Analysis of Climate Change for Drought Forecasting Using High-Resolution Data and Ensemble Learning with Optimized Pruning Model

Indirani M[1,*], Venketbabu T[2], Vinmathi M.S[3], Senduru Srinivasulu[4]

[1]Department of Information Technology, Adithya Institute of Technology, Coimbatore, India.

mindirani2024@gmail.com

[2]Department of Computer Science and Engineering, School of Computing, Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Avadi, Chennai 600062, India, tvenketbabu@gmail.com

[3]Department of Computer Science and Engineering, Panimalar Engineering College, Poonamallee, Chennai, Tamil Nadu 600123, vinmathis@gmail.com

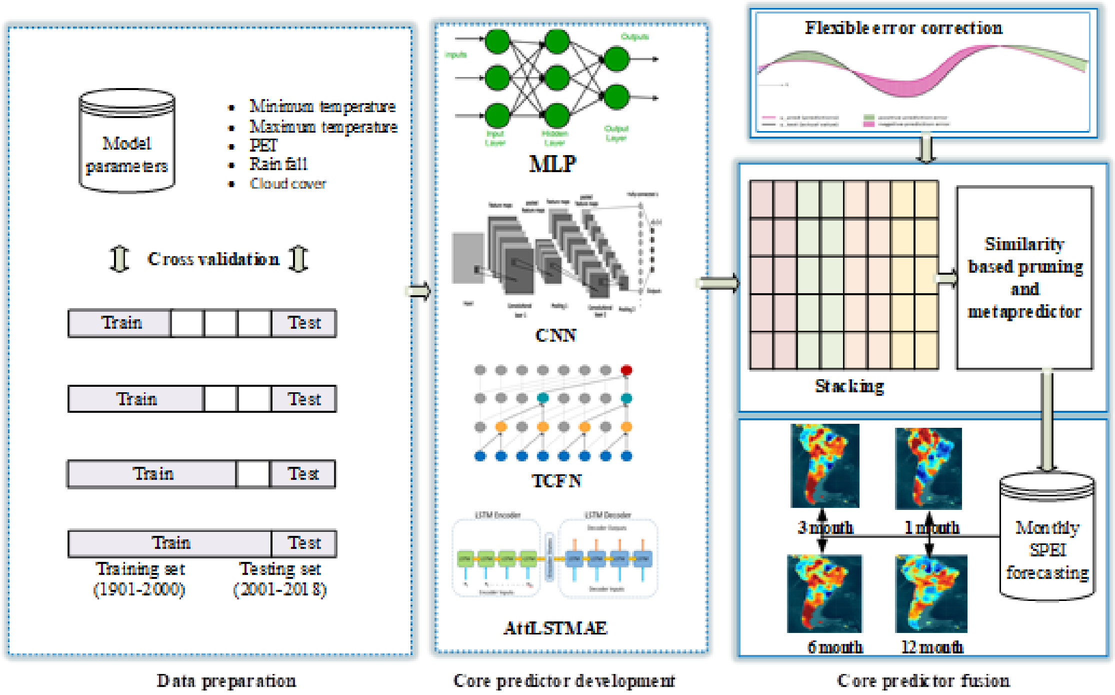
[4]School of Computing, Department of Computer Science and Engineering, Sathyabama Institute of science and Technology, Chennai, 600119, India.

sendurusrinivasulu.cse@sathyabama.ac.in

[*]Corresponding author:

E-mail: mindirani2024@gmail.com

Graphical Abstract



Abstract

Investigating the influence of climate change on drought in a dynamic environment is essential for human society, agriculture, and ecology. Draught forecasting often considers machine learning techniques that use climate-mode indices as predictor variables. However, forecasting in long lead times is still difficult due to the consequences of climate change and the difficulties associated with evaluating drought. In this study, a novel ensemble learning with optimized pruning model (EnsLOP) based on different deep learning models, i.e., multilayer perceptron (MLP), convolutional neural network (CNN), temporal convolutional feature network (TCFN) and Attention-driven LSTM Autoencoder (AttLSTMAE), is applied to improve the forecasting capability of draught index namely Standardised Precipitation Evapotranspiration Index (SPEI). This study collects the preceding lag memory of climatic mode indicators such as rainfall, temperature, precipitation, and cloud cover as predictor variables to achieve significantly accurate draught forecasts. Also, a new flexible error correction (FEC) is proposed to reduce the prediction errors of the core predictors. The simulation results demonstrate that the proposed EnsLOP model gained a distinct advantage in terms of SPEI prediction with comparatively low relative errors (RMSE =0.098 and R2 <0.098).

Keywords: Climate variables, draught, standardized precipitation evapotranspiration index (SPEI), ensemble learning, pruning, error correction and forecasting.

1. Introduction

Climate change is one of the most significant challenges of the twentieth century. India is one of many places in the world that are now undergoing climatic variability. Yang et al. (2023) developed an extreme events with more frequent and intense globally as a result of climate change, and tropical regions are particularly at risk of experiencing these occurrences. Kumar et al. (2023) introduce a Climate factors such as temperature and precipitation have a great influence on agricultural practices and water bodies. Yang et al. (2020) the hydrological cycle is mainly controlled by meteorological factors, and extreme variations in annual precipitation and temperature over long periods of time can cause natural disasters such as floods and droughts. Tamilvizhi T et al. (2022) implemented a methods for droughts occur in all climatic zones due to dry weather that can last for extended periods of time and cause a major imbalance in the water cycle.

Nguyen et al. (2023) executed a drought will get worse if the rate of precipitation varies. In light of this, scientists have employed a number of drought indexes; The Standard Precipitation Index (SPI) is the most widely used index for analyzing precipitation data. The SPI is used together with other indicators such as the Rainfall Anomaly Index (RAI) to identify severe drought evented by Santhanaraj R. K et al. (2023). Liu et al. (2021) introduced a SPI to determine the start and end of a drought event and assess the influence of draught at multiple time intervals (monthly to yearly). Surendran R et al. (2023) stated the Univariate indices may not adequately capture the essence of a drought episode because drought is caused by many variables (such as precipitation, evapotranspiration, and soil moisture). Sharma et al. (2022)

found the result, multivariate drought indices have been created including the US Drought Monitor (USDM), the Multivariate Standardized Drought Index (MSDI), and the Standardized Precipitation Evapotranspiration Index (SPEI). Wang et al. (2020) shared Information from two or more meteorological variables is integrated using multivariate indicators. Xu et al. (2021) discussed the use of drought index to analyze future drought estimates based on current levels is beneficial to climate policy and drought responses.

Rehana et al. (2020) developed the SPEI has gained popularity as a meteorological drought index because it includes atmospheric climatic need, such as the disparity between precipitation and potential evapotranspiration (PET). SPEI has been shown to be a more trustworthy metric than SPI because it incorporates both PET and precipitation. Ullah et al. (2023) discussed about temperature and the techniques used to calculate evapotranspiration have an impact on SPEI. Furthermore, it has been demonstrated that the drought characteristics calculated from SPI and SPEI in monsoon regions are equivalent at short time scales. Surendran R et al. (2023) analyzed in the literature, three different types of model, such as physical, data-driven, and hybrid models, are used for forecasting draught because draught is fundamentally nonlinear. Hu et al. (2021) said the application of data-driven models has received more attention because it has been shown to produce better forecasting results than physical-based models.

Dikshit et al. (2021) implements the artificial neural networks (ANN) method to predict draught with both short and long lead times. Guo et al. (2024) were able to quickly identify broad trends or differences between drought indices and weather information. Chao et al. (2020) executes the basic deep learning networks developed for sequence modelling are repeat (RNN) and convolutional neural networks (CNN) and are often preferred over multilayered perceptrons (MLPs) to forecast climate changes. Surendran R et al. (2023) developed a the long-short-term memory unit (LSTM) unit is a variant of the ordinary RNN design that incorporates gating approaches and skip connections to address the issue of vanishing or expanding gradients.

