Postprint: Machine Learning-Based Drought Prediction Study for the North Haihe River System

Authors: Zhao Meiyan, Hu Tao, Zhang Yuhu, Pu Xiao, peak, Peak

Date: 2020-11-19T00:00:00+00:00

Abstract

Improving drought prediction accuracy can provide reliable data support for watershed drought response and risk prevention, and constructing and comparing appropriate drought models is a current research hotspot. This study used the Standardized Precipitation Index (SPI) at four temporal scales (3, 6, 9, and 12 months) as the characterization indicator, and employed three machine learning algorithms—Wavelet Neural Network (WNN), Support Vector Regression (SVR), and Random Forest (RF)—to construct drought prediction models for the northern Hai River basin, respectively. Kendall, K-S, and MAE tests were used to evaluate model performance and stability. The results indicate that: (1) The performance of WNN and SVR models varies across different SPI temporal scales, with WNN being most suitable for 12-month SPI drought prediction and SVR being most suitable for 6-month SPI drought prediction. (2) For 3- and 12-month SPI, RF demonstrates the optimal prediction performance (Kendall $>$ 0.898, MAE $<$ 0.05); for 6- and 9-month SPI, SVR shows the optimal prediction performance (Kendall $>$ 0.95, MAE $<$ 0.04). (3) The stability of model prediction performance differs, with RF exhibiting the highest prediction stability, followed by SVR. (4) The differences and similarities in performance among the three constructed models are primarily due to SVR' s transformation into a convex optimization problem, which addresses WNN' s tendency to fall into local optima, thereby improving model prediction performance. RF integrates diverse regression trees, reducing the negative impact of weak learners and enhancing model prediction accuracy and stability. Additionally, RF has a stronger capability in handling precipitation data containing noise.

Full Text

Drought Prediction Based on Machine Learning Models in the Northern Part of Haihe River Basin

ZHAO Meiyan[1], HU Tao[1], ZHANG Yuhu[2], PU Xiao[2], GAO Feng[3]
[1]School of Mathematical Sciences, Capital Normal University, Beijing, China
[2]College of Resources, Environment & Tourism, Capital Normal University, Beijing, China [3]National Meteorological Information Center, Beijing, China

Abstract

Improving drought prediction accuracy provides reliable data support for basin drought response and risk prevention, and constructing suitable drought prediction models represents a current research hotspot. This study developed drought prediction models for the northern Haihe River Basin using three machine learning algorithms: Wavelet Neural Network (WNN), Support Vector Regression (SVR), and Random Forest (RF). The Standardized Precipitation Index (SPI) at four time scales (3, 6, 9, and 12 months) served as the drought indicator. Model performance and stability were evaluated using Kendall rank correlation coefficient, Mean Absolute Error (MAE), and Kolmogorov-Smirnov test. Results demonstrated that model performance varied across different SPI time scales: WNN performed optimally for 3-month and 12-month SPI (Kendall $>$ 0.898, MAE $<$ 0.05), while SVR excelled for 6-month and 9-month SPI (Kendall $>$ 0.95, MAE $<$ 0.04). The three models exhibited different prediction stabilities, with RF demonstrating the highest stability, followed by SVR. The differential performance among models primarily stems from: (1) SVR' s transformation into a convex optimization problem resolves WNN' s tendency to fall into local optima, thereby improving prediction performance; (2) RF integrates diverse regression trees, reducing negative impacts from weak learners and enhancing both prediction accuracy and stability; (3) RF demonstrates stronger capacity for handling noisy precipitation data. This study provides a comprehensive analysis of multi-model, multi-time-scale drought prediction performance and preliminary exploration of the underlying statistical mechanisms driving model differentiation, offering valuable insights for drought prediction in this region and beyond.

Keywords: drought; machine learning; SPI; northern Haihe River Basin

Introduction

Drought is one of the most common, complex, and severely impactful meteorological disasters on human society [1]. With climate warming, drought severity in the Haihe River Basin tends to increase and its occurrence range is expanding [2], leading to increasingly severe drought disasters. Early drought prediction enables timely establishment of drought warning mechanisms for effective prevention, reducing impacts on human life, property, and ecological environments.

As a combination of wavelet transform and neural networks, Wavelet Neural Network (WNN) possesses superior nonlinear processing capabilities and has been extensively applied in drought prediction research. For instance, Zhang et al. [3] used WNN to predict the Standardized Precipitation Index (SPI) in the northern Haihe River Basin, confirming its fitting capability for multi-month scale SPI. Support Vector Regression (SVR), extended from classification problems, employs structural risk minimization principles and is suitable for small-sample, nonlinear, and high-dimensional problems, also finding wide application in drought prediction [4]. Random Forest (RF), a combination model based on Classification and Regression Trees (CART), exhibits stable prediction performance while handling noisy predictor variables effectively, demonstrating strong performance in prediction studies [5].

