GU J Sci, Part A, 8(1): 43-57 (2021)

Gazi University

Journal of Science

PART A: ENGINEERING AND INNOVATION

http://dergipark.gov.tr/gujsa

Araştırma Makalesi Research Article

Automated Sensor Data Validation and Correction with Long Short-Term Memory Recurring Neural Network Model

Abraham Sudharson PONRAJ[1]*, T. VIGNESWARAN[1], J. Christy JACKSON[2]

[1]*School of Electronics Engineering, Vellore Institute of Technology, 600127, Chennai, India*
[2]*School of Computer Science Engineering, Vellore Institute of Technology, 600127, Chennai, India*

Keywords

Data Validation
LSTM
SARIMA
Reference Evapotranspiration ETo
Weather Data

Abstract

Proper irrigation planning by matching reference evapotranspiration (ETo) with active crop growth requirement leads to an improved water usage efficiency and thereby improving the crop yield. ETo is primarily influenced by the following weather parameters the air temperature, relative humidity, wind speed and solar radiation. To make the ETo estimation system fault tolerant it is important to validate the real time data from the weather station, since the sensors used in these weather stations are prone to error due to influence of various environmental factors. A Recurring Neural Network (RNN) based Data Validation and Correction (DVC) algorithm was proposed to identify the faulty data and to correct them. Long Short-Term Memory (LSTM) RNN model is used to forecast the weather data such as temperature, solar radiation, wind speed and relative humidity. It uses statistical significance test to identify faulty data and isolate them. Then the DVC approach corrects the faulty data by replacing them by LSTM forecasted data. The performance evaluation of this approach showed better forecasting ability when compared with Seasonal Autoregressive Integrated Moving Average (SARIMA) based DVC and thereby improving overall performance of the DVC approach.

Cite

Ponraj, A. S., T, Vigneswaran, T., & Jackson, J. C. (2021). Automated Sensor Data Validation and Correction with Long Short-Term Memory Recurring Neural Network Model. *GU J Sci, Part A*, *8*(1), 43-57.

Author ID (ORCID Number)

A. S. Ponraj, 0000-0002-3044-0985
T. Vigneswaran, 0000-0002-0478-6739
J. C. Jackson, 0000-0001-9468-7672

Article Process

Submission Date	05.07.2020
Revision Date	21.03.2021
Accepted Date	26.03.2021
Published Date	29.03.2021

1. INTRODUCTION

Global water intake doubles each 20 years, nearly twofold the rate of population increase. According to a Food and Agriculture Organisation (FAO) forecast, irrigation will have to meet 70-80% of the increase in food consumption between 2000 and 2030 (OECD, 2008). Though the practice of irrigation in agricultural farm is only around 300 million hectares or 20% of the cultivable land, but contributes significantly to over 40% of world food production (Fischer et al., 2012). Irrigation can mitigate the risk involved with rain-fed agriculture in dry regions. This helps protect farming from water shortages that are expected to occur more often. Efficient water use can maximize the diversity of agricultural crop, increase the yield, improve the economy and provide food at reduced price (Tiwari & Dinar, 2002). In view of the important role played by irrigation of agricultural farm in worldwide water usage, which significantly increase water availability for other human and environmental uses by better farm management practices that improve irrigation water production (Watson & Burnett, 2017).

*Corresponding author, e-mail: abrahamsudharson@gmail.com

A precise estimate of the amount of water lost to the atmosphere by combined evaporation and transpiration cycle called evapotranspiration (ET) is fundamentally mandatory for effective farm water management in order to improve agricultural crop water usage (Martin & Gilley, 1993; Tomar & Ranade, 2001). Reference evapotranspiration (ET_o) plays a vital role in solving the issues like soil water balance, irrigation system and water supply in the agro-ecosystem by providing a sustainable water management in these water starved regions. Hence proper irrigation planning by matching ET_o with active crop growth requirement leads to an improved water usage efficiency and thereby improving the crop yield (Allen et al., 1998) ET_o can be calculated by many empirical and non-empirical equation which depends on large amount of weather data (Cobaner, 2011). However, ET_o is primarily influenced by the following weather parameters the air temperature, relative humidity, wind speed and solar radiation (Allen et al., 1998; Ponraj & Vigneswaran, 2019). This research work contributes by creating a fault tolerant model for predicting the reference evapotranspiration ET_o using the daily weather data like the temperature, relative humidity, wind speed and solar radiation. There is a greater chance that the weather data used to estimate ET_o is incorrect.

