



Physics-guided interpretable ensemble learning for daily reference evapotranspiration estimation in arid regions

Babak Mohammadi¹ · Mingjie Chen² · Mohammad Reza Nikoo³ · Ali Al-Maktoumi² · Hussam Eldin Elzain²

Received: 13 February 2026 / Accepted: 13 September 2026

© The Author(s) 2026

Abstract

Dependable estimates of reference evapotranspiration (ET_o) are fundamental to managing water sustainably across arid regions, yet conventional empirical methods have limited capacity to represent the complex non-linear interactions among meteorological variables governing ET_o dynamics. This study develops a physics-guided interpretable ensemble learning framework that integrates rule-based (RuleFit) modeling, Support Vector Machines (SVMs), Generalized Simulated Annealing (GenSA) optimization, and SHapley Additive exPlanations (SHAP). Twenty-one RuleFit models were systematically developed across three variable groups, meteorological (air and moisture), solar and radiation, and derived thermodynamic (temperature-vapor), using ten years (2015–2024) of daily observations from two stations (Bahla and Sunaynah) in Oman. The best configuration from each group was integrated into a hierarchical ensemble in which an SVM fuses the group-level predictions as meta-features, with hyperparameters tuned via GenSA. The RuleFit-SVM-GenSA framework achieved exceptional test performance at Bahla (coefficient of determination (R^2)=0.982, Root Mean Square Error (RMSE)=0.216 mm/d) and Sunaynah (R^2 =0.989, RMSE=0.19 mm/d), outperforming benchmark Hargreaves-Samani method (R^2 =0.932–0.958, RMSE=0.379–0.42 mm/d). SHAP analysis identified maximum temperature, solar radiation, and saturation vapor pressure as dominant predictors, confirming physically consistent behavior. Results demonstrate that the developed framework simultaneously achieves high predictive accuracy while retaining a transparent, rule-based model structure at the base-learner level, providing water resource managers with a reliable tool for precision irrigation and hydrological applications in water-scarce arid environments.

Keywords Reference evapotranspiration · Generalized simulated annealing optimization · Meteorological data · Explainable machine learning · Time-series prediction

Extended author information available on the last page of the article

1 Introduction

Evapotranspiration (ET) constitutes one of the dominant terms in the water cycle and is the largest consumptive use of water in agricultural systems (Allen et al., 1998; Katul et al., 2012; Lingling Zhao et al., 2013). Reliable quantification of ET is indispensable for effective water resource planning, management, and regulation in agricultural landscapes, as it establishes the foundation for sustainable water balance, mitigates the impacts of water scarcity, and prevents the overuse and waste of limited freshwater resources (Djaman et al., 2018; Wanniarachchi & Sarukkalgige, 2022). In irrigated agriculture, precise ET estimation enables optimal irrigation scheduling, determination of crop water requirements, and assessment of irrigation system performance, thereby enhancing water use efficiency and agricultural productivity while minimizing environmental degradation (Minhas et al., 2020; Pereira et al., 2020). The significance of ET quantification becomes particularly acute in arid and semi-arid regions, where agriculture accounts for a predominant share of total freshwater consumption and where climate change-induced temperature increases are projected to further intensify evaporative demand (Chartzoulakis & Bertaki, 2015; FAO, 2020; Williams & Abatzoglou, 2025). Furthermore, accurate ET estimation provides critical information for drought monitoring, hydrological model validation, climate impact assessment, and the understanding of ecosystem processes, making it an essential parameter for addressing the dual challenges of ensuring food security and sustainable water management in the face of a growing global population and increasing climatic variability (Fisher et al., 2017; Zhang et al., 2016).

The Food and Agriculture Organization Penman–Monteith (FAO-PM) equation has been established as the international standard for estimating reference evapotranspiration (ETo), and it has performed reliably across a wide range of climate types (Ippolito et al., 2024; Vishwakarma et al., 2022; Yonaba et al., 2023). However, the primary limitation of the FAO-PM method is its substantial meteorological data requirements, which are often unavailable in data-scarce regions, particularly in developing countries (Córdova et al., 2015; Paredes et al., 2020). To address these data constraints, numerous simplified empirical models have been developed, including the Hargreaves-Samani (Hargreaves & Samani, 1985) equation, which requires only temperature data, the Priestley-Taylor (Priestley & Taylor, 1972) method, which utilizes radiation and temperature inputs, and other temperature-based or radiation-based approaches such as the Blaney-Criddle (Blaney & Criddle, 1962) and Makkink (Makkink, 1957) equations. While these empirical methods offer computational simplicity and reduced data requirements, they are inherently limited by their region-specific calibration requirements and empirical coefficients developed for particular climatic zones, often yielding inconsistent results when applied without calibration or beyond their original development contexts (Chia et al., 2020; Sharafi & Mohammadi Ghaleni, 2021). Furthermore, integrating empirical constants and simplified resistance parameterization into these models can lead to significant estimation errors, thereby limiting their reliability for precision agriculture applications and irrigation scheduling in regions with high climatic variability (Dong et al., 2024; Ghiat et al., 2021).

The proliferation of machine learning (ML) algorithms has revolutionized ETo estimation, offering data-driven alternatives that overcome some of the limitations of empirical methods and improve estimation accuracy (Huang et al., 2019; Mehdizadeh et al., 2021). Classical ML approaches, such as artificial neural networks, extreme learning machines,

random forests, and support vector machines (SVM), have consistently demonstrated superior ETo predictive accuracy and emerged as powerful tools for capturing complex non-linear relationships and temporal dependencies in meteorological time series (Amani et al., 2024; Bellido-Jiménez et al., 2021; Kiraga et al., 2024; Reis et al., 2019). Tree-based ensembles, particularly extreme gradient boosting (XGBoost), have gained substantial traction in hydrological modeling largely because of their exceptional capability to handle high-dimensional data, automatically capture feature interactions, and provide robust predictions even with limited meteorological inputs; studies demonstrate that XGBoost coupled with optimization algorithms can achieve coefficients of determination (R^2) exceeding 0.90 for daily ETo forecasting (Mohammadi et al., 2024; Long Zhao et al., 2024; Zhou et al., 2024). However, despite these advances, critical challenges persist, including the "black-box" nature of most ML models, which hinders physical interpretation and stakeholder trust, high computational costs and substantial data requirements for training complex architectures; limited transferability across geographical regions without extensive recalibration; and vulnerability to overfitting when applied to small datasets (Hassija et al., 2024; Rudin, 2019).

Friedman and Popescu (2008) proposed RuleFit as a paradigm-shifting approach to interpretable ML in which sparse linear models automatically integrate detected feature interactions via decision rules derived from tree ensembles. Unlike conventional ensemble methods that operate as opaque predictors, RuleFit extracts conjunctive decision rules from gradient boosted trees and fits a regularized linear model using the original features alongside these rule-based features, thereby maintaining the predictive power of complex tree ensembles while providing transparent, interpretable decision pathways (Bologna, 2021; Luo et al., 2022). Each rule consists of simple logical statements (conjunctions) concerning input variable values, enabling domain experts to understand the global model structure, local prediction influences, and variable importance patterns across different regions of the input space (Förnkrantz et al., 2020; Kozielski et al., 2025; Macha et al., 2022). The dual advantage of RuleFit lies in its ability to achieve predictive accuracy comparable to state-of-the-art ensemble methods while producing models with significantly fewer active features and interpretable rule structures, making it particularly suitable for environmental applications where understanding causal mechanisms and identifying dominant predictors are as critical as prediction accuracy itself (Fokkema, 2020; Li et al., 2024). In the context of ETo modeling, RuleFit's capacity to automatically discover meteorological variable interactions and generate human-readable prediction rules addresses the fundamental trade-off between model performance and interpretability that has long challenged operational agricultural water management and irrigation scheduling systems.

Most existing ensemble-XAI studies for ETo rely on black-box learners combined with post-hoc explanation tools such as SHapley Additive exPlanations (SHAP), where interpretability remains external to the model structure. In contrast, the proposed framework embeds interpretability intrinsically through rule-based learning, using RuleFit to generate sparse, human-readable decision rules that explicitly capture non-linear interactions among atmospheric variables. Moreover, unlike conventional ensemble approaches that stack predictors without physical differentiation, the proposed ensemble is physically structured, with each RuleFit model built on its own variable group and the meta-learner capturing interactions among these processes rather than raw inputs. This dual-level interpretability framework, consisting of structural (rule-based) and explanatory (SHAP-based) components, fundamentally distinguishes the present approach from prior ensemble-XAI studies and positions

the approach as a physics-guided interpretable ensemble rather than a fully inherently interpretable end-to-end model.

Despite substantial advances in ML-based ETo prediction, critical research gaps persist that limit the operational deployment of ML models in water-scarce regions. To address these gaps, the current study presents a physics-guided interpretable ensemble learning framework that integrates eXtreme RuleFit models with SVMs optimized using Generalized Simulated Annealing (GenSA), coupled with SHAP for comprehensive model interpretation. The novelty of this research lies in: (1) the systematic evaluation of 21 input scenarios spanning meteorological (air and moisture), solar and radiation, and derived thermodynamic variable groups using RuleFit models to establish optimal feature combinations for each category at two meteorological stations (Bahla and Sunaynah) in Oman for daily ETo prediction; (2) the innovative ensemble strategy wherein best-performing RuleFit models from each variable group serve as meta-features for SVM training, effectively capturing complex inter-group interactions while maintaining interpretability through the underlying rule structures; (3) the application of GenSA hyperparameter optimization to enhance SVM predictive capability beyond conventional tuning approaches; and (4) the comprehensive SHAP-based explainability analysis applied to the proposed RuleFit framework provides an interpretable view of feature importance and variable interactions.

2 Material and methods

2.1 Study area

The current study focuses on two meteorological stations in Oman's arid region: Bahla (22° 54' N, 57° 15' E, approximately 513 m above sea level) and Sunaynah (23° 34' N, 55° 54' E, approximately 264 m above sea level). Both stations experience a hot desert climate (Köppen classification BWh), characterized by extremely hot and arid conditions with intense sunshine. These locations were strategically selected due to their significant agricultural importance, as they host many farming projects (e.g., Mazoon). Precise estimation of ETo in these areas is crucial for enhancing irrigation strategies, conserving significant water reserves, and ensuring the sustainability of agricultural activities in an environment characterized by scarce and limited freshwater availability. Mean daily air temperature during the study period (2015–2024) averaged 28.97 °C at Bahla station and 29.54 °C at Sunaynah station, with maximum temperatures reaching 41.32 °C and 41.67 °C, respectively, and minimum temperatures dropping to 11.5 °C and 12.62 °C. Figure 1 illustrates the geographical locations of both meteorological stations within the Sultanate of Oman. The dataset encompassed daily meteorological observations spanning a ten-year period (2015–2024), with a systematic temporal split implemented to ensure robust model training and testing. The training dataset comprised observations from January 1, 2015, to July 2, 2022, representing approximately 75% of the total dataset, and was used for model development, hyperparameter tuning, and rule extraction in the RuleFit framework. The remaining 25% of the dataset, covering the period from July 3, 2022, to December 31, 2024, was reserved as an independent test set to evaluate the performance of the developed models.

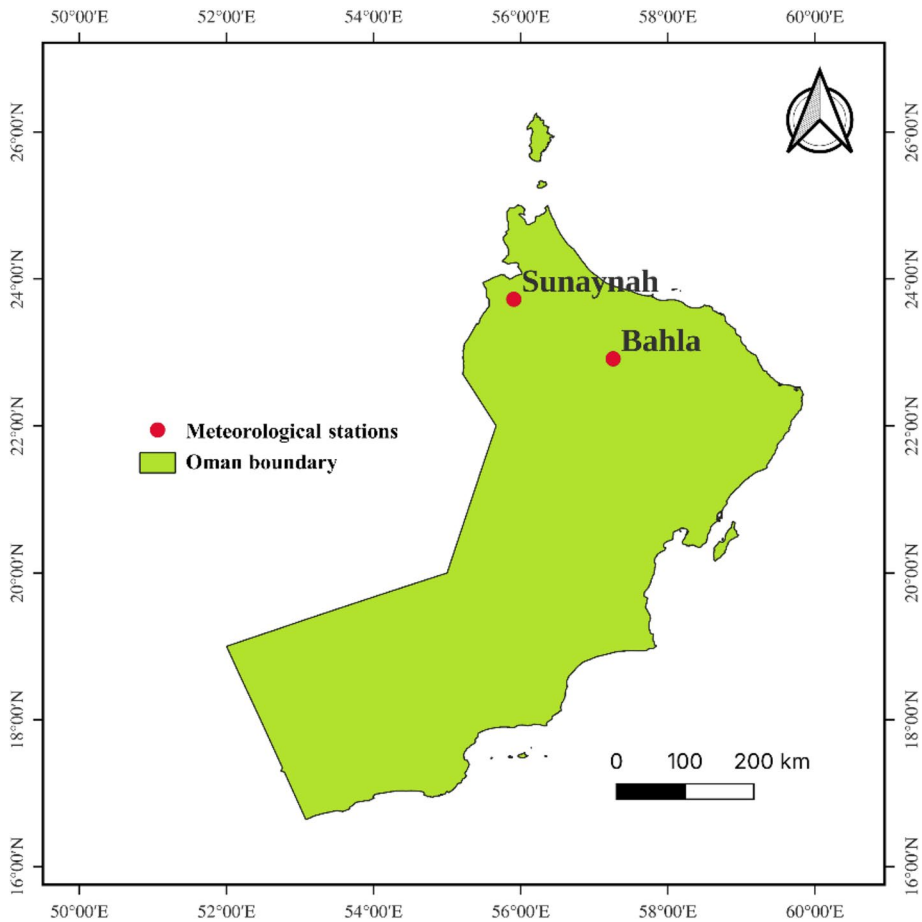


Fig. 1 Map of the study area and the locations of the meteorological stations

2.2 Reference evapotranspiration calculation and benchmark empirical method

For ETo calculation, the FAO-56 Penman–Monteith (PM) formulation is the accepted benchmark and is applied routinely by the research community and by international agencies (Allen et al., 1998). The FAO-56 PM method integrates both energy balance and aerodynamic components to provide reliable ETo estimates, making it the recommended approach by the Food and Agriculture Organization of the United Nations (Allen et al., 1998; Pereira et al., 2015). In this study, ETo values calculated using the FAO-56 PM equation (Eq. 1) were treated as observed data for model development and evaluation.

$$ET_o = \frac{0.408\Delta(R_n - G) + \gamma \times \left(\frac{900}{T+273}\right)u_2 \times (e_s - e_a)}{\Delta + \gamma(1 + 0.34u_2)} \quad (1)$$

In this equation, ETo is reference evapotranspiration (mm/d). The parameter Δ refers to the slope of the saturation vapor pressure curve (kPa/°C), while R_n denotes net radiation at

the land surface ($\text{MJ}/(\text{m}^2\cdot\text{d})$). The soil heat flux is expressed as G ($\text{MJ}/(\text{m}^2\cdot\text{d})$), and γ is the psychrometric constant ($\text{kPa}/^\circ\text{C}$). The variable T indicates the mean daily air temperature measured at a height of 2 m ($^\circ\text{C}$), whereas u_2 corresponds to wind speed at 2 m height (m/s). The variables e_a and e_s indicate the actual and saturated vapor pressures (kPa), and their difference ($e_s - e_a$) defines the vapor pressure deficit (kPa). To evaluate the performance of the developed ML-based frameworks, the Hargreaves-Samani (H-S) method (Hargreaves & Samani, 1985), a widely recognized empirical method, was employed as a benchmark (expressed by Eq. 2).

$$ET_o = 0.408 \times 0.0023 \times R_a \times (T + 17.8)(T_{\max} - T_{\min})^{0.5} \quad (2)$$

In the H-S equation, R_a represents extraterrestrial radiation ($\text{MJ}/(\text{m}^2\cdot\text{d})$), T refers to the mean daily air temperature ($^\circ\text{C}$), and 17.8 is the constant value, and $T_{\max}$ and $T_{\min}$ are the maximum and minimum daily air temperatures ($^\circ\text{C}$), respectively. The original empirical coefficient of 0.0023 (shown in Eq. 2) was calibrated to 0.00251 for Bahla and 0.00248 for Sunaynah based on observed FAO-56 ET_o data. Calibration of the H-S method was performed locally for each station using a non-linear least-squares procedure, in which the optimization minimized the residual sum of squares between observed and predicted daily ET_o . All calibrations were carried out using the training dataset (January 1, 2015, to July 2, 2022) to ensure a fair comparison with the ML framework, with validation performed on the independent test period (July 3, 2022, to December 31, 2024).

