Agricultural and Forest Meteorology 263 (2018) 225–241

Contents lists available at ScienceDirect

Agricultural and Forest Meteorology

journal homepage: www.elsevier.com/locate/agrformet

Evaluation of SVM, ELM and four tree-based ensemble models for predicting daily reference evapotranspiration using limited meteorological data in different climates of China

Junliang Fan[a], Wenjun Yue[b], Lifeng Wu[a,c,*], Fucang Zhang[a,d], Huanjie Cai[a], Xiukang Wang[e], Xianghui Lu[c], Youzhen Xiang[a]

[a] Key Laboratory of Agricultural Soil and Water Engineering in Arid and Semiarid Areas of Ministry of Education, Northwest A&F University, Yangling 712100, China
[b] College of Water Resources and Environmental Engineering, Zhejiang University of Water Resources and Electric Power, Hangzhou 310018, China
[c] School of Hydraulic and Ecological Engineering, Nanchang Institute of Technology, Nanchang 330099, China
[d] Institute of Water-saving Agriculture in Arid Areas of China, Northwest A&F University, Yangling 712100, China
[e] College of Life Sciences, Yan'an University, Yan'an, 716000, China

ARTICLE INFO

Keywords:
Reference evapotranspiration
Random forest
M5 model tree
Gradient boosting decision tree
Extreme gradient boosting

ABSTRACT

Accurate estimation of reference evapotranspiration (ET_0) is of great importance for the regional water resources planning and irrigation scheduling design. The FAO-56 Penman-Monteith model is recommended as the reference model to predict ET_0, but its application is commonly restricted by lack of complete meteorological data at many worldwide locations. This study evaluated the potential of machine learning models, particularly four relatively simple tree-based assemble algorithms (i.e. random forest (RF), M5 model tree (M5Tree), gradient boosting decision tree (GBDT) and extreme gradient boosting (XGBoost)), for estimating daily ET_0 with limited meteorological data using a K-fold cross-validation method. For assessment of the tree-based models in terms of prediction accuracy, stability and computational costs, these models were further compared with their corresponding support vector machine (SVM) and extreme learning machine (ELM) models. Four input combinations of daily maximum and maximum temperature (T_{max} and T_{min}), relative humidity (H_r), wind speed (U_2), global and extra-terrestrial solar radiation (R_s and R_a) with T_{max}, T_{min} and R_a as the base dataset were considered using meteorological data during 1961–2010 from eight representative weather stations in different climates of China. The results showed that, when lack of complete meteorological data, the machine learning models using T_{max}, T_{min}, H_r, U_2 and R_a obtained satisfactory ET_0 estimates in the temperate continental, mountain plateau and temperate monsoon zones of China (RMSE $< 0.5\,\text{mm d}^{-1}$). However, models with three input parameters of T_{max}, T_{min} and R_s were superior for daily ET_0 prediction in the tropical and subtropical zones. The ELM and SVM models offered the best combination of prediction accuracy and stability. The simple tree-based XGBoost and GBDT models showed comparable accuracy and stability to the SVM and ELM models, but exhibited much less computational costs. Considering the complexity level, prediction accuracy, stability and computational costs of the studied models, the XGBoost and GBDT models have been recommended for daily ET_0 estimation in different climatic zones of China and maybe elsewhere with similar climates around the world.

1. Introduction

Evapotranspiration (ET), the water loss to the atmosphere from soil evaporation and plant transpiration, is a significant factor in the soil-plant-atmosphere interactions and is an essential component of surface water budget and energy balance (Torres et al., 2011; Liu et al., 2013; Fan et al., 2014; Feng et al., 2017c; Fan et al., 2018a). Accurate estimation of crop water requirements (i.e. actual evapotranspiration, ET_a) is a prerequisite for irrigation scheduling design and planning (Perera et al., 2014; Kisi, 2016). Although ET_a can be directly measured using water vapor transfer methods (e.g. eddy covariance and Bowen ratio) or water budget measurements (e.g. weighting lysimeters) (Ding et al., 2010), their applications are largely restricted due to the high costs and technical complexities, particularly for developing countries. ET_a can be alternatively estimated by multiplying the reference evapotranspiration (ET_0) with the crop coefficient (K_c) (Jensen, 1968). The FAO-56