With weather data easily available these days and with data driven technologies used to estimate or predict ET_o makes is it even more vulnerable. The data used to develop these models and further for real time prediction with local sensor weather data or weather data from different source can be wrong or may contain spurious values. Mostly, sensors are subject to extreme environment condition and inherent sensor faults are the main drivers to these spurious data. This work aims at developing autonomous data validation and correction mechanism with weather data, which can be further used in creating a fault tolerant ET_o prediction system. This is accomplished by identifying anomalies with statistical methods for each data value. Anomalies can be effectively identified in the input data variable by using long short-term memory recurring neural network (LSTM) model. The incorrect data value is removed when the data error has been identified. The faulty data is corrected by replacing it with the reliable forecasted value. This paper is organized as follows: section 2 describes the related work, section 3 outlines the methodology followed, section 4 discusses the results and section 5 concludes the work.

2. REVIEW OF RELATED WORK

Exhaustive data causes serious issues for a security mechanism in data analytics to identify all irregularities in real-time. The detection of anomalies and outliers from sensor data and other sources can no longer be considered human work. Hence, a need to simplify and automated the anomaly detection process. Automatic anomaly detection based on machine learning, statistics, etc. has been investigated in various areas, including network intrusion detection, authentication, medical data validation, sensor fault detection and more (Chandola et al., 2009). Given the number of such analyses in recent years, many anomaly detection methods still have not been able to minimize abnormalities in certain conditions (Sharma et al., 2010; O'Reilly et al., 2014). It is not the only threat, but as several sensors produces different type of data which may or may not be time dependent, in a single system. For IoT devices with multiple sensors, it is ineffective to use independent faulty data detection algorithms for each sensor.

Presently time series approach has gained popularity among other data-driven approaches to analyses data with periodical trend. For instance, one can find this approach used in various applications like electricity load forecast, market analysis to predict air ticket demand, battery health, etc (Venugopal & Vigneswaran, 2019). All these application exhibits periodical trend and may include seasonality as well, which are the temporal correlation of time series model (Box et al., 1994). A time- approach is capable to interpret the weather data or the observed sensor data as a structure of time-invariant parameters and this is their advantage. Further, this invariance is the key to any prediction model and thereby providing vital information to enhance the model. Furthermore, their efficacy for both linear and nonlinear models contributes to their competency for data fault analysis. Such facts reflect their need for improving the effectiveness of fault diagnostic application (Chatfield, 2000).

There have been several studies on various time series forecasting methods. The variable data is known to choose independent random step series in the random walk RW model (Simmons, 1986). This approach suggests that previous knowledge of data is not relevant and only new observations are valuable. The simple exponential smoothing method (SES model) has been widely applied in seasonal data for forecasting applications (Gardner, 2006). The autoregressive integrated moving average (ARIMA) model is a most commonly used approach for forecasting in the last two decades amongst the other time series forecasting approaches (Mondal et al., 2014). The Sensor Fault Detection and Faulty Data Accommodation (SFDFDA) method were used in sewage monitoring system. It combines seasonal autoregressive moving average (SARIMA) system with a statistical significance test to allow correction of faulty temperature sensor data with forecasted temperature sensor data (Thiyagarajan et al., 2018). ADSaS (Anomaly Detection system using SARIMA and STL) combines SARIMA and Seasonal and Trend decomposition using Loess (STL) for improved performance in anomaly detection (Lee & Kim, 2018). It can be used for both periodic and non-periodic data.

SARIMA based model for forecasting has shown lesser accuracy and is capable of generating inconsistent forecast, though it is largely a flexible model. A recurring neural network (RNN) models is used to predict weather parameters, like the wind speed, air temperature and pressure and wind speed. Its flexibility enables the designer to update the network work design with ease (Roesch & Günther, 2019). An approach based RNN self-learning optimization for rain forecasting using weather dataset was successfully developed (Salman et al., 2015). ConvLSTM is a LSTM version, which comprises a convolution function within the LSTM cell. A model with the help of weather radar data is used for predicting rainfall by convLSTM data. It further reveals that its version of ConvLSTM decreased the RMSE by 23.0% when compared to linear regression models (Kim et al., 2017). LSTM based RNN has proven particularly effective for time series based faulty data identification (Malhotra et al., 2015). A new anomaly detection system was successfully developed to identify cyber-attacks using the Recurrent Neural Network (Goh et al., 2017). RNN and LSTM find its application in medical data and have been investigated in identifying anomalies in them (Chauhan & Vig, 2015).

