Big data based architecture for drought forecasting using LSTM, ARIMA, and Prophet: Case study of the Jiangsu Province, China

Hanen Balti, Ali Ben Abbes, Nedra Mellouli, Imed Riadh Farah, Yan-Fang Sang, Imed Riadh Farah, Myriam Lamolle, Yanxin Zhu

Big data based architecture for drought forecasting using LSTM, ARIMA, and Prophet: Case study of the Jiangsu Province, China

1st Hanen Balti
Riadi Laboratory
University of Manouba
Manouba, Tunisia
hanen.balti@ensi-uma.tn

2nd Ali Ben Abbes
Riadi Laboratory
University of Manouba
Manouba, Tunisia
Ali.benabbes@yahoo.fr

3rd Nedra Mellouli
LIASD Laboratory
Universiy of Paris 8
Paris, France
n.mellouli@iut-univ-paris8.fr

4th Yanfang Sang
Key Lab. of Water Cycle
and Related Land Surface Processes
Inst. of Geographic Sciences
and Natural Resources Research
Chinese Academy of Sciences
Beijing, China
sangyf@igsnrr.ac.cn

5th Imed Riadh Farah
Riadi Laboratory
University of Manouba
Manouba, Tunisia
imedriadh.farah@isamm.uma.tn

6th Myriam Lamolle
LIASD Laboratory
University of Paris 8
Paris, France
m.lamolle@iut.univ-paris8.fr

7th Yanxin Zhu
Key Lab. of Water Cycle
and Related Land Surface Processes
Inst. of Geographic Sciences
and Natural Resources Research
Chinese Academy of Sciences
Beijing, China
zhuyx.18s@igsnrr.ac.cn

***Abstract*—Drought disasters significantly affected human life and water resources. Therefore, forecasting methods like statistical models, machine learning, and deep learning architectures help scientists to take effective decisions to decrease the effects of natural disasters by providing decision-making plans. Droughts can be forecasted using meteorological indices like the standardized precipitation evapotranspiration index (SPEI), which aid governments in taking drought-prevention steps. In this paper, we present a big drought architecture for drought modeling and forecasting. The proposed architecture is composed of 5 layers: Data collection, data preprocessing, data storage, data processing and interpretation, and decision making. Besides, we present a comparative study between three different methods ARIMA, PROPHET, and LSTM for drought forecasting. Three different metrics are used for the performance evaluation Root Mean Squared Error (RMSE), coefficient of determination (R2), and Mean Squared Error (MAE). Experiments are carried out using data from the province of Jiangsu. Results revealed that LSTM outperformed the other models, and ARIMA outperformed the PROPHET model.**

***Index Terms*—Data analytics, Big data, Drought, Long-Short Term Memory, ARIMA, PROPHET, SPEI**

I. INTRODUCTION

Drought is an important type of natural disaster, and its frequent occurrences cause massive socio-economic losses worldwide. In China, almost all regions are vulnerable to drought disasters [1]. For example, a severe drought disaster in the Guangxi province in 2004 caused withered crops and massive power losses, due to a complete lack of water to produce hydropower. From Autumn of 2009 to Spring of 2010, South China experienced severe droughts, causing the lack of drinking water for approximately 21 million people [2]. In 2011, drought swept the Chinese territory from northeast to southwest [3], causing a water shortage for about 2.2 million people [4]. Overall, the regions that frequently encounter drought disasters include the Huang-Huai-Hai plain in North China, the Liao River Basin in Northeast China, the Yunnan province in Southwest China, the Guangdong and Fujian provinces in South China, and Northwest China but except the northern part of the Xinjiang province [10]. Besides, there are higher (shorter) durations of drought than the average level in arid and semi-arid (humid) regions. However, the periods of drought have a difference in these regions. The droughts in Northeast China and Southwest China mainly occurred in Spring, but they more occurred in Summer in the lower and middle reaches of the Yangtze River Basin. In autumn, droughts more likely occurred in the lower reach of the Yangtze River Basin and across the Southeastern China coast. Winter droughts mainly occurred in the upper reaches of the Yangtze River Basin and the Yellow River Basin, the Liao River Basin, and Southwest China [5], [6]. Timely and accurate drought monitoring/forecasting is an important approach for the prevention and mitigation of drought disasters [7]. However, with the rapid growth rate of the data volume, traditional models like physical models, for example, the geomorphology-based hydrological model (GBHM) [8] and the Xinanjiang model [9], [10] encountered

978-1-6654-1224-7/21/$31.00 ©2021 IEEE

some limitations which are mainly caused by the high dimensionality, heterogeneity and the non-linearity of data [11]. For solving these problems, artificial intelligence (i.e., big data technologies, machine learning, and deep learning (DL) methods) gains growing popularity in drought monitoring/forecasting studies. Especially, the DL architectures are increasingly used recently for the issue. The DL is an ML technique that has attracted broad attention in several fields. Deep neural network-based learning has lately become one of the fastest-growing and most exciting fields of Big Data science [12]. Neural networks are a collection of models derived from neural biological networks composed of interconnected neurons whose associations can be modified and adapted to the inputs [13]. The DL techniques have been widely applied in analyzing, estimating, designing, filtering, processing, recognition, and detection tasks [14]. It has achieved great success in many applications including drought monitoring/forecasting. Several DL architectures were used in the literature to monitor or forecast droughts in China such as the Deep Neural Network (DNN) [14] (Fig.1.a), Convolutional Neural Networks (CNN) (Fig. 1.b), Deep Belief Networks (DBN) (Fig. 1.c), Long Short-Term Memory (Fig. 1.d), and Recurrent Neural Networks (RNN) (Fig. 1.e).

