metrics include the coefficient of determination (R^2), root mean square error (RMSE) and cross-validated predictive ability (Q^2) (Wold *et al.,* 2001).

Recent research increasingly integrates these multivariate techniques with machine learning (ML) algorithms. In such hybrid frameworks, PCR or PLSR performs feature extraction and the resulting components are then used as inputs for ML models to enhance predictive performance (Liakos *et al.,* 2018; Wei *et al.,* 2022). These integrated approaches are particularly useful in irrigation science for applications such as soil moisture monitoring, crop water stress assessment and irrigation decision-support systems (Hussein *et al.,* 2024; Liakos *et al.,* 2018).

4.4 Composite Indices and Multivariate Decision Metrics

Composite indices provide an effective approach for integrating multiple soil and irrigation water quality parameters into a single indicator for irrigation planning. These indices simplify complex multivariate datasets into interpretable scores that support decision-making.

Statistical techniques play an important role in index development, particularly in variable selection, normalization and weighting. Methods such as PCA and FA are commonly used to derive objective weights for variables included in composite indices, thereby reducing subjectivity in index construction (Sutadian *et al.,* 2016). Examples include irrigation water quality indices and soil–water suitability indices used for irrigation management. Despite their usefulness, composite indices require proper validation to ensure that they accurately represent underlying soil-water conditions relevant to irrigation systems.

4.4.1 Methodological Considerations in Constructing and Validating Composite Indices

The development of composite indices typically involves four major steps: variable selection, normalization, weighting and aggregation. Statistical techniques are often applied to determine objective weights for variables, which improves the reliability of the index (Becker *et al.,* 2017).

However, composite indices can be sensitive to methodological choices such as weighting methods and normalization procedures. Therefore, sensitivity analysis and validation are necessary to ensure the robustness of the indices and their usefulness for irrigation planning.

4.5 Integration of Multivariate Methods in Irrigation Decision Support

The integration of multivariate statistical techniques provides a comprehensive analytical framework for understanding complex soil-water systems and supporting irrigation management decisions. Dimension reduction methods help identify dominant variables, clustering techniques enable irrigation zone classification, regression models support predictive analysis and composite indices facilitate decision-making.

Recent studies demonstrate that combining these techniques within decision-support frameworks can effectively translate complex environmental data into actionable irrigation strategies, including risk assessment, management zone prioritization and adaptive water resource management (De Girolamo *et al.,* 2017; Mukherjee *et al.,* 2022; Radmehr *et al.,* 2022; Ronco *et al.,* 2017).

The overall workflow of multivariate statistical techniques applied in irrigation data analysis is illustrated in Figure 5. The framework highlights how dimension reduction techniques such as PCA and FA are first used to identify key variables within soil–water datasets. These results are then used in Cluster Analysis for spatial classification and in regression models such as PCR and PLSR for predictive modeling. Finally, the outputs are integrated into composite indices and decision-support tools that support irrigation planning and agricultural management. A summary of the major multivariate statistical techniques used in soil-water -irrigation studies, along with their primary objectives, applications and outputs, is presented in Table 6.

Table 6. Multivariate statistical techniques used in soil-water-irrigation studies.

Statistical method	Main purpose	Typical application in irrigation studies	Key outputs	Representative references
Principal Component Analysis (PCA) / Factor Analysis (FA)	Dimension reduction and identification of dominant variables	Identification of major soil and irrigation water quality factors; handling multicollinearity among correlated parameters	Principal components, factor loadings, explained variance	(Liu *et al.,* 2003; Singh *et al.,* 2004)
Cluster Analysis (CA)	Classification and grouping	Delineation of irrigation zones; grouping soil or water samples based on similarity	Clusters, dendrograms, group memberships	Cloutier *et al.,* 2008; Simeonov *et al.,* 2003; Yidana *et al.,* 2008
Principal Component Regression (PCR) / Partial Least Squares Regression (PLSR)	Prediction and relationship modeling	Modeling relationships between soil-water variables and irrigation performance indicators such as yield or water-use efficiency	Regression coefficients, predicted values	Hastie *et al.,* 2009; Wold *et al.,* 2001

Composite indices (e.g., irrigation water quality indices)	Decision support and simplification	Integration of multiple soil and water parameters into a single suitability or quality score for irrigation planning	Index values, suitability classes	de Andrade Costa *et al.,* 2020; Sutadian *et al.,* 2016

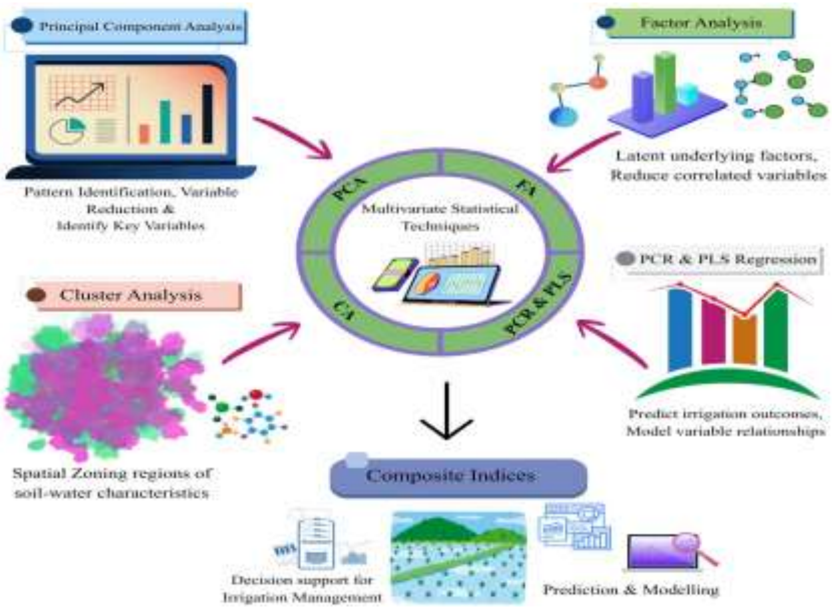


Figure 5. Multivariate statistical framework for irrigation data analysis and management.