3. METHODOLOGY

Forecasting is a method to predict future data trends with the help of mathematical model, based on past and historical data trends collected. The weather data measured by the sensor from the local automatic weather station are interpreted as time series data. The future results of the respective weather parameters may be estimated by a correct mathematical model using the historical dynamics of the various weather parameters. These data serve as a substitute sensor data to the actual sensor data and used to detect abnormalities and possible system failures by comparing both the data. Fault data detection: Data anomalies are abnormalities in data that do not align with typical behavioural patterns, which indicated fault in the data. So if the weather data unexpectedly deviates from or has any unusual occurrence in the usual behaviour pattern is labeled as anomalies. Identifying and correcting these anomalies are very vital. Finally, once the data is forecasted and anomaly detection is completed, it's essential to correct the weather data which seems to be incorrect. If not corrected, the entire system will be deemed a failure. Therefore, the forecasted data of that instant replaces the faulty data. Many more simultaneous instants of abnormalities can be termed as sensor failure in the system. The proposed approach is neatly described in Figure 1.

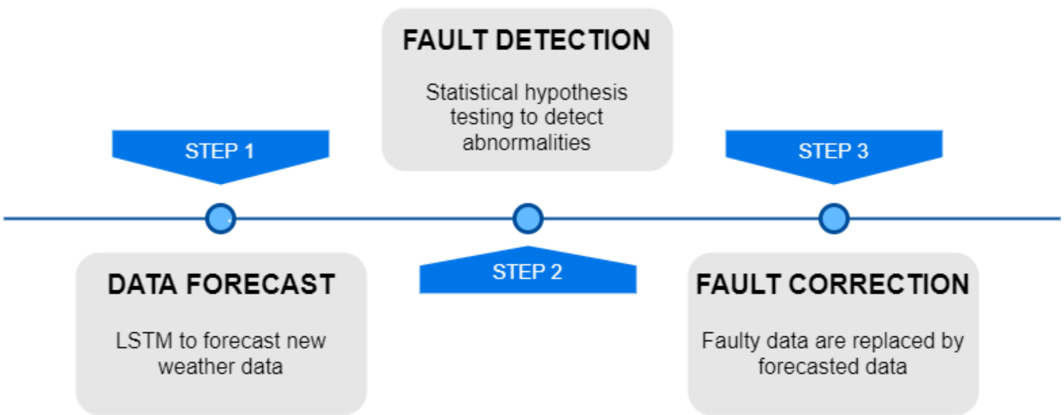


Figure 1. *Data Correction and Validation Methodology*

Step 1: Data forecast

The first step is to forecast weather data with the help of historical weather data. This step is vital as it sets the tone for the data validation and correction (DVC) model. This model makes use of the LSTM or SARIMA model to forecast the new weather data. The new forecasted data is used to replace the faulty data if an abnormality is detected in the data.

Step 2: Fault Detection

This step helps in identifying the faulty data. In this model, statistical method, like the hypothetical significance test, is used to validate the correctness of the data. The abnormalities detected by this test are classified as faulty data, and the faulty data are isolated. This is a simple but very effective approach.

Step 3: Fault Correction

The third and the final step is faulty data correction. When dealing with real time data, it is inevitable to have a faulty data because of various well known reasons as the harsh environmental condition or system failures. This step is an essential approach, as it contributes to making the system fault-tolerant. Once the faulty data is detected they are replaced by the forecasted weather data from step 1. A prolonged exhibition of faulty data ends in classifying it as a system failure.

3.1. Data Description

Localized weather data from the Tamil Nadu Agriculture University (TNAU), Coimbatore automated weather station were used for the study. It is located at 11.01°N latitude and 79.93°E longitude with an elevation of 431m above sea level. TNAU is in the city of Coimbatore located in Tamil Nadu. Coimbatore has a tropical wet and dry climate with an average maximum temperature of 31.5°C and minimum temperature of 22.13°C. The test data for the model was considered from the 1st of January 2017 to the 31st of December 2018. Similar weather parameters such as minimum air temperature AT_{min} in °c, maximum air temperature AT_{max} in °c, wind speed U_2 in km/hr, relative humidity RH in % and solar radiation in cal/cm2/day are used. Table 1 provides the statistical data parameters of the TNAU weather data.