2.3 Developed frameworks

The methodological framework employed in the current study (shown in Fig. 2) comprised a multi-stage ML approach designed to systematically evaluate and optimize ET_o estimation under varying data availability scenarios. Twenty-one distinct RuleFit models were developed across three input variable groups: (1) meteorological variables (air and moisture parameters), (2) solar and radiation variables, and (3) derived thermodynamic variables (temperature-vapor relationships). The RuleFit algorithm (Friedman & Popescu, 2008) was selected as the base modeling approach due to its capacity to capture complex non-linear relationships while maintaining model transparency. Performance evaluation using multiple error metrics on the independent test set identified the optimal model for each group, which were subsequently used as inputs to an ensemble framework. These three best-performing models were integrated using an SVM with a radial basis function (RBF) kernel (Cortes & Vapnik, 1995; Vapnik, 2013), yielding an ensemble approach (RuleFit-SVM) that leverages complementary information across different variable categories. To further enhance predictive performance, hyperparameter optimization was performed using the GenSA algorithm (Tsallis & Stariolo, 1996; Xiang et al., 2013), yielding the final RuleFit-SVM-GenSA model. Model interpretability was ensured through SHAP analysis (Lundberg & Lee, 2017), which measures how each predictor influences the model's output, enabling the identification of the most influential meteorological and radiation factors governing ET_o dynamics at both study stations. The framework is interpretable at two levels, with different guarantees at each. At the base-learner level, every RuleFit model is an additive combination of linear terms and sparse conjunctive IF-THEN rules, so each group-level prediction decomposes exactly into its active rules and is directly readable by a domain expert. Interpretability at

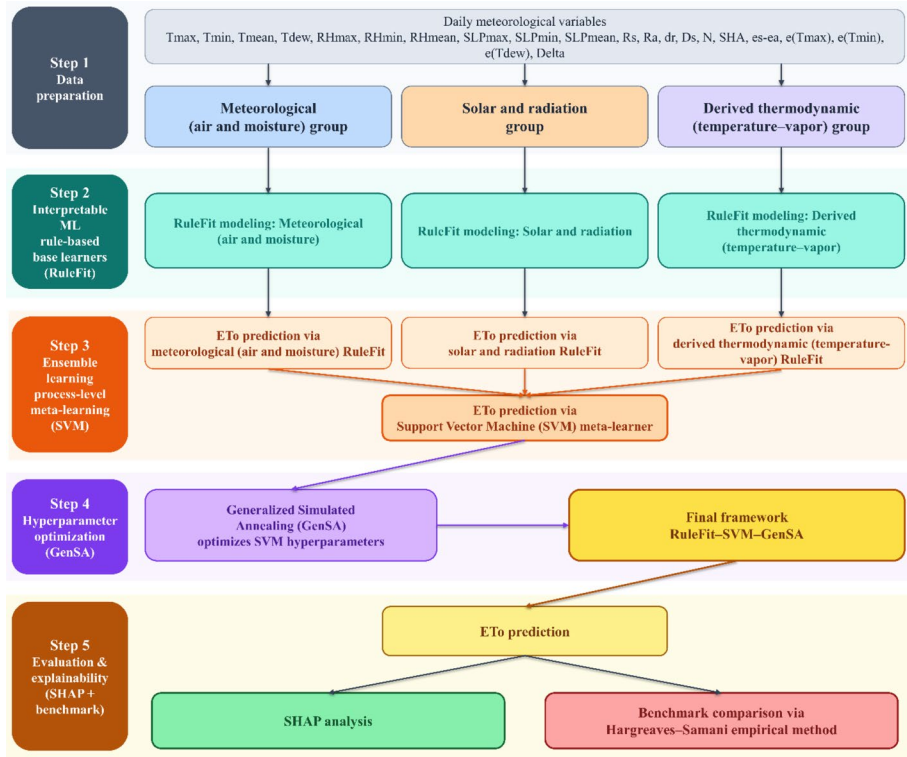


Fig. 2 Flowchart of applied framework, including modeling inputs and outputs of each step

this level is intrinsic rather than reconstructed. At the fusion level, however, the SVM meta-learner combines the three group-level predictions nonlinearly, and this stage is not inherently transparent; attribution across process groups therefore depends on SHAP applied post hoc. The system as a whole is accordingly a hybrid, comprising an intrinsically interpretable base layer whose outputs are fused by an opaque, post-hoc-explained meta-learner, and we make no claim of end-to-end inherent interpretability.

The partition of predictors into meteorological, radiative and thermodynamic groups is likewise a constraint imposed a priori from the structure of the FAO-56 equation, not a decomposition the model recovered from data. Restricting each base learner to one group ensures that its rules involve variables belonging to a single physical process, but it does not establish that the learner identified that process independently, nor that the fitted thresholds correspond to causal mechanisms. The rules remain data-driven empirical relationships estimated from one record.

The architecture described below was designed to overcome the limitations of monolithic learners and homogeneous ensembles for ETo estimation by enforcing a physically-structured decomposition of the ETo problem and coupling intrinsic rule-based modeling with SVM-based meta-learning and SHAP-based explanation. Unlike conventional ensembles that combine predictors operating on the same raw feature space, each base learner in the proposed framework is restricted to a physically coherent variable group, so the generated rules describe a physically coherent subset of variables rather than a mixture of

correlated drivers across processes; this grouping does not imply that each learner independently identifies a causal physical subprocess. The SVM meta-learner then operates on these process-level predictions instead of the raw inputs, while SHAP provides an independent verification layer that confirms the consistency of the learned relationships with established ETo physics.

2.4 Input variable classification and scenario development

Input variables for the RuleFit modeling framework were systematically classified into three distinct groups based on their physical characteristics and data acquisition requirements, as detailed in Table 1. The first group (meteorological (air and moisture) variables) encompassed ten variables, including station-level pressure measurements (SLPmean, SLPmax, SLPmin), air temperature indicators (Tmean, Tmax, Tmin), relative humidity values (RHmean, RHmax, RHmin), and dewpoint temperature (Tdew). The second group (solar and radiation variables) comprised six variables related to energy input and astronomical factors, including solar radiation (Rs), extraterrestrial radiation (Ra), solar declination (Ds), inverse Earth-Sun distance (dr), sunset hour angle (SHA), and daylight hours (N). The third group (derived thermodynamic (temperature-vapor) variables) consisted of five computationally derived variables that characterize the vapor pressure dynamics and energy partitioning in the atmosphere, including the slope of the saturation vapor pressure curve (Delta), saturation vapor pressures at different temperature points ($e(T_{min})$, $e(T_{max})$, $e(T_{dew})$), and vapor pressure deficit ($es-ea$). This classification scheme enabled a systematic evaluation of model performance across different data availability scenarios while maintaining physical consistency within each variable group. The design of the 21 RuleFit models followed a hybrid theory-driven and data-driven rationale. The partitioning of input variables into three groups was theory-driven, reflecting the physical structure of the PM equation, in which ETo is governed by several variables. This grouping ensures that each subsequent base learner captures a physically coherent process rather than a mixture of correlated but mechanistically different drivers. Within each group, a progressive variable-reduction strategy was applied to generate the individual model configurations. This systematic reduction scheme provides a controlled framework for assessing how predictive skill responds to decreasing information content. The resulting 21 configurations were subsequently compared using independent test-set metrics, and the best-performing model within each group was short-listed in a data-driven manner for integration into the ensemble meta-learner. Among the variables listed in Table 1, SLPmean, SLPmax, SLPmin, Tmean, Tmax, Tmin, RHmean, RHmax, and RHmin are directly obtained from the meteorological stations as routinely measured observations. All remaining variables in Table 1 are computationally derived using standard empirical, astronomical, and thermodynamic relationships. It should be noted that the derived thermodynamic-group variables (Delta, $e(T_{min})$, $e(T_{max})$, $e(T_{dew})$, $es-ea$) are deterministic functions of the meteorological group's raw temperature and humidity observations, and therefore do not constitute an independent source of information relative to the meteorological group; they represent a physically transformed version of the same underlying temperature-humidity process.

A progressive variable reduction strategy was implemented within each group to develop twenty-one distinct modeling scenarios, as presented in Table 2. For the meteorological group, ten scenarios (RuleFit1 through RuleFit10) were constructed by sequentially elimi-

Table 1 Classification and description of input variables organized into three variable groups for RuleFit model development

Group	Variable Description	Variable ID	Unit
Meteorological (air and moisture)	Mean station level pressure	SLPmean	kPa
	Maximum station level pressure	SLPmax	kPa
	Minimum station level pressure	SLPmin	kPa
	Mean air temperature	Tmean	°C
	Maximum air temperature	Tmax	°C
	Minimum air temperature	Tmin	°C
	Mean relative humidity	RHmean	%
	Maximum relative humidity	RHmax	%
	Minimum relative humidity	RHmin	%
	Dewpoint temperature	Tdew	°C
Solar and radiation	Extraterrestrial radiation	Ra	MJ/(m ² ·d)
	Solar radiation	Rs	MJ/(m ² ·d)
	Solar declination	Ds	radians
	Inverse Earth-Sun distance	dr	-
	Sunset hour angle	SHA	radians
	Daylight hours	N	hours
Derived thermodynamic (temperature-vapor)	Slope of the saturation vapor pressure curve	Delta	kPa/°C
	Saturation vapor pressure at minimum temperature	e(Tmin)	kPa
	Saturation vapor pressure at maximum temperature	e(Tmax)	kPa
	Saturation vapor pressure at the dewpoint temperature	e(Tdew)	kPa
	Vapor pressure deficit	es-ea	kPa

Table 2 Progressive variable reduction scenarios for RuleFit model development across three input variable groups

No	Group	Input(s)	Model
1	Meteorological (air and moisture)	Tmax, Tmin, Tmean, Tdew, RHmax, RHmin, RHmean, SLPmax, SLPmin, SLPmean	RuleFit1
2		Tmax, Tmin, Tmean, Tdew, RHmax, RHmin, RHmean, SLPmax, SLPmin	RuleFit2
3		Tmax, Tmin, Tmean, Tdew, RHmax, RHmin, RHmean, SLPmax	RuleFit3
4		Tmax, Tmin, Tmean, Tdew, RHmax, RHmin, RHmean	RuleFit4
5		Tmax, Tmin, Tmean, Tdew, RHmax, RHmin	RuleFit5
6		Tmax, Tmin, Tmean, Tdew, RHmax	RuleFit6
7		Tmax, Tmin, Tmean, Tdew	RuleFit7
8		Tmax, Tmin, Tmean	RuleFit8
9		Tmax, Tmin	RuleFit9
10		Tmax	RuleFit10
11	Solar and radiation	Rs, Ra, dr, Ds, N, SHA	RuleFit11
12		Rs, Ra, dr, Ds, N	RuleFit12
13		Rs, Ra, dr, Ds	RuleFit13
14		Rs, Ra, dr	RuleFit14
15		Rs, Ra	RuleFit15
16		Rs	RuleFit16
17	Derived thermodynamic (temperature-vapor)	es-ea, e(Tmax), e(Tmin), e(Tdew), Delta	RuleFit17
18		es-ea, e(Tmax), e(Tmin), e(Tdew)	RuleFit18
19		es-ea, e(Tmax), e(Tmin)	RuleFit19
20		es-ea, e(Tmax)	RuleFit20
21		es-ea	RuleFit21

nating variables, starting with the complete set of ten variables in RuleFit1 and progressively reducing to a single maximum temperature variable (Tmax) in RuleFit10. This hierarchical reduction followed a logical sequence that prioritized retaining temperature variables (which are universally available) while systematically removing humidity and pressure measurements, which are often unavailable in data-scarce regions. Similarly, the solar and radiation group included six scenarios (RuleFit11 through RuleFit16), progressing from all six radiation-related variables to only solar radiation in RuleFit16, reflecting the varying complexity of radiation measurements across different monitoring networks. The derived thermodynamic group comprised five scenarios (RuleFit17 through RuleFit21), reducing from the full complement of vapor pressure variables to vapor pressure deficit only in RuleFit21. The rationale for selecting the input variables in Table 2 follows directly from the structure of the FAO-56 PM equation, which explicitly requires temperature, humidity, pressure, and radiation measurements, together with their thermodynamic derivatives, and from the associated astronomical parameters needed to estimate extraterrestrial radiation. Accordingly, each variable listed in Table 2 corresponds to either a direct FAO-56 input or

an astronomical factor needed for radiation estimation. The progressive variable-reduction design was not constructed to exhaustively enumerate all possible variable combinations, but to define a physically meaningful reduction pathway in which variables are eliminated in descending order of data-acquisition difficulty, thereby generating a sequence of configurations that mirror realistic data-availability scenarios in arid-region monitoring networks.

2.5 RuleFit modeling framework

The RuleFit algorithm represents a sophisticated ensemble learning approach that synergistically combines the interpretability of rule-based models with the predictive power of modern ML techniques. Originally proposed by Friedman and Popescu (Friedman & Popescu, 2008), RuleFit automatically constructs a sparse linear model from a large collection of decision rules extracted from an ensemble of decision trees, thereby bridging the gap between model accuracy and interpretability. The fundamental principle underlying RuleFit is to transform complex tree ensembles into a transparent set of interpretable rules and linear terms, where each rule represents a conjunction of conditions on the input variables. Mathematically, the RuleFit model can be expressed as:

$$\hat{f}(x) = a_0 + \sum_{j=1}^p a_j x_j + \sum_{k=1}^K b_k r_k(x) \quad (3)$$

where $\hat{f}(x)$ denotes the predicted output, a_0 is the intercept term, a_j is the coefficient for the j -th linear term of the original input variables x_j , b_k denotes the coefficient for the k -th rule $r_k(x)$, p refers to the number of input variables, and K is the total number of extracted rules. Each rule $r_k(x)$ is a binary indicator function that equals 1 when all conditions in the rule are satisfied and 0 otherwise. The rules are generated by traversing all possible paths from the root to the leaves of an ensemble of decision trees, with each path defining a unique combination of variable thresholds (Friedman & Popescu, 2008). The model coefficients are estimated using regularized linear regression, specifically Lasso regularization (Tibshirani, 1996), which performs automatic variable selection by shrinking less important coefficients to zero, thereby producing a sparse and interpretable model. The hyperparameter configurations of the RuleFit models are listed in Table 3.

In this study, the eXtreme RuleFit (xrf) implementation was employed, which leverages the computational efficiency and enhanced predictive performance of the XGBoost algorithm (Chen & Guestrin, 2016). This study used the xrf package in the R programming environment to implement the RuleFit framework, exploiting the advantages of gradient-boosted trees via the XGBoost approach (Holub, 2022). The algorithm builds trees sequentially, where each new tree attempts to correct the residual errors made by the existing ensemble (Friedman, 2001, 2002).