4.6 Comparative Evaluation of Multivariate Techniques

Different multivariate statistical techniques provide distinct analytical advantages when applied to soil and irrigation water datasets. The selection of an appropriate method depends on the objectives of the study, the structure of the dataset and the level of interpretability required. Some techniques are mainly used for dimensionality reduction, while others are applied for classification, pattern recognition, or visualization of complex relationships. A comparative overview of commonly used multivariate techniques and their roles in irrigation-related studies is presented in Table 7.

Table 7. Comparison of multivariate techniques in irrigation studies.

Method	**What it does well**	**Limitation**	**Best use in irrigation studies**
PCA	Reduces dimensionality	Low physical interpretability	Identifying dominant hydro-chemical variables
FA	Reveals latent processes	Requires assumptions	Salinity and nutrient source identification
CA	Groups similar samples	Sensitive to distance metric	Irrigation zone classification
ICA	Captures independent signals	Difficult to interpret	Emerging patterns in mixed water sources
UMAP	Non-linear visualization	Not yet standardized	Exploratory irrigation data patterns

PCA and FA are still among the most popular approaches used in irrigation-related research, although the suitability of these approaches depends on the specific objective of the study. Principal Component Analysis can be used to reduce dimensionality, while Factor Analysis can be used to gain more in-depth knowledge of the hydrogeochemical processes. In addition, Cluster Analysis can be used to classify the irrigation zones, although it is sensitive to data scaling. New approaches, such as Independent Component Analysis and Uniform Manifold Approximation and Projection, have shown promise in capturing non-linear relationships, although they have not been used in irrigation-related research to any great extent, primarily because of difficulties associated with interpretability and replicability.

5. Research Gaps and Limitations in Statistical Soil-Water-Irrigation Research

Despite advancements in multivariate statistics, machine learning and remote sensing, some important aspects remain salient in statistical studies on soil, water and irrigation. Although data-based approaches have improved our understanding of irrigation systems, their integration with effective irrigation management models still faces limitations, especially under data scarcity, uncertainty and changing climatic patterns (Loucks and Van Beek, 2017).

5.1 Data Availability, Quality and Integration Challenges

The major limitation in statistical irrigation management models is related to the quality and availability of irrigation-related data sets, including estimates of soil moisture, irrigation and water use, which are often derived from heterogeneous sources with variable resolutions, making statistical inference and prediction challenging (Horsburgh *et al.,* 2016; Little and Rubin, 2019). Uncertainty levels experienced during data integration and assimilation also play a significant role in irrigation management models, resulting in considerable variability in irrigation impacts and making statistical models less transferable between irrigation systems.

According to a recent global assessment reported in 2024, although the area under irrigation has increased considerably during the early part of the twenty-first century, more than half of this increase has occurred in areas that were already under some level of water stress. This again reflects the well-recognized problems in the quality of irrigation data, monitoring systems and sustainability assessment methodologies (Adawe *et al.,* 2024; Mehta *et al.,* 2024).

Another issue related to data availability is the emerging issue of governance. The absence of standardized metadata, data structures and workflows makes it difficult to reproduce and transfer statistical and machine learning models from one region to another and this is becoming an emerging issue in large-scale irrigation analytics.

5.2 Methodological and Validation Gaps

Exploratory multivariate analysis, with principal component analysis at its forefront, still finds widespread application; however, interpretability still lacks sufficient robustness and sensitivity assessment, despite its reliance on data preprocessing and scaling methodologies (Jolliffe and Cadima, 2016). Machine learning algorithms, including ensemble methods, have been shown to offer excellent predictive capabilities, but their lack of interpretability means that these models fail to offer sufficient insights for irrigation decision-making, which requires a transparent approach (Breiman, 2001; Kamilaris *et al.,* 2017).

Validation methodologies still pose a major problem, with conventional methodologies failing to account for the inherent dependency structures that exist within irrigation datasets, thereby providing overly optimistic estimates with limited generalizability potential(Roberts *et al.,* 2017). This is particularly evident with machine learning models, which, when applied to irrigation time series, suffer from a non-stationarity that arises due to climate change, policy interventions and management adaptations, which eventually impact model performance over time if not adequately considered. Therefore, conclusions derived from methodologies appear highly specific, with limited generalizability potential across irrigation systems with heterogeneous soil, climate and management regimes.

5.3 Limited Translation into Decision Support

An important limitation has also been the lack of translation of statistical outputs into irrigation management strategies. A large number of research studies have presented their results using statistical analysis, but few have attempted to incorporate their research within a framework of a decision support system, mainly because statistical outputs rarely translate into management-relevant indicators that align with the priorities of stakeholders. Although the potential for improving the relevance of research using participatory approaches to modeling has been presented, their integration into irrigation-focused statistical research has remained low (Loucks and Van Beek, 2017; Voinov *et al.,* 2016).