However, previous drought prediction research predominantly constructed single-algorithm models focusing on single time-scale drought indicators [6], lacking comprehensive comparative analysis across multiple models and time scales. Literature comparing WNN, SVR, and RF models within the same study area across different time scales remains scarce. Furthermore, most previous studies did not analyze model stability or explore the intrinsic statistical mechanisms underlying performance differences among algorithms. Addressing these gaps, this study employed SPI at four time scales (3, 6, 9, and 12 months) to construct and evaluate drought prediction models for the northern Haihe River Basin using WNN, SVR, and RF, preliminarily investigating the internal mechanisms of model differentiation and identifying optimal drought prediction models. The results provide valuable attempts for drought prediction in this region and other areas.

1. Study Area and Data

1.1 Study Area

The northern Haihe River Basin is located in the upper reaches of Beijing and Tianjin, primarily including the Jiyun River, Chaobai River, Beiyun River, and Yongding River (Fig. 1 [Figure 1: see original paper]). The basin area is 8.34$×10^{4}$ km^2, with mountainous and plain areas accounting for 62.5% and 37.5%, respectively. The region features a temperate east Asian monsoon climate with an average annual precipitation of approximately 490 mm. In recent years, precipitation in the northern Haihe River Basin has been generally low [7], with both drought severity and extent showing upward trends [8]. Historical records indicate up to 20 drought occurrences in the basin.

1.2 Data

This study utilized daily precipitation data from 1960 to 2010 obtained from the China Meteorological Administration (CMA) for eight national benchmark meteorological stations in the northern Haihe River Basin. Data underwent strict revision and quality control, with missing precipitation values replaced

by nearby station averages to ensure continuous, complete records. Using the daily precipitation data, SPI series were calculated at 3-month, 6-month, 9-month, and 12-month scales. The SPI calculation method follows references [9,10]. Station information is detailed in Table 1 .

2. Model Algorithms

2.1 Wavelet Neural Network (WNN)

WNN integrates wavelet transform theory with neural networks, representing a novel feedforward neural model that employs wavelet functions as activation functions for hidden layer neurons [11]. For input variables x (i = 1,⋯,k), let ω denote connection weights between input and hidden layers, ω represent weights between hidden and output layers, h denote wavelet basis functions, b represent translation factors, and a denote scaling factors. The hidden layer output is:

$$h_j = h\left(\frac{\sum_{i=1}^{k}\omega_{ij}x_i - b_j}{a_j}\right), \quad j = 1, ..., l$$

where l is the number of hidden neurons. For hidden neuron output h(i) and m output neurons, the output layer result is:

$$y(k) = \sum_{j=1}^{l}\omega_{jk}h(i), \quad k = 1, ..., m$$

This study employed a three-layer WNN with Morlet wavelet basis function:

$$y = \cos(1.75x)e^{-x^2/2}$$

The training process minimizes mean square error using gradient descent to adjust network weights and wavelet parameters. Training involves: (1) Setting learning rate and hidden neuron count l, randomizing weights ω , ω and wavelet parameters a , b ; (2) Splitting data into training and testing sets; (3) Inputting training samples sequentially, calculating network output and prediction error, and updating parameters via backpropagation; (4) Iterating until termination criteria are met.

2.2 Support Vector Regression (SVR)

SVR addresses regression problems by establishing relationships between predicted vectors and support vectors in training data based on statistical learning theory. Given training set D = {(x , y) | x ^{n}, y , i = 1,⋯,l}, as input space, and nonlinear mapping ϕ(x), SVR aims to find function f(x) = ω ϕ(x) + b that best approximates observed values by minimizing:

$$R(f) = \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{l} L_\varepsilon(y_i, f(x_i))$$

where L_ is the -insensitive loss function:

$$L_\varepsilon(y_i, f(x_i)) = \begin{cases} 0 & \text{if } |f(x_i) - y_i| < \varepsilon \\ |f(x_i) - y_i| - \varepsilon & \text{if } |f(x_i) - y_i| \geq \varepsilon \end{cases}$$

Introducing slack variables , * and applying structural risk minimization transforms the problem into convex optimization:

$$\min_{\omega,\xi,\xi^*} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*)$$

$$\text{s.t.} \begin{cases} y_i - \omega^T\phi(x_i) - b \leq \varepsilon + \xi_i \\ \omega^T\phi(x_i) + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases}$$

The Lagrangian formulation yields the dual optimization problem. Solving via Sequential Minimal Optimization (SMO) produces the decision function:

$$f(x) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)k(x_i, x) + b$$

where k(x , x) is the kernel function.