2.6 Development of an ensemble modeling strategy: RuleFit-SVM-GenSA framework

This study develops a novel hybrid ensemble modeling framework that integrates the interpretable rule-based predictions from optimal RuleFit models with the powerful non-linear

Table 3 Hyperparameter configuration for RuleFit models implemented using the eXtreme RuleFit (xrf) package in R

Hyperparameter	Value	Description
XGBoost Tree Ensemble Parameters		
nrounds	500	Number of boosting rounds (trees) in the ensemble; controls the total number of decision trees generated for rule extraction
max_depth	3	Maximum depth of individual trees; limited to 3 levels to ensure simple, interpretable rules with at most three conditions per rule
eta	0.01	Learning rate (shrinkage parameter); conservative step size in gradient descent optimization to enable fine-grained model refinement and reduce overfitting
subsample	0.5	Fraction of training data sampled for each tree (50%); implements stochastic gradient boosting to enhance generalization and reduce tree correlation
RuleFit Model Structure Parameters		
type	both	Model type specifying the inclusion of both original input variables (linear terms) and extracted rules (interaction terms) in the final model
deoverlap	TRUE	De-overlap rules parameter; enables post-processing to identify and remove redundant rules with substantial coverage overlap, enhancing interpretability
family	Gaussian	Distribution family for the response variable; specifies continuous regression with a squared error loss function, appropriate for ETo estimation

mapping capabilities of SVM. SVM, originally developed for classification tasks and later extended to regression problems as Support Vector Regression (SVR) (Vapnik & Chervonenkis, 1974), operates by mapping input data into a high-dimensional feature space where linear relationships can be established, even when the original input–output relationship is highly non-linear. The fundamental optimization problem in ε -SVR seeks to minimize:

$$\min_{\omega, b} \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*) \quad (4)$$

subject to the constraint that prediction errors remain within the ε -insensitive tube, where ω is the weight vector, b represents the bias term, and C serves as the regularization parameter, determining the balance between model simplicity and fitting accuracy (e.g., training error). The variables ξ_i and ξ_i^* serve as slack variables that account for points outside the ε -tube. For this study, the RBF kernel was selected for its superior ability to handle non-

linear relationships and proven effectiveness in hydrological applications (Raghavendra & Deka, 2014). The ensemble architecture receives predictions from the three best-performing RuleFit models (one from each variable group) as input features, effectively creating a meta-learning framework in which the SVM learns to optimally combine the specialized knowledge captured by each RuleFit model's interpretation of the different physical processes governing ETo.

To optimize the performance of the SVM ensemble model, the GenSA algorithm is employed for comprehensive hyperparameter optimization, representing a significant methodological advancement over conventional grid search or random search approaches. GenSA extends the classical simulated annealing framework by implementing a generalized visiting distribution based on Tsallis statistics, enabling more efficient exploration of the hyperparameter search space and improved convergence to global optima (Tsallis & Stariolo, 1996; Xiang et al., 2013). The algorithm mimics the physical process of annealing in metallurgy, analogously searching for the minimum of the objective function by probabilistically accepting worse solutions early in the search (at high temperatures) to escape local minima, then gradually reducing this acceptance probability as the search progresses toward convergence. The GenSA acceptance mechanism employs a temperature-dependent probability criterion based on the generalized Boltzmann distribution, enabling the algorithm to avoid local minima while systematically guiding the search toward the global optimum. The search ranges of the SVM hyperparameters and the GenSA control settings are summarized in Table 4.

2.7 Model performance evaluation metrics and SHAP analysis

A comprehensive set of statistical performance metrics was employed to evaluate the predictive accuracy and reliability of the models. These metrics included Mean Absolute Error (MAE), Root Mean Square Error (RMSE), the ratio of RMSE to the standard deviation of observations (RSR), and the coefficient of determination (R^2), which are presented in Eqs. 5, 6, 7, 8.

$$MAE = \frac{1}{n} \sum_{i=1}^n |ET_i^p - ET_i^o| \quad (5)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (ET_i^p - ET_i^o)^2} \quad (6)$$

$$RSR = \frac{RMSE}{STDEV_{obs}} \quad (7)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (ET_i^o - ET_i^p)^2}{\sum_{i=1}^n (ET_i^o - \overline{ET^o})^2} \quad (8)$$

In Eqs. 5, 6, 7, 8, n is the total number of observations, ET_i^o is the observed ETo value calculated using the FAO-56 PM method for the i -th day, ET_i^p is the corresponding ETo

Table 4 SVM hyperparameter search ranges and GenSA control settings used in the RuleFit-SVM-GenSA ensemble

Component	Parameter	Search Range/Value	Description
SVM Hyperparameters (Optimization Target)			
Cost	C	[0.1, 1000]	Regularization parameter controlling the trade-off between model complexity and training error; determines penalty for prediction errors exceeding the ε -insensitive tube
Gamma	σ (gamma)	[0.0001, 10]	RBF kernel parameter controlling the influence radius of individual training samples; governs model smoothness and capacity for capturing local patterns
Epsilon	ε	[0.01, 1]	Width of the ε -insensitive zone where no penalty is incurred for prediction errors; defines the tolerance margin around predictions
GenSA Optimization Control Parameters			
Maximum iterations	maxit	500	Maximum number of iterations allowed for the optimization algorithm; controls the extent of search exploration
Maximum function evaluations	max.call	1000	Maximum number of objective function evaluations (SVM training and validation); limits computational budget
Initial temperature	temperature	5	Starting temperature for the simulated annealing process; controls initial acceptance probability for worse solutions to escape local minima

value predicted by the model, $\overline{ET^o}$ represents the mean observed ETo, and $STDEV_{Obs}$ is the standard deviation of the observed ETo values. SHAP is an interpretable ML framework based on Shapley values derived from cooperative game theory that evaluates the influence of each input feature on the model's predictions (Lundberg & Lee, 2017). Positive SHAP values indicate that a feature increases the predicted ETo, whereas negative values indicate a decrease, with the magnitude representing the strength of the feature's contribution.

3 Results

3.1 Meteorological (air and moisture)-based ETo prediction

The performance evaluation of the ten meteorological-based RuleFit models at Bahla station (listed in Table 5) revealed consistently high predictive accuracy across all input configurations. During the training phase, all models demonstrated excellent goodness-of-fit, achieving R^2 values of 0.972 (RuleFit1) to 0.941 (RuleFit10), RMSE values spanning 0.281 mm/d to 0.406 mm/d, and MAE values between 0.225 mm/d and 0.327 mm/d. The comprehensive RuleFit1 model, incorporating all ten meteorological variables, achieved strong training performance ($R^2=0.972$, RMSE=0.281 mm/d, RSR=0.167). The progressive reduction in model complexity from RuleFit1 to RuleFit10 resulted in a steady decline in performance, with the single-variable model (RuleFit10 using only Tmax) achieving the weakest training metrics ($R^2=0.941$, RMSE=0.406 mm/d, RSR=0.242), though it still maintained acceptable accuracy levels for practical applications. Test set performance metrics showed the expected degradation relative to training results, with R^2 values of 0.955 (RuleFit1) to 0.911 (RuleFit10), RMSE values spanning 0.342 mm/d to 0.482 mm/d, and RSR values ranging from 0.209 to 0.298. Notably, RuleFit3 emerged as the optimal configuration for this group, achieving the best test set performance with RSR=0.209, $R^2=0.956$, RMSE=0.337 mm/d, and MAE=0.269 mm/d, marginally outperforming the more complex RuleFit1 (RSR=0.212, $R^2=0.955$) and RuleFit2 (RSR=0.21, $R^2=0.956$) models. The identification of RuleFit3 as the optimal model within the meteorological group provides valuable practical guidance for ETo estimation in data-limited scenarios, demonstrating that accurate predictions can be achieved with only eight carefully selected variables, without the need for complete meteorological measurements.

Figure 3 presents scatter plots comparing observed and predicted daily ETo values for the ten meteorological-based RuleFit models at Bahla station. The plots reveal minimal scatter and negligible systematic bias for models RuleFit1 through RuleFit7, with data points densely concentrated along the 1:1 (perfect prediction) line and no evident heteroscedasticity across the full range of ETo values. RuleFit3, identified as the optimal model in this group, exhibits particularly tight clustering with virtually no outliers, confirming its superior predictive capability as indicated by the quantitative metrics (test $R^2=0.956$). As model complexity decreases from RuleFit7 to RuleFit10, the scatter plots show progressively

Table 5 Performance evaluation metrics for the meteorological (air and moisture) group RuleFit models at the Bahla station

Train	RuleFit1	RuleFit2	RuleFit3	RuleFit4	RuleFit5	RuleFit6	RuleFit7	RuleFit8	RuleFit9	RuleFit10
MAE	0.225	0.227	0.228	0.239	0.244	0.26	0.280	0.280	0.292	0.327
RMSE	0.281	0.283	0.285	0.295	0.300	0.324	0.349	0.349	0.364	0.406
RSR	0.167	0.169	0.170	0.176	0.179	0.193	0.208	0.208	0.217	0.242
R^2	0.972	0.971	0.971	0.969	0.968	0.963	0.957	0.957	0.953	0.941
Test	RuleFit1	RuleFit2	RuleFit3	RuleFit4	RuleFit5	RuleFit6	RuleFit7	RuleFit8	RuleFit9	RuleFit10
MAE	0.272	0.271	0.269	0.293	0.297	0.327	0.356	0.356	0.359	0.391
RMSE	0.342	0.340	0.337	0.367	0.369	0.407	0.451	0.451	0.449	0.482
RSR	0.212	0.210	0.209	0.228	0.229	0.252	0.279	0.279	0.278	0.298
R^2	0.955	0.956	0.956	0.948	0.948	0.936	0.922	0.922	0.923	0.911

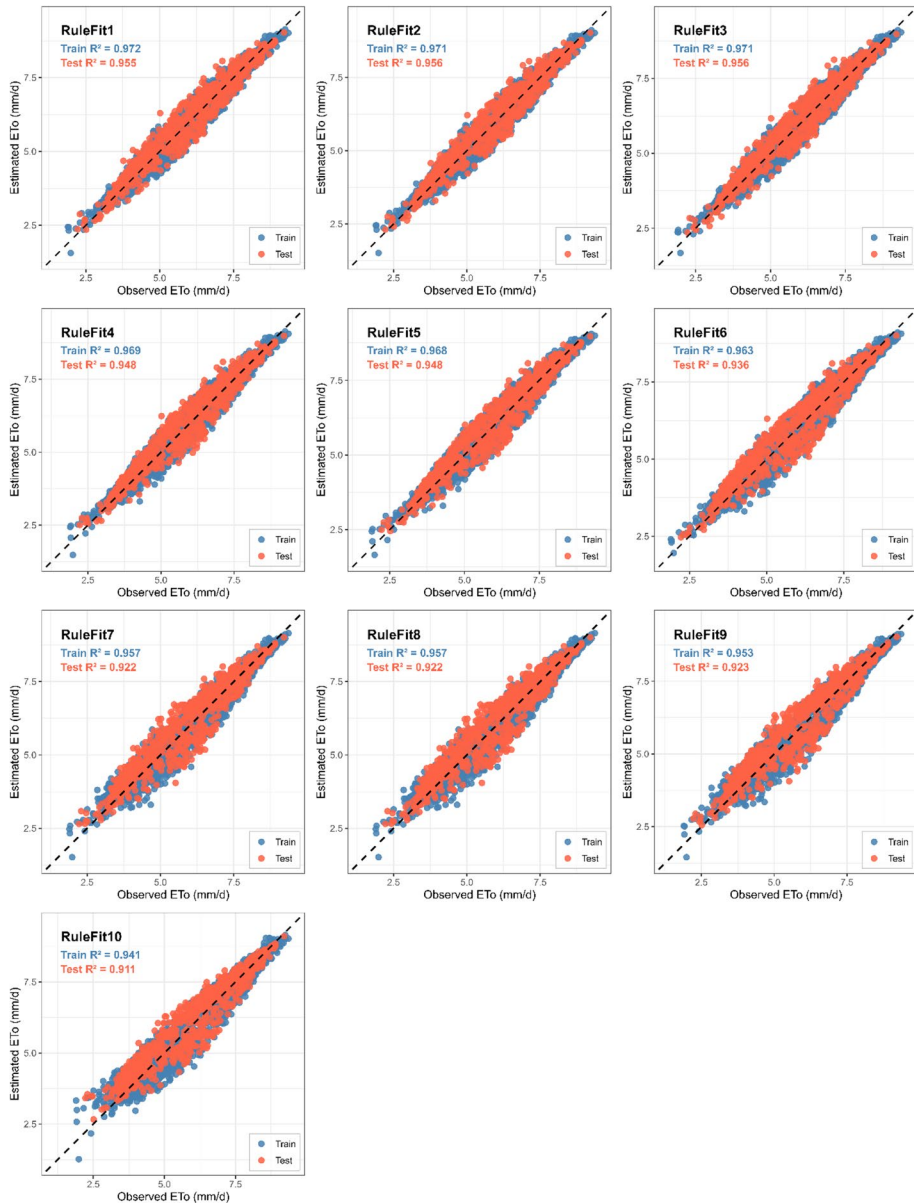


Fig. 3 Comparison of observed and predicted daily ETo for meteorological (air and moisture) group RuleFit models at Bahla station. The 1:1 perfect prediction line is shown by the dashed black line

increased dispersion, reflecting the diminished capacity of temperature-only inputs to capture the full complexity of atmospheric evaporative demand. Notably, even the minimalist RuleFit10 model (using only T_{max}) maintains a reasonable predictive structure without systematic over- or under-prediction patterns, albeit with noticeably greater scatter compared to the more comprehensive models.

Table 6 lists the performance metrics of meteorological-based RuleFit models at Sunaynah station. Training-phase results demonstrated R^2 values ranging from 0.975 (RuleFit1) to 0.937 (RuleFit10), RMSE values between 0.305 mm/d and 0.488 mm/d, and MAE values between 0.245 mm/d and 0.406 mm/d. The comprehensive RuleFit1 model (incorporating all ten meteorological variables) achieved excellent training metrics ($R^2=0.975$, RMSE=0.305 mm/d, RSR=0.157), closely followed by RuleFit2 ($R^2=0.975$, RMSE=0.307 mm/d, RSR=0.158) and RuleFit3 ($R^2=0.975$, RMSE=0.31 mm/d, RSR=0.16). A systematic degradation in model performance was observed with progressive variable reduction followed an expected pattern, with RSR increasing from 0.157 (RuleFit1) to 0.251 (RuleFit10) during training, though the single-variable model (RuleFit10) still achieved acceptable accuracy ($R^2=0.937$, RMSE=0.488 mm/d). Independent test set evaluation revealed RuleFit2 as the optimal configuration for Sunaynah station, retaining nine meteorological variables (excluding only SLPmean) and achieving the best generalization performance with RSR=0.191, $R^2=0.963$, RMSE=0.355 mm/d, and MAE=0.285 mm/d. This result differs notably from the Bahla station, where RuleFit3 emerged as optimal. RuleFit1 demonstrated nearly equivalent test performance (RSR=0.192, $R^2=0.963$), while RuleFit3 showed marginally higher error (RSR=0.195, $R^2=0.962$).

Figure 4 presents scatter plots comparing observed and predicted daily ETo values for the meteorological-based RuleFit models at Sunaynah station. Models RuleFit1 through RuleFit3 exhibit remarkably tight clustering along the 1:1 (perfect prediction) line, with minimal scatter and no systematic bias across the entire ETo range, visually confirming their nearly equivalent test performance ($R^2=0.963$ for RuleFit1 and RuleFit2, $R^2=0.962$ for RuleFit3). As model complexity decreases from RuleFit4 to RuleFit10, progressive increases in scatter become evident, though the dispersion remains comparable to that of the corresponding models at Bahla station. Even the minimalist RuleFit10 configuration (using only Tmax) maintains a reasonable predictive structure without severe outliers or systematic over- or under-prediction bias, achieving a respectable test R^2 of 0.911. The strong visual coherence between training- and test-set predictions across all panels confirms the absence of overfitting and validates the robust generalization capability of the RuleFit framework at Sunaynah station.