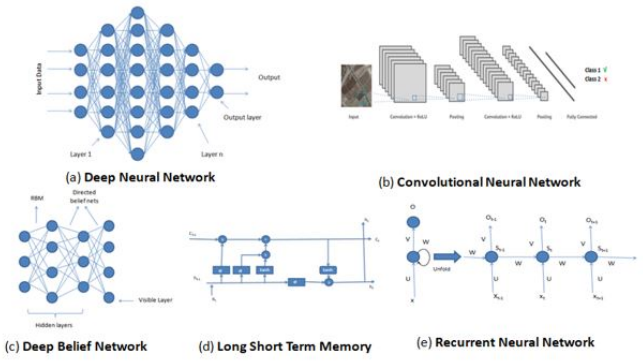

Fig. 1. Deep learning architectures for drought monitoring/forecasting.

These architectures are about learning multiple levels of representations and abstractions that make better sense of data such as images, time series, and texts, and thus indicate better performances [15]. Drought data are highly heterogeneous data. They may include remote sensing data, climate data, biophysical data, and agricultural data. Therefore, it is important to choose the most effective models to process these data to forecast or monitor this phenomenon. In this paper, we present a big data-based architecture for drought forecasting, and further propose a comparative study between three different models for drought forecasting i.e., Auto-Regressive Integrated Moving Average (ARIMA), Prophet, and Long short-term memory (LSTM). The rest of this paper is organized as follows: Section 2 presents a literature review for drought monitoring and forecasting in China, and Section 3 presents the proposed methodology; in Section 4, we experimentally evaluate the proposed methodology and give the conclusion and future works in Section 5.

II. LITERATURE REVIEW

Considering the statistical-based methods, ML-based methods, and DL-based methods that are used for drought monitoring/forecasting over China, an overview of them is presented here, and their ability in processing a massive volume of heterogeneous data is especially discussed.

A. Statistical models

Statistical models and ML-based methods are widely used for drought monitoring/forecasting over China. [16] proposed an approach for drought forecasting in the Sanjiang Plain. They evaluated the performance of ARIMA, Wavelet Neural Network (WNN), and Support Vector Machines (SVM) and proved that ARIMA performs the best. [17] used three models (ARIMA, ANN, and WNN) for drought forecasting in the Hai River Basin. Their results revealed that the WANN is the most suitable for SPI-6 and SPI-12 forecasting. [18] used AR(1), Seasonal Autoregressive Integrated Moving Average (SARIMA), ARIMA for drought forecasting in Guanzhong Plain of China. They proved that the SARIMA can detect changes better than ARIMA and AR(1), and ARIMA gave the best results in drought forecasting. [19] proposed a study for the evaluation and agreement analysis of the drought forecasting results of the AR and SARIMA using the Kappa coefficient. The results showed that the accuracy of SARIMA in forecasting is higher than the AR accuracy. [20] proposed a statistical model, for 1-6 month lead drought forecasting in China, where a statistical component refers to climate signal weighting using support vector regression (SVR), and dynamic modules include the Ensemble Mean (EM) and Bayesian Model Averaging (BMA) of the North American Multi-Model Ensemble (NMME) climatological model. The results proved that the statistical and hybrid models are more suitable than EM and BMA for drought forecasting.

B. Machine Learning methods

[21]developed drought forecasting models based on the Ordinary Least Squares (OLS), penalized linear regression (PLR), Decision Trees (DT), AdaBoost, and Random forest (RF) to predict the SPEI at different timescales of 3, 6, 12, and 24 months in Heilongjiang Province, Liaoning Province, Jilin Province, and the eastern part of the Inner Mongolia Autonomous, Northeast China. The results showed that the PLR model is the best in forecasting SPEI at different timescales. [22] used Distributed lag nonlinear model (DLNM), ANN, XGBoost to predict SPEI based on the Oceanic Niño Index (ONI), Southern Oscillation Index (SOI), Pacific Decadal Oscillation (PDO), North Atlantic Oscillation (NAO), Atlantic Multidecadal Oscillation (AMO) and Interdecadal Pacific Oscillation (IPO). They proved that XGBoost is the best for SPEI predicting. Another ML model used in drought forecasting/monitoring across China is the Artificial Neural Network (NN). In fact, [23] analyzed the spatial and temporal patterns of drought based on model simulation. An ANN model for drought warning was developed using monthly temperature and precipitation data from 1949 to 2015. [24]

proposed the Integrated Agricultural Drought Index as a new drought index (IDI). This index explains the relationship between agricultural drought conditions and a multitude of variables. The IDI is calculated using Remote Sensing data and the BPNN. It can detect drought conditions with a non-stationary relationship. Precipitation, Land Surface Temperature (LST), Normalized Difference Vegetation Index (NDVI), soil water power, and elevation are among the metro-hydrological variables included in IDI. The results indicate that the IDI based on ML algorithms can relax the assumption used in many existing indices that the input and output data are linearly correlated.