Dikshit et al. (2021) established a deep learning models are known to have high variances and low biases. Bentsen et al. (2023) maintaining high prediction accuracy and durability is challenging for a single deep learning model in a complex and dynamic application environment. Barzkar et al. (2022) provide an ensemble learning has proven to be a successful approach to resolving this problem. It does this by utilizing multiple distinct individual models, as well as specific ensemble procedures, to enhance the generality of the complete model. These points motivate us to propose a novel ensemble learning to forecast the draught by analysing the climate variables. The scope of this research work are listed as follows:

- Analyze the climate variables at variable lead times for monthly SPEI predictions.
- Introduce a robust temporal convolutional feature network (TCFN) to extract adequate local features.
- Minimize the prediction errors of the core predictors using the flexible error correction (FEC) approach.

The Objectives of the proposed work are listed as follows:

- To transform the original input data from high-dimensional to low-dimensional while retaining important features through the introduction of a novel Attention-Driven LSTM Autoencoder (AttLSTMAE).
- To propose a new ensemble deep learning model based on MLP, CNN, TCFN and AttLSTMAE for SPEI forecasting. It can improve the generality and resilience of the entire model based on the idea of adaptable extraction of inherent features within the climate variables.
- To eliminate redundant learners according to the similarity and diversity-based pruning method.

The structure of this paper is organized as follows. Section 2 describes recent draught forecasting methods. Section 3 explains in detail the proposed forecasting model. Section 4 validates the performance of the proposed method through simulation. Finally, the paper is concluded in Section 5.

2. Related Works

The most susceptible societies can be warned of impending droughts and prepared for their negative effects with the help of drought predictions. Wan et al. (2023) analysed the temporal and spatial patterns of the drought period and harshness using Theil-Sen and Mann-Kendall (M-K) tests. Furthermore, the association between drought characteristics and climate parameters has been investigated using partial correlation analysis. According to this study, decision makers could develop an efficient measure to mitigate the negative social and ecological impacts related to climate change by knowing the primary climate elements that cause drought episodes. The author examined the possibility of creating drought prediction models using different machine learning methods: Support Vector Machine (SVM), Artificial Neural Network (ANN), and k-Nearest Neighbour (KNN). These models were used to estimate three classes of droughts: moderate, severe, and extreme, taking into account different cropping cycles. Furthermore, a unique feature selection method was applied for the first time in drought modelling to find the best possible set of predictors.

Al Moteri et al. (2024) introduced a hybrid Convolutional long short-term memory with self-attention for forecasting the shoreline drought because of its ability to capture intricate interactions between climate parameters. The effectiveness of the LSTM model on the prediction of draughts has been validated by considering several drought factors, including the severity of the drought, the classification of the drought, or geographical variation. Dikshit et al. (2023) aimed to predict the widely used drought measure, SPEI, using a stacked LSTM model. Here, the hydroclimatic indicators including temperature, PET, rainfall, and cloud cover were used along with some meteorological measurement. The results of this paper showed that the prediction abilities at an extended forecasting horizon can be improved by lagged climatic variables.

The work used Gene Expression Programming, Model Tree, and Multivariate Adaptive Regression Spline models to compute SPEI values for different climates. These models were executed using meteorological data such as wind speed, rainfall, relative humidity, maximum, lowest temperatures, and average temperatures. CNN-LSTM is a new hybrid intelligence model that has been developed and verified for short-term climate-based drought projection. This model was used to anticipate multiple time-scale drought indicators, specifically three- and six-month SPEI. The effectiveness of this model was verified using statistical accuracy measurements and graphical examinations. According to the results, CNN-LSTM performed better than all the benchmarks.

This investigation demonstrates that data-driven models are generally chosen for forecasting weather-related water and parameters. In the previous ten years, several researches investigated the use of numerous intelligent data models, including SVM, ANN, and kNN, to considerably forecast the draught. But these independent machine learning techniques lead to overfitting for large datasets due to the intricate and non-linear interactions between the predictors. To overcome the limitations of individual models, deep learning (DL) techniques such as CNN and LSTM have been developed and have been shown to produce greater precision. Although these individual learners provide ease of use and computational speed, their generalizability, robustness, and scalability are frequently constrained. Nevertheless, ensemble learning approaches can reduce these drawbacks and improve overall model performance through the aggregation of predictions from several models. The final ensemble performance is greatly influenced by the modelling and optimization techniques used at each step. Therefore, the objective of the current work was to improve the performance of the ensemble by applying error correction and pruning techniques.

3. Materials and methods

3.1. Data and Selection of Drought Index

In this paper, the daily weather records given by the CRU TS v 4.03 dataset are utilized. This data set has been used in a number of research projects, including agricultural, ancient climate, and climatic variation investigations. The dataset offers ten distinct principal and auxiliary variables. The variables used in this study are divided into three categories: principal variables, which include mean temperature and precipitation; auxiliary variables, which include cloud cover and vapour pressure; and derivative variables, which include minimum and maximum temperatures and potential evapotranspiration. Drought indices are valuable measurements for identifying, tracking and measuring drought occurrences. The most commonly utilized draught measurement is SPEI. It depends on both rainfall and temperature data, while SPI depends only on rainfall data. The factors used to calculate SPEI are precipitation and PET as given below:

$$D' = P' - PET \tag{1}$$

where P' denotes the precipitation (millimetre) and PETdenotes the potential evapotranspiration (millimeter). The D' series is fitted with different log-logistic distributions to compute SPEI. The PDF and CDF of different log-logistic distributions are defined as follows:

$$g(x) = \frac{\alpha}{\delta}\left(\frac{X-\vartheta}{\delta}\right)^{\alpha-1}\left[1+\left(\frac{X-\vartheta}{\delta}\right)^{\alpha}\right]^{-2} \tag{2}$$

where δ, α and ϑ are the magnitude, silhouette and source variables, respectively. Also, the L-moment process is used to obtain the log-logistic distribution variables as provided as follows:

$$\alpha = \frac{2\omega_1 - \omega_0}{6\omega_1 - \omega_0 - 6\omega_2} \tag{3}$$

$$\delta = \frac{(\omega_0 - 2\omega_1)\alpha}{\Gamma\left(1+\frac{1}{\alpha}\right)\Gamma\left(1+\frac{1}{\alpha}\right)} \tag{4}$$

$$\vartheta = \omega_0 - \delta\Gamma\left(1+\frac{1}{\alpha}\right)\Gamma\left(1+\frac{1}{\alpha}\right) \tag{5}$$

the following expression:

where ω_0, ω_1 and ω_2 denotes the weighted probability statistics and are computed usi

$$W_k = \frac{1}{m}\left(\frac{m-1}{k}\right)^{-1}\sum_{i-1}^{m-r}\square\left(\frac{m-1}{k}\right)x_i, \qquad k = 0{,}1{,}2 \tag{6}$$

where m denotes the dimension of the sample and x_i represents the descending ordered vector descending of the data points. Then, the predictable Pearson-III distribution parameters are used to compute the CDF of log-logistic dissemination.

$$G(x) = \left[1+\left(\frac{X-\vartheta}{\delta}\right)^{-\alpha}\right]^{-1} \tag{7}$$

Here, the time series of $(P' - PET)$ is modelled across many time scales using log-logistic distribution variables. Moreover, the Kolmogorov-Smirnov (K-S) test is used to validate the fitted log-logistic distribution variables for hydrological water balance data of D'. With the values of (x) , the SPEI values were calculated as given below:

$$SPEI = W - \frac{C_0+C_1W+C_2W^2}{1+d_1W+d_2W^2+d_3W^3} \tag{8}$$

Where $W = \sqrt{-2ln(p)}\ for\ W \leq 0.5$, p denotes the likelihood of surpassing a detected D' value, $p = 1 - G(x)$. When $p > 0.5$, p is substituted by $1 - p$ and the sign of the resulting SPEI is inverted. The coefficients are $C_0 = 2.55155170$, $C_1 = 0.8028530$, $C_2 = 0.0103280$, $d_1 = 1.4327880$, $d_2 = 0.1892690$, $d_2 = 0.0013080$.