Table 6 Performance evaluation metrics for meteorological (air and moisture) group RuleFit models at Sunaynah station

Train	RuleFit1	RuleFit2	RuleFit3	RuleFit4	RuleFit5	RuleFit6	RuleFit7	RuleFit8	RuleFit9	RuleFit10
MAE	0.245	0.247	0.250	0.269	0.271	0.292	0.321	0.321	0.330	0.406
RMSE	0.305	0.307	0.310	0.333	0.336	0.364	0.398	0.398	0.409	0.488
RSR	0.157	0.158	0.160	0.172	0.173	0.188	0.205	0.205	0.211	0.251
R^2	0.975	0.975	0.975	0.971	0.970	0.965	0.958	0.958	0.956	0.937
Test	RuleFit1	RuleFit2	RuleFit3	RuleFit4	RuleFit5	RuleFit6	RuleFit7	RuleFit8	RuleFit9	RuleFit10
MAE	0.286	0.285	0.292	0.338	0.339	0.369	0.390	0.390	0.390	0.467
RMSE	0.357	0.355	0.362	0.422	0.423	0.456	0.481	0.481	0.478	0.553
RSR	0.192	0.191	0.195	0.227	0.228	0.245	0.259	0.259	0.257	0.298
R^2	0.963	0.963	0.962	0.948	0.948	0.940	0.933	0.933	0.934	0.911

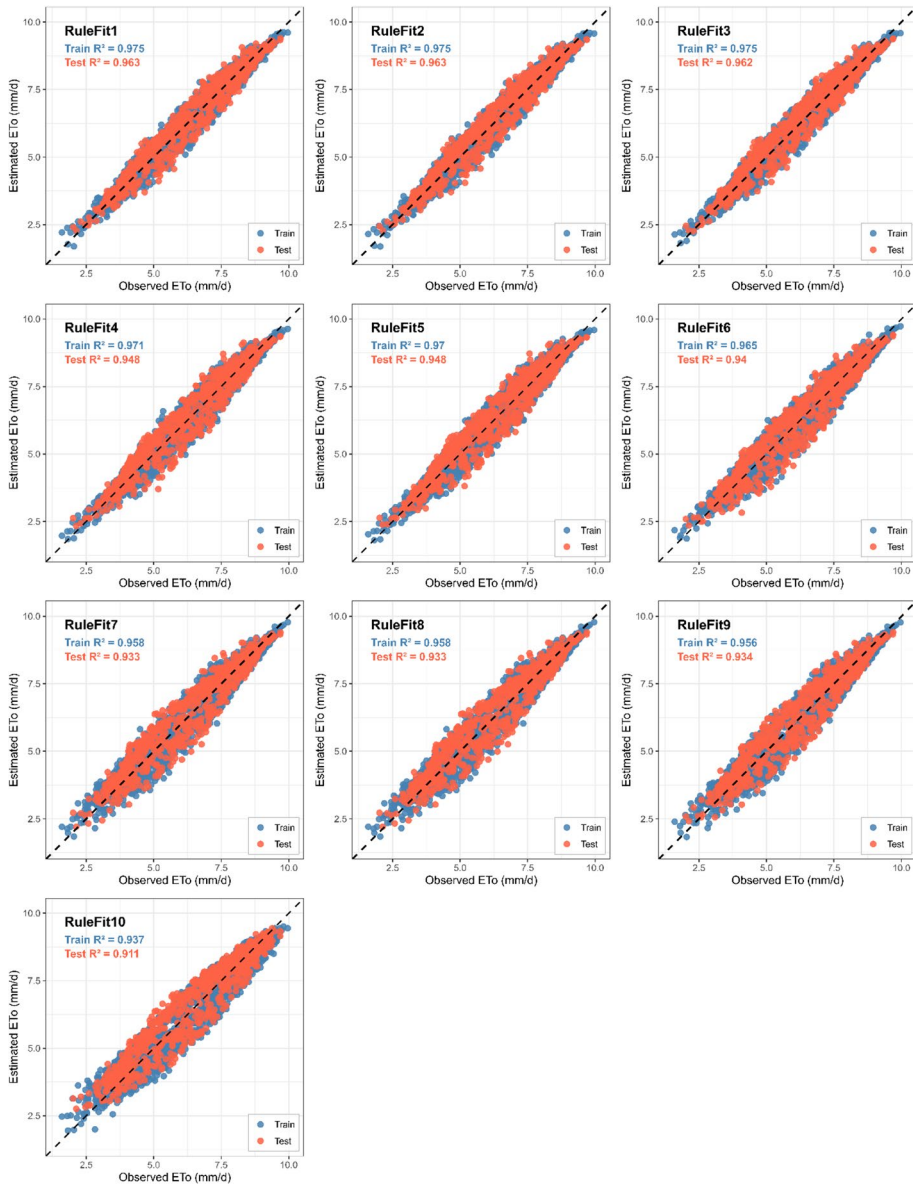


Fig. 4 Comparison of observed and predicted daily ETo for meteorological (air and moisture) group RuleFit models at Sunaynah station. The 1:1 perfect prediction line is shown by the dashed black line

3.2 Solar and radiation-based ETo prediction

Table 7 provides a performance comparison of the solar and radiation group models at the Bahla station. During the training phase, RuleFit11 through RuleFit13 demonstrated identical performance metrics ($R^2=0.96$, $RMSE=0.336$ mm/d, $MAE=0.26$ mm/d, $RSR=0.2$), despite progressively removing variables from the full six-variable set (RuleFit11: Rs, Ra, dr,

Table 7 Performance evaluation metrics for solar and radiation group RuleFit models at the Bahla station

Train	RuleFit11	RuleFit12	RuleFit13	RuleFit14	RuleFit15	RuleFit16
MAE	0.260	0.260	0.260	0.261	0.332	0.416
RMSE	0.336	0.336	0.336	0.338	0.409	0.520
RSR	0.200	0.200	0.200	0.201	0.244	0.310
R ²	0.960	0.960	0.960	0.960	0.941	0.904
Test	RuleFit11	RuleFit12	RuleFit13	RuleFit14	RuleFit15	RuleFit16
MAE	0.337	0.337	0.337	0.338	0.412	0.439
RMSE	0.468	0.468	0.468	0.467	0.533	0.566
RSR	0.290	0.290	0.290	0.289	0.330	0.350
R ²	0.916	0.916	0.916	0.916	0.891	0.877

Ds, N, SHA) to the four-variable configuration (RuleFit13: Rs, Ra, dr, Ds). In contrast, further simplification to RuleFit15 (Rs and Ra only) resulted in noticeable performance decline ($R^2=0.941$, RMSE=0.409 mm/d, RSR=0.244), while the single-variable RuleFit16 model (Rs only) showed substantial degradation ($R^2=0.904$, RMSE=0.52 mm/d, RSR=0.31), although still maintaining acceptable accuracy for many practical applications. Independent test set evaluation confirmed the training-phase patterns, with RuleFit11 through RuleFit13 again achieving virtually identical performance ($R^2=0.916$, RMSE=0.468 mm/d, RSR=0.29), validating the robust generalization capability of these radiation-based configurations. RuleFit14 emerged as the optimal model for this group, achieving nearly equal predictive accuracy compared to more complex models while requiring fewer input variables than the six-variable RuleFit11 configuration. The test-set performance of RuleFit15 ($R^2=0.891$, RMSE=0.533 mm/d, RSR=0.330) and RuleFit16 ($R^2=0.877$, RMSE=0.566 mm/d, RSR=0.350) remained acceptable but was substantially inferior to that of RuleFit14.

Figure 5 depicts observed versus predicted scatter plots of daily ETo values for the solar and radiation group models at Bahla station. RuleFit11 through RuleFit14 exhibit visually indistinguishable scatter patterns, confirming their equivalent performance metrics and validating that the progressive removal of SHA, N, and Ds variables does not compromise prediction accuracy. The optimal RuleFit14 model demonstrates a strong linear relationship across the full ETo range with minimal systematic bias, reflecting the ability of key radiation-related variables to adequately represent evaporative demand, despite the absence of moisture-related inputs. The transition from RuleFit14 to RuleFit15 shows noticeable increases in dispersion, while RuleFit16 exhibits the most pronounced scatter among all radiation group models with several outliers departing from the 1:1 (perfect prediction) line. Nonetheless, even RuleFit16 maintains reasonable predictive coherence without severe systematic over- or under-prediction patterns, demonstrating the strong explanatory power of solar radiation as a primary driver of ETo, even under highly simplified input conditions.

Table 8 lists the performance metrics of the solar and radiation group models at Sunaynah station. During the training phase, RuleFit11 through RuleFit14 achieved identical performance metrics ($R^2=0.965$, RMSE=0.361 mm/d, MAE=0.278 mm/d, RSR=0.186), despite the progressive elimination of SHA, N, and Ds from the full six-variable configuration. The training performance of Sunaynah's radiation models (RSR=0.186) notably was better than that of Bahla's equivalent models (RSR=0.2), suggesting potentially more consistent radiation-ETo relationships at Sunaynah, possibly attributable to differences in local atmospheric conditions, aerosol loading, or cloud cover patterns. The performance degradation at RuleFit15 ($R^2=0.943$, RMSE=0.463 mm/d, RSR=0.239) and RuleFit16 ($R^2=0.9$,

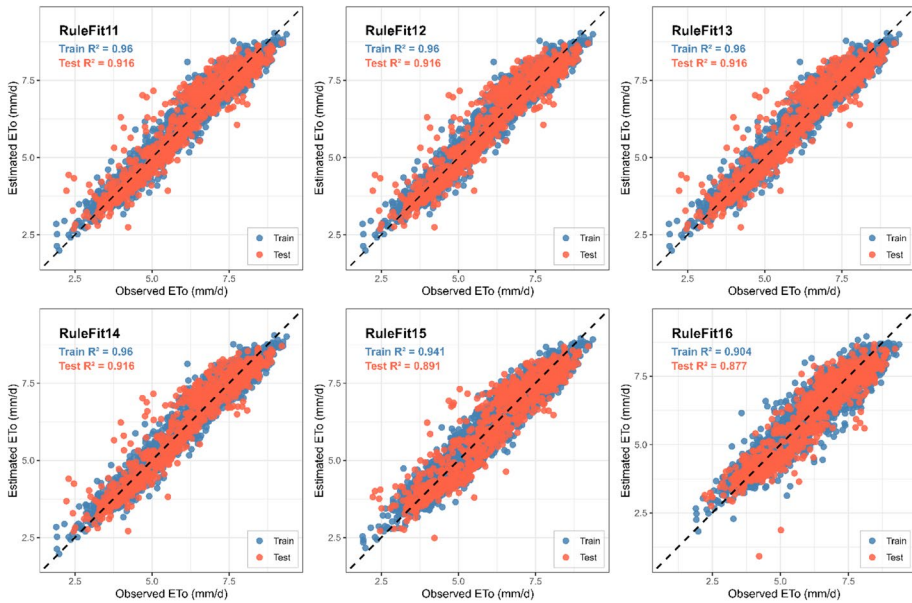


Fig. 5 Comparison of observed and predicted daily ETo for solar and radiation group RuleFit models at Bahla station. The 1:1 perfect prediction line is shown by the dashed black line

Table 8 Performance evaluation metrics for solar and radiation group RuleFit models at Sunaynah station

Train	RuleFit11	RuleFit12	RuleFit13	RuleFit14	RuleFit15	RuleFit16
MAE	0.278	0.278	0.278	0.278	0.372	0.488
RMSE	0.361	0.361	0.361	0.361	0.463	0.614
RSR	0.186	0.186	0.186	0.186	0.239	0.316
R ²	0.965	0.965	0.965	0.965	0.943	0.900
Test	RuleFit11	RuleFit12	RuleFit13	RuleFit14	RuleFit15	RuleFit16
MAE	0.344	0.344	0.344	0.345	0.486	0.601
RMSE	0.476	0.476	0.476	0.477	0.614	0.767
RSR	0.256	0.256	0.256	0.257	0.331	0.413
R ²	0.934	0.934	0.934	0.934	0.891	0.829

RMSE=0.614 mm/d, RSR=0.316) followed expected trends, with the single-variable model showing substantially reduced but still acceptable accuracy for preliminary ETo estimation in data-scarce scenarios. For the test set, RuleFit11 through RuleFit14 maintained nearly identical performance ($R^2=0.934$, RMSE=0.476–0.477 mm/d, RSR=0.256–0.257). RuleFit13 emerged as the optimal configuration for this group based on the parsimony principle, achieving equivalent accuracy to more complex models while reducing data requirements. The test performance degradation at RuleFit15 ($R^2=0.891$, RMSE=0.614 mm/d, RSR=0.331) and particularly at RuleFit16 ($R^2=0.829$, RMSE=0.767 mm/d, RSR=0.413) was more pronounced at Sunaynah compared to Bahla.

The scatter plots (shown in Fig. 6) show the observed versus predicted daily ETo values for the solar and radiation group models at Sunaynah station. RuleFit11 through RuleFit14 exhibit visually indistinguishable scatter patterns with remarkably consistent clustering

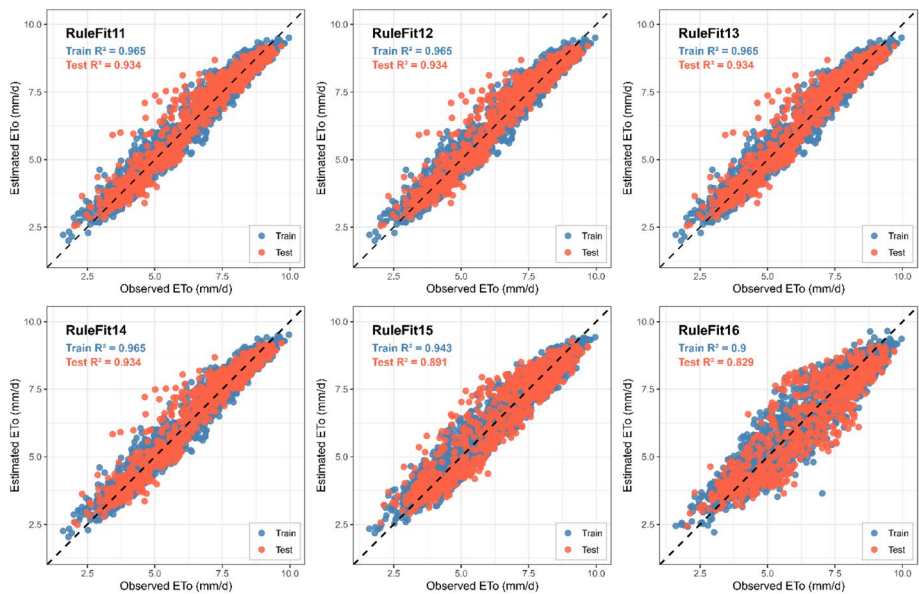


Fig. 6 Comparison of observed and predicted daily ETo for Solar and Radiation group RuleFit models at Sunaynah station. The 1:1 perfect prediction line is shown by the dashed black line

across the entire ETo range. The transition from RuleFit14 to RuleFit15 reveals noticeable increases in data point dispersion, accompanied by subtle tendencies toward under-prediction at higher ETo values, indicating the importance of D_s and d_r for maintaining seasonal prediction accuracy. RuleFit16 exhibits substantially increased scatter with several pronounced outliers departing from the 1:1 line and a wider prediction envelope across all ETo ranges.