C. Deep Learning models

[25] proposed a drought forecast model based on Deep Belief Network (DBN) for forecasting the Standardized Precipitation Index (SPI). Four different scale SPI series were computed at Xiqiao station in Yunnan Province. They proved that the DBN is more suitable for drought forecasting comparing to BPNN and ARIMA. [26] used DBN for precipitation forecasting. The DBN transforms the data feature representation from the original space to another new feature space, in addition to semantic features to improve the accuracy of the forecasting performance. The approach consists of 3 steps; Importing data (reading data from a database, data format conversion), Building a DBN model (initialize model parameter, pre-training model, fine-tuning the model), and finally testing the DBN model. This model was compared to other forecasting algorithms such as SVM with particle swarm optimization (PSO-SVM), SVM with mesh optimization, and SVM with genetic algorithm optimization. The results showed that SVM could give efficient results with small datasets while DBN gives good results when dealing with large-scale datasets. [27] proposed a big data-based approach for precipitation forecasting based on deep belief nets, called DBNPF (Deep Belief Network for Precipitation Forecast). The data used for the experiments are the daily hydrological multivariate time series data of four areas (Zun Yi of Guizhou Province, Hezuo of Gansu Province, Jinan of Shandong Province, and Changchun of Jilin Province) of China from 1956 to 2015. These multivariate time series include the 17 environmental factors such as mean site air pressure, daily maximum pressure, daily minimum pressure, average temperature, max wind speed, sunshine hours, maximum wind direction, maximum wind speed, average wind speed, max wind direction, large-scale water evaporation, small water evaporation, minimum relative humidity, average relative humidity, daily maximum temperature, average water vapor pressure, and daily minimum temperature. The results showed that the DBNPF gave the best results compared to other models such as ARIMA, and SVM. [28] deployed LSTM network models for predicting the precipitation based on meteorological data from 2008 to 2018 in Jingdezhen City. They used different climate variables such as temperature, dew point temperature, minimum temperature, maximum temperature, atmospheric pressure, pressure tendency, relative humidity, wind speed, wind direction, maximum wind, total cloud cover, the height of the lowest cloud, and the amount of cloud. The researchers used different numbers of neurons. The results revealed that using a large number of hidden neurons doesn't always lead to better performance. [29] used convolutional- LSTM (C-LSTM) to estimate precipitation based on well-resolved atmospheric dynamical fields. They compared C-LSTM against the general circulation models (GCM) precipitation product and classical downscaling methods such as SVM in the Xiangjiang River Basin in South China. C-LSTM gave the best performance comparing to SVM, CNN, and Quantile Mapping Method. [30] deployed ANN and LSTM network models for simulating the rainfall-runoff process based on flood events from 1971 to 2013. The results revealed that the two architectures are efficient for rainfall-runoff models and better than conceptual and physical-based models. LSTM models outperformed the ANN models with the values of R2 and NSE beyond 0.9, respectively. [31]proposed the use of a deep CNN (DCNN) to identify and classify maize drought stress. The dataset used in this study contains 656 Gray and RGB (Red Green Blue) outdoor maize images taken from Panasonic camera; including 219 optimum moisture images, 218 light drought stress images, and 219 moderate droughts. The DCNN models used are ResNet50 and ResNet152. There were three treatments in the experiment: optimal moisture, light drought, and moderate drought stress. The results demonstrated a significant performance of the proposed method. The accuracy of identifying and classifying drought stress was 98.14 percent and 95.95 %, respectively, for the entire dataset. The results of the comparison experiments on the same dataset showed that DCNN outperformed the Gradient Boosting Decision Tree (GBDT) model. [32] proposed an automatic detection system for drought stress in the middle growth stage of maize based on CNN architecture. The architecture combines the Gabor Filter and the Sppnet called G-Net. The data was acquired from Zhengzhou, Henan province, China. The dataset contains 1,391 samples, which involve 266 suitable moisture (in whole growth stages) images, 286 mild drought stress images, 283 moderate drought stress images, 292 severe drought stress images, and 264 super drought stress images. They used different directions and wavelengths of the Gabor filter to obtain the texture feature and then constitute a feature matrix after blocking and condensing features. Finally, the data were fed to CNN for secondary feature extraction and classification. The average recognition rate of the experiment is 98.84%.