Different types of draught can be illustrated by calculating SPEI at varying time scales from one month to twenty-four months. In general, meteorological drought is best described by shorter time scales (1-3 months), agrarian drought is best described by longer time scales (3-6 months), and hydrological drought is best described by longer time scales (12-24 months). Access to the global SPEI database using the CRU data set on various monthly scales is available at https://spei.csic.es/database.html. After the data are calculated, they might be utilized to comprehend various aspects of the drought. The categories of different draughts based on SPEI values are provided in Table 1.

Table 1. Draught Classes Based on SPEI Range

SPEI range	Classes
≤ -2.00	Extremely Dry (ED)
$-1.99 \sim -1.50$	Severely Dry (SD)
$-1.49 \sim -1.00$	Moderately Dry (MD)
$-0.99 \sim 0.99$	Near Normal (NN)
$1.00 \sim 1.49$	Moderately Wet (MW)
$1.5 \sim 1.99$	Severely Wet (SW)
≥ 2.00	Extremely Wet (EW)

3.2 Model development

After the SPEI data collection, predictor values were gathered from related sources. This study uses high-resolution hydroclimatic predictors such as temperature (minimum, maximum, and mean), precipitation, cloud cover, and PET. In this work, a new ensemble learning model with optimized pruning (EnsLOP) is proposed to anticipate month-wise SPEI at various lead times. This paper mainly aims to offer an adequate testing dataset in addition to the largest possible input dataset for training. Thus, the parameter obtained from 1901 to 1990 is used for training, and the remaining data from 1990 to 2018 is used for testing purposes. Figure 1 shows the structural configuration of the proposed EnsLOP model for the prediction of SPEI. The three phases of the model are data preparation, core predictor development (CPD), and core predictor fusion (CPF). In the initial phase, the raw high-resolution data are partitioned using the cross-validation method. During the CPD phase, the training and testing set are utilized to construct a sequence of core predictors. In addition, a flexible error correction (FEC) technique is proposed to address all the core predictor predictions. The final ensemble model is created in the CPF phase by combining the core predictors using a stacking-basis ensemble approach. Finally, a similarity index and a divergence-based ensemble pruning approach is introduced to improve the accuracy and steadyness of the ensemble model.

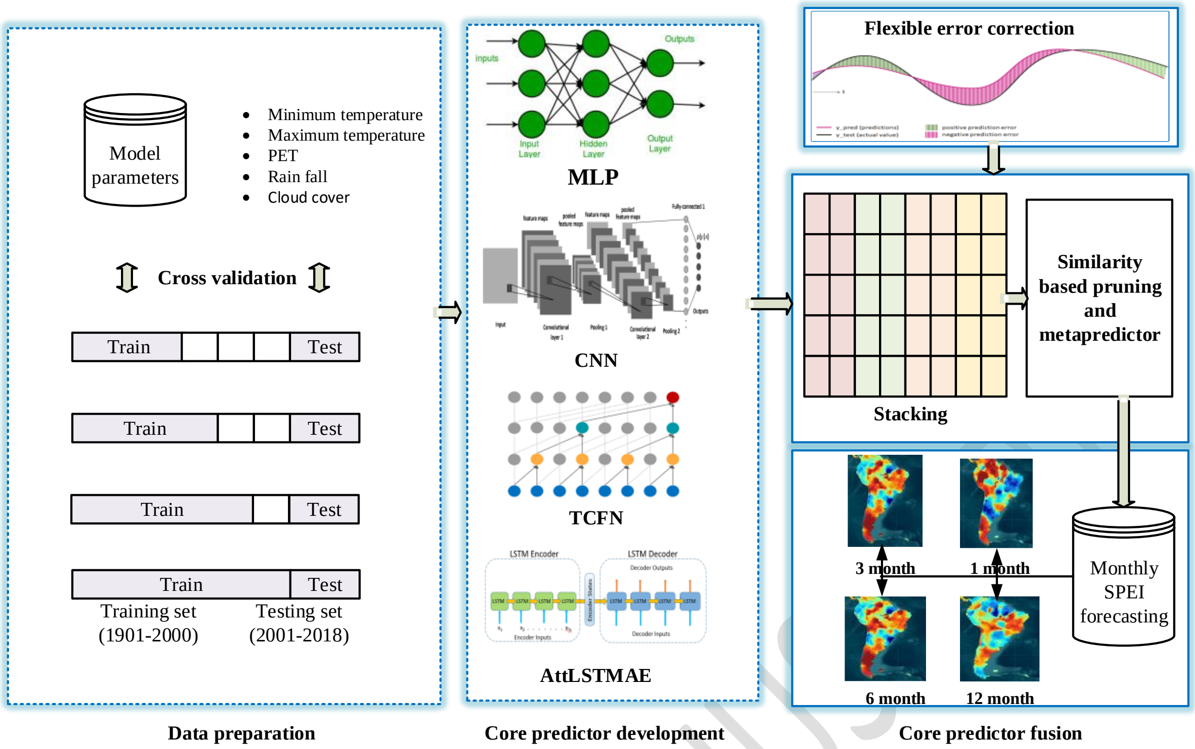


Figure 1. Proposed architecture for draught forecasting

3.2.1 Development of the core predictor

Different kinds of data features can be effectively extracted by specific kinds of Deep neural networks (DNNs). The ensemble learning uses different machine learning methods (i.e., core learners) for solving a problem and then merged to provide superior outcomes. In this work, a MLP, CNN, temporal convolutional feature network (TCFN) and an Attention-driven LSTM autoencoder (AttLSTMAE) are proposed as the core predictors for the ensemble model. MLP denotes a feedforward artificial neural network. Its network structure is more straightforward than that of other neural networks; it consists primarily of three layers: an input layer, a hidden layer, and an output layer. CNN uses a structure similar to a grid to signify and extract data features. A CNN applies a number of distinct convolution kernels (weight matrices) to the local data for creating feature maps with various feature information. After that, more abstract data features are extracted by convolving these feature maps.

3.2.2 Temporal convolutional feature network

As seen in Figure 2, the fundamental building block of TCFN is a "Conv1D," which is responsible for extracting local characteristics from the input. Additionally, TCFN uses 2 multiple head convolutional neural layers, each made up of three Conv1D blocks to find higher-level multiple scale features from the previously extracted low-level features. In addition, a self-attention layer is placed between the two multiple head layers to correlating the locations of the local features acquired from the first multiple head layer and enhancing the input features of the second multiple head layer.

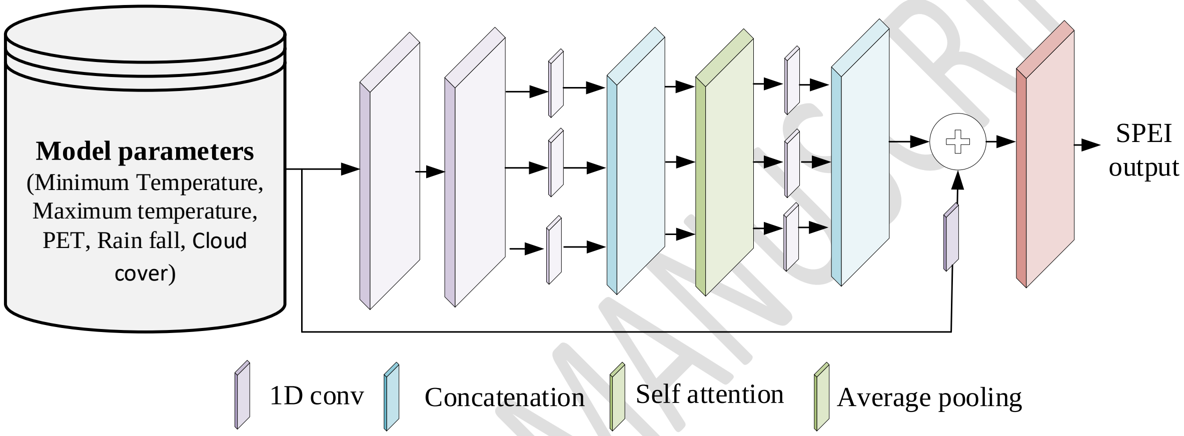


Figure 2. Architecture of TCFN

A 1DConv unit contains a 1D convolution, a batch normalization (BN) and a leaky rectified linear unit (Lky-ReLU) activation function, as given as follows:

$$Out_{1DC} = A_{L-ReLU}\left(A_{BN}\left(A_{conv}(s)\right)\right) \qquad (9)$$

where, Out_{1DC} and s denote the outcome and income of the 1DConv unit correspondingly. A_{L-ReLU}, A_{BN}, and A_{conv} represent the Lky-ReLU activation, BN, and convolution functions of Lky-ReLU, respectively. The convolution unit is utilized to explore the local features from the input as given below:

$$A_{conv}(s) = \varpi_{cnn} \otimes s + B_{cnn} \qquad (10)$$

where, ϖ_{cnn} and B_{cnn} denote the weight and bias values of CNN correspondingly. $\otimes$ represents the convolution operator. Consider $s_{BN} = \{x_1, x_2, \dots x_M\}$as the input of BN unit. Here, x_j and

Mdenote the j -th sample and batch dimension. Also, $\ddot{\mu} = \frac{1}{M}\sum_{j=1}^{M} \square\, x_j$and $\varepsilon = \sqrt{\frac{1}{M}\sum_{j=1}^{M} \square\, x_j - \ddot{\mu}}$ represent the mean and standard deviation of s_{BN}, respectively. $A_{BN}(s_{BN})$ is described as:

$$A_{BN}(x_1, x_2, \dots x_M) = \left(\vartheta \frac{x_1 - \ddot{\mu}}{\varepsilon + \rho} + \tau, \vartheta \frac{x_2 - \ddot{\mu}}{\varepsilon + \rho} + \tau, \dots \vartheta \frac{x_M - \ddot{\mu}}{\varepsilon + \rho} + \tau\right) \tag{11}$$

where, $\vartheta \epsilon R^{+}$ and $\tau \epsilon R$ denote the learning parameters and $\rho > 0$ represents small random value.

The BN unit guarantees a quicker training process by eliminating the internal covariate shift. Furthermore, the capability of extracting local features is enhanced by regularizing the proposed model using BN. In contrast to the RELU, which takes only positive values into account, the Lky-ReLU incorporates both positive and negative values. As a result, the loss of characteristics throughout the data transfer process is minimized. The Lky-ReLU activation can be mathematically modelled as

$$(s_{actvn}) = \{\varsigma s_{actvn} \quad , s_{actvn} < 0 \; s_{actvn} \quad , s_{actvn} \geq 0 \tag{12}$$

where, s_{actvn}denotes the input of the Lky-ReLU and ς represents the negative number's coefficient.

3.2.3 Attention-Driven LSTM Autoencoder (AttLSTMAE)

This model integrates an LSTM with auto encoder (AE), and the resultant model is improved using an attention method to selectively focus on the input data. Figure 3 illustrates the structure of the proposed AttLSTMAE. Here, the LSTM networks are used as the encoder and decoder model of the AE. The high-dimensional input data series is used as static vector input to the encoder. The data handled by the encoder technique maintain dependences between different data points within a time-series sequence through the usage of LSTM memory cells. It also continuously reduces the higher-dimensional input vector into a lower-dimensional vector until it grasps the latent space. The output vector is reconstructed from the compact representation of the input data in the latent space using the LSTM decoder. Additionally, it employs reconstruction error rates to detect SPEI.

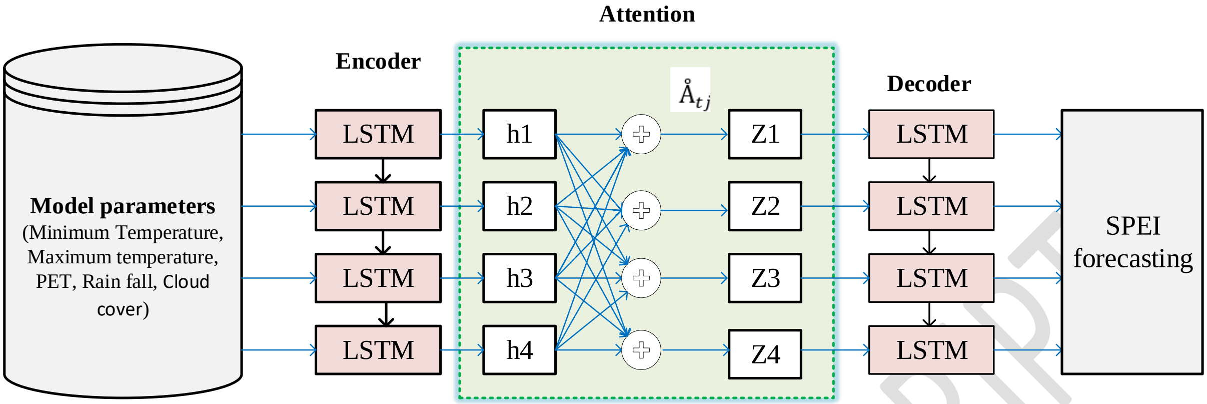


Figure 3. Structure of AttLSTMAE

Step1: Input Sequence Data

The input data is a time series data$\{x_1, x_2, \dots x_N\}$. From this data, a static V-length time window data $\{x_1, x_2, \dots x_V\}$is initially generated where $x_t \in R^n$ denotes a n -features input at time t. Then, the proposed model reshapes these data as a 2-dimensional matrix.

Step2: LSTM Encoder with Attention Mechanism

The LSTMAE interacts with several LSTM units to recognize the most significant features of the input. In the LSTM model, the memory units replace the RNN summation units. The gating mechanism used by LSTM memory blocks allows the network to retain and access data for extended periods of time. These gates decide whether the cell state data should be updated, maintained, or removed by the LSTM unit. Although the output of the LSTM unit is a function of all previous time steps, it may not be able to efficiently collect data on long-term inputs due to its small memory. The effectiveness of LSTM can be improved using an attention method while handing long-term input data. This attention method allowed the neural network to focus on the more crucial information in the input data. In the proposed model the attention method is placed in the space between the two layers of the LSTM network to give discerning significance to the input data. Initially, the attention method computes g_{tj}at every time step according to the hidden state $\hbar_t$ of the LSTM encoder as provided as follows:

$$g_{tj} = tanh\left(\varpi_a\left[u_{t-1}, \hbar_j\right]\right), \qquad g_j \in [-1,1] \tag{13}$$

Where u_{t-1}denotes the LSTM unit's hidden state at one time step earlier and ϖ_a denotes a weight matrix that is fine-tuned throughout the training procedure. Also, g_{tj}stands for a placement model score, describing the relations between an input at location j and an output at location t. Subsequently, this score undergoes normalization through a Softmax function as

$$Å_{tj}=\frac{exp(g_{tj})}{\sum_{i=1}^{T} exp(ti)} \tag{14}$$

Next, the semantic vector is formed using the normalized score provided as follows.

$$Z_t = \sum_{t=1}^{T} Å_{tj}\hbar_j \tag{15}$$

The above semantic vector is utilized for the calculation of the hidden state of the subsequent layer as:

$$u_t = tanh(\varpi_b[u_{t-1}, y'_{t-1}, Z_t]) \tag{16}$$

where y'_{t-1}represents one step earlier output.

Step3: LSTM Decoder

The primary function of the LSTM decoder is to function as a series unfolding layer, recovering the output time series structure after series folding.