3.3 Derived thermodynamic (temperature-vapor)-based ETo prediction

Table 9 demonstrates the performance metrics of the derived thermodynamic group models at the Bahla station. During the training phase, the comprehensive RuleFit17 model (incorporating e_s - e_a , $e(T_{max})$, $e(T_{min})$, $e(T_{dew})$, Δ) achieved excellent performance metrics ($R^2=0.967$, $RMSE=0.304$ mm/d, $MAE=0.248$ mm/d, $RSR=0.181$), closely fol-

Table 9 Performance evaluation metrics for derived thermodynamic (temperature-vapor) group RuleFit models at Bahla station

Train	RuleFit17	RuleFit18	RuleFit19	RuleFit20	RuleFit21
MAE	0.248	0.252	0.252	0.264	0.337
RMSE	0.304	0.309	0.309	0.323	0.428
RSR	0.181	0.184	0.184	0.192	0.255
R^2	0.967	0.966	0.966	0.963	0.935
Test	RuleFit17	RuleFit18	RuleFit19	RuleFit20	RuleFit21
MAE	0.307	0.310	0.310	0.317	0.419
RMSE	0.380	0.384	0.384	0.390	0.538
RSR	0.235	0.238	0.238	0.242	0.333
R^2	0.945	0.943	0.943	0.942	0.889

lowed by RuleFit18 and RuleFit19 which showed nearly equivalent accuracy ($R^2=0.966$, $RSR=0.184$). This near-equivalence suggests that Delta provides minimal additional information when es-ea and saturation vapor pressures at different temperature points are already included in the model. The removal of $e(T_{dew})$ in RuleFit19 similarly had a negligible impact, indicating redundancy among the saturation vapor pressure terms. Performance degradation became evident at RuleFit20, which retained only es-ea and $e(T_{max})$ ($R^2=0.963$, $RSR=0.192$), while the single-variable RuleFit21 model relying exclusively on es-ea showed a substantial decline ($R^2=0.935$, $RMSE=0.428$ mm/d, $RSR=0.255$), though still maintaining acceptable accuracy levels for preliminary ETo estimation. In the test set, evaluation confirmed RuleFit17 as the optimal configuration within this group, achieving the best generalization performance with $R^2=0.945$, $RMSE=0.38$ mm/d, $MAE=0.307$ mm/d, and $RSR=0.235$. RuleFit18 and RuleFit19 demonstrated marginally lower test accuracy ($R^2=0.943$, $RSR=0.238$), while RuleFit20 showed further degradation ($R^2=0.942$, $RSR=0.242$). The test performance of RuleFit21 ($R^2=0.889$, $RMSE=0.538$ mm/d, $RSR=0.333$) remained within acceptable limits but was considerably lower than that of the more comprehensive models.

Figure 7 presents scatter plots comparing observed and predicted daily ETo values for the derived thermodynamic group models at Bahla station. RuleFit17 through RuleFit19 exhibit visually similar scatter patterns with consistent data point concentration across the full ETo range, confirming their nearly equivalent performance metrics and demonstrating that Delta and $e(T_{dew})$ contribute minimally to prediction accuracy when other thermodynamic variables are present. The optimal RuleFit17 model shows excellent agreement between observed and predicted values with minimal systematic bias. RuleFit21 (relying exclusively on (es-ea)) exhibits noticeably increased scatter with wider prediction envelopes and several outliers departing from the 1:1 line, demonstrating the limitations of single-variable ther-

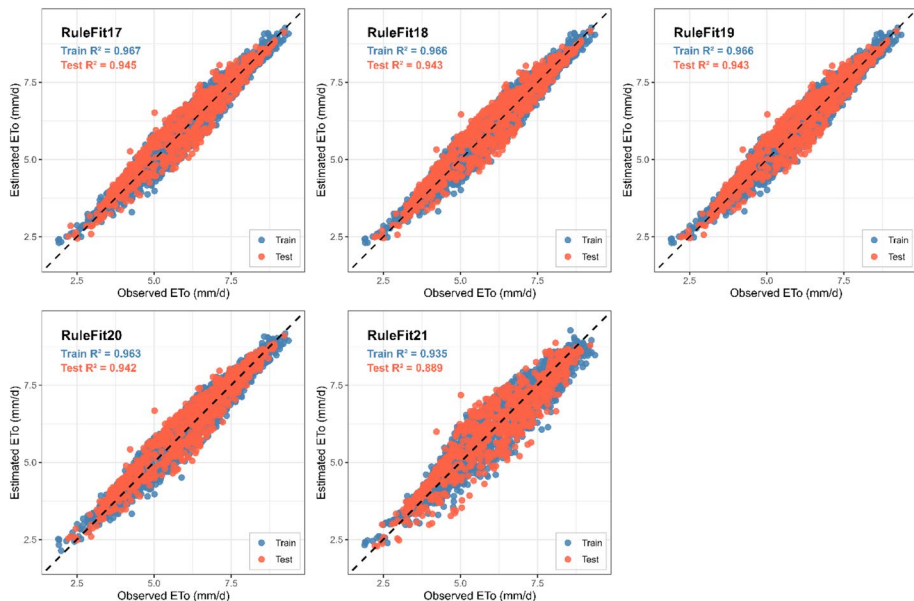


Fig. 7 Comparison of observed and predicted daily ETo for derived thermodynamic (temperature-vapor) group RuleFit models at Bahla station. The 1:1 perfect prediction line is shown by the dashed black line

modynamic approaches despite es-ea being a fundamental driver of evaporation. Nonetheless, even RuleFit21 maintains a reasonable predictive structure without severe systematic over- or under-prediction patterns across the entire range, confirming that vapor pressure deficit alone captures substantial information about atmospheric evaporative demand in arid climates.

Table 10 lists the performance metrics of the derived thermodynamic group models at Sunaynah station. During the training phase, RuleFit17 model (incorporating es-ea, $e(T_{\max})$, $e(T_{\min})$, $e(T_{\text{dew}})$, and Delta) achieved superior performance metrics ($R^2=0.969$, RMSE=0.343 mm/d, MAE=0.28 mm/d, RSR=0.177), representing the best training accuracy within the derived thermodynamic group at Sunaynah. RuleFit18 and RuleFit19 demonstrated virtually identical performance ($R^2=0.967$, RMSE=0.35 mm/d, RSR=0.18), confirming the redundancy of these variables when vapor pressure deficit and saturation vapor pressures at $T_{\max}$ and $T_{\min}$ are already included. Performance degradation became more pronounced at RuleFit20, which retained only vapor pressure deficit and $e(T_{\max})$ ($R^2=0.957$, RSR=0.208), while the single-variable RuleFit21 model relying exclusively on vapor pressure deficit showed a substantial decline ($R^2=0.946$, RMSE=0.45 mm/d, RSR=0.232). Results of the test section confirmed RuleFit17 as the optimal configuration within the thermodynamic group, achieving excellent generalization performance with $R^2=0.947$, RMSE=0.429 mm/d, MAE=0.349 mm/d, and RSR=0.231. RuleFit18 and RuleFit19 exhibited marginally lower test accuracy ($R^2=0.945$, RSR=0.233) in the test phase. RuleFit20 showed noticeable degradation ($R^2=0.93$, RSR=0.264), while RuleFit21 demonstrated further reduced but still acceptable performance ($R^2=0.921$, RMSE=0.521 mm/d, RSR=0.28) in the test phase.

Figure 8 demonstrates the scatter plots of observed versus predicted daily ETo values for the derived thermodynamic group models at Sunaynah station. RuleFit17 through RuleFit19 exhibit visually indistinguishable scatter patterns with highly consistent data point concentration across the entire ETo range, confirming their nearly equivalent performance metrics and demonstrating superior clustering compared to the radiation group models at this station. The optimal RuleFit17 model shows excellent agreement between observed and predicted values with minimal systematic bias. The transition from RuleFit19 to RuleFit20 reveals subtle but discernible increases in dispersion. RuleFit21 (relying solely on es-ea) exhibits slightly increased scatter with wider prediction envelopes from the 1:1 line, though the overall linear structure remains substantially superior to the single-variable radiation model (RuleFit16).

Table 10 Performance evaluation metrics for derived thermodynamic (temperature-vapor) group RuleFit models at Sunaynah station

Train	RuleFit17	RuleFit18	RuleFit19	RuleFit20	RuleFit21
MAE	0.280	0.286	0.286	0.331	0.363
RMSE	0.343	0.350	0.350	0.403	0.450
RSR	0.177	0.180	0.180	0.208	0.232
R^2	0.969	0.967	0.967	0.957	0.946
Test	RuleFit17	RuleFit18	RuleFit19	RuleFit20	RuleFit21
MAE	0.349	0.354	0.354	0.413	0.429
RMSE	0.429	0.433	0.433	0.491	0.521
RSR	0.231	0.233	0.233	0.264	0.280
R^2	0.947	0.945	0.945	0.930	0.921

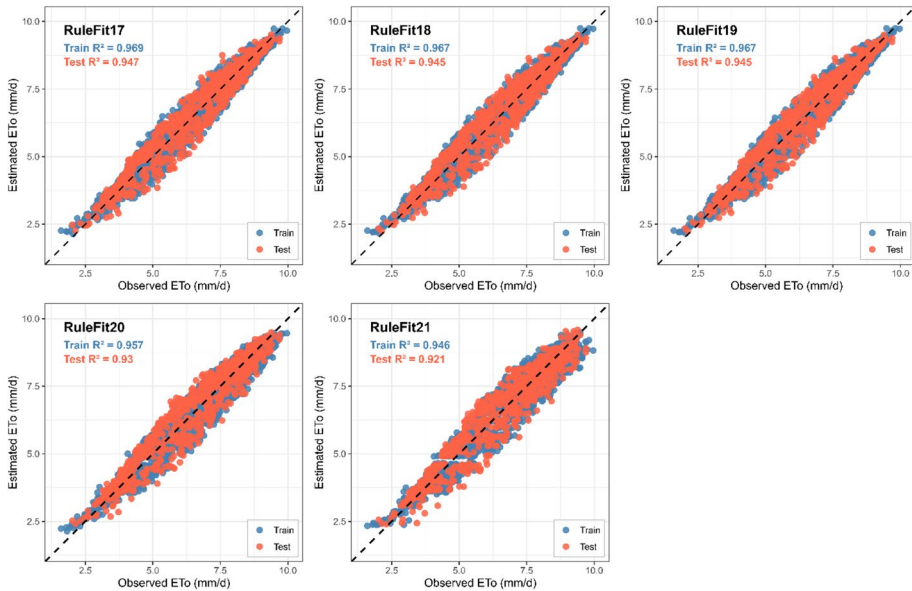


Fig. 8 Comparison of observed and predicted daily ETo for derived thermodynamic (temperature-vapor) group RuleFit models at Sunaynah station. The 1:1 perfect prediction line is shown by the dashed black line

3.4 Ensemble model development and performance comparison with the benchmark empirical method

The systematic evaluation of all twenty-one RuleFit model configurations across three variable groups enabled the identification of optimal representatives for ensemble integration, with station-specific selections reflecting local climatic characteristics and variable importance patterns. At Bahla station, the selected candidate models comprised RuleFit3 from the meteorological group (8 variables: Tmax, Tmin, Tmean, Tdew, RHmax, RHmin, RHmean, SLPmax), RuleFit14 from the solar and radiation group (3 variables: Rs, Ra, dr), and RuleFit17 from the derived thermodynamic group (5 variables: es-ea, e(Tmax), e(Tmin), e(Tdew), Delta). At Sunaynah station, the optimal configurations differed slightly, with RuleFit2 selected from the meteorological group (9 variables: Tmax, Tmin, Tmean, Tdew, RHmax, RHmin, RHmean, SLPmax, SLPmin), RuleFit13 from the solar and radiation group (4 variables: Rs, Ra, dr, Ds), and RuleFit17 from the derived thermodynamic group (5 variables, identical to Bahla). The predictions generated by these three candidate models at each station were subsequently used as meta-features to train two ensemble variants (listed in Table 11): RuleFit-SVM, which employs a standard SVM with default hyperparameters, and RuleFit-SVM-GenSA, which uses hyperparameters optimized via the GenSA technique.

Table 12 presents the results of the ensemble modeling framework at Bahla station. During the training phase, RuleFit-SVM-GenSA achieved outstanding performance with $R^2=0.992$, RMSE=0.149 mm/d, MAE=0.113 mm/d, and RSR=0.089, representing the best training performance among all evaluated approaches. The RuleFit-SVM ensemble also demonstrated excellent results ($R^2=0.99$, RMSE=0.171 mm/d, RSR=0.102), confirming

Table 11 Summary of optimal RuleFit models selected from each variable group and their integration into ensemble modeling frameworks at Bahla and Sunaynah stations. The candidate models represent the best-performing configurations within each group, as determined by test-phase evaluation metrics

Station	Group	Candidate	Model	
Bahla	Meteorological (air and moisture)	RuleFit3	RuleFit-SVM	RuleFit-SVM-GenSA
	Solar and radiation	RuleFit14		
	Derived thermodynamic (temperature-vapor)	RuleFit17		
Sunaynah	Meteorological (air and moisture)	RuleFit2	RuleFit-SVM	RuleFit-SVM-GenSA
	Solar and radiation	RuleFit13		
	Derived thermodynamic (temperature-vapor)	RuleFit17		

Table 12 Performance comparison of ensemble models (RuleFit-SVM and RuleFit-SVM-GenSA) against H-S empirical method at Bahla station

Train	RuleFit-SVM	RuleFit-SVM-GenSA	H-S Method
MAE	0.135	0.113	0.312
RMSE	0.171	0.149	0.383
RSR	0.102	0.089	0.228
R ²	0.990	0.992	0.948
Test	RuleFit-SVM	RuleFit-SVM-GenSA	H-S Method
MAE	0.181	0.154	0.336
RMSE	0.248	0.216	0.420
RSR	0.154	0.134	0.260
R ²	0.976	0.982	0.932

the effectiveness of the meta-learning integration strategy. The H-S method achieved intermediate accuracy ($R^2=0.948$, $RMSE=0.383$ mm/d, $RSR=0.228$). The test set evaluation confirmed the robust generalization capability of the ensemble models, with RuleFit-SVM-GenSA achieving the best independent validation performance ($R^2=0.982$, $RMSE=0.216$ mm/d, $MAE=0.154$ mm/d, $RSR=0.134$). The GenSA optimization provided improvements over the standard ensemble, reducing test RMSE from 0.248 to 0.216 mm/d and RSR from 0.154 to 0.134. The superior performance of RuleFit-SVM-GenSA validates the novel ensemble strategy, demonstrating a synergistic integration of specialized RuleFit models from different variable groups, combined with optimized SVM via the GenSA technique.

Figure 9 presents scatter plots of observed versus predicted daily ETo values for ensemble ML models and empirical benchmark methods at the Bahla station. The RuleFit-SVM-GenSA panel exhibits exceptionally tight agreement with the 1:1 perfect-prediction line, with minimal dispersion throughout the observed ETo values, demonstrating homoscedastic residual patterns and the absence of systematic bias even at extreme values. The RuleFit-SVM panel shows similar strong performance, with slightly higher dispersion than the GenSA-optimized variant. The H-S method shows intermediate scatter, with reasonable overall structure but noticeably wider prediction envelopes than the ML ensembles.