5.4 Synthesis of Key Gaps

Based on the discussion above, the major gaps in irrigation-focused statistical research on soil, water and irrigation can be synthesized as data quality, emerging issues related to governance, reusability, non-stationary/non-independent data and the lack of a linkage with irrigation management decisions.

6. Future Research Directions in Statistical Soil-Water-Irrigation Studies

Soil-water-irrigation research is increasingly shaped by climate variability, agricultural intensification and advances in sensing and data analytics. While earlier studies relied mainly on exploratory statistical approaches, future research will require integrated and predictive statistical frameworks capable of supporting sustainable irrigation management (El-Rawy *et al.,* 2024).

6.1 Integration of Multisource and High-Resolution Data

Future irrigation studies will increasingly rely on integrating multisource datasets, including soil physical and chemical properties, irrigation water quality parameters, meteorological variables and remotely sensed indicators such as vegetation indices, evapotranspiration and soil moisture. Advances in remote sensing and digital soil mapping have made it possible to obtain spatially explicit environmental information at finer scales, improving the monitoring of soil–water systems (Atzberger, 2013; Grunwald *et al.,* 2011; Hengl *et al.,* 2017).

However, integrating heterogeneous datasets presents statistical challenges due to differences in spatial resolution, temporal frequency and uncertainty. Future research should therefore focus on hierarchical modeling and scale-matching approaches to improve statistical reliability and reduce uncertainty in irrigation management decisions (Beven and Binley, 2014).

6.1.1 Integration of Remote Sensing, IoT and Ground Observations

Remote sensing data, IoT sensor networks and ground observations provide complementary information about soil–water–plant interactions. Remote sensing offers broad spatial coverage, while IoT sensors can continuously monitor soil moisture, salinity and irrigation performance at high temporal resolution (Kamienski *et al.,* 2019; O'Shaughnessy *et al.,* 2013).

Combining these datasets using multivariate statistical frameworks can improve the estimation of soil moisture, salinity dynamics and crop water stress. Such integrated monitoring systems are particularly valuable for precision irrigation management (Jin *et al.,* 2018; Reichle *et al.,* 2017; Zeng *et al.,* 2017).

6.2 Advancing Predictive Multivariate Analysis

Exploratory techniques such as Principal Component Analysis (PCA) and Cluster Analysis (CA) remain useful for identifying patterns in irrigation datasets. However, future research should increasingly emphasize predictive multivariate modeling, especially for high-dimensional datasets characterized by multicollinearity. Methods such as Partial Least Squares Regression (PLSR) provide effective solutions for modeling relationships between soil–water variables and irrigation performance indicators (Abdi, 2010; Wold *et al.,* 2001).

Robust validation strategies are also essential for reliable modeling. Conventional random cross-validation may lead to biased performance estimates when spatial or temporal autocorrelation is present. Therefore, validation methods that explicitly account for spatial and temporal dependence should be adopted in future irrigation studies (Dormann *et al.,* 2013; Meyer *et al.,* 2018).

6.2.1 Dynamic Statistical Modeling for Adaptive Irrigation

The increasing availability of real-time environmental data from soil sensors, weather stations and irrigation systems enables the development of dynamic statistical models for irrigation management. Approaches such as state-space models and time-varying parameter models allow continuous updating of predictions as new observations become available. These methods can support adaptive irrigation scheduling by responding to short-term changes in rainfall, soil moisture and crop water demand, thereby improving irrigation efficiency and water resource management (Abioye *et al.,* 2020; Hatfield and Dold, 2019).

6.3 Hybrid Multivariate Statistics and Interpretable Machine Learning

Machine learning techniques are increasingly applied in irrigation research because they can capture complex non-linear relationships in environmental datasets. However, many machine learning models lack interpretability, which can limit their practical use in agricultural decision-making.

Future research should therefore focus on hybrid frameworks that combine multivariate statistical methods with interpretable machine learning models. In such approaches, multivariate statistics can be used for dimensionality reduction and variable selection, while machine learning algorithms improve predictive accuracy. Interpretable models allow researchers and practitioners to understand the relationships between environmental variables and irrigation outcomes (Jordan and Mitchell, 2015; Molnar, 2020; Olden *et al.,* 2004).

6.3.1 Balancing Interpretability and Predictive Performance

A major challenge in irrigation modeling is achieving a balance between model interpretability and predictive performance. Hybrid modeling approaches that integrate statistical preprocessing with explainable machine learning techniques provide a promising direction for addressing this challenge and improving decision-making in irrigation management.

6.4 Strengthening Links Between Statistical Analysis and Irrigation Decision Support

For statistical research to have practical impact, analytical outputs must be translated into decision-support tools. Future studies should focus on converting statistical results into actionable irrigation management strategies, such as irrigation zoning derived from clustering results, scheduling thresholds based on multivariate risk analysis and water allocation strategies using probabilistic or optimization-based models (Loucks and Van Beek, 2017; Voinov *et al.,* 2016).

6.5 Transferable and Region-Specific Statistical Frameworks

Another challenge in irrigation research is developing statistical frameworks that are both transferable across regions and adaptable to local environmental conditions. Standardized approaches for variable definitions, preprocessing methods and performance metrics can improve the comparability of irrigation studies. Cross-regional comparisons can help identify universal indicators and region-specific drivers of irrigation performance, particularly under changing climatic conditions (Montanari and Koutsoyiannis, 2014).