D. Discussion

In the literature, various models and methodologies were used for drought forecasting like statistical models, ML models, and DL architecture. ARIMA was the most used statistical model. It revealed high performance for forecasting univariate and multivariate time series. For the ML based-models, ANN-based models outperformed the other ML algorithms. The neural networks were overtaking the

earth science applications especially drought monitoring and forecasting. Drought data can be uncertain, and high resolution. Also, these kinds of data are multi-source, multi-scale, multi-resolution, and multi-temporal .i.e. with high complexity. As neural network-based architectures, DL architectures gave accurate results when dealing with massive heterogeneous data. The most used architectures in drought monitoring/forecasting are CNNs, DBNs, DNNs, and RNNs. CNN's are commonly used for image processing. The RNNs, DNNs, DBNs, and the LSTMs architectures are commonly used for time series processing. These architectures have several advantages such as the possibility of extracting features on their own from a massive amount of data. The deep neural networks can discover new, more complex features that other machine learning algorithms. This study presents a comparative study between ARIMA and LSTM for drought forecasting. Besides, PROPHET, as a forecasting model that to the best of our knowledge was never used in drought forecasting in China, will be compared to the other models.

III. STUDY AREA

The Jiangsu province is located in the coastal area of East China. The province is between the latitudes 30″451 N and 35″201 N and the longitudes 116″181 E and 121″571 E, covering around 102,600 km2 (2). Most parts of the province lie below 50 meters above sea level. Jiangsu's climate is a humid subtropical climate. Jiangsu has around 80.4 million population. This province observed the worst drought in the last 50 years occurred in the spring of 2011 [33].

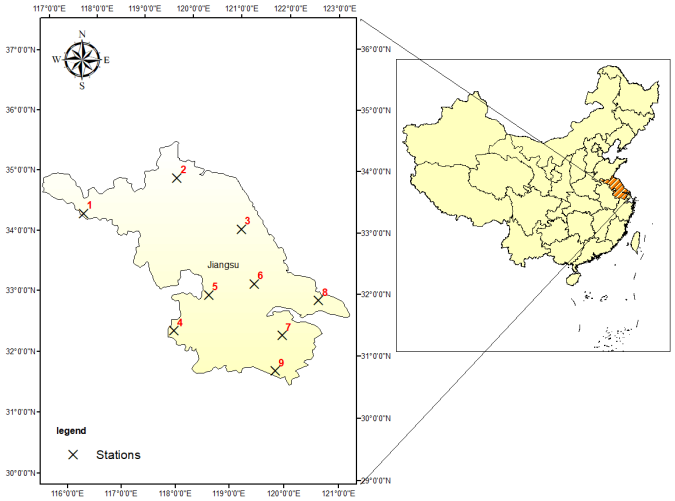

Fig. 2. Study Area.

IV. METHODOLOGY

The proposed methodology consists of 5 main steps: Data collection, data preprocessing, data storage, data processing, and interpretation for decision making.

A. Data collection

Data collection consists of generating and gathering data from different resources. These data are massive and heterogeneous. These data are remote sensing data (e.g., Normalized Difference Vegetation Index (NDVI), Land Surface Temperature (LST)), Climate data (.e.g. Standardized Precipitation Evapotranspiration Index (SPEI), Evapotranspiration (ETP), humidity, precipitation, wind speed, pressure), biophysical data (e.g. Soil Moisture). These data are a range of structured, semi-structured, or unstructured data they are also characterized by their multidimensionality (e.g. multispectral, multi-resolution, multi-temporal data). Thus, every day, Gigabytes of data are generated from different sources. TABLE I describes the data used in this study.

TABLE I
DATA DESCRIPTION

Data	Temporal resolution	Spatial Resolution
NDVI	16 days	1-km
LST	8 days	1-km
SPEI	Monthly	
ETP	Daily	
Soil Moisture	Monthly	0.25°x0.25°
Climate Variable	Daily	

B. Data preprocessing

The used data are collected from different sources and at different time scales, so the pre-processing step is very important to guarantee effective results.

Raw-data retrieving process in which object data will be transformed to raw-data. For example, NDVI and LST data are extracted from satellite images. This step must be well thought out.

Identifying and filtering missing values. Every raw data should be checked and any missed value must be verified or corrected.

Data correction: Here, all the records must be verified. For example, the SPEI must be numerical values, so any character represents an anomaly. For satellite images, mosaicking, atmospheric and geometric correction are performed.

C. Data storage

The data were stored in a Hadoop-based data warehouse (DWH). Apache Hive is used for the conception of the DWH. A snowflake schema was adopted for data modeling. This schema is composed of one Fact table named OperationFact and 13 dimensional tables named (Product_Dimension, Sensor_Dimension, Image_Dimension, Product_Dimension, SatelliteFeature_Dimension, Drought_Index_Dimension, ClimateStation_Dimension, Date_Dimension, Climate-Feature_Dimension, BiophysicalFeature_Dimension, BiophysicalStation_Dimension, Location_Dimension, Country, and Province). To mine the stored data, HiveQL (HQL) was used. HQL is a SQL-like language for DWH mining using Apache Hive.