3.2.4 Flexible error correction method

The forecast accuracy of the final ensemble model is derived from the forecast accuracy of the constituent core predictors. As a result, the forecast series of the core predictors must be thoroughly examined and corrected before combining the models. It needs an error correction model to correct future prediction results. This work introduced a new FEC approach to correct the errors of the core models. Initially, the error (E_n) sequence at l time points prior to time

$n+1$ is gathered in order to anticipate the SPEI value. In this case, l is found to be three by using the trial-and-error method.

$$E_n = SPEI_n - \underline{SPEI_n} \tag{17}$$

where$SPEI_n$, $\underline{SPEI_n}$and E_n denote the actual SPEI, forecasted SPEI, and error, respectively. Then the proposed algorithm corrects the forecasting value $SPEI_{n+1}$ as:

$$\underline{SPEI'_{n+1}} = \underline{SPEI_{n+1}} + \sum_{n=1}^{l} \square \left(SPEI_n - \underline{SPEI_n}\right) \times \frac{\left|SPEI_n - \underline{SPEI_n}\right|}{MAE} \times exp(-n) \tag{18}$$

where MAE denotes the maximum absolute, and $\underline{SPEI'_{n+1}}$ is the adjusted SPEI value. The above expression consists of two terms to correct the error series. The term$\left(SPEI_n - \underline{SPEI_n}\right) \times \frac{\left|SPEI_n - \underline{SPEI_n}\right|}{MAE}$ denotes the size and directional tendency of error, while the term $exp(-n)$ denotes the time decaying parameter. This shows that the impact of the past forecast inaccuracy on the forecasting value at time $n+1$ slowly decreases with increasing time. Therefore, the proposed FEC uses past error series data flexibly to achieve real-time error correction during SPEI forecasting.

3.2.5 Core predictor fusion using an optimized pruning method

One of the crucial steps in the ensemble building process is core predictor fusion. The primary idea behind stacking is to combine the output of each core predictor to create new features. These features are then sent to the next-stage metapredictor to construct the correspondence between the output of the core predictor and the actual SPEI. The existing accuracy-based pruning removes the effective east members of the ensemble. In this work, an optimized ensemble pruning process is introduced for producing the best fusion model and lowering the generalization error of the ensemble model. This method takes into account both the diversity of the core predictors and their predictive outcome for pruning. Figure 4 illustrates the incorporation of a two-stage pruning technique.

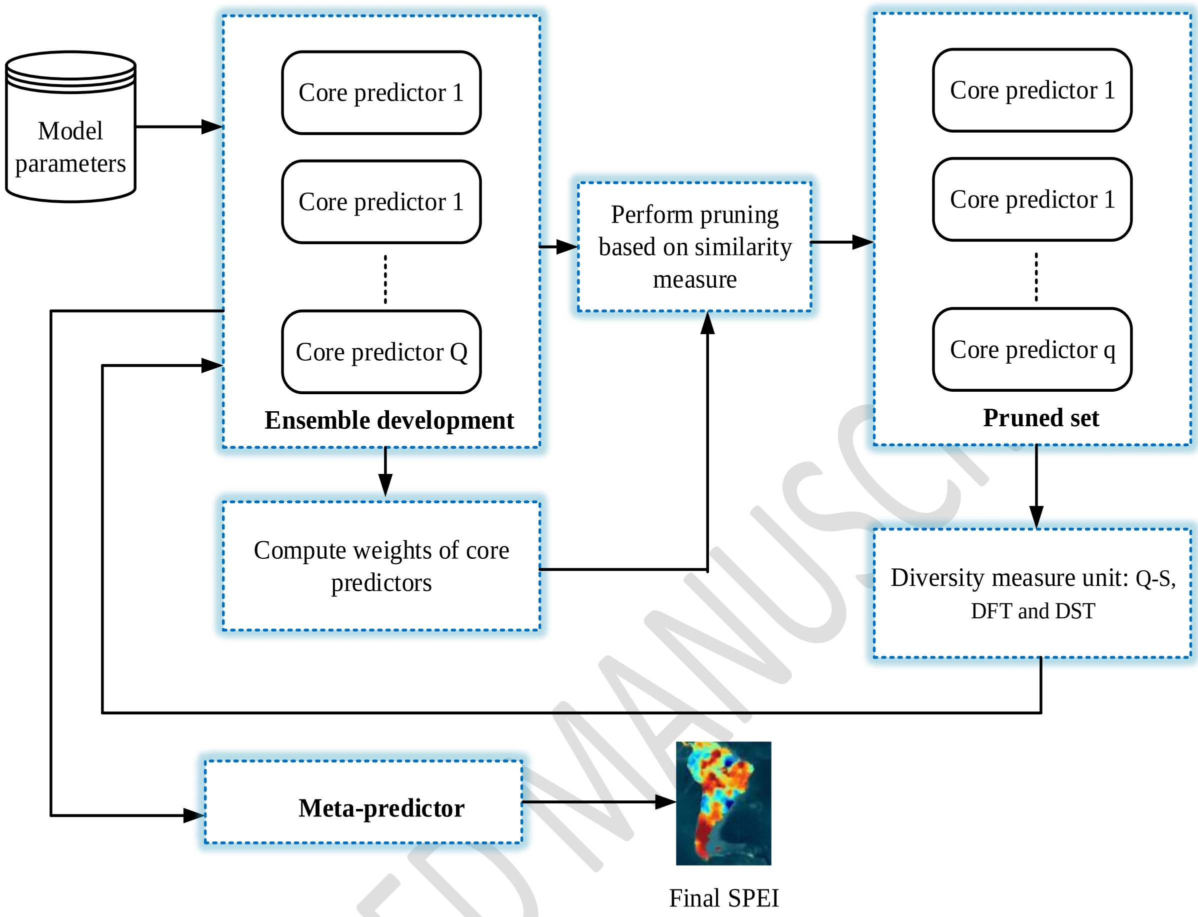


Figure 4. Optimized ensemble pruning

Initially, a collection of $q$poor predictors (L^{q}_{dive}) are chosen from Qtotal predictors in the ensemble (E). These selection processes are carried out through the computation of weights of corresponding predictors on present input. The predictor with the lowest weight is regarded the worst in terms of precision. At the initial level of pruning, the predictors are compared to one another, and the conceptual equivalency is used to identify similar predictors. The prediction value of two predictors that are considered for comparison across all the $X_i = \{x_1, x_2, \dots x_N\}$ is represented as L_m and $L_n$respectively. Subsequently, the similarity index is computed using the following expression:

$$P_v(X_i, L_m, L_n) = \{1, \quad if L_m(X) = L_n(X)\ 0 \qquad otherwise \tag{19}$$

$$SI(X_c, L_m , L_n) = \frac{\sum_{\square}^{\square} \square P_v(X, L_m , L_n)}{\sum_{i=1}^{X_c} \square X_i} \tag{20}$$

Where, P_v, SI and $X_c$stands for predicted value, similarity index, and present correspondingly. When the predictors L_m and L_n exhibit similarity, one of them is excluded from the ensemble, because they have been trained in the same way and represent the same notion. Subsequently, the poor predictors $L_{dive}^{q}$are chosen from Qtotal predictors. Then $L_{dive}^{q}$is sent to diversity checking unit that calculates the diversity of the ensemble E for removing the predictor $L_q \in [L_{dive}^{q}]$. Here the predictor q being the one whose elimination results in the greatest diversity within the system.

Three widely used diversity measures including disagreement (DST), double fault (DFT), and Q-static (Q-S) have been examined in the second level of the pruning stage. When the predictor L_q forecasts the SPEI of input series accurately then $SPEI(X) = 1$. Alternatively, $SPEI(X) = 0$ when it forecasts SPEI wrongly. Let V^{ij}be the amount of training data whose forecast is i and j $(i, j \in (0,1))$ for predictors L_m and L_n correspondingly. The diversity measures Q-S, DFT and DST can be computed using

$$Q\_S(L_m, L_n) = \frac{V^{00}V^{11} - V^{01}V^{10}}{V^{00}V^{11} + V^{10}V^{01}} \tag{21}$$

$$DFT(L_m, L_n) = \frac{V^{00}}{V^{00}V^{11} + V^{10}V^{01}} \tag{22}$$

$$DST(L_m, L_n) = \frac{V^{01} + V^{10}}{V^{00}V^{11} + V^{10}V^{01}} \tag{23}$$

Finally, a meta-predictor is constructed over the current input for SPEI prediction.