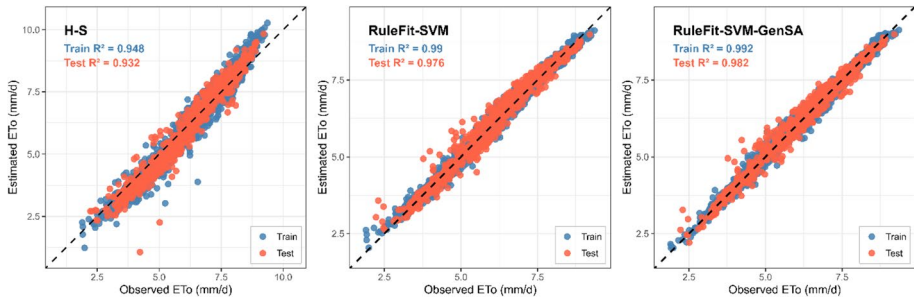


Fig. 9 Comparison of observed and predicted daily ETo for ensemble models and benchmark empirical method at Bahla station. The 1:1 perfect prediction line is shown by the dashed black line

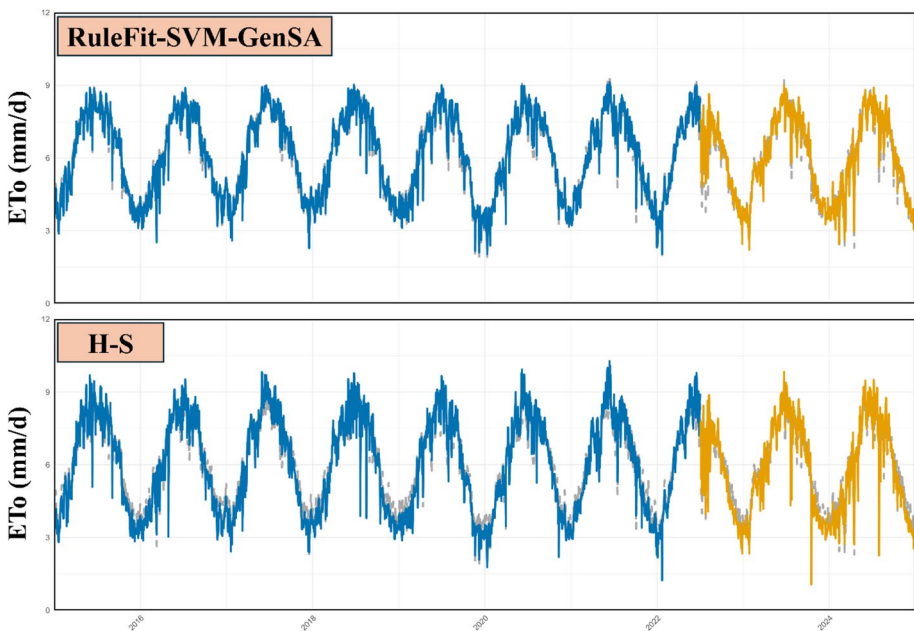


Fig. 10 Time series of observed and predicted daily ETo using the proposed RuleFit-SVM-GenSA model and H-S method at Bahla station from January 2015 to December 2024. The gray color represents observed ETo, the blue color represents predicted ETo during the training phase, and the orange color represents predicted ETo during the testing phase

Figure 10 shows the time series plots of observed and predicted daily ETo using the RuleFit-SVM-GenSA and H-S methods at the Bahla station. The RuleFit-SVM-GenSA predictions track the observed values with remarkable fidelity throughout both training and test periods, accurately capturing seasonal trends, inter-annual variations, and high-frequency fluctuations without phase lags. The time-series visualization confirms that the ensemble ML framework achieves superior statistical metrics and maintains temporal consistency and physically realistic behavior across multiple annual cycles.

Table 13 Performance comparison of ensemble models (RuleFit-SVM and RuleFit-SVM-GenSA) against H-S empirical method at Sunaynah station

Train	RuleFit-SVM	RuleFit-SVM-GenSA	H-S Method
MAE	0.144	0.132	0.308
RMSE	0.182	0.166	0.386
RSR	0.094	0.085	0.199
R^2	0.991	0.993	0.961
Test	RuleFit-SVM	RuleFit-SVM-GenSA	H-S Method
MAE	0.160	0.144	0.311
RMSE	0.216	0.190	0.379
RSR	0.116	0.102	0.204
R^2	0.986	0.989	0.958

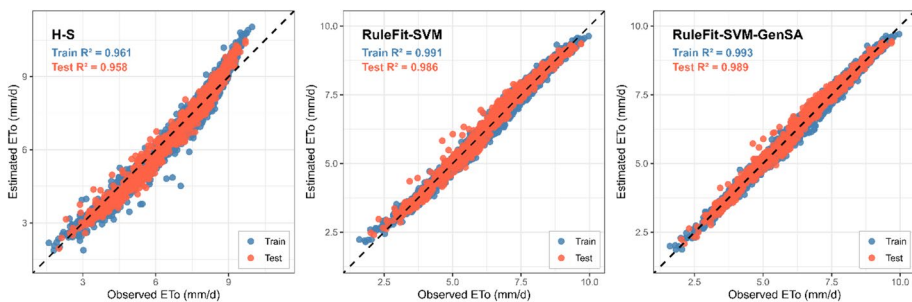


Fig. 11 Comparison of observed and predicted daily ETo for ensemble models and benchmark empirical method at Sunaynah station. The 1:1 perfect prediction line is shown by the dashed black line

The ensemble modeling framework at Sunaynah station achieved exceptional predictive performance (see Table 13), surpassing results at Bahla station and substantially outperforming the benchmark H-S method across all evaluation metrics. During the training phase, RuleFit-SVM-GenSA demonstrated outstanding accuracy with $R^2=0.993$, $RMSE=0.166$ mm/d, $MAE=0.132$ mm/d, and $RSR=0.085$. The RuleFit-SVM ensemble also achieved excellent results ($R^2=0.991$, $RMSE=0.182$ mm/d, $RSR=0.094$), confirming the effectiveness of the meta-learning integration strategy. The H-S method achieved notably better test-phase results at Sunaynah than at Bahla ($R^2=0.958$, $RMSE=0.379$ mm/d, $RSR=0.204$). The test set evaluation confirmed the superior generalization capability of the ensemble models, with RuleFit-SVM-GenSA achieving the best independent validation performance ($R^2=0.989$, $RMSE=0.19$ mm/d, $MAE=0.144$ mm/d, $RSR=0.102$). The GenSA optimization provided substantial improvements over the standard ensemble at Sunaynah, reducing test RMSE from 0.216 to 0.19 mm/d and RSR from 0.116 to 0.102. The H-S's improved performance at Sunaynah (test $R^2=0.958$) compared to Bahla (test $R^2=0.932$) suggests that the temperature range may be a more reliable predictor of ETo at this location.

Figure 11 shows scatter plots of observed versus predicted daily ETo values from ensemble ML models and empirical benchmark methods at the Sunaynah station. The RuleFit-SVM-GenSA panel exhibits remarkably tight agreement with the 1:1 perfect-prediction line, with minimal dispersion throughout the observed ETo values, showing even tighter concentration than the equivalent Bahla results and demonstrating near-perfect homoscedastic residual patterns without systematic bias. The RuleFit-SVM panel shows similarly excellent performance, with only marginally higher dispersion than the GenSA-optimized

variant. The H-S method demonstrates notably improved scatter characteristics at Sunaynah compared to Bahla, maintaining reasonable structure around the 1:1 line with moderate dispersion and reduced systematic bias.

Figure 12 shows time series plots of observed and predicted daily ETo values from the RuleFit-SVM-GenSA and H-S methods at the Sunaynah station. The RuleFit-SVM-GenSA predictions track the observed values with remarkable precision throughout both training and test periods, accurately capturing seasonal amplitudes, inter-annual variations, transitional spring and autumn dynamics, and short-term meteorological events without perceptible phase lags or magnitude biases. The seamless transition across the training-test boundary demonstrates robust model generalization with no degradation in tracking fidelity during the independent test period, confirming that the ensemble framework has learned generalizable physical relationships rather than memorized training-specific patterns. In contrast to the RuleFit-SVM-GenSA predictions, empirical methods would exhibit systematic temporal patterns, and H-S would demonstrate seasonal bias.

Four configurations were evaluated at both stations: (i) the best-performing individual RuleFit model; (ii) an ensemble of the three shortlisted RuleFit models from the meteorological, radiative, and thermodynamic groups; (iii) the RuleFit-SVM ensemble with default SVM hyperparameters; and (iv) the complete RuleFit-SVM-GenSA framework (representing the fully integrated architecture). The progression of test-set accuracy across these four configurations demonstrates that each module contributes incrementally to the overall performance. The transition from the best single RuleFit to the ensemble yields a modest but consistent reduction in error metrics (e.g., RMSE), reflecting the benefit of combin-

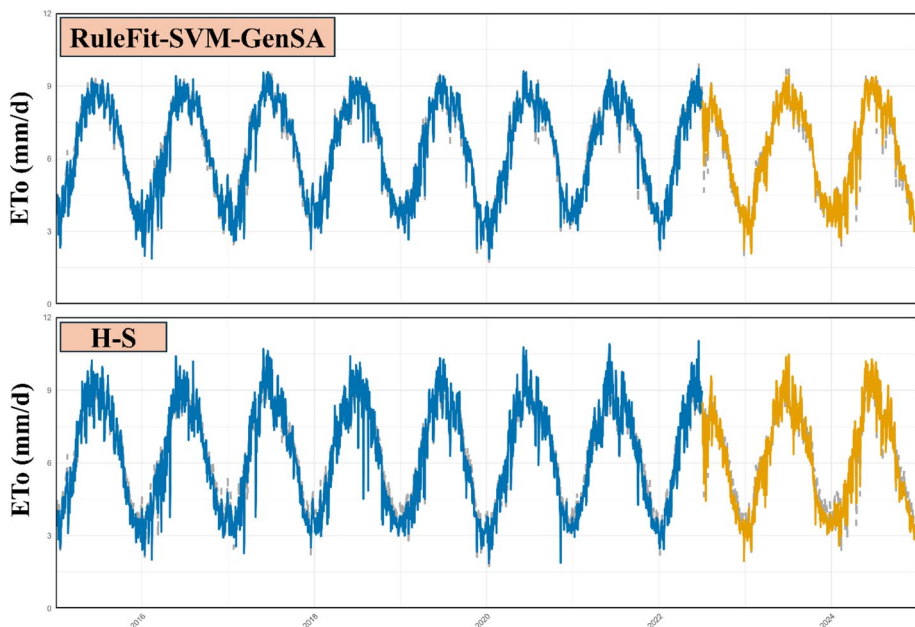


Fig. 12 Time series of observed and predicted daily ETo via the proposed RuleFit-SVM-GenSA model and H-S method at Sunaynah station from January 2015 to December 2024. The gray color represents observed ETo, the blue color represents predicted ETo during the training phase, and the orange color represents predicted ETo during the testing phase

ing complementary process-specific information across variable groups. The SVM acts as a meta-learner that intelligently combines the predictions from the three RuleFit models. The GenSA algorithm then performs the SVM hyperparameter tuning, ensuring that the proposed framework operates at its highest possible accuracy. To further strengthen the benchmark comparison, a standalone SVM with an RBF kernel was additionally trained and evaluated using the optimal RuleFit configurations identified within each variable group and applied independently at both stations. The input configurations and performance metrics are reported in the Appendix (Tables 14, 15 and 16). Across three variable groups and both stations, the standalone SVM consistently achieved lower accuracy than the corresponding RuleFit configurations and the RuleFit-SVM-GenSA ensemble during the test section.

4 Discussion

4.1 SHAP-based feature importance and model interpretability analysis

The SHAP feature importance analysis at Bahla station (Fig. 13) reveals distinct hierarchies of variable influence within each modeling group while confirming physically consistent driv-

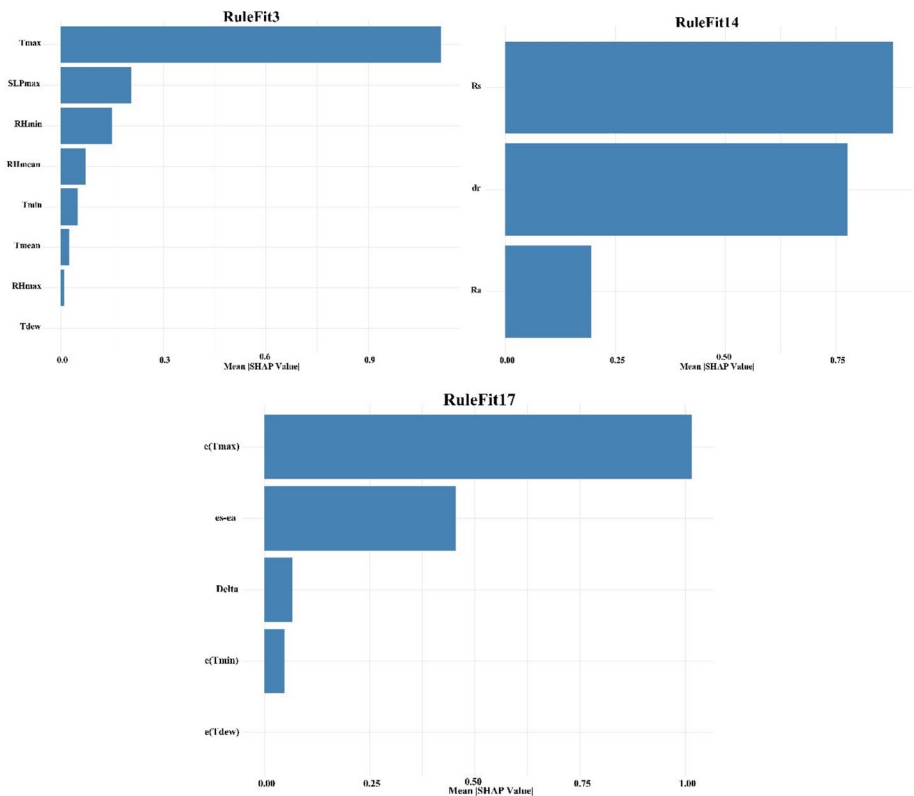


Fig. 13 SHAP feature importance analysis for optimal RuleFit models at Bahla station across three variable groups: (a) RuleFit3 for meteorological (air and moisture) group, (b) RuleFit14 for solar and radiation group, (c) RuleFit17 for derived thermodynamic (temperature-vapor) group

ers of ETo dynamics in arid environments. Within the meteorological group (RuleFit3 as the best model), Tmax is the dominant predictor, and SLPmax and RHmin rank second and third, respectively. Within the solar and radiation group (RuleFit14 as the best model), Rs dominates the importance ranking, as expected given its direct representation of the energy available for latent heat exchange, while dr and Ra show substantially lower but non-negligible importance, primarily contributing seasonal baseline adjustments and astronomical corrections. The derived thermodynamic group (RuleFit17 as the best model in the derived thermodynamic group) reveals $e(T_{max})$ as the most critical predictor, which captures the upper bound of daily vapor pressure dynamics, followed by $e_s - e_a$ as the second-ranked important predictor, directly quantifying atmospheric evaporative demand through the vapor pressure gradient between saturated and actual conditions. Because the derived thermodynamic variables are deterministic transformations of raw temperature/humidity observations, SHAP importance rankings among these correlated raw and derived variables should be interpreted with caution, as attributions can be sensitive to feature correlation and the choice of background distribution.