6.6 Key Priorities for Future Statistical Research

Future advances in soil-water-irrigation research will depend on statistically robust and decision-oriented analytical frameworks. Priority areas include multisource data integration, predictive and dynamic multivariate modeling and the integration of statistical methods with interpretable machine learning approaches. Strengthening the connection between statistical analysis and irrigation decision support will be essential for addressing water sustainability challenges under increasing environmental uncertainty. The major research priorities and their expected contributions are summarized in Table 8.

Table 8. Future research priorities in statistical soil-water-irrigation studies.

Priority area	Statistical focus	Expected contribution	Representative references
Multisource data integration	Hierarchical modeling, uncertainty-aware data fusion	Robust integration of soil, climate and remote sensing data	Atzberger, 2013; Beven and Binley, 2014; Grunwald *et al.,* 2011; Hengl *et al.,* 2017
Remote sensing & IoT integration	Multivariate and spatio-temporal statistics	Improved soil moisture and crop water status estimation	Jin *et al.,* 2018; Kamienski *et al.,* 2019; Zeng *et al.,* 2017
Predictive multivariate modeling	PLS regression, regularized methods	Development of irrigation efficiency and risk indicators	Abdi, 2010; Dormann *et al.,* 2013; Wold *et al.,* 2001
Dynamic modeling	State-space and time-varying models	Adaptive irrigation scheduling	Abioye *et al.,* 2020; Hatfield and Dold, 2019
Hybrid statistical-ML approaches	Interpretable machine learning	Transparent modeling of nonlinear soil-water interactions	Jordan and Mitchell, 2015; Molnar, 2020; Olden *et al.,* 2004; Van Klompenburg *et al.,* 2020
Decision-support integration	Probabilistic and optimization frameworks	Translation of analysis into irrigation decisions	Loucks and Van Beek, 2017; Voinov *et al.,* 2016
Transferability	Standardized statistical protocols	Cross-regional applicability of irrigation insights	Montanari and Koutsoyiannis, 2014)

6.7 Conceptual Framework for Statistical Soil-Water-Irrigation Assessment

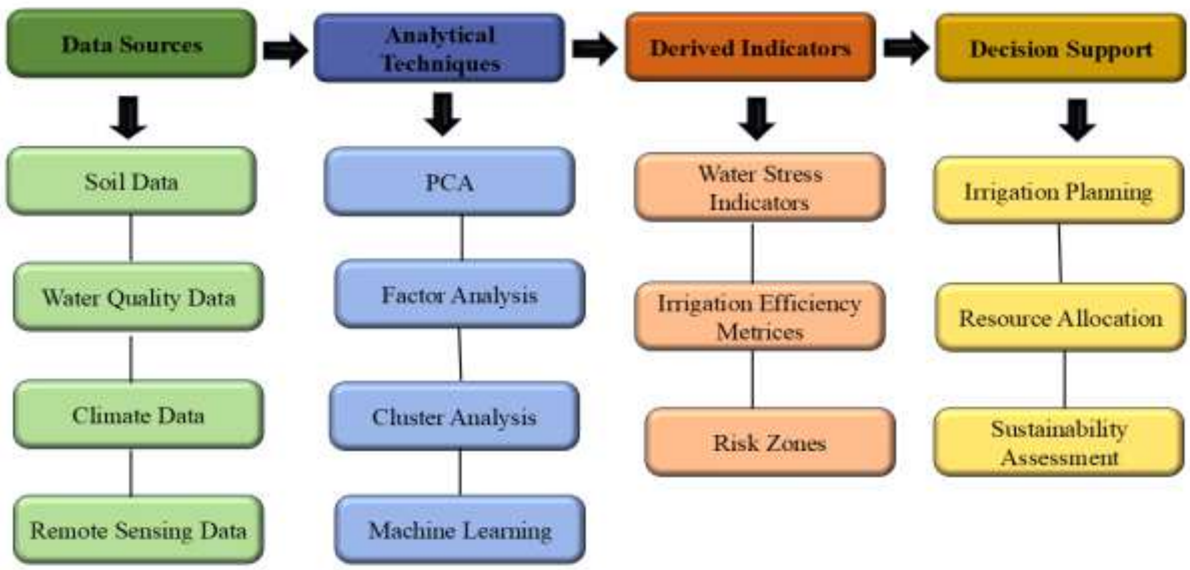


Figure 7. Conceptual framework linking data sources, multivariate techniques and decision-oriented outputs in statistical soil-water-irrigation studies.

CONCLUSION

This article aims to provide an in-depth review of the application of multivariate statistical techniques in the analysis of the interactions of soils and water in irrigation systems, including dimension reduction, classification, regression and development of composite indices. These techniques have the potential to transform complex and multidimensional data into useful information, which can be effectively used in the efficient management of irrigation systems.

Despite the remarkable progress made in the development of the methodology, there are still limitations in the application of the techniques, especially in the areas of data quality, model validation, regional applicability and the effective use of statistical results in decision-making processes. Therefore, the development of effective and efficient analytical frameworks in addressing the challenges in the application of the techniques in irrigation management is important. Multivariate statistical analysis plays a crucial role in improving the efficiency, flexibility and sustainability of the irrigation system in response to the dynamics of climatic conditions and the concepts and potential areas of research in the application of the techniques in the management of the system have been discussed in the article.