As an example of query, here we would extract SPEI-1

data from 1990 to 2019 of the Jiangsu province to forecast drought in the next step. Example:
SELECT ID.IndexValue FROM OperationFact OF, Date_Dimension D, Index_Dimension ID, Province Pr
WHERE D.ID_Date=OF.ID_Date
and OF.ID_Index= ID.ID_Index
and $D.Year >= 1990$ and $D.Year <= 2019$
and ID.IndexName= "SPEI-1"
and Pr.Name= "Jiangsu";

D. Drought Forecasting

After extracting SPEI data, in the previous step, drought will be forecasted using three different methodologies ARIMA, PROPHET, and LSTM. We aim to give an accurate prediction of SPEI for the year 2019 using a 1-month timescale.

1) ARIMA: ARIMA is considered to be one of the most effective prediction methods for univariate time series models. ARIMA models are generally applied where time series show non-stationarity in their data. To remove the non-stationarity, an initial differencing step must be performed one or more times. The evolving variable of interest is regressed on its own lagged (prior) values, which is referred to as AR. The regression error is a linear combination of error values that occurred at the same time, as shown by the moving average (MA). Given a time series of data X(t) where t is an integer index and the X(t) are real numbers, an ARMA (p', q) model is given by:

$$(1 - \sum_{i=1}^{p'} \alpha L^i) X_t = (1 + \sum_{i=1}^{q} \Theta L^i) \epsilon_t$$

where L is the lag operator, the α^i are the parameters of the autoregressive part of the model, the h_i are the parameters of the moving average part and ϵ_t are error terms. The error terms ϵ_t are usually considered to be independently distributed, identically distributed variables measured from a normal distribution with zero means. Seasonal and nonseasonal ARIMA models have variant parameters. Three parameters are available for describing the seasonal ARIMA model:

- P = number of seasonal autoregressive terms
- D = number of seasonal differences
- Q = number of seasonal moving-average terms

The following three parameters can be used to define a nonseasonal ARIMA model:

- p = number of autoregressive terms
- d = number of nonseasonal differences
- q = number of moving-average terms

2) PROPHET: PROPHET is open source software for forecasting time series data that is available in Python and R. PROPHET has a high sensitivity to missing data, catching pattern changes, and significant outliers. Furthermore, it obtains a fair estimation of the mixed data without requiring manual intervention. PROPHET has its unique data frame that makes it simple to manage time series and seasonality. Two simple columns are required in the data frame. The "ds" column is one of these columns, and it stores the date-time series. The corresponding values of the time series in the data frame are stored in the "y" column. As a result, the system works well with seasonal time series and offers some choices for dealing with seasonality in the dataset. Seasonality may be set on an annual, weekly, or regular basis. A data analyst should select the available time granularity for the forecast model on the dataset since these options are available [34].

E. LSTM

LSTM is an RNN architecture. The LSTM network is well-fitting if there are unknown time lags and connections between important events to draw from experience to classification processes and to prediction time series. LSTM uses a memory unit called cell state to model long-term dependencies. It has a chain-like structure, having three gates that are implemented using the logistic function. The structure of LSTM is shown in Fig. 3 . There are three types of gates that determine the cell state, which include an input, forget gate, and output gate. The gates analyze and control the quantity of information. The working mechanism of the gates and their information flow is expressed using these equations:

$$f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f)$$

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i)$$

$$C_t^{'} = tanh(W_C.[h_{t-1}, x_t] + b_C)$$

$$C_t = f_t.C_{t-1} + i_t.C_t^{'}$$

$$o_t = \sigma(W_o.[h_{t-1}, x_t] + b_o$$

$$h_t = \sigma_t.tanh(C_t)$$

For LSTM tuning, Adam Optimizer is used, the learning rate

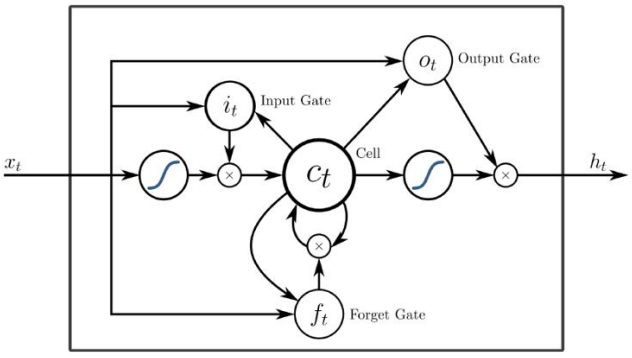

Fig. 3. LSTM architecture.

is 0.001, and the batch size is 50.

F. Interpretation and decision making

Due to the danger of drought phenomena and its impact on different fields, decision-makers need to take precautions. Therefore, they need to use big data to further develop the traditional decision-making process. Therefore, this step aims to present the final results in form of representations that help them to understand and deduce potential insights using curves or maps.