2.3 Random Forest (RF)

Proposed by Leo Breiman in 2001 [12], RF is an ensemble learning algorithm based on CART that handles both classification and regression. Using bootstrap resampling, RF generates K training subsets by sampling with replacement from the original dataset, with each subset constructing one decision tree. At each node, random feature selection enables splitting based on minimum impurity. Trees grow to maximum depth without pruning, forming a multivariate nonlinear ensemble. For regression, the final prediction averages all tree predictions; for classification, majority voting determines the outcome.

The RF regression algorithm construction follows a similar process to SVR, with primary optimization focused on lag order. Analysis revealed lag order significantly impacts prediction performance. The optimal lag order range was investigated, with MAE reaching minimum values at specific lag orders, establishing the optimal lag order for the Beijing station model.

3. Model Construction and Validation

3.1 Model Construction

Modeling processes were similar across stations; thus, the Beijing station serves as an exemplar. Models utilized only SPI data, with historical SPI values as inputs and current SPI as output. Data from 1960–2000 were used for model development, while 2001–2010 data served for validation.

WNN Parameter Tuning: Hyperparameters include input nodes (lag order), hidden nodes, and learning rate. The development data were split into training (80%) and validation (20%) sets. Grid search identified optimal hyperparameters minimizing MAE. Table 2 presents the tuning results.

SVR and RF Lag Order Selection: For SVR and RF, lag order selection followed the same methodology. Figure 3 [Figure 3: see original paper] and Figure 4 [Figure 4: see original paper] illustrate the lag order selection process, showing MAE minima at specific lag values.

3.2 Model Evaluation

Three metrics assessed model performance: MAE, Kendall rank correlation coefficient, and Kolmogorov-Smirnov (KS) test. MAE measures proximity between predicted and observed values (lower values indicate better fit). Kendall correlation assesses monotonic relationships (values approaching 1 indicate stronger correlation). The KS test, implemented in R, calculates the maximum absolute distance D between empirical distribution functions; larger D values increase the likelihood that samples derive from different distributions.

3.3 Model Testing

3.3.1 Comparative Analysis Models were compared across all stations for 3-month SPI prediction (Table 3). Results show: - **WNN**: Achieved Kendall > 0.898 and MAE < 0.05 at most stations, though KS test values exceeded 0.05 at select stations. Overall performance was optimal for 3-month SPI. - **SVR**: Demonstrated superior performance for 6-month SPI (Table 4) with Kendall $>$ 0.95 and MAE < 0.04, though KS test values were elevated at Fengning, Huailai, Zunhua, and Beijing stations. - **RF**: Showed excellent performance across all metrics, with KS test values ≤ 0.05. For 9-month SPI (Table 5), RF achieved optimal performance (Kendall > 0.95, MAE < 0.04). For 12-month SPI (Table 6), WNN again performed optimally.

3.3.2 Stability Analysis Model stability was assessed by calculating average evaluation metrics across stations to examine SPI time scale impacts (Figure 5 [Figure 5: see original paper]). Key findings: - **WNN** exhibited the greatest sensitivity to SPI time scale variation, with metric ranges of 0.15 for Kendall, 0.06 for KS test, and 0.04 for MAE. Performance was optimal at 3-month and 12-month scales. - **SVR** showed moderate sensitivity, with metric ranges of

0.08 (Kendall), 0.03 (KS test), and 0.02 (MAE), performing best at 6-month and 9-month scales. - **RF** displayed the highest stability, with minimal metric variation across time scales (Kendall range: 0.05; KS test range: 0.02; MAE range: 0.01), showing no significant performance differences across SPI scales.

4. Discussion

Using daily precipitation data from 1960–2010 and SPI as the drought indicator, this study developed and evaluated drought prediction models for the northern Haihe River Basin. WNN and SVR demonstrated varying applicability across time scales: WNN excelled at 3-month and 12-month SPI, while SVR outperformed at 6-month and 9-month scales. RF consistently delivered stable, high-performance predictions across all scales.

