Journal of Electronic Research and Application, 2026, Volume 10, Issue 4
http://ojs.bbwpublisher.com/index.php/JERA
Online ISSN: 2208-3510
Print ISSN: 2208-3502

Tree Sap Flow Prediction Based on the Fusion of CEEMDAN-Copula Entropy-LSTM

Zixiang Wang, Kai Xu

Anhui Institute of Information Technology, Wuhu 241199, Anhui, China

Abstract: Tree trunk sap flow is jointly affected by environmental factors and physiological mechanisms, showing nonlinear and random characteristics, which makes it difficult for traditional methods to achieve high-precision prediction. To address this problem, this paper introduces CEEMDAN to decompose the sap flow sequence at multiple scales, combines Copula entropy and signal energy to construct a modal component reconstruction strategy, and further uses LSTM to realize prediction. Experimental results show that the proposed model achieves 0.6759 and 0.9755 in MAPE and R^2 indicators respectively, which is superior to the comparison models, providing a new idea for sap flow prediction and transpiration flux estimation.

Keywords: Sap flow prediction; Copula entropy; Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN); Environmental factors

Online publication: May 21, 2026

1. Introduction

Terrestrial water cycle regulates the spatial and temporal distribution of surface water resources and is closely related to human production and life [1]. At present, the problem of surface water shortage in nearly half of the world is worsening, and reasonable water resource management has become an important challenge for sustainable development [2]. Therefore, it is crucial to understand the changes and effects of the terrestrial water cycle [3]. Transpiration is the main surface water flux, and stem sap flow accounts for more than 99% of transpiration water consumption, which can directly reflect the transpiration water consumption of forest trees [4–6].

At present, most tree sap flow predictions adopt univariate analysis, linear correlation measurement or empirical equations, which have theoretical limitations [7–9]. Moreover, the nonlinearity and high randomness between sap flow and environmental factors increase the difficulty of prediction. Commonly used signal decomposition methods, such as EMD and EEMD, have defects such as mode mixing and incomplete decomposition [10,11].

To this end, this paper proposes a fusion model: CEEMDAN is used to decompose the sap flow sequence, Copula Entropy (CE) and signal energy are combined to reconstruct modal components, CE is then used to screen environmental factors, and LSTM is introduced to realize sap flow prediction.

2. Basic principles

2.1. CEEMDAN method

Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) is an improved method based on EMD [12]. By introducing adaptive white noise and performing multiple decompositions, it effectively alleviates the mode mixing problem and ensures the completeness of the decomposition results.

Its basic process is add multiple groups of adaptive white noise with a mean of 0 to the original signal x(t) to obtain noisy signals, perform EMD decomposition on them and take the average to obtain the first-order modal component C_1; then subtract the extracted component from the original signal to get the residual r_1, and repeat the above process on the residual; until the residual meets the monotonic condition, the signal is finally expressed as:

$$x(t) = \sum_{k=1}^{m} C_k + r_m(t) \tag{1}$$

where $r_m(t)$ is the residual term, also considered as the trend term, and $\sum_{k=1}^{m} C_k$ is the sum of m-order modal components.

2.2. Copula entropy

Copula theory is used to describe the dependence structure between multiple random variables and can characterize the nonlinear correlation between variables [13,14]. Let $u = F_X(x)$ and $v = F_Y(y)$, then the joint density can be expressed as $C(u, v)$. According to Copula entropy theory, the joint entropy of multiple variables can be expressed as the sum of marginal entropy and Copula entropy. Copula entropy is defined as:

$$H_C(U, V) = -\iint C(u, v) \log C(u, v) d_u d_v \tag{2}$$

2.3. Long Short-Term Memory Network (LSTM)

Long Short-Term Memory (LSTM) is an improved model based on Recurrent Neural Network (RNN) [15]. By introducing input gate, forget gate, output gate and cell state, it realizes effective storage and update of historical information. Its activation function tanh is used for nonlinear mapping of inputs, and the cell states C_t and C_{t-1} are transmitted in an approximately linear manner, thereby alleviating the problems of gradient disappearance and gradient explosion. Therefore, LSTM has good long-term dependence modeling ability and nonlinear feature expression ability, and is suitable for the modeling and prediction of complex time series data.

3. CEEMDAN-Copula Entropy-LSTM

Aiming at the nonlinear and non-stationary characteristics of tree trunk sap flow sequences, this paper constructs a CEEMDAN-Copula Entropy-LSTM fusion prediction model. First, CEEMDAN is used to decompose the sap flow sequence to obtain multi-scale modal components and trend terms, so as to reduce non-stationarity.

4. Experimental verification

4.1. Data description

Experimental data are from the SAPFLUXNET project, provided by Cate Macinnis-Ng and Julia Kaplick of the University of Auckland [16]. To avoid errors caused by missing value imputation, complete one-year data from 12:30 on December 31, 2011 to 11:30 on December 31, 2012 were selected, with a total of 8784 records. The data include sap flow values of a single kauri tree and 9 environmental factors: temperature (°C), vapor pressure deficit (kPa), wind speed ($m \cdot s^{-1}$), photosynthetic photon flux density ($\mu mol \cdot m^{-2} \cdot s^{-1}$), shallow and deep soil water content (SWC, $cm^3 \cdot cm^{-3}$), relative humidity (RH, %), shortwave incident radiation (SW, $W \cdot m^{-2}$), and net radiation (Rad, $W \cdot m^{-2}$).