V. EXPERIMENTATION AND VALIDATION

A. Data

Drought indices have a valuable role in analyzing drought. The SPEI is used in multiple studies. SPEI is therefore capable of fulfilling the requirements of a drought index as it is flexible enough to be implemented in different scientific disciplines. This index could measure drought severity by correlating time and space. A major advantage of this index is the inclusion of potential evapotranspiration (PET) in its calculation. Therefore, it would be able to reflect the effect of PET on drought. TABLE II represents different classes of drought based on SPEI values.

TABLE II
SPEI CLASSIFICATION

Class	Value
Extreme wet	$SPEI >= 2.0$
Severe wet	$1.5 <= SPEI < 2.0$
Moderete wet	$1 <= SPEI < 1.5$
Normal	$-1 <= SPEI < 1$
Moderate dry	$-1.5 < SPEI <= -1.0$
Severe dry	$-2.0 < SPEI <= -1.5$
Extreme dry	$SPEI <= -2.0$

TABLE III represents the statistical evaluation of the SPEI-1 values for each station.

TABLE III
STATISTICAL EVALUATION OF THE SPEI-1 VALUES OF THE NINE STATIONS

Station	Max	Min	Avg	STD
Station 1	2.58	-2.89	0.00	0.995
Station 2	3.07	-2.91	0.00	1.00
Station 3	2.81	-3.20	0.00	0.991
Station 4	2.76	-2.90	0.00	0.993
Station 5	2.64	-2.90	0.00	0.989
Station 6	2.84	-2.84	0.00	0.993
Station 7	2.36	-2.90	0.00	0.988
Station 8	3.24	-2.90	0.00	0.992
Station 9	3.12	-2.89	0.00	0.990

B. Performance metrics

The performance of the used forecasting models for each of the 9 stations was calculated using statistical indices like R2, RMSE, and MAE. These metrics express the degree of the models' certainty. The R2 quantifies the level of the linear correlation between the forecast and observed data. The deviation of the total and absolute error is evaluated using the RMSE and the MAE indices, respectively.

The Coefficient of determination (R^2):

$$R^2 = \left(\frac{\sum_{i=1}^{N}[(Y_0 - \bar{Y}_0).(Y_p - (\bar{Y}_p)]}{\sqrt{\sum_{i=1}^{N}(Y_0 - \bar{Y}_0)^2).\sum_{i=1}^{N}(Y_{p-p})^2}}\right)^2$$

Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Y_0 - Y_p)^2}{N}}$$

Mean Absolute Error (MAE):

$$MAE = \frac{\sum_{i=1}^{N}|Y_o - Y_p|}{N}$$

Where Y_o represents the observed value; Y_p represents the predicted value and N is the number of data points.

C. Results and discussion

Based on the observed and forecasted SPEI values from ARIMA, PROPHET, and LSTM models, drought severity maps for the month of August 2019 were created for the Jiangsu Province. Fig. 4 illustrates the results of drought's spatial distribution over Jiangsu Province.

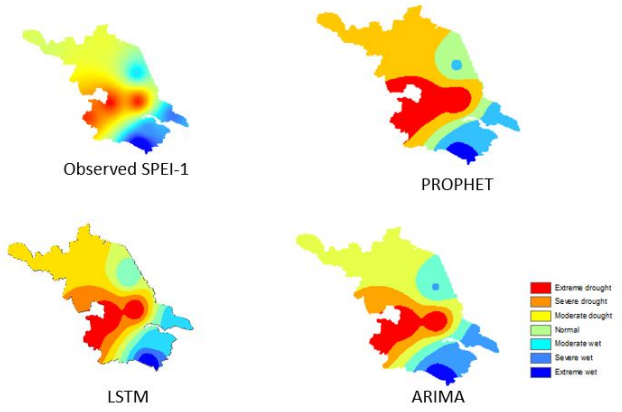

Fig. 4. Drought severity maps based on SPEI-1 of 6-month lead forecasting.

The maps provide a spatial distribution of drought classes over the study area. While most parts of the province showed moderate drought, the LSTM and ARIMA maps showed slightly wet conditions.

The results of ARIMA, PROPHET and LSTM performance are shown in TABLE IV, TABLE V, and TABLE VI, respectively.

For LSTM tuning, Adam Optimizer is used, the learning rate is 0.001, and the batch size is 50. Besides, the data were split 60% for the learning, 20% for the test, and 20% for the validation. The results revealed that LSTM gave the best performance for almost all the stations. However, ARIMA and PROPHET showed lower performance in predicting the SPEI-1 series. For example, for the first station, the RMSE value was 0.49 for the LSTM, and 0.79 for ARIMA and

TABLE IV
ARIMA PERFORMANCE

Station	R^2	RMSE	MAE
Station 1	0.63	0.76	0.6
Station 2	0.67	0.53	0.48
Station 3	0.68	0.44	0.35
Station 4	0.79	0.59	0.46
Station 5	0.66	0.57	0.37
Station 6	0.74	0.58	0.42
Station 7	0.62	0.70	0.58
Station 8	0.60	0.70	0.57
Station 9	0.65	0.59	0.44