4. Results and discussion

The proposed drought forecasting model is simulated using Python programming language. Figure 5 analyzes the predicted results with respect to several drought features for four different lead times: one month, three months, six months, and twelve months. It is not possible to depict all the anticipated outcomes during the testing phase. Consequently, the first sample of the predicted SPEI sequence was shown. For example, a one-month lead time is displayed for

January 2016. Similarly, a lead time of three months is shown for March 2016, a lead time of six months for June 2016, and a lead time of twelve months for December 2016.

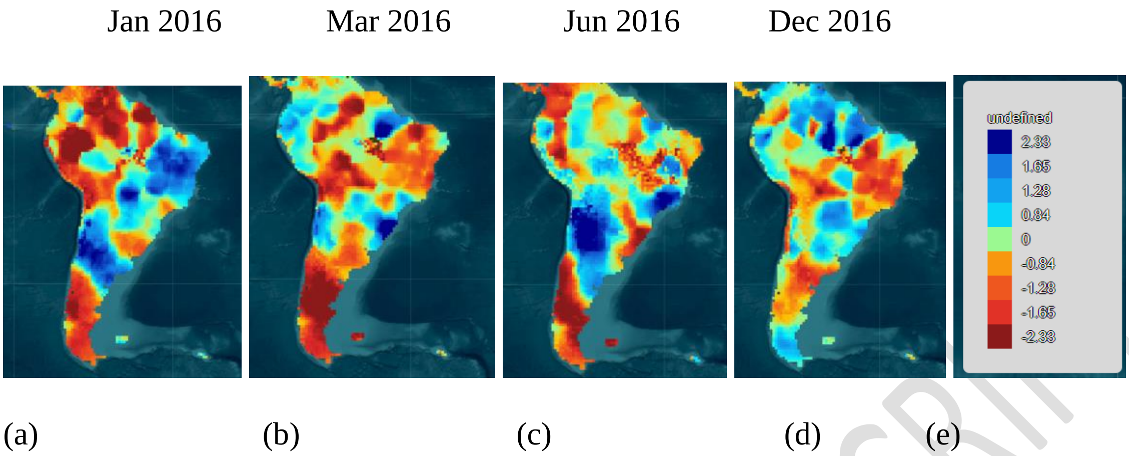
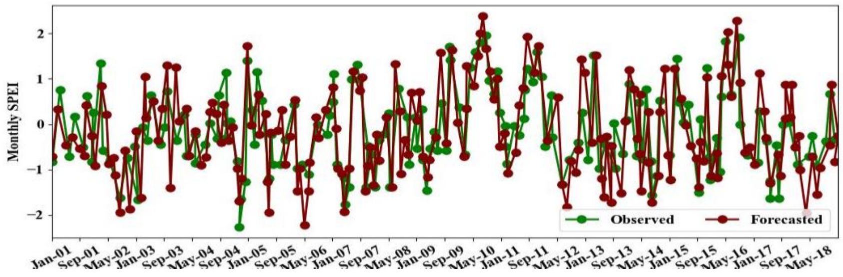


Figure 5. Predicted SPEI 1 values with lead times of a) 1 month, b) 3 months, c) 6 months, and d) 12 months (e) SPEI scale

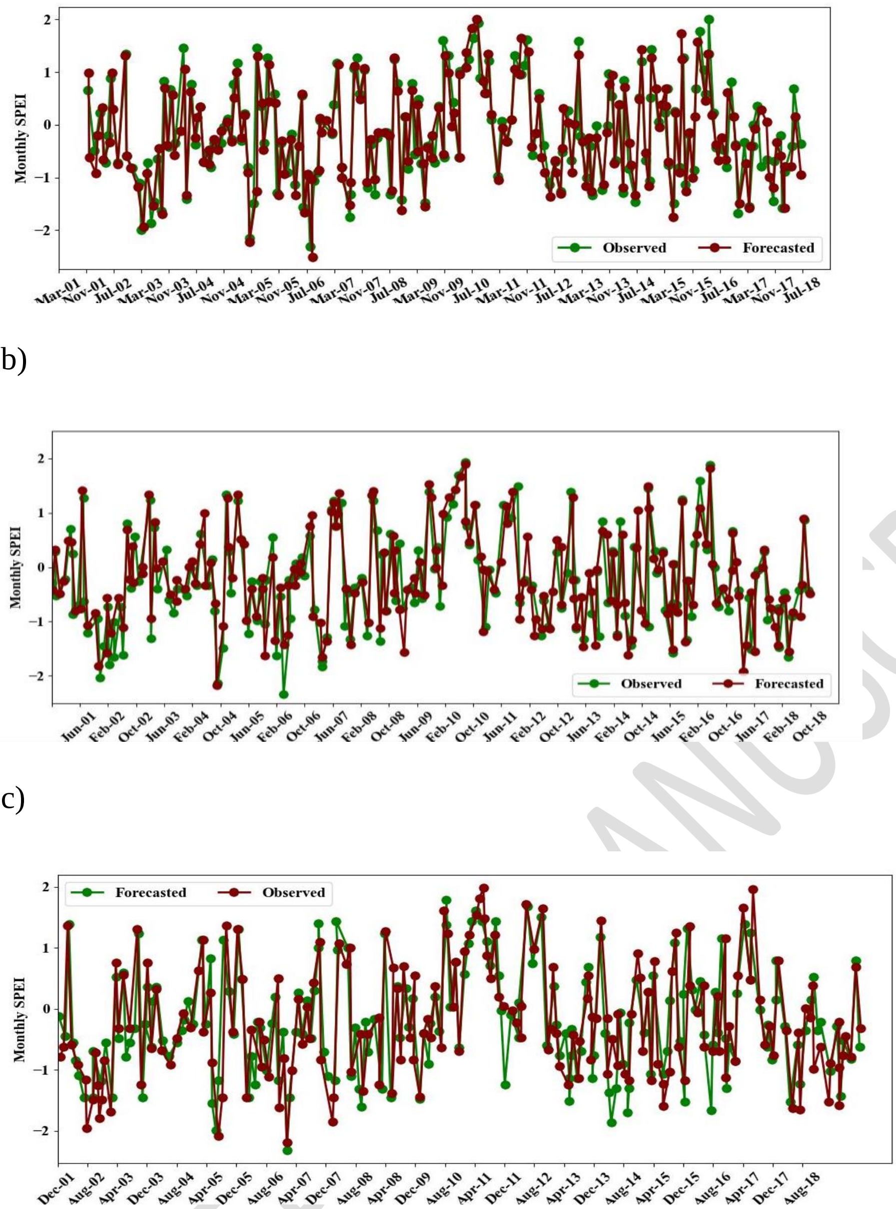
Figures 6 (a) and (d) show the changes in the intensity values of the drought at various lead times. A helpful statistical measure called a threat score (TS) was employed to understand the predicted outcomes in relation to the observed values. TS uses the subsequent expression to calculate the percentage of accurately anticipated results with respect to the observed values:

$$\tau_s = \frac{H}{H+M+FA} \tag{24}$$

Where, $H, M$and FA represents hit, miss and false alarm respectively. TS has a value between 0 and 1, where 0 denotes no talent and 1 represents the best score. According to the findings, TS was 0.97 for a one-month lead time, 0.95 for a three-month lead time, 0.90 for a six-month lead time, and 0.85 for a twelve-month lead time. These findings demonstrate that the model can predict monthly SPEI values with sufficient precision.