Figure 14 demonstrates the SHAP feature importance analysis at Sunaynah station. Within the meteorological group (RuleFit2 as the best-performing model), Tmax dominates the importance hierarchy, reflecting its fundamental control over vapor-pressure gradients and energy availability for evaporation. The retention of both SLPmax and SLPmin (as the

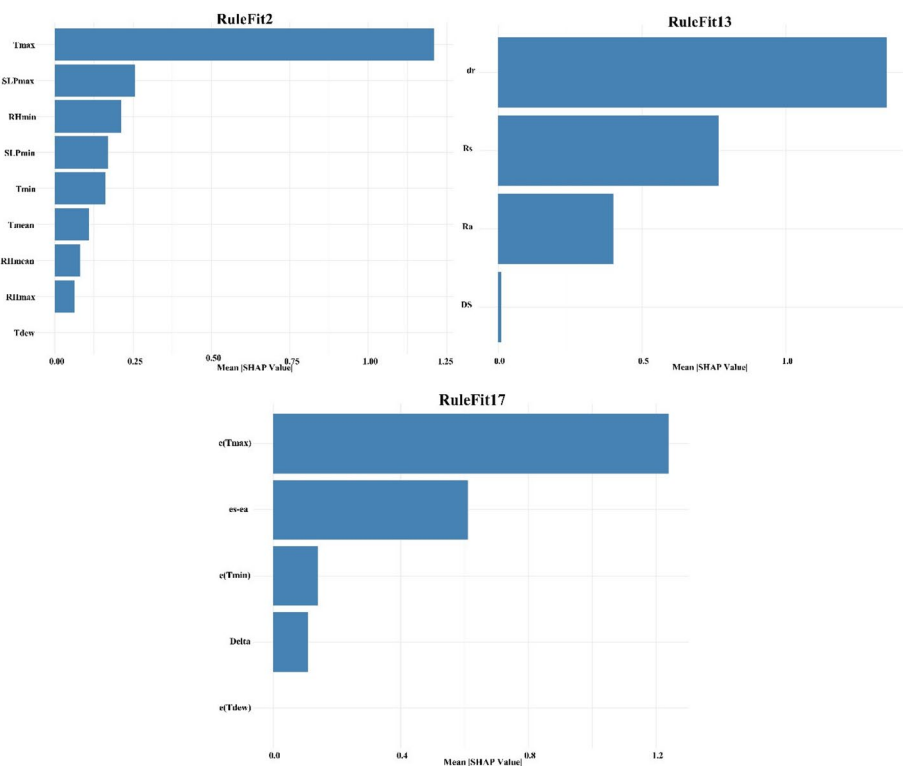


Fig. 14 SHAP feature importance analysis for optimal RuleFit models at Sunaynah station across three variable groups: (a) RuleFit2 for meteorological (air and moisture) group, (b) RuleFit13 for solar and radiation group, (c) RuleFit17 for derived thermodynamic (temperature-vapor) group

second and fourth most important variables) in the optimal Sunaynah configuration, despite their relatively lower SHAP importance values (compared to $T_{\max}$), suggests that these pressure variables contribute subtle but meaningful information for capturing synoptic-scale atmospheric circulation influences on ETo at this location. Within the solar and radiation group (RuleFit13 as the best model), dr dominates as the predictor with substantially higher SHAP importance than all other radiation variables. Also, the derived thermodynamic group (RuleFit17, the best model in the derived thermodynamic group) shows that $e(T_{\max})$ and $es-ea$ are the first and second most important variables, respectively.

4.2 SHAP dependence analysis: feature-response relationships and interaction effects

Figure 15 shows the SHAP dependence plots for the optimal RuleFit3 model at Bahla station. $T_{\max}$ exhibits a strongly positive non-linear relationship with SHAP values, showing that increasing $T_{\max}$ consistently drives higher predicted ETo across its entire range. $RH_{\min}$ and RH_{mean} show predominantly negative SHAP contributions that become more strongly negative as $RH_{\min}$ and RH_{mean} increase, confirming the physically expected inverse relationship where high humidity suppresses ETo by reducing atmospheric drying power. These SHAP dependence relationships closely align with established ETo physics

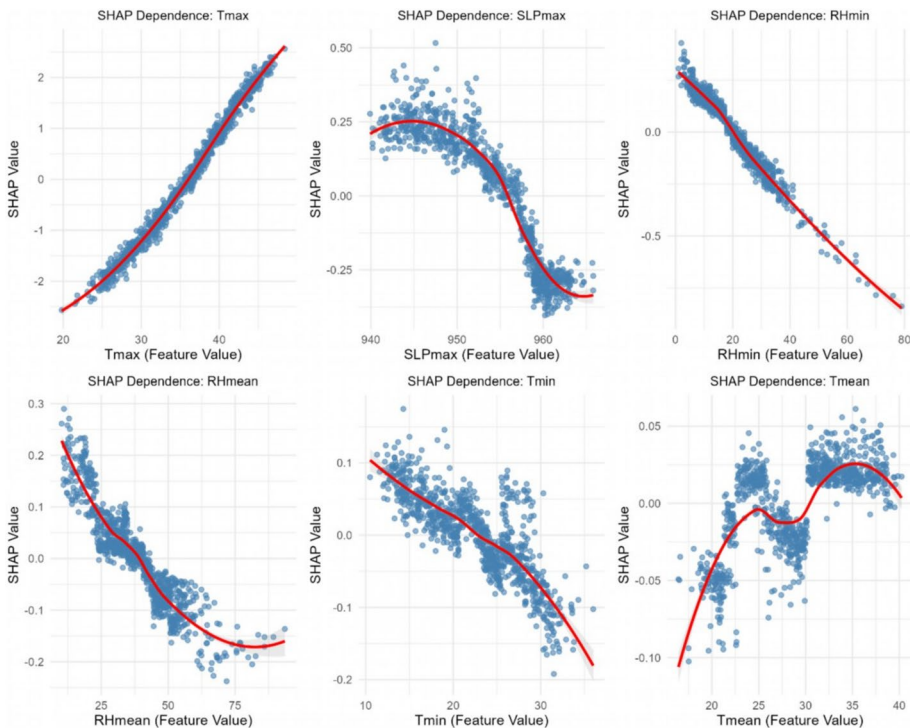


Fig. 15 SHAP dependence plots for the top six features in the RuleFit3 model at Bahla station, including $T_{\max}$, $SLP_{\max}$, $RH_{\min}$, RH_{mean} , $T_{\min}$, and T_{mean} . The spread and trends reveal how each feature influences model outputs across its observed range

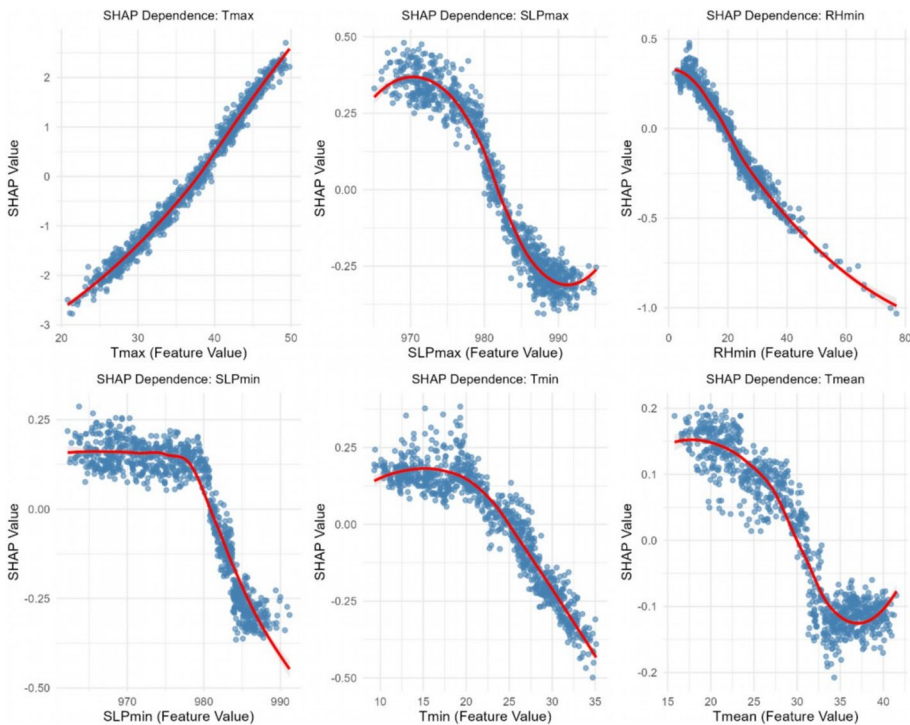


Fig. 16 SHAP dependence plots for the top six features in RuleFit2 model at Sunaynah station, including Tmax, SLPmax, RHmin, SLPmin, Tmin, and Tmean. The spread and trends reveal how each feature influences model outputs across its observed range

from the PM equation: ETo increases with temperature and solar radiation (energy supply) and decreases with relative humidity. Figure 16 shows the SHAP dependence plots for the optimal RuleFit2 model at Sunaynah station. Tmax demonstrates a strongly positive monotonic relationship with SHAP values across its observed range, with a remarkably consistent gradient indicating stable temperature- ETo coupling. RHmin shows predominantly negative SHAP values that intensify as humidity increases, confirming the expected inverse relationship: higher atmospheric moisture content reduces evaporative demand. The complex non-linear pattern of SLPmax suggests that atmospheric pressure influences ETo through multiple competing mechanisms: moderate pressures may coincide with favorable synoptic conditions for evaporation, while very high pressures may indicate subsiding air masses with reduced vertical mixing, suppressing evaporative flux despite clear skies.

4.3 Limitations of the current study and future research directions

Although the RuleFit-SVM-GenSA framework for ETo estimation performed well for ETo estimation, several limitations warrant acknowledgment and suggest directions for model refinement and extended validation. First, the framework was developed and validated using data from two meteorological stations in arid regions of Oman, limiting its conclusions regarding geographic transferability to other arid and semi-arid regions with different

elevation ranges, proximity to water bodies, aerosol loading characteristics, or atmospheric circulation patterns. Second, the ensemble framework requires comprehensive meteorological data from all three variable groups for optimal performance, potentially limiting its applicability in data-scarce regions where incomplete monitoring networks or sensor failures compromise data availability. Although the progressive variable reduction experiments provide guidance for degraded-data scenarios, operational implementation under sustained data gaps remains untested. Third, the computational requirements of the RuleFit-SVM-GenSA framework (particularly during the GenSA hyperparameter optimization phase) exceed those of traditional empirical methods, potentially constraining real-time operational deployment in resource-limited settings; however, this limitation is substantially reduced once the models are trained and deployed for routine prediction.

Arid environments can differ substantially in elevation, latitude, aerosol concentration, proximity to oceanic or continental moisture sources, synoptic circulation regimes, and the seasonal dynamics of temperature and humidity extremes. These differences can shift the relative importance of the meteorological, radiative, and thermodynamic drivers of ETo and, consequently, alter the optimal variable subsets and rule structures identified. The structural components of the framework (the physically grouped base learners, the process-level meta-learning strategy, and the dual-level interpretability design) are transferable in principle; however, the specific rule coefficients, station-specific optimal configurations, and GenSA-tuned SVM hyperparameters should be regarded as locally calibrated quantities rather than universal parameters. To illustrate the expected behavior under contrasting arid conditions, three representative cases can be considered. In hyper-arid, low-latitude coastal deserts, intermittent maritime moisture intrusions and elevated aerosol loading are likely to increase the relative importance of humidity-related variables, such as RH_{min} and $e(T_{dew})$. In high-elevation continental arid regions, thinner atmospheres and stronger diurnal temperature amplitudes are expected to elevate the role of the radiative and thermodynamic groups, particularly R_s , R_a , $e(T_{max})$, and $es-ea$. In semi-arid transitional zones, stronger non-linear coupling among temperature, humidity, and radiation is expected to make the SVM meta-learner especially beneficial relative to single-process or linear-combination approaches. These considerations confirm that the framework's architecture remains portable across diverse arid contexts, while highlighting that re-tuning of the model hyperparameters is necessary for each new climatic setting.

Future research should prioritize several key directions to enhance the framework's generalizability, operational utility, and scientific understanding of ETo dynamics in arid environments. First, expanding the geographic scope to include diverse arid and semi-arid stations across multiple continents, with elevation gradients, latitude bands, and proximity to moisture sources, would enable a robust assessment of model transferability and the identification of location-specific calibration requirements. Second, integrating satellite-based remote sensing products (including land surface temperature, vegetation indices, atmospheric water vapor retrievals, and downwelling solar radiation estimates) could facilitate spatial scaling from point-based station predictions to continuous ETo mapping across agricultural regions, supporting landscape-scale water resource planning and spatially distributed hydrological modeling. Third, future work should extend the framework with a fourth variable group representing aerodynamic transport, using wind speed at stations equipped with reliable anemometry. This would complete the correspondence between the group structure and the terms of the FAO-56 equation, in which aerodynamic transport is a term the present configu-

ration does not represent explicitly. Fourth, a comparative evaluation against additional ML architectures, including gradient boosting variants and alternative ensemble strategies (e.g., stacking), would clarify the capability of RuleFit-SVM-GenSA across various scenarios.

5 Conclusion

This study successfully developed and validated a RuleFit-based ensemble framework with integrated explainability for ETo estimation in arid environments, evaluated at Bahla and Sunaynah stations in Oman. The comprehensive evaluation of twenty-one RuleFit model configurations across three variable groups provided the basis for developing the SVM-GenSA-based ensemble model. At Bahla station, RuleFit3 (eight meteorological variables), RuleFit14 (three radiation variables), and RuleFit17 (five thermodynamic variables) emerged as optimal configurations, achieving test set RSR values of 0.209, 0.289, and 0.235, respectively. In contrast, at Sunaynah station, RuleFit2 (nine meteorological variables), RuleFit13 (four radiation variables), and RuleFit17 achieved superior performance, with test RSR values of 0.191, 0.256, and 0.231, respectively, confirming station-specific differences in predictor importance. The hierarchical ensemble strategy, which integrates predictions from these optimal models through GenSA-optimized SVM (RuleFit-SVM-GenSA), achieved exceptional predictive accuracy, with test $R^2=0.982$ (RMSE=0.216 mm/d, RSR=0.134) at Bahla and $R^2=0.989$ (RMSE=0.19 mm/d, RSR=0.102) at Sunaynah. The pronounced performance superiority of the RuleFit-SVM-GenSA framework over benchmark approaches validates the effectiveness of advanced ML techniques for ETo estimation in data-rich environments, while the progressive variable-reduction experiments provide practical guidance for operational implementations under varying data-availability constraints.

The integration of SHAP analysis addressed the black-box limitation of complex ensembles by quantifying feature contributions, attributing predictions to individual inputs, and checking that to individual inputs and to check that the resulting dependencies agree with established ETo physics. SHAP feature-importance analysis revealed that Tmax, Rs, dr, e(Tmax), and es-ea accounted for the greatest contributions to ETo predictions at both stations. Tmax emerged as the dominant predictor within the meteorological group at both stations, while dr provided essential seasonal adjustments, and derived thermodynamic variables, particularly es-ea and e(Tmax), captured the coupled temperature-moisture dynamics governing latent heat exchange. SHAP dependence analyses revealed physically consistent non-linear relationships, with Tmax exhibiting strong positive effects on ETo predictions across the observed range, and relative humidity showing the expected negative contributions. Overall, the RuleFit-SVM-GenSA framework provides water resource managers, irrigation engineers, and agricultural planners with a highly accurate, physically interpretable tool for daily ETo estimation. The station-specific optimal model selections (RuleFit3 vs. RuleFit2 for meteorological models, RuleFit14 vs. RuleFit13 for radiation models) highlight the importance of local calibration and variable-importance assessment, offering evidence-based justification for investing in comprehensive meteorological monitoring networks at agricultural and hydrological stations rather than relying on simplified data collection protocols. The framework's demonstrated ability to achieve near-perfect ETo estimation accuracy ($R^2>0.98$) while maintaining interpretability through SHAP analysis establishes a new benchmark for ML applications in agricultural water management, effectively bridging the gap between predictive performance and physical understanding.