TABLE V
PROPHET PERFOEMANCE

Station	R^2	RMSE	MAE
Station 1	0.54	0.79	0.6
Station 2	0.64	0.53	0.42
Station 3	0.7	0.43	0.35
Station 4	0.7	0.58	0.36
Station 5	0.74	0.51	0.34
Station 6	0.73	0.55	0.37
Station 7	0.60	0.68	0.57
Station 8	0.6	0.75	0.45
Station 9	0.66	0.58	0.37

PROPHET. Similarly, for the MAE metric LSTM gave 0.38 in contrast ARIMA and prophet both gave 0.6. Contrariwise, the R2 measure revealed that Prophet was the less performing model having 0.54 with 0.63, and 0.83 for ARIMA and LSTM, respectively.

VI. CONCLUSION

The Chinese Territory is highly threatened by the droughts. This is due to both climatic causes and anthropogenic causes. The repetitive droughts impacted many fields in China such as agriculture, water resources, and human being. Consequently, drought monitoring and forecasting are very important to take precautions against this natural disaster. This paper presents an overview of the drought in China; the causes, the impacts, and a review of drought monitoring/forecasting approaches using the statistical, DL models is presented. A big data architecture for drought is presented. The architecture is composed of five layers; data collection, data preprocessing, data storage, data processing, and interpretation and decision making. A comparative study was provided between three models (ARIMA, PROPHET, and LSTM). Results revealed that LSTM outperformed the two other models. The evaluation is done using three different metrics R2, RMSE, MAE. For future works, a tool will be developed based on big data frameworks and DL architectures for drought monitoring and forecasting.

TABLE VI
LSTM PERFOEMANCE

Station	R^2	RMSE	MAE
Station 1	0.83	0.49	0.38
Station 2	0.83	0.4	0.33
Station 3	0.84	0.30	0.23
Station 4	0.91	0.35	0.27
Station 5	0.81	0.44	0.33
Station 6	0.84	0.47	0.36
Station 7	0.80	0.46	0.38
Station 8	0.79	0.55	0.39
Station 9	0.82	0.47	0.32

References

[1] J. Hays, "Drought in china," 2008. [Online]. Available: http://factsanddetails.com/china/cat10/sub64/item1879.html#chapter-1. [Accessed 27 January 2020].

[2] BCC, «Beijing climate center,» 2020. [Online]. Available: https://cmdp.ncc-cma.net/en/.

[3] Lu, E., Cai, W., Jiang, Z., Zhang, Q., Zhang, C., Higgins, R. W., Halpert, M. S. , «The day-to-day monitoring of the 2011 severe drought in China.,» Climate Dynamics, 43(1–2), pp. 1-9, 2014.

[4] FAO, «A severe winter drought in the North China Plain may put wheat production at risk,» FAO, 2011.

[5] Barriopedro, D., Gouveia, C. M., Trigo, R. M., Wang, L. , «The 2009/10 drought in China: possible causes and impacts on vegetation,» Journal of Hydrometeorology, 13(4), pp. 1251-1267, 2012.

[6] He, J., Yang, X., Li, Z., Zhang, X., Tang, Q. , « Spatiotemporal variations of meteorological droughts in China during 1961–2014: An investigation based on multi-threshold identification.,» International Journal of Disaster Risk Science, 7(1), pp. 63-75, 2016.

[7] Balti, H., Abbes, A. B., Mellouli, N., Farah, I. R., Sang, Y., Lamolle, M., «A review of drought monitoring with big data: Issues, methods, challenges and research directions.,» Ecological Informatics, 60, 101136., 2020.

[8] Yang, D, Distributed Hydrological Model Using Hillslope Discretization Based On Catchment Area Function: Development and Applications, Tokyo, 1998.

[9] Zhao Ren-Jun, Yi-Lin, Z., Fang Le-Run, Liu Xin-Ren, Zhang Quan-Sheng. , «The Xinanjiang model.,» Hydrological Forecasting. Proc. Oxford Symposium, April 1980 , p. 351–356, 1980.

[10] Z. Ren-Jun, «The Xinanjiang model applied in China.,» Journal of hydrology, 135(1-4), pp. 371-381, 1992.

[11] Haider, S. A., Naqvi, S. R., Akram, T., Umar, G. A., Shahzad, A., Sial, M. R., ... Kamran, M., « LSTM neural network based forecasting model for wheat production in Pakistan.,» 2019.

[12] Zhou, L., Pan, S., Wang, J., Vasilakos, A. V. , «Machine learning on big data: Opportunities and challenges,» 2017.

[13] Inoubli, R., Abbes, A. B., Farah, I. R., Singh, V., Tadesse, T., Sattari, M. T., «A review of drought monitoring using remote sensing and data mining methods,» chez In 2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP) (pp. 1-6). IEEE., 2020.