(a)



(b)

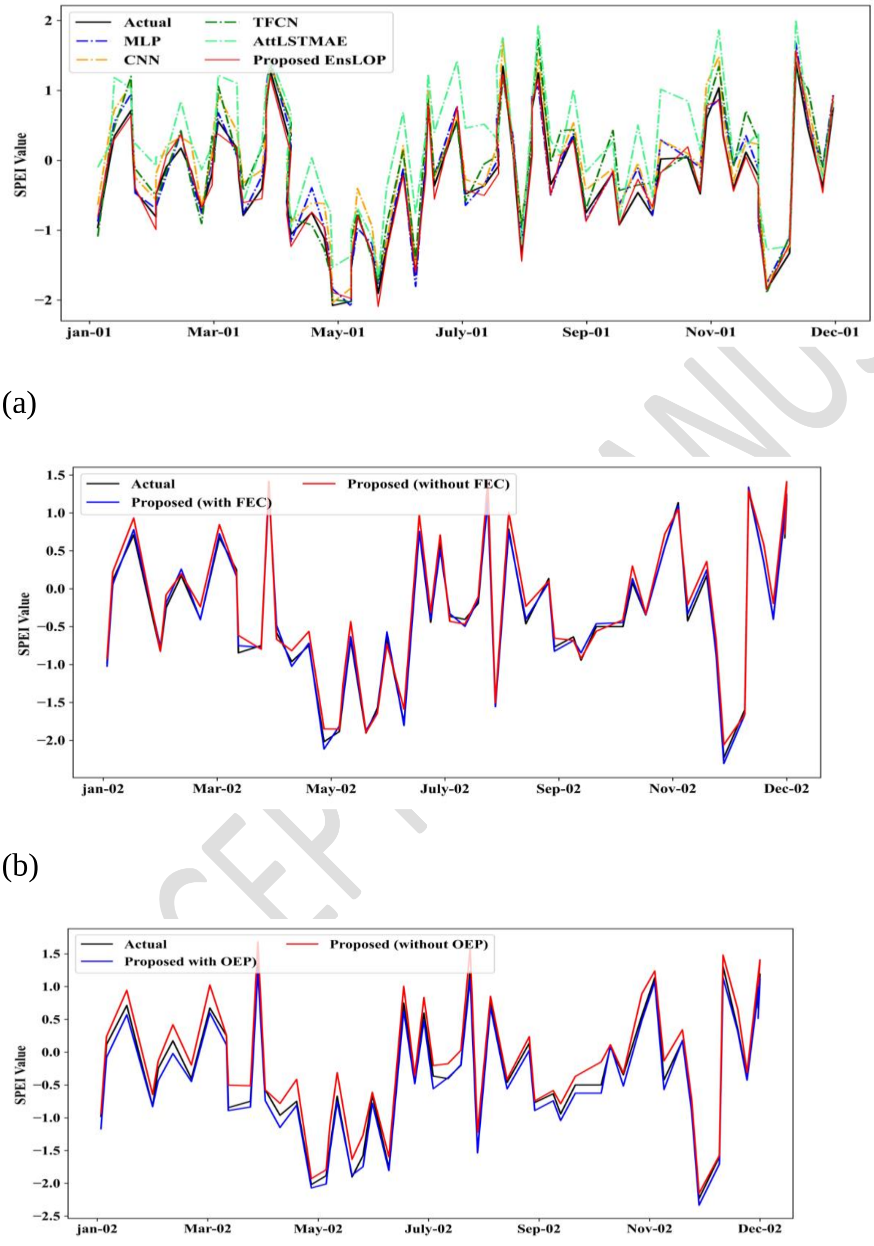
(c)

(d)

Figure 6. Drought intensity values at various lead times (a) 1 month (b) 3 months (c) 6 months (d) 12 months

4.1 Ablation study

This section examines the impact of individual deep learning models in the ensemble. Figure 7 (a) demonstrates that the EnsLOP prediction curve is more closely aligned with the actual value compared to the other four distinct deep learning models. Additionally, Figure 7 (b) illustrates the impact of the proposed FEC method in EnsLOP. It demonstrates that the ensemble model's forecasting curve is more similar to the real curve, while using the FEC processing. In Figure 7 (c), the forecast curve of EnsLOP is closer to the actual curve, while the forecast curve of EnsL without optimized ensemble pruning (OEP) shows poor performance. These analyses verify the usefulness of each technique used in the suggested model, allowing the dominance of the suggested EnsLOP approach.



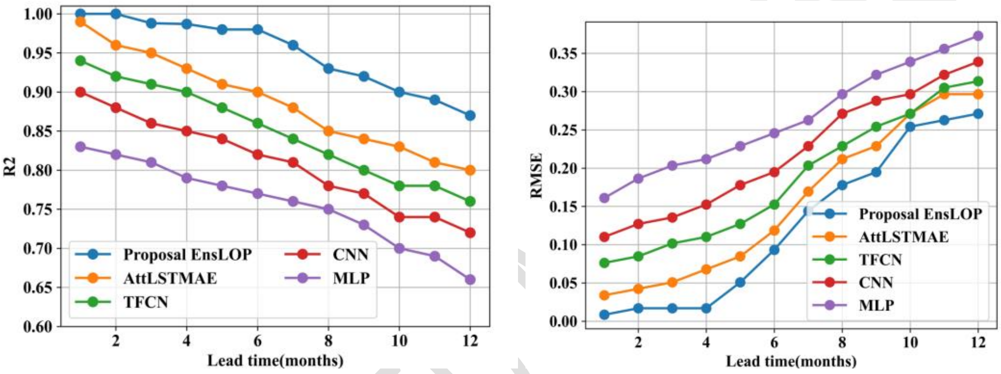
(a)

(b)

(c)

Figure 7. Ablation study with (a) individual core predictors, (b) FEC approach, and (c) optimized ensemble pruning (OEP)

The root mean square error (RMSE) and the coefficient of determination (R2) were used as statistical metrics to analyze the effectiveness of the proposed framework. RMSE is a useful metric for forecasting and penalizes big errors. The degree of correlation between the predicted and observed values is shown by the R2 value. This R2 value is a number between 0 and 1, where 0 means there is no relation and 1 represents an accurate match. However, a smaller RMSE score indicates better performance. The performance of the suggested EnsLOP is compared with that of the core predictors, CNN, TCFN, MLP and AttLSTMAE in Figure 8. It shows that the forecasting results of individual core predictors perform poorly compared to the suggested EnsLOP.



(a) (b)

Figure 8. Comparative analysis (a) R2 and (b) RMSE

The performance of the suggested draught prediction is compared with the state-of-the-art methods in Table 2. It is evident that the suggested EnsLOP performs better than any of the current models by achieving RMSE of 0.098 and R2 of 0.98. The reason for this is that the large-scale variance of climate data may be too much for traditional CNN, SVM, and KNN models to handle when trying to extract features with varying scales. The stacked LSTMs do not have the capability for handling temporal dependences which are lengthier than a certain step. When trained on a long-term dependency dataset (for example, 100 steps), the network encountered difficulties in learning the task. As a result, the effectiveness of these methods is not good enough for draught forecasting. However, the proposed model integrates the benefits of different models to achieve the best results.

Table 2 Comparative analysis with state-of-the-art methods

References	**Datasets: Climate variables**	**Model**	**RMSE**	**R2**
Surendran R., et al (2023)	PGF- version 3/ Air temperature, geopotential height, relative humidity, wind, Sea level pressure	SVM	0.33 @ 6-month lead	0.96 @ 6-month lead
		ANN	0.49 @ 6-month lead	0.95 @ 6-month lead
		KNN	0.62 @ 6-month lead	0.75 @ 6-month lead
Dikshit et al (2021)	CRU: Temperature, PET, rainfall, cloud cover and climatic indices	Stacked LSTM	0.11 @ 6-month lead	0.92@ 6month lead
Bentsen et al (2023)	synoptic stations at Iran: Temperature, humidity, rainfall and wind speed	GEP	0.250	-
		MT	0.107	-
		MARS	0.148	-
Proposed	CRU: Temperature, PET, rainfall, cloud cover	EnsLOP	0.098 @ 6-month lead	0.98 @ 6-month lead

5. Conclusions

This study presented a novel ensemble learning with an optimized pruning model for draught forecasting. High-resolution hydroclimatic variables, including temperature (minimum, maximum and mean), precipitation, cloud cover, and PET, were employed to validate the effectiveness of the techniques in the suggested model and the dominance of the general ensemble draught forecasting model. During the testing time, the suggested EnsLOP model predicts SPEI at varied lead periods and performs better than other baseline models. The proposed FEC approach aimed to reduce the forecasting error by taking into account the comparative pattern and temporal deterioration in the error series. The predicted results were evaluated using statistical metrics and looking at various aspects of the drought. The results demonstrated that the EnsLOP model performs better than conventional data-driven models in terms of statistical metrics. Regional drought management planners may find this study highly beneficial in planning for future potential drought conditions. The suggested approach can produce good forecasting results; however, it is not directly scalable to multidimensional data. Extending the suggested model to multivariate and multistep time series forecasting is one potential avenue for future research.