4.2. Modal component reconstruction strategy based on Copula entropy and signal energy

The tree trunk sap flow sequence is regarded as a discrete time signal. The sequence of length 8784 is decomposed by CEEMDAN to obtain 11-order modal components (IMF1–IMF11) and a trend term Res. Among them, each IMF reflects characteristics of different time scales (from high frequency to low frequency), and Res represents the long-term trend. Then, the correlation between each modal component and the original sequence is calculated based on Copula Entropy (CE), resulting in a vector $ce=[ce_1, ce_2, \ldots, ce_{11}]$. According to Parseval's theorem, the energy of each component is calculated to obtain an energy vector $e=[e_1, e_2, \ldots, e_{11}]$ [17]. Further, the correlation and energy are multiplied correspondingly to construct an indicator $m_i=c_i\, ce_i$, resulting in a vector $M=[m_1, m_2, \ldots, m_n]$, which is normalized and mapped to the [0,1] space, where m_{max} and m_{min} represent the maximum and minimum values in the M vector respectively:

$$J_i = \frac{m_i - m_{min}}{m_{max} - m_{min}} \tag{3}$$

Finally, a screening indicator vector $J=[j_1, j_2, \ldots, j_n]$ is obtained, which describes the interpretation degree of each order modal to the original sap flow signal. A larger j value indicates that the corresponding component signal has a greater interpretation degree to the original sequence.

According to the J values of each component in **Table 1**, IMF1 and IMF3–IMF10 have weak interpretation ability for the original sequence and are eliminated. Although the Copula entropy of the residual term Res is high, its energy is extremely small, so it is not retained according to the reconstruction strategy. Finally, IMF2 and IMF11 are selected for reconstruction, and the sum of their J values accounts for about 91.2% of the total, which can effectively retain the main characteristics of the original sequence while removing redundant information.

Table 1. J values of each component signal

Component Signal	CE	Signal Energy	J	Selected
IMF1	0.1160	1.9906E+03	0.0250	No
IMF2	0.4486	2.0568E+04	1.0000	Yes
IMF3	0.1064	4.5633E+03	0.0526	No
IMF4	0.0378	2.7027E+03	0.0111	No
IMF5	0.0886	3.2949E+03	0.0316	No

Component Signal	CE	Signal Energy	J	Selected
IMF6	0.0816	1.4749E+03	0.0131	No
IMF7	0.0782	4.2137E+02	0.0036	No
IMF8	0.0918	1.8629E+02	0.0019	No
IMF9	0.1354	1.1826E+02	0.0017	No
IMF10	0.1843	2.1535E+03	0.0430	No
IMF11	0.2102	4.0238E+04	0.9168	Yes
RES	0.6146	1.5780E-27	0.0000	No

4.3. Environmental variable selection based on Copula entropy

CE is used to select environmental factors for the reconstructed sequence composed of IMF2 and IMF11 component signals. According to the CE values of the nine environmental factors, five environmental variables are finally selected as the independent variables of the reconstructed sequence: vapor pressure deficit (VPD), wind speed (WS), deep soil water content (SWC_deep), relative humidity (RH), and net radiation (Rad). The sum of the CE values of the selected environmental variables accounts for 93.9% of the total CE values of all environmental factors.

4.4. Prediction results

In model construction, considering the temporal transmission of environmental factor characteristics and data scale, this study adopts a two-layer LSTM structure to model the reconstructed sequence. Experiments show that too many layers not only increase computational overhead but also may lead to a decline in convergence performance, so a deeper network is not adopted. The specific structure of the model is: the first LSTM layer has 16 units and outputs sequence features; the second LSTM layer has 64 units, extracts high-level temporal features, and completes prediction by combining Dropout and fully connected layers.

In addition, the time step is a key parameter representing the length of the input historical sequence. Through multiple comparative experiments, it is found that an excessively small or large time window will lead to obvious deviations between the prediction results and the true values. Finally, the time step is determined to be 22 to obtain a better prediction effect. The network prediction process is shown in **Figure 1**.