* Corresponding author at: School of Hydraulic and Ecological Engineering, Nanchang Institute of Technology, Nanchang 330099, China.
E-mail address: china.sw@163.com (L. Wu).

https://doi.org/10.1016/j.agrformet.2018.08.019
Received 21 February 2018; Received in revised form 21 July 2018; Accepted 22 August 2018
0168-1923/

Penman–Monteith (FAO-56 PM) model which incorporates the thermodynamic and aerodynamic effects has proved to be more accurate than the other existing empirical models. Thus, it has been highly recommended by FAO (Food and Agriculture Organization of the United Nations) as the reference model for ET_0 calculation around the world (Allen et al., 1998). Nevertheless, the FAO-56 PM model requires a large number of input meteorological variables for its utilization, e.g., maximum/minimum air temperature, wind speed, relative humidity and solar radiation (Feng et al., 2017c; Fan et al., 2018c, d), which is a major drawback of this model. Application of ET_0 models with fewer meteorological variable inputs is thus required where lack of incomplete meteorological data (Wen et al., 2015). Over the past few decades, many efforts have been made to predict ET_0 from simple empirical models with limited data inputs (Mehdizadeh et al., 2017), e.g. temperature-based models (Hargreaves and Samani, 1985; Oudin et al., 2005), mass transfer-based models (Trabert, 1896; Romanenko, 1961) and radiation-based models (Priestley and Taylor, 1972; Xu et al., 2000; Tabari et al., 2013). However, these simplified empirical models are considered to be most suited to estimate ET_0 on a weekly or monthly basis but less suitable for daily ET_0 estimation (Torres et al., 2011).

The calculation of ET_0 can be considered as a complex and non-linear regression process depending on a large number of meteorological variables. It is hard to develop accurate empirical models to represent all the complex processes. Therefore, researchers have put forward to machine learning algorithms for ET_0 estimation because they require no knowledge of internal variables and offer simple solutions for non-linear and multi-variable functions (Kisi, 2015; Wang et al., 2017). The prediction accuracy of ET_0 in data scarce regions can be significantly improved by the machine learning models due to their excellent capability of tackling non-linear relationships between the dependent and independent variables. Various machine learning techniques have been proposed to predict ET_0, including (1) artificial neural networks (ANNs), e.g. multi-layer perceptron (MLP) (Torres et al., 2011; Ladlani et al., 2014; Traore et al., 2016), generalized regression neural networks (Kisi, 2006; Ladlani et al., 2012; Feng et al., 2017a, c), radial basis function neural networks (RBF) (Trajkovic, 2005; Ladlani et al., 2012; Petković et al., 2016) and extreme learning machine (ELM) (Abdullah et al., 2015; Feng et al., 2016; Gocic et al., 2016); (2) kernel-based algorithms, e.g. supportvector machine (SVM) (Eslamian et al., 2009; Torres et al., 2011; Tabari et al., 2012; Shiri et al., 2014; Wen et al., 2015; Shrestha and Shukla, 2015) and least-squares support vector machine (Guo et al., 2011; Kisi, 2013a, 2016); (3) tree-based assemble models, e.g., M5 model tree (M5Tree) (Pal and Deswal, 2009; Rahimikhoob, 2014; Kisi and Kilic, 2016) and random forest (RF) (Feng et al., 2017a); and (4) other machine learning models, e.g. adaptive neuro fuzzy inference system (ANFIS) (Dogan, 2009; Cobaner, 2011; Baba et al., 2013; Shiri et al., 2014; Petković et al., 2015), multivariate adaptive regression spline (MARS) (Kisi and Parmar, 2016; Deo et al., 2016), genetic programming (GP) (Shiri et al., 2012; Gocić et al., 2015; Feng et al., 2016) and fuzzy genetic models (Kisi and Cengiz, 2013; Kisi, 2013b).