[14] Várkonyi-Kóczy, A. R. (Ed.), Engineering for Sustainable Future, Lecture Notes in Networks and Systems, 2020.

[15] Gheisari, M., Wang, G., Bhuiyan, M. Z. A. , « A Survey on Deep Learning in Big Data.,» 2017.

[16] Zhang, Y., Yang, H., Cui, H., Chen, Q. , «Comparison of the Ability of ARIMA, WNN and SVM Models for Drought Forecasting in the Sanjiang Plain, China,» Natural Resources Research, pp. 1-8, 2019.

[17] Zhang, Y., Li, W., Chen, Q., Pu, X., Xiang, L. , «Multi-models for SPI drought forecasting in the north of Haihe River Basin, China,» Stochastic environmental research and risk assessment, 31(10), pp. 2471-2481, 2017.

[18] Tian, M., Wang, P., Khan, J. , «Drought forecasting with vegetation temperature condition index using arima models in the guanzhong plain,» Remote Sensing, 8(9), p. 690, 2016.

[19] Tian, M., Wang, P., Yan, T., Liu, C., «Adjustment of Kappa coefficient and its application in precision and agreement evaluation of drought forecasting models,» Nongye Gongcheng Xuebao/Transactions of the Chinese Society of Agricultural Engineering, 28(24), pp. 1-7, 2012.

[20] Xu, L., Chen, N., Zhang, X. , « A comparison of large-scale climate signals and the North American Multi-Model Ensemble (NMME) for drought prediction in China.,» Journal of hydrology, 557, pp. 378-390, 2018.

[21] Li, Z., Chen, T., Wu, Q., Xia, G., Chi, D. , « Application of penalized linear regression and ensemble methods for drought forecasting in Northeast China,» Meteorology and Atmospheric Physics, pp. 1-18, 2019.
[22] Zhang, R., Chen, Z. Y., Xu, L. J., Ou, C. Q., «Meteorological drought forecasting based on a statistical model with machine learning techniques in Shaanxi province, China.,» Science of The Total Environment, 665, pp. 338-346, 2019.
[23] Yang, M., Mou, Y., Meng, Y., Liu, S., Peng, C., Zhou, X, « Modeling the effects of precipitation and temperature patterns on agricultural drought in China from 1949 to 2015.,» Science of the Total Environment., 2019.
[24] Liu, X., Zhu, X., Zhang, Q., Pan, Y., Yang, T., Sun, P. , « A remote sensing and artificial neural network-based integrated agricultural drought index: Index development and applications.,» Catena, 186. , 2020.
[25] Junfei, C., Zeyuan, H., Qiongji, J. , « SPI-based drought characteristics analysis and prediction for Xiqiao Station in Yunnan Province, China,» Disaster Advances, 5(4), pp. 1260-1268, 2012.
[26] Du, J., Liu, Y., Liu, Z. , « Study of Precipitation Forecast Based on Deep Belief Networks.,» Algorithms, 11(9), 132, 2018.
[27] Zhang, P., Jia, Y., Zhang, L., Gao, J., Leung, H. , « A deep belief network based precipitation forecast approach using multiple environmental factors,» Intelligent Data Analysis, 22(4), pp. 843-866, 2018.
[28] Kang, J., Wang, H., Yuan, F., Wang, Z., Huang, J., Qiu, T., « Prediction of Precipitation Based on Recurrent Neural Networks in Jingdezhen, Jiangxi Province, China,» Atmosphere, 11(3), 246, 2020.
[29] Miao, Q., Pan, B., Wang, H., Hsu, K., Sorooshian, S. , « Improving monsoon precipitation prediction using combined convolutional and long short term memory neural network,» Water, 11(5), 977, 2019.
[30] Hu, C., Wu, Q., Li, H., Jian, S., Li, N., Lou, Z., « Deep learning with a long short-term memory networks approach for rainfall-runoff simulation.,» Water, 10(11), 1543, 2018.
[31] An, J., Li, W., Li, M., Cui, S., Yue, H. , « Identification and classification of maize drought stress using deep convolutional neural network.,» Symmetry, 11(2), 256., 2019.
[32] Jiang, B., Wang, P., Zhuang, S., Li, M., Gong, Z. , « Drought Stress Detection in the Middle Growth Stage Of Maize Based On Gabor Filter and Deep Learning.,» In 2019 Chinese Control Conference (CCC) , pp. 7751-7756, 2019.
[33] Tao, H., Fischer, T., Zeng, Y., Fraedrich, K., «Evaluation of TRMM 3B43 precipitation data for drought monitoring in Jiangsu Province, China,» Water, 8(6), 221, 2016.
[34] F. Reasearch, «Prophet Forecasting At Scale.,» [Online]. Available: https://research.fb.com/ prophet-forecasting-at-scale/. [Accès le 02 2021].
[35] Chen, X., Jiang, J., Li, H. , « Drought and flood monitoring of the Liao River Basin in Northeast China using extended GRACE data.,» Remote Sensing, 10(8), 1168., 2018.