Appendix 1

Table 14 Input configurations used for the standalone SVM benchmark at Bahla and Sunaynah stations. The inputs correspond to the optimal RuleFit configurations identified within each variable group

Station	Group	Input variables	Model ID
Bahla	Meteorological (air and moisture)	Tmax, Tmin, Tmean, Tdew, RHmax, RHmin, RHmean, SLPmax	SVM-Met (corresponding RuleFit3)
	Solar and radiation	Rs, Ra, dr	SVM-Rad (corresponding RuleFit14)
	Derived thermodynamic (temperature–vapor)	es-ea, e(Tmax), e(Tmin), e(Tdew), Delta	SVM-Thermo (corresponding RuleFit17)
Sunaynah	Meteorological (air and moisture)	Tmax, Tmin, Tmean, Tdew, RHmax, RHmin, RHmean, SLPmax, SLPmin	SVM-Met (corresponding RuleFit2)
	Solar and radiation	Rs, Ra, dr, Ds	SVM-Rad (corresponding RuleFit13)
	Derived thermodynamic (temperature–vapor)	es-ea, e(Tmax), e(Tmin), e(Tdew), Delta	SVM-Thermo (corresponding RuleFit17)

Table 15 Performance metrics of the standalone SVM benchmark at Bahla station across the three variable groups

Phase	Metrics	SVM-Met	SVM-Rad	SVM-Thermo
Train	MAE	0.243	0.292	0.268
	RMSE	0.317	0.397	0.343
	RSR	0.189	0.236	0.205
	R ²	0.964	0.944	0.958
Test	MAE	0.276	0.345	0.310
	RMSE	0.348	0.497	0.390
	RSR	0.215	0.308	0.241
	R ²	0.954	0.905	0.942

Table 16 Performance metrics of the standalone SVM benchmark at Sunaynah station across the three variable groups

Phase	Metrics	SVM-Met	SVM-Rad	SVM-Thermo
Train	MAE	0.271	0.300	0.306
	RMSE	0.358	0.402	0.394
	RSR	0.184	0.207	0.203
	R ²	0.966	0.957	0.959
Test	MAE	0.297	0.355	0.350
	RMSE	0.378	0.501	0.443
	RSR	0.204	0.270	0.239
	R ²	0.959	0.927	0.943

Acknowledgements The study is financially supported by grants #CL/SQU-IGGCAS/WRC/23/01, #IG/DVC/WRC/26/436, and IG/DVC/WRC/25/134 from Sultan Qaboos University (SQU). Support from SQU group DR/RG/17's is appreciated.

Data availability Data will be made available on reasonable request.

Declarations

Conflict of interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

References

- Allen, R. G., Pereira, L. S., Raes, D., & Smith, M. (1998). *Crop evapotranspiration: Guidelines for computing crop water requirements*. Food and Agriculture Organization of the United Nation. <https://www.fao.org/3/x0490e/x0490e00.htm>
- Amani, S., Shafizadeh-Moghadam, H., & Morid, S. (2024). Utilizing machine learning models with limited meteorological data as alternatives for the FAO-56PM model in estimating reference evapotranspiration. *Water Resources Management*, 38(6), 1921–1942. <https://doi.org/10.1007/s11269-023-03670-2>
- Bellido-Jiménez, J. A., Estévez, J., & García-Marín, A. P. (2021). New machine learning approaches to improve reference evapotranspiration estimates using intra-daily temperature-based variables in a semi-arid region of Spain. *Agricultural Water Management*, 245, Article 106558. <https://doi.org/10.1016/j.agwat.2020.106558>
- Blaney, H. F., & Criddle, W. D. (1962). *Determining consumptive use and irrigation water requirements*. U.S. Department of Agriculture.
- Bologna, G. (2021). A rule extraction technique applied to ensembles of neural networks, random forests, and gradient-boosted trees. *Algorithms*, 14(12), 339. <https://doi.org/10.3390/a14120339>
- Chartzoulakis, K., & Bertaki, M. (2015). Sustainable water management in agriculture under climate change. *Agriculture and Agricultural Science Procedia*, 4, 88–98. <https://doi.org/10.1016/j.aaspro.2015.03.011>
- Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). New York, NY, USA: Association for Computing Machinery. <https://doi.org/10.1145/2939672.2939785>
- Chia, M. Y., Huang, Y. F., & Koo, C. H. (2020). Support vector machine enhanced empirical reference evapotranspiration estimation with limited meteorological parameters. *Computers and Electronics in Agriculture*, 175, Article 105577. <https://doi.org/10.1016/j.compag.2020.105577>
- Córdova, M., Carrillo-Rojas, G., Crespo, P., Wilcox, B., & Céleri, R. (2015). Evaluation of the Penman-Monteith (FAO 56 PM) method for calculating reference evapotranspiration using limited data. *Mountain Research and Development*, 35(3), 230–239. <https://doi.org/10.1659/MRD-JOURNAL-D-14-0024.1>
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273–297. <https://doi.org/10.1007/BF00994018>
- Djaman, K., O'Neill, M., Owen, C. K., Smeal, D., Koudahe, K., West, M., et al. (2018). Crop evapotranspiration, irrigation water requirement and water productivity of maize from meteorological data under semiarid climate. *Water*, 10(4), 405. <https://doi.org/10.3390/w10040405>

- Dong, J., Xing, L., Cui, N., Guo, L., Liang, C., Zhao, L., et al. (2024). Estimating reference crop evapotranspiration using optimized empirical methods with a novel improved Grey Wolf Algorithm in four climatic regions of China. *Agricultural Water Management*, 291, Article 108620. <https://doi.org/10.1016/j.agwat.2023.108620>
- FAO. (2020). *The state of food and agriculture 2020. Overcoming water challenges in agriculture*. Food and Agriculture Organization of the United Nations.
- Fisher, J. B., Melton, F., Middleton, E., Hain, C., Anderson, M., Allen, R., et al. (2017). The future of evapotranspiration: Global requirements for ecosystem functioning, carbon and climate feedbacks, agricultural management, and water resources. *Water Resources Research*, 53(4), 2618–2626. <https://doi.org/10.1002/2016WR020175>
- Fokkema, M. (2020). Fitting prediction rule ensembles with R package pre. *Journal of Statistical Software*, 92, 1–30. <https://doi.org/10.18637/jss.v092.i12>
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232.
- Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4), 367–378. [https://doi.org/10.1016/S0167-9473\(01\)00065-2](https://doi.org/10.1016/S0167-9473(01)00065-2)
- Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2(3), 916–954. <https://doi.org/10.1214/07-AOAS148>
- Fürnkranz, J., Kliegr, T., & Paulheim, H. (2020). On cognitive preferences and the plausibility of rule-based models. *Machine Learning*, 109(4), 853–898. <https://doi.org/10.1007/s10994-019-05856-5>
- Ghiat, I., Mackey, H. R., & Al-Ansari, T. (2021). A review of evapotranspiration measurement models, techniques and methods for open and closed agricultural field applications. *Water*, 13(18), Article 2523. <https://doi.org/10.3390/w13182523>
- Hargreaves, G. H., & Samani, Z. A. (1985). Reference Crop Evapotranspiration from Temperature. *Applied engineering in agriculture*, 1(2), 96–99. <https://doi.org/10.13031/2013.2677>
- Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., et al. (2024). Interpreting black-box models: A review on explainable artificial intelligence. *Cognitive Computation*, 16(1), 45–74. <https://doi.org/10.1007/s12559-023-10179-8>
- Holub, K. (2022, October 4). xrf: eXtreme RuleFit. <https://cran.r-project.org/web/packages/xrf/index.html>. Accessed 30 November 2025
- Huang, G., Wu, L., Ma, X., Zhang, W., Fan, J., Yu, X., et al. (2019). Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions. *Journal of Hydrology*, 574, 1029–1041. <https://doi.org/10.1016/j.jhydrol.2019.04.085>
- Ippolito, M., De Caro, D., Cannarozzo, M., Provenzano, G., & Ciraolo, G. (2024). Evaluation of daily crop reference evapotranspiration and sensitivity analysis of FAO Penman-Monteith equation using ERA5-Land reanalysis database in Sicily, Italy. *Agricultural Water Management*, 295, Article 108732. <https://doi.org/10.1016/j.agwat.2024.108732>
- Katul, G. G., Oren, R., Manzoni, S., Higgins, C., & Parlange, M. B. (2012). Evapotranspiration: A process driving mass transport and energy exchange in the soil-plant-atmosphere-climate system. *Reviews of Geophysics*. <https://doi.org/10.1029/2011RG000366>
- Kiraga, S., Peters, R. T., Molaei, B., Evett, S. R., & Marek, G. (2024). Reference evapotranspiration estimation using genetic algorithm-optimized machine learning models and standardized Penman-Monteith equation in a highly advective environment. *Water*, 16(1), 12. <https://doi.org/10.3390/w16010012>
- Kozielski, M., Sikora, M., & Wawrowski, L. (2025). Towards consistency of rule-based explainer and black box model — Fusion of rule induction and XAI-based feature importance. *Knowledge-Based Systems*, 311, Article 113092. <https://doi.org/10.1016/j.knosys.2025.113092>
- Li, X., Liu, Z., Yang, Z., Meng, F., & Song, T. (2024). A high-precision interpretable framework for marine dissolved oxygen concentration inversion. *Frontiers in Marine Science*. <https://doi.org/10.3389/fmars.2024.1396277>
- Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>. Accessed 29 November 2025
- Luo, C., Li, S., Zhao, Q., Ou, Q., Huang, W., Ruan, G., et al. (2022). RuleFit-based nomogram using inflammatory indicators for predicting survival in nasopharyngeal carcinoma, a bi-center study. *Journal of Inflammation Research*, 15, 4803–4815. <https://doi.org/10.2147/JIR.S366922>
- Macha, D., Kozielski, M., Wróbel, Ł., & Sikora, M. (2022). RuleXAI—A package for rule-based explanations of machine learning model. *SoftwareX*, 20, Article 101209. <https://doi.org/10.1016/j.softx.2022.101209>
- Makkink, G. F. (1957). Testing the Penman formula by means of lysimeters. *Journal of the Institution of Water Engineers*, 11, 277–288.

- Mehdizadeh, S., Mohammadi, B., Pham, Q. B., & Duan, Z. (2021). Development of boosted machine learning models for estimating daily reference evapotranspiration and comparison with empirical approaches. *Water*, 13(24), Article 3489. <https://doi.org/10.3390/w13243489>
- Minhas, P. S., 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, Article 105832. <https://doi.org/10.1016/j.agwat.2019.105832>
- Mohammadi, B., Chen, M., Reza Nikoo, M., Cheraghalizadeh, M., Yu, Y., Zhang, H., & Yu, R. (2024). An explainable hybrid framework for estimating daily reference evapotranspiration: Combining extreme gradient boosting with Nelder-Mead method. *Journal of Hydrology*, 644, Article 132130. <https://doi.org/10.1016/j.jhydrol.2024.132130>
- Paredes, P., Pereira, L. S., Almorox, J., & Darouich, H. (2020). Reference grass evapotranspiration with reduced data sets: Parameterization of the FAO Penman-Monteith temperature approach and the Hargreaves-Samani equation using local climatic variables. *Agricultural Water Management*, 240, Article 106210. <https://doi.org/10.1016/j.agwat.2020.106210>
- Pereira, L. S., Allen, R. G., Smith, M., & Raes, D. (2015). Crop evapotranspiration estimation with FAO56: Past and future. *Agricultural Water Management*, 147, 4–20. <https://doi.org/10.1016/j.agwat.2014.07.031>
- Pereira, L. S., Paredes, P., & Jovanovic, N. (2020). Soil water balance models for determining crop water and irrigation requirements and irrigation scheduling focusing on the FAO56 method and the dual Kc approach. *Agricultural Water Management*, 241, Article 106357. <https://doi.org/10.1016/j.agwat.2020.106357>
- Priestley, C. H. B., & Taylor, R. J. (1972). On the Assessment of Surface Heat Flux and Evaporation Using Large-Scale Parameters. *Monthly Weather Review*, 100(2), 81–92. [https://doi.org/10.1175/1520-0493\(1972\)100<3C0081:OTAOSH>3E2.3.CO;2](https://doi.org/10.1175/1520-0493(1972)100<3C0081:OTAOSH>3E2.3.CO;2)
- Raghavendra, N. S., & Deka, P. C. (2014). Support vector machine applications in the field of hydrology: A review. *Applied Soft Computing*, 19, 372–386. <https://doi.org/10.1016/j.asoc.2014.02.002>
- Reis, M. M., da Silva, A. J., Zullo Junior, J., Tuffi Santos, L. D., Azevedo, A. M., & Lopes, É. M. G. (2019). Empirical and learning machine approaches to estimating reference evapotranspiration based on temperature data. *Computers and Electronics in Agriculture*, 165, Article 104937. <https://doi.org/10.1016/j.compag.2019.104937>
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42526-019-0048-x>
- Sharafi, S., & Mohammadi Ghaleni, M. (2021). Calibration of empirical equations for estimating reference evapotranspiration in different climates of Iran. *Theoretical and Applied Climatology*, 145(3), 925–939. <https://doi.org/10.1007/s00704-021-03654-5>
- Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso | Journal of the Royal Statistical Society Series B: Statistical Methodology | Oxford Academic, 58(1), 267–288.
- Tsallis, C., & Stariolo, D. A. (1996). Generalized simulated annealing. *Physica A: Statistical Mechanics and its Applications*, 233(1–2), 395–406. [https://doi.org/10.1016/S0378-4371\(96\)00271-3](https://doi.org/10.1016/S0378-4371(96)00271-3)
- Vapnik, V. N. (2013). *The nature of statistical learning theory*. Springer Science & Business Media.
- Vapnik, V., & Chervonenkis, A. (1974). *Theory of pattern recognition*. Nauka.
- Vishwakarma, D. K., Pandey, K., Kaur, A., Kushwaha, N. L., Kumar, R., Ali, R., et al. (2022). Methods to estimate evapotranspiration in humid and subtropical climate conditions. *Agricultural Water Management*, 261, Article 107378. <https://doi.org/10.1016/j.agwat.2021.107378>
- Wanniarachchi, S., & Sarukkalgige, R. (2022). A review on evapotranspiration estimation in agricultural water management: Past, present, and future. *Hydrology*, 9(7), Article 123. <https://doi.org/10.3390/hydrology9070123>
- Williams, E. L., & Abatzoglou, J. T. (2025). Climate Change Increases Evaporative and Crop Irrigation Demand in North America. *Earth's Future*, 13(7), e2025EF005931. <https://doi.org/10.1029/2025EF005931>
- Xiang, Y., Gubian, S., Suomela, B., & Hoeng, J. (2013). *Generalized simulated annealing for global optimization: the GenSA package*. [https://doi.org/10.1016/S0375-9601\(97\)00474-X](https://doi.org/10.1016/S0375-9601(97)00474-X)
- Yonaba, R., Tazen, F., Cissé, M., Mounirou, L. A., Belemtougri, A., Ouedraogo, V. A., et al. (2023). Trends, sensitivity and estimation of daily reference evapotranspiration ET0 using limited climate data: Regional focus on Burkina Faso in the West African Sahel. *Theoretical and Applied Climatology*, 153(1), 947–974. <https://doi.org/10.1007/s00704-023-04507-z>
- Zhang, Y., Peña-Arancibia, J. L., McVicar, T. R., Chiew, F. H. S., Vaze, J., Liu, C., et al. (2016). Multi-decadal trends in global terrestrial evapotranspiration and its components. *Scientific Reports*, 6(1), Article 19124. <https://doi.org/10.1038/srep19124>

- Zhao, L., Xia, J., Xu, C., Wang, Z., Sobkowiak, L., & Long, C. (2013). Evapotranspiration estimation methods in hydrological models. *Journal of Geographical Sciences*, 23(2), 359–369. <https://doi.org/10.1007/s11442-013-1015-9>
- Zhao, L., Wang, Y., Shi, Y., Zhao, X., Cui, N., & Zhang, S. (2024). Selecting essential factors for predicting reference crop evapotranspiration through tree-based machine learning and Bayesian optimization. *Theoretical and Applied Climatology*, 155(4), 2953–2972. <https://doi.org/10.1007/s00704-023-04760-2>
- Zhou, H., Ma, L., Niu, X., Xiang, Y., Chen, J., Su, Y., et al. (2024). A novel hybrid model combined with ensemble embedded feature selection method for estimating reference evapotranspiration in the North China Plain. *Agricultural Water Management*, 296, Article 108807. <https://doi.org/10.1016/j.agwat.2024.108807>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Babak Mohammadi¹ · Mingjie Chen²  · Mohammad Reza Nikoo³ · Ali Al-Maktoumi² · Hussam Eldin Elzain²

✉ Mingjie Chen
mingjie@squ.edu.om

¹ Hydrology Research Unit, Swedish Meteorological and Hydrological Institute, Norrköping, Sweden

² Water Research Center, Sultan Qaboos University, Muscat, Oman

³ Department of Civil and Architectural Engineering, Sultan Qaboos University, Muscat, Oman