Data availability
No data was used for the research described in the article.
Declaration of competing interest
The authors declare that they have no conflict of interest
Funding
Not applicable

REFERENCES
1. Abdel-Fattah, M. K., Abd-Elmabod, S. K., Aldosari, A. A., Elrys, A. S., & Mohamed, E. S. (2020). Multivariate analysis for assessing irrigation water quality: A case study of the Bahr Mouise Canal, Eastern Nile Delta. *Water*, *12*(9), 2537. https://doi.org/10.3390/w12092537

2. Abdi, H. (2010). Partial least squares regression and projection on latent structure regression (PLS Regression). Wiley interdisciplinary reviews: computational statistics, 2(1),97-106. https://doi.org/10.1002/wics.51

3. 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
4. Abidi, J. H., Elzain, H. E., Sabarathinam, C., Selmane, T., Selvam, S., Farhat, B., Mammou, A. B., & Senapathi, V. (2024). Evaluation of groundwater quality indices using multi-criteria decision-making techniques and a fuzzy logic model in an irrigated area. Groundwater for Sustainable Development, 25, 101122. https://doi.org/10.1016/j.gsd.2024.101122
5. Abioye, E. A., Abidin, M. S. Z., Mahmud, M. S. A., Buyamin, S., Ishak, M. H. I., Abd Rahman, M. K. I., Otuoze, A. O., Onotu, P., & Ramli, M. S. A. (2020). A review on monitoring and advanced control strategies for precision irrigation. Computers and Electronics in Agriculture, 173, 105441. https://doi.org/10.1016/j.compag.2020.105441
6. Adawe, A. M., Abdi, S. J., Abdi, A. M., & Omar, A. D. (2024). Water Quality Assessment of Groundwater Using Multivariate Statistical Techniques: A Case Study of Mogadishu, Banadir Region, Somalia. American Journal of Environmental Protection, 13(1), 19-29. https://doi.org/10.11648/j.ajep.20241301.13
7. Ali, A., Rondelli, V., Martelli, R., Falsone, G., Lupia, F., & Barbanti, L. (2022). *Management zones delineation through clustering techniques based on soil traits, NDVI data and multiple year crop yields.* Agriculture**,** 12(2), 231. https://doi.org/10.3390/agriculture12020231
8. Anil, C., Thiyagarajan, G., Selvakumar, S., Velmurugan, M., & Kannan, B. (2022). Effect of Partial Root Drying and Deficit Irrigation on Yield and Water Use Efficiency of Tomato Grown Under Greenhouse. International Journal of Environment and Climate Change, 12(11), 306-313. https://doi.org/10.9734/ijecc/2022/v12i1130975
9. Asma, B., & Şener, Ş. (2024). Appraisal of groundwater suitability and hydro-chemical characteristics by using various water quality indices and statistical analyses in the Wadi Righ area, Algeria. Water Supply, 24(5), 1938-1957. https://doi.org/10.2166/ws.2024.103
10. Atzberger, C. (2013). Advances in remote sensing of agriculture: Context description, existing operational monitoring systems and major information needs. Remote Sensing, 5(2), 949-981. https://doi.org/10.3390/rs5020949
11. Becker, W., Saisana, M., Paruolo, P., & Vandecasteele, I. (2017). Weights and importance in composite indicators: Closing the gap. Ecological indicators, 80, 12-22. https://doi.org/10.1016/j.ecolind.2017.03.056
12. Beven, K., & Binley, A. (2014). GLUE: 20 years on. Hydrological processes, 28(24), 5897-5918. https://doi.org/10.1002/hyp.10082
13. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324
14. Bro, R., & Smilde, A. K. (2014). Principal component analysis. Analytical methods, 6(9), 2812-2831. https://doi.org/10.1039/C3AY41907J
15. Chaudhary, S., Singh, G., Gupta, D., Singh Maunas, S., & Mishra, V. K. (2024). Characterization of groundwater potability and irrigation potential in Uttar Pradesh, India using water quality index and multivariate statistics. Journal of Hydroinformatics, 26(5), 1100-1121. https://doi.org/10.2166/hydro.2024.291
16. Cloutier, V., Lefebvre, R., Therrien, R., & Savard, M. M. (2008). Multivariate statistical analysis of geochemical data as indicative of the hydrogeochemical evolution of groundwater in a sedimentary rock aquifer system. Journal of Hydrology, 353(3-4), 294-313. https://doi.org/10.1016/j.jhydrol.2008.02.015
17. Coifman, R. R., & Lafon, S. (2006). Diffusion maps. Applied and computational harmonic analysis, 21(1), 5-30. https://doi.org/10.1016/j.acha.2006.05.011