Table 1. *Statistical Parameter of the TNAU Coimbatore Weather Dataset*

Dataset	Statistics	AT_{max} in °C	AT_{min} in °C	U_2 in km/h	SR in cal/cm^2/day	RH_{max} in %	RH_{min} in %
Training	Minimum	23.5	13.5	1.6	60.2	59	15
	Maximum	37.5	26.8	22.8	492	98	91
	Mean	31.88	22.39	6.53	344.72	85.70	54.05
	Std Deviation	2.51	2.20	3.47	78.34	5.03	12.27
	Skewness	-0.07	-1.09	1.62	-0.97	-1.58	0.10
Testing	Minimum	23.5	16	1.6	60.2	68	32
	Maximum	35	26	19.7	442.5	96	88
	Mean	30.47	22.43	7.17	317.52	86.59	60.59
	Std Deviation	2.00	1.58	4.21	82.33	4.69	11.04
	Skewness	-0.59	-1.23	1.17	-1.08	-1.57	0.08

3.2. Weather Data Forecast

For forecasting the weather data SARIMA and LSTM models were used in this study. To make the ARIMA even more effective, the seasonality parameter was included to model a wide range of seasonal data. It is termed as the SARIMA model with the order (p,d,q)(P,D,Q)m where the new P,Q and D are the parameter for seasonality and the m is periods per season as shown in Equation (3) (Box et al., 2015). For example, if monthly data are modeled then the value of m will be 12. The predicted value for an instant t can be obtained just by taking the product of the seasonal and non-seasonal parameters.

$$(1-\emptyset_1 B)(1-\Phi_1 B^{12})(1-B)(1-B^{12})z_t = (1+\theta_1 B)(1+\Theta_1 B^{12})e_t \tag{1}$$

RNN's problem of the vanishing gradient was tackled with the introduction of LSTM, which stands for Long Short Term Memory, introduced by Hochreiter and Schmidhuber in the year 1997 (Hochreiter & Schmidhuber, 1997). LSTMs retain information for a longer time and allow for RNNs to learn more efficiently from the past data. A LSTM consists of recurrently connected memory block each consisting of an input cell, a forget gate and an output gate. At first the input x_t is processed with the previous state h_{t-1} in the input gate and passed over to the forget gate. Then, the output s_t of the forget gate is passed through a tanh activation function. Finally, again an element-wise multiplied is done by output o of an output gate, which is similar to the input gate and the final output h_t is produced (Husein & Chung, 2019). The final output h_t is described in Equation (2).

$$h_t = tanh(s_t) \cdot o \tag{2}$$

3.3. Statistical Hypothesis Testing

The statistical hypothesis test is used to determine the fault in the observed data and the predicted sensor in this study. The probability value (p-value) is nothing but the probability of getting the actual value of a test, when the null-hypothesis is true. Here, it is determined with the help of the chi-square test. It is clearly described in Equation (3).

$$X^2 = \sum_{i=1}^{i=W_L} \frac{\left[(A_t)_i - (P_{t+f})_i\right]^2}{(P_{t+f})_i} \tag{3}$$

Where, X^2 is the chi-square, i is the instantaneous time, W_L is the width of the window, A_t is the actual observed or actual weather value, and P_{t+f} is the predicted output from LSTM model.

The goodness of fit for the LSTM model depends in the chi-square distribution of X_{df}^2 of the X^2. Degree of freedom df, it is the total number of the data points at a given window W_L minus one ($df = W_L - 1$) and it influence the distribution X_{df}^2. The width of the window remains constant for the analysis but it keeps moving with the time t. The significance value often denoted by α is a value which is chosen beforehand. It is used to estimate the p-value at a given instant in a particular window. The significance value is determined by the confident interval value (Sedgwick, 2014). If the confident interval is 95% then the significance value α will be 5% or 0.05.With the df and α value available the p-value can be determined by the following relationship shown in Equation (3).

$$p - value = P\left(X^2_{df,\alpha} \geq X^2\right) \tag{4}$$

From Equation (4) it can be said that a p-value less than the chosen significance value rejects the null hypothesis. The value is considered to be significant only if the above case is met, that is X^2_{df} should be greater than the X^2.

3.4. Data Validation and Correction

The data validation and correction approach heavily depends on the LSTM forecast model and the statistical significance test, to detect or identify and correct any sensor faults or abnormalities. The first and foremost in this approach is to forecast the weather parameter with the help of the time series model LSTM. The parameters like the epochs, activation, optimiser, hidden layers and learning rate should be chosen appropriately.