Model performance differences arise from intrinsic algorithmic characteristics: 1. **Optimization Nature**: WNN, as a neural network, risks converging to local optima, compromising prediction accuracy. SVR transforms the problem into convex optimization, avoiding local minima and enhancing precision. 2. **Ensemble Learning**: RF integrates diverse regression trees, mitigating weak learner impacts and improving both accuracy and stability. Its robustness to noise provides advantages when handling precipitation data containing measurement errors. 3. **Computational Efficiency**: WNN parameter tuning is most complex and computationally slow. SVR requires only lag order optimization, offering faster computation. RF demonstrates intermediate complexity with stable outcomes.

For operational drought prediction and early warning, model selection should be flexible: choose WNN for 3-month and 12-month SPI, SVR for 6-month and 9-month SPI, and RF when stable performance across multiple time scales is prioritized. Future research should explore these models' applicability at longer time scales and in other regions, while further investigating the statistical mechanisms underlying divergent prediction trajectories.

5. Conclusion

1. **Multi-scale Performance**: The three machine learning models exhibited scale-dependent performance. WNN performed optimally for 3-month and 12-month SPI; SVR excelled for 6-month and 9-month SPI; RF showed consistent performance across all scales without significant variation.
2. **Within-scale Comparison**: For 3-month SPI, WNN achieved optimal performance (Kendall > 0.898, MAE < 0.05), accurately reflecting SPI fluctuations. For 6-month and 9-month SPI, SVR delivered optimal results (Kendall > 0.95, MAE < 0.04). For 12-month SPI, WNN again proved superior. RF maintained the highest stability across all evaluations.
3. **Stability Assessment**: RF exhibited the highest stability against SPI time scale variation, with minimal metric ranges (Kendall: 0.05; KS test:

0.02; MAE: 0.01). WNN showed the lowest stability, while SVR demonstrated intermediate stability.

References

[1] GAO Taotao, YIN Shuyan, WANG Shuixia. Spatial and temporal variations of drought in northern and southern regions of Qinling Mountains based on standardized precipitation evapotranspiration index[J]. Arid Land Geography, 2018, 41(4): 85-94.

[2] WANG Wenjing, YAN Junping, LIU Yonglin, et al. Characteristics of droughts in the Haihe Basin based on meteorological drought composite index[J]. Arid Land Geography, 2016, 39(2): 334-336.

[3] NI Haishen, GU Yin, PENG Yuejin. Spatio-temporal pattern and evolution trend of drought disaster in China in recent seventy years[J]. Journal of Natural Disasters, 2019, 28(6): 176-181.

[4] ZHU S, LUO X, CHEN S, et al. Improved hidden markov model incorporated with Copula for probabilistic seasonal drought forecasting[J]. Journal of Hydrologic Engineering, 2020, 25(6).

[5] WANG Zhicheng. Regional drought prediction based on improved Markov chain[J]. Water Resources Development and Management, 2018, (2): 55-57.

[6] MA Qiyun, ZHANG Jiquan, WANG Yongfang, et al. Characteristics and prediction of drought in growing season in Inner Mongolia pastoral area[J]. Journal of Arid Land Resources and Environment, 2016, 30(7): 157-163.

[7] HAN Huiming, LIU Zheyue, LIU Chenglin, et al. Improvement of grey model and its application in forecast of meteorological drought[J]. South-to-North Water Transfers and Water Science & Technology, 2019, 17(6): 62-68.

[8] GU Hongbo, LIU Zhiyu. Time regularity analysis and trend prediction of agricultural drought disaster in Hunan Province[J]. Journal of Hunan University of Science & Technology (Social Science Edition), 2016, 19(5): 110-116.

[9] YANG Huirong, ZHANG Yuhu, CUI Hengjian, et al. Applicability of ARIMA and ANN models for drought forecasting[J]. Arid Land Geography, 2018, 41(5): 47-55.

[10] ZHANG Y, YANG H, CUI H, et al. Comparison of the ability of ARIMA, WNN and SVM models for drought forecasting in the Sanjiang Plain, China[J]. Natural Resources Research, 2019, 29: 1447-1464.

[11] YANG Haimin, PAN Zhisong, BAI Wei. A survey of time series prediction methods[J]. Computer Science, 2019, 46(1): 21-28.

[12] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[13] SHU Xingsheng, WANG Ziru, LI Fuwei. Short-term rainfall multi-mode integrated forecasting based on machine learning models[J]. South-to-North Water

Transfers and Water Science & Technology, 2020, 18(1): 42-50.

[14] CUO Mu, JIAYONG Cicheng, HONG Mei. Using data mining methods to explore meteorological drought forecasts at river basin scales[J]. Sichuan Environment, 2018, 37(4): 65-70.