18. de Andrade Costa, D., Soares de Azevedo, J. P., Dos Santos, M. A., & dos Santos Facchetti Vinhaes Assumpção, R. (2020). Water quality assessment based on multivariate statistics and water quality index of a strategic river in the Brazilian Atlantic Forest. Scientific Reports, 10(1), 22038. https://doi.org/10.1038/s41598-020-78563-0
19. De Girolamo, A. M., Barca, E., Pappagallo, G., & Porto, A. L. (2017). Simulating ecologically relevant hydrological indicators in a temporary river system. Agricultural Water Management, 180, 194-204. https://doi.org/10.1016/j.agwat.2016.05.034
20. Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., & Leitão, P. J. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography, 36(1), 27-46. https://doi.org/10.1111/j.1600-0587.2012.07348.x
21. Ebstu, E. T., Hatiye, S. D., Goshime, D. W., Dingemanse, J. D., Dugassa, D. D., Fitensa, T., Enssa, G., Chare, C. D., Areru, D. A., & Demeke, Y. G. (2025). Development and Testing of a Low-Cost Soil Moisture Sensor for Real-Time Irrigation Scheduling. Irrigation and Drainage. https://doi.org/10.1002/ird.70026
22. El-Rawy, M., Wahba, M., Fathi, H., Alshehri, F., Abdalla, F., & El Attar, R. M. (2024). Assessment of groundwater quality in arid regions utilizing principal component analysis, GIS and machine learning techniques. Marine Pollution Bulletin, 205, 116645. https://doi.org/10.1016/j.marpolbul.2024.116645
23. García, S., Luengo, J., & Herrera, F. (2015). Data preprocessing in data mining (Vol. 72). Springer. https://doi.org/10.1007/978-3-319-10247-4
24. Gautam, V. K., Kothari, M., Al-Ramadan, B., Singh, P. K., Upadhyay, H., Pande, C. B., Alshehri, F., & Yaseen, Z. M. (2024). Groundwater quality characterization using an integrated water quality index and multivariate statistical techniques. PloS one, 19(2), e0294533. https://doi.org/10.1371/journal.pone.0294533
25. Ghoto, S. M., Abbasi, H., Memon, S. A., Brohi, K. M., Chhachhar, R., Ghanghlo, A. A., & Tunio, I. A. (2025). Multivariate hydro-chemical assessment of groundwater quality for irrigation purposes and identifying soil interaction effects and dynamics of NDVI and urban development in alluvial region. Environmental Advances, 19, 100621. https://doi.org/10.1016/j.envadv.2025.100621
26. Grunwald, S., Thompson, J., & Boettinger, J. (2011). Digital soil mapping and modeling at continental scales: finding solutions for global issues. Soil Science Society of America Journal, 75(4), 1201-1213. https://doi.org/10.2136/sssaj2011.0025
27. Hagage, M., Hewaidy, A. G. A., & Abdulaziz, A. M. (2025). Groundwater quality assessment for drinking, irrigation, aquaculture and industrial uses in the waterlogged northeastern Nile Delta, Egypt: a multivariate statistical approach and water quality indices. Modeling Earth Systems and Environment, 11(1), 59. https://doi.org/10.1007/s40808-024-02242-6
28. Hallouz, F., Meddi, M., & Ali Rahmani, S. (2022). Multivariate analysis to assess the quality of irrigation water in a semi-arid region of north west of Algeria: case of Ghrib reservoir. Environmental Earth Sciences, 81(5), 158. https://doi.org/10.1007/s12665-022-10272-5
29. Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference and prediction (Vol. 2). Springer. https://doi.org/10.1007/978-0-387-84858-7
30. Hatfield, J. L., & Dold, C. (2019). Water-use efficiency: advances and challenges in a changing climate. Frontiers in plant science, 10, 103. https://doi.org/10.3389/fpls.2019.00103
31. Helena, B., Pardo, R., Vega, M., Barrado, E., Fernandez, J. M., & Fernandez, L. (2000). Temporal evolution of groundwater composition in an alluvial aquifer (Pisuerga River, Spain) by principal component analysis. Water research, 34(3), 807-816. https://doi.org/10.1016/S0043-1354(99)00225-0
32. Helfenstein, A., Mulder, V. L., Hack-ten Broeke, M. J., van Doorn, M., Teuling, K., Walvoort, D. J., & Heuvelink, G. (2024). BIS-4D: mapping soil properties and their uncertainties at 25 m resolution in the Netherlands. Earth System Science Data, 16(6), 2941-2970. https://doi.org/10.5194/essd-16-2941-2024
33. Hengl, T., Mendes de Jesus, J., Heuvelink, G. B., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., & Bauer-Marschallinger, B. (2017). SoilGrids250m: Global gridded soil information based on machine learning. PloS one, 12(2), e0169748. https://doi.org/10.1371/journal.pone.0169748
34. Heru, N., Priya, U. N., & Kridanto, S. (2021). Normalization and outlier removal in class center-based firefly algorithm for missing value imputation. Journal of Big Data, 8(1). https://doi.org/10.1186/s40537-021-00518-7
35. Horsburgh, J. S., Aufdenkampe, A. K., Mayorga, E., Lehnert, K. A., Hsu, L., Song, L., Jones, A. S., Damiano, S. G., Tarboton, D. G., & Valentine, D. (2016). Observations Data Model 2: A community information model for spatially discrete Earth observations. Environmental modeling & software, 79, 55-74. https://doi.org/10.1016/j.envsoft.2016.01.010
36. Höskuldsson, A. (1988). PLS regression methods. Journal of chemometrics, 2(3), 211-228. https://doi.org/10.1002/cem.1180020306
37. Hussein, E. E., Derdour, A., Zerouali, B., Almaliki, A., Wong, Y. J., Ballesta-de los Santos, M., Minh Ngoc, P., Hashim, M. A., & Elbeltagi, A. (2024). Groundwater quality assessment and irrigation water quality index prediction using machine learning algorithms. Water, 16(2), 264. https://doi.org/10.3390/w16020264