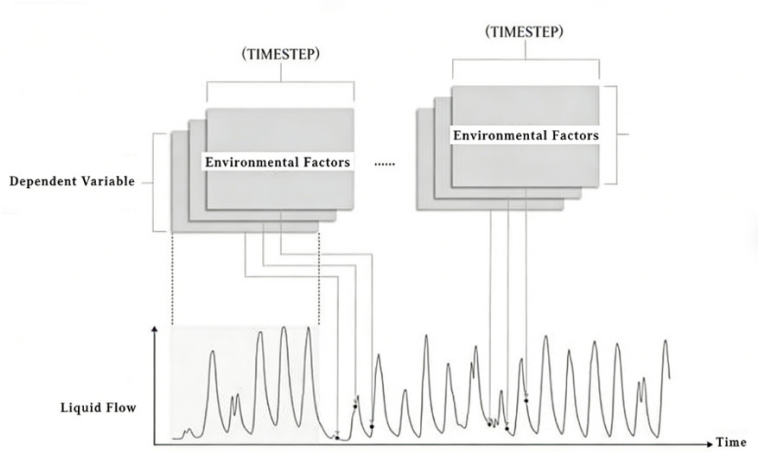

Figure 1. Schematic diagram of network prediction process.

The model parameters are determined through multiple experiments as follows: learning rate 0.0001, number of iterations 200, batch size 128, and time step 22. Under this configuration, the model can effectively learn the relationship between environmental factors and the target sequence and achieve good prediction results. Further, the proposed fusion model is compared with Multiple Linear Regression (MLR), Random Forest (RF), BP network, Convolutional Neural Network (CNN) and single LSTM, as shown in **Table 2**.

Table 2. Model performance comparison

Model	MSE	MAPE (%)	R^2
MLR	0.01262	102.4	0.7172
RF	0.009952	115.9	0.7740
BP	0.01155	39.56	0.7205
CNN	0.0037	179.1	0.9167
LSTM	0.0523	1.1331	0.9476
CEEMDAN-Copula Entropy-LSTM	0.0229	0.6759	0.9755

5. Conclusion

Compared with other models, the fusion CEEMDAN-Copula Entropy-LSTM model proposed in this paper achieves the best performance in two model evaluation indicators: MAPE and R^2. Among them, MAPE has decreased by 45.72% compared with the single LSTM, with a significant improvement, providing a useful idea for transpiration flux estimation based on tree sap flow.

Disclosure statement

The authors declare no conflict of interest.

References

[1] Xie Z, et al., 2019, Climate Feedback of Human Water Use Activities and Its Impact on Terrestrial Water Cycle: Progress and Challenges. Advances in Earth Science, 34(8): 801–813.

[2] Sun S, et al., 2021, Unraveling the Effect of Inter-Basin Water Transfer on Reducing Water Scarcity and its Inequality in China. Water Research, 2021(194): 116931.

[3] Tang Q, et al., 2019, Comprehensive Integration and Simulation of Terrestrial Water Cycle Processes. Advances in Earth Science, 34(2): 115–123.

[4] Gharbia S, et al., 2018, Spatially Distributed Potential Evapotranspiration Modeling and Climate Projections. Science of The Total Environment, 2018(633): 571–592.

[5] Huang Y, et al., 2021, Dynamics of *Populus euphratica* Sap Flow Rate and Its Relationship with Meteorological Factors in the Growing Season in the Southeast of the Kumtag Desert. Journal of Plant Sciences, 39(3): 247–257.

[6] Yao Y, et al., 2021, Comparison and Application of Calculation Methods for Forest Transpiration Water Consumption. Chinese Journal of Applied Ecology, 32(8): 2989–2998.

[7] Teng H, et al., 2020, Analysis of Water Consumption Characteristics and Construction of Prediction Model of

Ziziphus jujuba. Journal of Central South University of Forestry and Technology, 40(1): 22–29.

[8] Bai Z, et al., 2016, Dynamic Changes of Stem Sap Flow of *Larix sibirica*. Journal of Hebei Agricultural University, 39(3): 49–54.

[9] Li X, et al., 2014, Relationship Between Transpiration Water Consumption of Ginkgo biloba and Environmental Factors. Journal of Beijing Forestry University, 36(4): 23–29.

[10] Lü S, Sun Y, 2022, Short-Term Load Forecasting Method Based on EMD-RVM Model. Microcomputer Applications, 38(10): 22–24+31.

[11] Fu Z, et al., 2022, Research on Degradation Degree Prediction Method of Hydropower Units Based on EEMD and LSTM. Acta Energiae Solaris Sinica, 43(2): 75–81.

[12] Torres M, et al., 2011, A Complete Ensemble Empirical Mode Decomposition with Adaptive Noise, IEEE International Conference on Acoustics, Speech and Signal Processing, 4144–4147.

[13] Ma J, 2022, Copula Entropy: Theory and Applications.

[14] Yanling L, Yunpeng Z, Ling Z, 2017, Variations Detection of Bivariate Dependence Based on Copulas Model. International Journal of Applied Mathematics, 47(2): 255–260.

[15] Hochreiter S, Schmidhuber J, 1997, Long Short-Term Memory. Neural Computation, 9(8): 1735–1780.

[16] Poyatos R, et al., 2016, SAPFLUXNET: Towards a Global Database of Sap Flow Measurements. Tree Physiology, 36(12): 1449–1455.

[17] Iwasaki A, 2020, Deriving the Variance of the Discrete Fourier Transform Test Using Parseval's Theorem. IEEE Transactions on Information Theory, 66(2): 1164–1170.