The next step is to validate the sensor data or any given data to detect any possible abnormalities. The significance test is used in the approach to detect the abnormalities. The first step here is to determine the chi-square value X^2, it in turn help in determining the p-value. The pre-chosen critical significance value α is 0.05. Hence in the DVC algorithm the p-value is compared against the value 0.95. Such that, p-value greater than 0.95 indicates a significant value and a value less than 0.95 indicates a fault. The significant values can be used further to train and develop the forecasting model.

The final step is to replace the value. The actual value or the observed sensor value is compared with the forecasted value to check it lies with the predictive limit. If it lies within the predictive limit then there is no need to disturb the value. But when the value lies outside the predictive limit then there arises a concern to correct the value at that particular instant. The forecasted value from the LSTM model is inserted in the place of the actual or observed value at that particular instant. A warning is signaled by the system if there are three or more consecutive faulty data. Algorithm 1 neatly describes the DVC approach to fault detection and replacement. In algorithm 1 both SARIMA based and LSTM based weather prediction is given, however the DVC approach with LSTM is preferred since it has better accuracy.

4. RESULTS AND DISCUSSION

The TNAU, Coimbatore weather data collected from January 2017 to December 2018 were used to develop the forecasting model using LSTM. Mean absolute error (MAE), root mean square error (RMSE) and mean absolute percentage error (MAPE) are used to evaluate the performance of the algorithm. They are given Equation (5), (6) and (7) respectively.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|P_i - A_i| \tag{5}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(P_i - A_i)^2} \tag{6}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|P_i - A_i|}{A_i} \times 100 \tag{7}$$

Where, P_i is the predicted value at instant i, A_i is the actual observed value and N is the maximum number of data points.

Algorithm 1 Data Validation and Correction (DVC)

1. **LSTM_RNN for weather prediction**

for all i ϵ1 : length[A(t)$_i$] do
  predicting [P_{t+f}]$_i$
  estimate predictive limit [P_{t+f} (-)]$_i$ and [P_{t+f} (+)]$_i$
  i = i + 365
end

(or) **SARIMA for weather prediction**

for all i ϵ1 : length[A(t)$_i$] do
  determine (p, d and q) and (P, D and Q)
  predicting [P_{t+f}]$_i$
  estimate predictive limit [P_{t+f} (-)]$_i$ and [P_{t+f} (+)]$_i$
  i = i + 365
end

2. **Calculate CHI square**

$$X^2 = \sum_{i=1}^{i=W_L} \frac{\left[(A_t)_i - (P_{t+f})_i\right]^2}{(P_{t+f})_i}$$

Calculate p-value

p-value = $P\left(X^2_{df,\alpha} \geq X^2\right)$

Faulty Data Detection

if p-value > 0.95,
  Actual data is error free and can be used for training the model further
End
if p-value < 0.95,
  Possibility of abnormalities in the actual data and has to be investigated
End

3. **Faulty Data Correction**

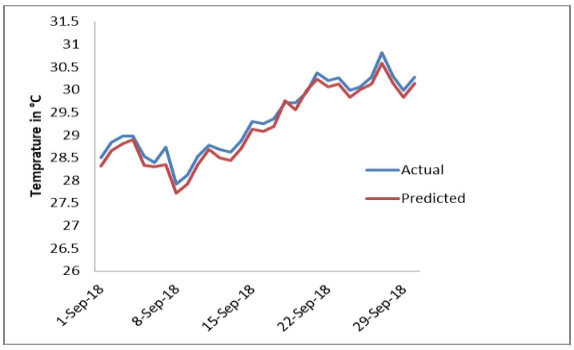
if $[(P_{t+f}(-))_i < (A_t)_i < (P_{t+f}(+))_i]$ then
  no correction
end
if $(A_t)_i$ lie outside $(P_{t+f}(-))_i$ and $(P_{t+f}(+))_i$ tShen
  value is replaced with forecasted value of that instant
end
if more than 3 $(A_t)_i$ lie outside $(P_{t+f}(-))_i$ and $(P_{t+f}(+))_i$ then
  WARNING is indicated by the system and
  value is replaced with forecasted value of that instant
end

The weather parameters used in this study were the air temperature AT, relative humidity RH, wind speed U_2 and solar radiation SR. Further, one month of data was selected to validate the DVC algorithm. Random errors were introduced in the selected one month data and using statistical hypothesis test the errors were identified. Once they were identified they were subsequently replaced by the forecasted data of that particular time instant. Table 2 shows the performance evaluation of the model for all the four weather parameter.