Data Availability Statement:

The source of data sets are download from the following link

https://www.ncdc.noaa.gov/cdo-web/datasets and

http://apdrc.soest.hawaii.edu/data/data.php

References

Yang, Shuai, Mou Leong Tan, Qixuan Song, Jian He, Nan Yao, Xiaogang Li, and Xiaoying Yang (2023), Coupling SWAT and Bi-LSTM for improving daily-scale hydro-climatic simulation and climate change impact assessment in a tropical river basin, *Journal of environmental management* **330**, 117244.

Kumar, Nikhil, Vikas Poonia, B. B. Gupta, and Manish Kumar Goyal (2021), A novel framework for risk assessment and resilience of critical infrastructure towards climate change, *Technological Forecasting and Social Change* **165,** 120532.

Yang, Mingxia, Yuling Mou, Yanrong Meng, Shan Liu, Changhui Peng, and Xiaolu Zhou (2020), Modeling the effects of precipitation and temperature patterns on agricultural drought in China from 1949 to 2015, *Science of the total environment* **711,** 135139.

Tamilvizhi T., Surendran R., Romero C.A.T. and Sadish M. (2022), Privacy preserving reliable data transmission in cluster based vehicular adhoc networks, *Intelligent Automation & Soft Computing*, **34**, 1265–1279.

Nguyen, Duy Thao, Saqib Ashraf, Minhhuy Le, and Mustajab Ali. (2023), Projection of climate variables by general circulation and deep learning model for Lahore, Pakistan, *Ecological Informatics* **75** 102077.

Santhanaraj R. K., Rajendran S., Romero C. A. T., and Murugaraj, S. S. (2023). Internet of Things Enabled Energy Aware Metaheuristic Clustering for Real Time Disaster Management. *Comput. Syst. Sci. Eng.*, **45**, 1561-1576.

Liu, Changhong, Cuiping Yang, Qi Yang, and Jiao Wang. (2021), Spatiotemporal drought analysis by the standardized precipitation index (SPI) and standardized precipitation evapotranspiration index (SPEI) in Sichuan Province, China. *Scientific Reports*, **11**, 1280.

Surendran R., Tamilvizhi T., and Lakshmi, S. (2021), Integrating the Meteorological Data into a Smart City Service Using Cloud of Things (CoT). In Emerging Technologies in Computing: 4th EAI/IAER International Conference, iCETiC 2021, Virtual Event, August 18–19, 2021, Springer International Publishing, 4, 94-111.

Sharma, Aditya, Devesh Sharma, and S. K. Panda (2022), Assessment of spatiotemporal trend of precipitation indices and meteorological drought characteristics in the Mahi River basin, India. *Journal of Hydrology*, **605**, 127314.

Wang, Fei, Zongmin Wang, Haibo Yang, Danyang Di, Yong Zhao, and Qiuhua Liang, (2020), A new copula-based standardized precipitation evapotranspiration streamflow index for drought monitoring, *Journal of Hydrology*, **585,** 124793.

Xu, Lei, Nengcheng Chen, Chao Yang, Chong Zhang, and Hongchu Yu. (2021), A parametric multivariate drought index for drought monitoring and assessment under climate change, *Agricultural and Forest Meteorology*, **310,** 108657.

Rehana, Shaik, and G. Sireesha Naidu. (2021), Development of hydro-meteorological drought index under climate change–Semi-arid river basin of Peninsular India. Journal of Hydrology **594,** 125973.

Ullah, Irfan, Xieyao Ma, Jun Yin, Abubaker Omer, Birhanu Asmerom Habtemicheal, Farhan Saleem, Vedaste Iyakaremye, Sidra Syed, Muhammad Arshad, and Mengyang Liu. (2023), Spatiotemporal characteristics of meteorological drought variability and trends (1981–2020) over South Asia and the associated large-scale circulation patterns, *Climate Dynamics*, **60,** 2261-2284.

Surendran R., Alotaibi Y. and Subahi A.F. (2023), Wind Speed Prediction Using Chicken Swarm Optimization with Deep Learning Model. *Computer Systems Science & Engineering,* **46,** 3.

Hu, Xiaolong, Liangsheng Shi, Guang Lin, and Lin Lin (2021), Comparison of physical-based, data-driven and hybrid modeling approaches for evapotranspiration estimation, *Journal of Hydrology*, 601126592.

Dikshit, Abhirup, Biswajeet Pradhan, and M. Santosh. (2022), Artificial neural networks in drought prediction in the 21st century–A scientometric analysis. Applied Soft Computing, **114,** 108080.

Guo, Ning, Hao Chen, Qiong Han, and Tiejun Wang (2024). Evaluating data-driven and hybrid modeling of terrestrial actual evapotranspiration based on an automatic machine learning approach, *Journal of Hydrology* **628**, 130594.

Chao, Wei-Ting, Chih-Chieh Young, Tai-Wen Hsu, Wen-Cheng Liu, and Chian-Yi Liu. "Long-lead-time prediction of storm surge using artificial neural networks and effective typhoon parameters: Revisit and deeper insight, *Water* **12,** (2020): 2394.

Surendran R., Alotaibi Y. and Subahi, A.F. (2023), Lens-Oppositional Wild Geese Optimization Based Clustering Scheme for Wireless Sensor Networks Assists Real Time Disaster Management. *Comput. Syst. Sci. Eng*.**, 46,** 835-851.

Dikshit, Abhirup, Biswajeet Pradhan, and Abdullah M. Alamri. (2021), Long lead time drought forecasting using lagged climate variables and a stacked long short-term memory model, *Science of the Total Environment, **755,*** 142638.

Bentsen, Lars Ødegaard, Narada Dilp Warakagoda, Roy Stenbro, and Paal Engelstad (2023). Spatio-temporal wind speed forecasting using graph networks and novel Transformer architectures, *Applied Energy*, **333,** 120565.

Barzkar, Ali, Mohammad Najafzadeh, and Farshad Homaei. (2022), Evaluation of drought events in various climatic conditions using data-driven models and a reliability-based probabilistic model, *Natural Hazards*, **110,** 1931-1952.

Wan, Lingling, Virgílio A. Bento, Yanping Qu, Jianxiu Qiu, Hongquan Song, RongRong Zhang, Xiaoping Wu, Feng Xu, Jinkuo Lu, and Qianfeng Wang. (2023), Drought characteristics and dominant factors across China: Insights from high-resolution daily SPEI dataset between 1979 and 2018, *Science of The Total Environment,* **901**, 166362.

Al Moteri, Moteeb, Fadwa Alrowais, Wafa Mtouaa, Nojood O. Aljehane, Saud S. Alotaibi, Radwa Marzouk, Anwer Mustafa Hilal, and Noura Abdelaziz Ahmed. (2024), An enhanced drought forecasting in coastal arid regions using deep learning approach with evaporation index, *Environmental Research,* **246,** 118171.

Danandeh Mehr, Ali, Amir Rikhtehgar Ghiasi, Zaher Mundher Yaseen, Ali Unal Sorman, and Laith Abualigah (2023). A novel intelligent deep learning predictive model for meteorological drought forecasting, *Journal of Ambient Intelligence and Humanized Computing*, **14**, 10441-10455.