38. Jin, X., Kumar, L., Li, Z., Feng, H., Xu, X., Yang, G., & Wang, J. (2018). A review of data assimilation of remote sensing and crop models. European journal of agronomy, 92, 141-152. https://doi.org/10.1016/j.eja.2017.11.002
39. 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), 20150202. https://doi.org/10.1098/rsta.2015.0202
40. Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives and prospects. Science, 349(6245), 255-260. https://doi.org/10.1126/science.aaa8415
41. Kamienski, C., Soininen, J.-P., Taumberger, M., Dantas, R., Toscano, A., Salmon Cinotti, T., Filev Maia, R., & Torre Neto, A. (2019). Smart water management platform: IoT-based precision irrigation for agriculture. Sensors, 19(2), 276. https://doi.org/10.3390/s19020276
42. Kamilaris, A., Kartakoullis, A., & Prenafeta-Boldú, F. X. (2017). A review on the practice of big data analysis in agriculture. Computers and Electronics in Agriculture, 143, 23-37. https://doi.org/10.1016/j.compag.2017.09.037
43. Kannel, P. R., Lee, S., Lee, Y.-S., Kanel, S. R., & Khan, S. P. (2007). Application of water quality indices and dissolved oxygen as indicators for river water classification and urban impact assessment. Environmental monitoring and assessment, 132(1), 93-110. https://doi.org/10.1007/s10661-006-9505-1
44. Li, P., & Ren, L. (2021). Evaluating the saline water irrigation schemes using a distributed agro-hydrological model. Journal of Hydrology, 594, 125688. https://doi.org/10.1016/j.jhydrol.2020.125688
45. Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. Sensors, 18(8), 2674. https://doi.org/10.3390/s18082674
46. Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data*. John Wiley & Sons. https://doi.org/10.1002/9781119482260
47. Liu, C.-W., Lin, K.-H., & Kuo, Y.-M. (2003). Application of factor analysis in the assessment of groundwater quality in a blackfoot disease area in Taiwan. Science of the Total Environment, 313(1-3), 77-89. https://doi.org/10.1016/S0048-9697(02)00683-6
48. Lobell, D. B., & Burke, M. B. (2010). On the use of statistical models to predict crop yield responses to climate change. Agricultural and Forest Meteorology, 150(11), 1443-1452. https://doi.org/10.1016/j.agrformet.2010.07.008
49. Loucks, D. P., & Van Beek, E. (2017). Water resource systems planning and management: An introduction to methods, models and applications. Springer. https://doi.org/10.1007/978-3-319-44234-1
50. Mainali, J., Chang, H., & Chun, Y. (2019). A review of spatial statistical approaches to modeling water quality. Progress in Physical Geography: Earth and Environment, 43(6), 801-826. https://doi.org/10.1177/0309133319852003
51. McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426. https://doi.org/10.21105/joss.00861
52. Mehta, P., Siebert, S., Kummu, M., Deng, Q., Ali, T., Marston, L., Xie, W., & Davis, K. F. (2024). Half of twenty-first century global irrigation expansion has been in water-stressed regions. Nature Water, 2(3), 254-261. https://doi.org/10.1038/s44221-024-00206-9
53. Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., & Nauss, T. (2018). Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. Environmental modeling & software, 101, 1-9. https://doi.org/10.1016/j.envsoft.2017.12.001
54. Minhas, P., Ramos, T. B., Ben-Gal, A., & Pereira, L. S. (2020). Coping with salinity in irrigated agriculture: Crop evapotranspiration and water management issues. Agricultural Water Management, 227, 105832. https://doi.org/10.1016/j.agwat.2019.105832
55. Miralha, L., & Kim, D. (2018). Accounting for and predicting the influence of spatial autocorrelation in water quality modeling. ISPRS international journal of geo-information, 7(2), 64. https://doi.org/10.3390/ijgi7020064
56. Molnar, C. (2020). Interpretable machine learning. Lulu. com. https://christophm.github.io/interpretable-ml-book/
57. Montanari, A., & Koutsoyiannis, D. (2014). Modeling and mitigating natural hazards: Stationarity is immortal! Water resources research, 50(12), 9748-9756. https://doi.org/10.1002/2014WR016092
58. Mukherjee, I., Singh, U. K., & Chakma, S. (2022). Evaluation of groundwater quality for irrigation water supply using multi-criteria decision-making techniques and GIS in an agroeconomic tract of Lower Ganga basin, India. Journal of Environmental Management, 309, 114691. https://doi.org/10.1016/j.jenvman.2022.114691
59. Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: an overview. Wiley interdisciplinary reviews: data mining and knowledge discovery, 2(1), 86-97. https://doi.org/10.1002/widm.53