[15] WU Jing, CHEN Yuanfang, YU Shengnan. Research on drought prediction based on random forest model[J]. China Rural Water and Hydropower, 2016, (11): 17-22.

[16] ZHANG Y, LI W, CHEN Q, et al. Multi-models for SPI drought forecasting in the north of Haihe River Basin, China[J]. Advances in Meteorology, 2015, 2015: 1-15.

[17] ZHANG Jiao, TIAN Qi, WANG Meiping. Heating load prediction for heating systems based on support vector regression with cross validation[J]. Journal of North University of China (Natural Science Edition), 2014, 35(5): 189-206.

[18] WANG Jinan, LI Fei. Review of inverse optimal algorithm of in-situ stress field and new achievement[J]. Journal of China University of Mining & Technology, 2015, 44(2): 189-205.

[19] AHMADEBRAHIMPOUR E, AMINNEJAD B, KHALILI K. Application of global precipitation dataset for drought monitoring and forecasting over the Lake Urmia Basin with the GA-SVR model[J]. International Journal of Water, 2018, 12(3): 262-277.

[20] GE Qiang. Evaluation of sustainable utilization of water resources in Kuitun River based on random forest[J]. Pearl River, 2019, 40(1): 79-83.

[21] TYRALIS H, PAPACHARALAMPOUS G, LANGOUSIS A. A brief review of random forests for water scientists and practitioners and their recent history in water resources[J]. Water, 2019, 11(5): 910.

[22] SHEN Runping, GUO Jia, ZHANG Jingxian, et al. Construction of a drought monitoring model using the random forest based remote sensing[J]. Journal of Geo-information Science, 2017, 19(1): 125-133.

[23] ZHANG Yuhu, XIANG Liu, SUN Qing, et al. Bayesian probabilistic forecasting of seasonal hydrological drought based on Copula function[J]. Scientia Geographica Sinica, 2016, 36(9): 1437-1444.

[24] CAI W, ZHANG Y, YAO Y, et al. Probabilistic analysis of drought spatiotemporal characteristics in the Beijing-Tianjin-Hebei metropolitan area in China[J]. Atmosphere, 2015, 6(4): 534-551.

[25] ZHANG Y, XIE P, PU X, et al. Spatial and temporal variability of drought and precipitation using cluster analysis in Xinjiang, northwest China[J]. Asia Pacific Journal of Atmospheric Sciences, 2019, 55: 155-164.

[26] ZHANG Y, CAO W, CHEN Q, et al. Analysis of changes in precipitation and drought in Aksu River Basin, northwest China[J]. Stochastic Environmental

Research & Risk Assessment, 2017, 31(10): 2471-2481.

[27] ZHANG Shuyu, WANG Jianhua, ZHAI Jiaqi. Characteristics analysis of time serial of rainfall in the northern part of Haihe River Basin from 1956 to 2012[J]. South-to-North Water Transfers and Water Science & Technology, 2016, 14(3): 36-42.

[28] ZONG Yan, WANG Yanjun, ZHAI Jianqing. Spatial and temporal characteristics of meteorological drought in the Haihe River Basin based on standardized precipitation index[J]. Journal of Arid Land Resources and Environment, 2013, 27(12): 198-202.

[29] HE J, YANG X, LI J, et al. Spatiotemporal variation of meteorological droughts based on the daily comprehensive drought index in the Haihe River Basin, China[J]. Natural Hazards, 2015, 75(S2): 199-217.

[30] LI Wenqing, JIANG Yuan, ZHAO Shoudong, et al. Response of tree-ring width chronology of pinus tabulaeformis to multi-scale standardized precipitation index (SPIn) in the Liupan Mountain area[J]. Acta Ecologica Sinica, 2017, 37(10): 3365-3374.

[31] WANG Yu, LU Wenxi, BIAN Jianmin, et al. Surrogate model of numerical simulation model of groundwater based on wavelet neural network[J]. China Environmental Science, 2015, 35(1): 139-146.

[32] WANG Xia, WANG Zhanqi, JIN Gui, et al. Land reserve prediction using different kernel based support vector regression[J]. Transactions of the Chinese Society of Agricultural Engineering, 2014, (4): 204-211.

[33] BREIMAN L. Statistical modeling: The two cultures (with comments and a rejoinder by the author)[J]. Statistical Science, 2001, 16(3): 199-231.

[34] WANG Yisen, XIA Shutao. A survey of random forests algorithms[J]. Information and Communications Technologies, 2018, 12(1): 49-55.

Note: Figure translations are in progress. See original paper for figures.