Among these machine learning models, the SVM and ELM models exhibited generally better prediction accuracy than the other models (Abdullah et al., 2015; Wen et al., 2015; Feng et al., 2016; Gocic et al., 2016; Feng et al., 2017c; Yin et al., 2017a). Tabari et al. (2012) evaluated the performances of SVM, ANFIS, multiple linear regression (MLR), multiple non-linear regression (MNLR) and empirical models for ET_0 estimation in a semi-arid environment of Iran. It was found that the SVM and ANFIS models were superior to those of the regression and empirical models. Wen et al. (2015) estimated daily ET_0 using the SVM model with limited meteorological data in the extreme arid regions of China. The results indicated that the SVM model produced more accurate ET_0 estimates than the ANN and empirical models. Kisi (2016) compared the accuracies of LSSVM, MARS and M5Tree for predicting ET_0 in the Mediterranean Region of Turkey. It was concluded that the LSSVM model outperformed the MARS and M5Tree models. Abdullah et al. (2015) firstly identified the good efficiency and generalization performance of the ELM models for ET_0 estimation in Iraq. Feng et al. (2017c) also predicted ET_0 using the ELM and GRNN models with only temperature data in southwest China. The results showed that the ELM model performed better than the GRNN and Hargreaves models. The SVM and ELM models have also been hybridized with other algorithms, e.g. genetic algorithm (GA) (Yin et al., 2017b), wavelet transform (WT) (Gocić et al., 2015) and firefly algorithm (FFA) (Gocić et al., 2015) to optimize the calibration process and improve the prediction accuracy.

Most of the well-established machine learning models, however, are complex and require high computational costs during training phase (Hassan et al., 2017). The tree-based ensemble models, e.g. M5Tree and RF models, have recently begun to attract people's attention, because they are relatively simple but still powerful algorithms for classification and regression problems (Alipour et al., 2014; Hassan et al., 2017; Feng et al., 2017b). Pal and Deswal (2009) investigated the potential of the M5Tree model to estimate daily ET_0 in California, USA. The results suggested that the M5Tree model could be successfully applied in modeling ET_0. Rahimikhoob (2014) compared the M5Tree and feed-forward ANN to estimate ET_0 in an arid climate. They found the estimated ET_0 values by the M5Tree and ANN models were in good agreement with those obtained by the FAO-56 PM model. Kisi and Kilic (2016) also explored the generalization performance of the M5Tree and ANN models for estimating ET_0 in two different areas of the USA. It was concluded that the M5Tree and ANN models outperformed the empirical models and the M5Tree model was a better choice than ANN for ET_0 prediction when lack of local input and output data. Feng et al. (2017a) have recently applied the RF model and compared it with the GRNN model for daily ET_0 estimation in southwest China. The results indicated that both the RF and GRNN models performed satisfactorily for estimating daily ET_0, and the RF model performed slightly better than the GRNN model. Shiri (2018) also employed the RF model coupled with the wavelet algorithm to estimate daily ET_0 at five weather stations in Southern Iran. It was found that the novel coupled RF models greatly improved the performance of the conventional RF model and the empirical models. Recently, two improved versions of gradient boosting models named gradient boosting decision tree (GBDT) (Friedman, 2002) and extreme gradient boosting (XGBoost) (Chen et al., 2015) have been widely used in many other fields (Son et al., 2015; Wang et al., 2016; Sheridan et al., 2016; Babajide Mustapha and Saeed, 2016; Song et al., 2016; Fan et al., 2018b) because they showed higher computational efficiency and better ability to deal with over-fitting problems. However, to the best of the authors' knowledge, these two models have not yet been applied in ET_0 studies.

It is apparent from the related reviews that the ANNs, SVM and ELM models have been frequently used for modeling ET_0, while the tree-based ensemble models, especially the GBDT and XGBoost models have been very minimal. Additionally, the comparison of these simple tree-based models with the commonly used SVM and ELM models has not been comprehensively conducted yet, particularly their applicability for estimating ET_0 with various input combinations of meteorological data under different climatic conditions. Although the high prediction accuracy is primarily considered when employing the machine learning models, the good stability and less computational effort are also essential to consider (Hassan et al., 2017). Some models are inherently unstable and may yield less accurate estimates when new dataset is used for prediction. The machine learning models are also highly time-consuming compared with the empirical models, especially when long-term series meteorological data from multiple sites were used for model development. Therefore, the aims of the present study were to: (1) to determine the effects of different input combination of meteorological data on the accuracy of daily ET_0 prediction in different climates of China, (2) to develop four tree-based ensemble models, i.e. RF, M5Tree, GBDT and XGBoost, for modeling daily ET_0 using limited meteorological data, and (3) to compare the prediction accuracy, stability as