60. Nugroho, H., Utama, N. P., & Surendro, K. (2021). Normalization and outlier removal in class center-based firefly algorithm for missing value imputation. Journal of Big Data, 8(1), 129. https://doi.org/10.1186/s40537-021-00518-7
61. O'Shaughnessy, S. A., Evett, S. R., Colaizzi, P. D., & Howell, T. A. (2013). Wireless sensor network effectively controls center pivot irrigation of sorghum. Applied Engineering in Agriculture, 29(6), 853-864. https://doi.org/10.13031/aea.29.9921
62. Olden, J. D., Joy, M. K., & Death, R. G. (2004). An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. Ecological modeling, 178(3-4), 389-397.https://doi.org/10.1016/j.ecolmodel.2004.03.013
63. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71
64. Pullanagari, R., & Cavalli, D. (2023). Advances and applications of multivariate statistics and soil-crop sensing to improve nutrient use efficiency and monitor carbon cycling. Nutrient cycling in agroecosystems, 127(1), 97-99.
https://doi.org/10.1007/s10705-023-10310-z
65. Radmehr, A., Bozorg-Haddad, O., & Loáiciga, H. A. (2022). Integrated strategic planning and multi-criteria decision-making framework with its application to agricultural water management. Scientific Reports, 12(1), 8406.
https://doi.org/10.1038/s41598-022-12194-5
66. Rasheed, M. W., Tang, J., Sarwar, A., Shah, S., Saddique, N., Khan, M. U., Imran Khan, M., Nawaz, S., Shamshiri, R. R., & Aziz, M. (2022). Soil moisture measuring techniques and factors affecting the moisture dynamics: A comprehensive review. Sustainability, 14(18), 11538. https://doi.org/10.3390/su141811538
67. Reichle, R. H., De Lannoy, G. J., Liu, Q., Ardizzone, J. V., Colliander, A., Conaty, A., Crow, W., Jackson, T. J., Jones, L. A., & Kimball, J. S. (2017). Assessment of the SMAP level-4 surface and root-zone soil moisture product using in situ measurements. Journal of hydrometeorology, 18(10), 2621-2645. https://doi.org/10.1175/JHM-D-17-0063.1
68. Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., & Thuiller, W. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 40(8), 913-929. https://doi.org/10.1111/ecog.02881
69. Ronco, P., Zennaro, F., Torresan, S., Critto, A., Santini, M., Trabucco, A., Zollo, A., Galluccio, G., & Marcomini, A. (2017). A risk assessment framework for irrigated agriculture under climate change. Advances in Water Resources, 110, 562-578. https://doi.org/10.1016/j.advwatres.2017.08.003
70. Sayel, J., Khalaf, A., & Najm, A. (2026). Multivariate analysis of seasonal variability in surface water quality for irrigation in Western Iraq. European Journal of Sustainable Development Research, 10(1). https://doi.org/10.29333/ejosdr/17444
71. Shrestha, S., & Kazama, F. (2007). Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river basin, Japan. Environmental modeling & software, 22(4), 464-475. https://doi.org/10.1016/j.envsoft.2006.02.001
72. Simeonov, V., Stratis, J., Samara, C., Zachariadis, G., Voutsa, D., Anthemidis, A., Sofoniou, M., & Kouimtzis, T. (2003). Assessment of the surface water quality in Northern Greece. Water research, 37(17), 4119-4124. https://doi.org/10.1016/S0043-1354(03)00398-1
73. Sudhakaran, S., Mahadevan, H., Arun, V., Krishnakumar, A. P., & Krishnan, K. A. (2020). A multivariate statistical approach in assessing the quality of potable and irrigation water environs of the Netravati River basin (India). Groundwater for Sustainable Development, 11, 100462. https://doi.org/10.1016/j.gsd.2020.100462
74. Singh, K. P., Malik, A., Mohan, D., & Sinha, S. (2004). Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India)-a case study. Water research, 38(18), 3980-3992.
75. https://doi.org/10.1016/j.watres.2004.06.011
76. Sutadian, A. D., Muttil, N., Yilmaz, A. G., & Perera, B. (2016). Development of river water quality indices-a review. Environmental monitoring and assessment, 188(1), 58. https://doi.org/10.1007/s10661-015-5050-0
77. Tejashvini, A., Subbarayappa, C. T., -, M., Chowdappa, H. D., & Ramamurthy, V. (2024). Assessment of irrigation water quality for groundwater in Semi-Arid Region, Bangalore, Karnataka. Water Science, 38(1), 548-568. https://doi.org/10.1080/23570008.2024.2414131
78. Van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. Computers and Electronics in Agriculture, 177, 105709. https://doi.org/10.1016/j.compag.2020.105709
79. Vega, M., Pardo, R., Barrado, E., & Debán, L. (1998). Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis. Water research, 32(12), 3581-3592. https://doi.org/10.1016/S0043-1354(98)00138-9

80. Voinov, A., Kolagani, N., McCall, M. K., Glynn, P. D., Kragt, M. E., Ostermann, F. O., Pierce, S. A., & Ramu, P. (2016). Modeling with stakeholders-next generation. Environmental modeling & software, 77, 196-220. https://doi.org/10.1016/j.envsoft.2015.11.016
81. Wei, S., Xu, T., Niu, G.-Y., & Zeng, R. (2022). Estimating irrigation water consumption using machine learning and remote sensing data in Kansas High Plains. *Remote Sensing, 14*(13), 3004. https://doi.org/10.3390/rs14133004
82. Wiafe, S., Mohammed, H., Kusi, K. A., Gyabaah, D., & Bandoh, T. (2025). An integrated multi-statistic evaluation of surface water quality for irrigation in the northern region of Ghana. Discover Water, 5(1), 80. https://doi.org/10.1007/s43832-025-00263-5
83. Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. Chemometrics and intelligent laboratory systems, 58(2), 109-130. https://doi.org/10.1016/S0169-7439(01)00155-1
84. Xiao, T., Shui, J., Yu, F., & Zhong, Y. (2025). *Research status of winter irrigation based on bibliometric analysis.* Agricultural Water Management, 322, 109974. https://doi.org/10.1016/j.agwat.2025.109974
85. Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. IEEE Transactions on neural networks, 16(3), 645-678. https://doi.org/10.1109/TNN.2005.845141
86. Yidana, S. M., Ophori, D., & Banoeng-Yakubo, B. (2008). A multivariate statistical analysis of surface water chemistry data-The Ankobra Basin, Ghana. Journal of Environmental Management, 86(1), 80-87. https://doi.org/10.1016/j.jenvman.2006.11.023