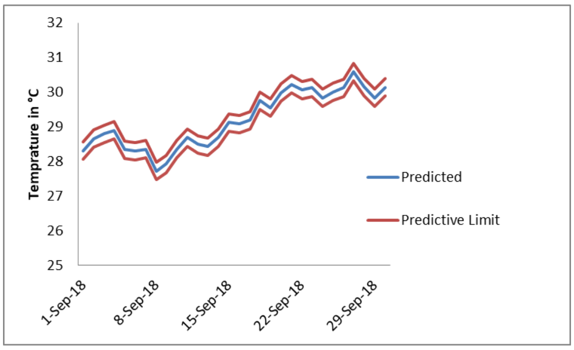
Table 2. *The Performance Values Obtained from the Data Validation Correction Model*

	DVC with LSTM			**DVC with SARIMA**		
Weather Parameter	MAE	RMSE	MAPE	MAE	RMSE	MAPE
Temperature	0.15	0.16	0.51	0.26	0.26	0.91
Relative Humidity	1.45	1.66	2.11	3.1	3.43	4.46
Wind Speed	0.45	0.72	7.18	0.66	0.92	10.89
Solar Radiation	21.99	25.68	6.56	42.25	46.75	12.33

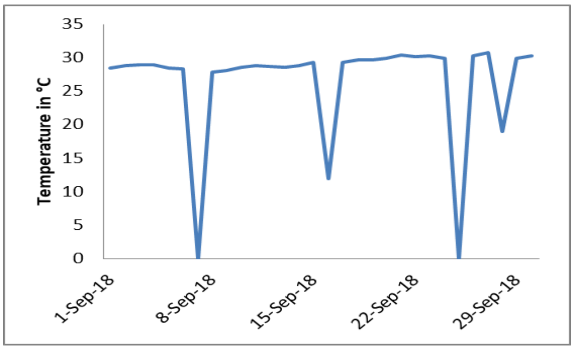
For the air temperature, the first step was to forecast the new temperature at instant *i* with the LSTM model. From the Table 2, MAE for temperature prediction is 0.15, RMSE is 0.16 and the MAPE obtained a value of 0.51. It is also clear that LSTM based approach shows better performance when compared to SARIMA based approach. The next step was to determine the predictive limit $(P_{t+f}(-))_i$ and $(P_{t+f}(+))_i$. After which random temperature values as error were introduced and then the p-value was determined for each instant. Then the p-values were compared with the critical value, which is set as 0.95 in this approach. Figure 2 (a) shows the plot of actual and the predicted temprature value, Figure 2 (b) is the predicted temprature values with predictive limit of the LSTM approach. In Figure 2 (c) introduced random temprature error values can be seen and in Figure 2 (d) p-value for the temperature value with induced error were plotted with the critical value for the LSTM based DVC. So anyting below the critcal value is considered a faulty data. Similarly the Figure 3 (a) shows the plot of actual and predicted temprature values using the SARIMA based DVC model. Figure 3 (b) is plot of the predicted temprature with predictive limit. Figure 3 (c) and (d), shows the temprature with induced temprature and plot which shows the p-value and critical limit respectively. Figure 2 (a) and Figure 3 (a) indicated the performance of the forecast of the LSTM and SARIMA model but it is the LSTM that performace better. Hence, the LSTM based DVC outshines the SARIMA based DVC in finding the faulty data and replacing them.



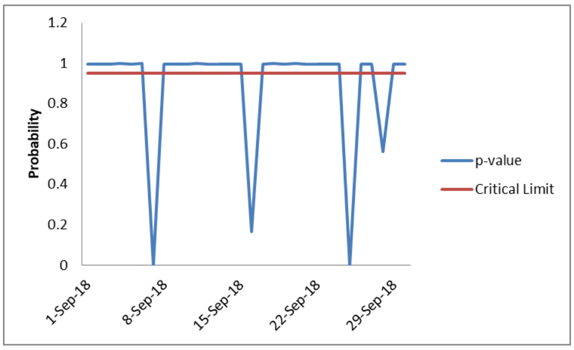
(a) *Predicted temperature by LSTM model*



(b) *Predicted temperature with predictive limit*

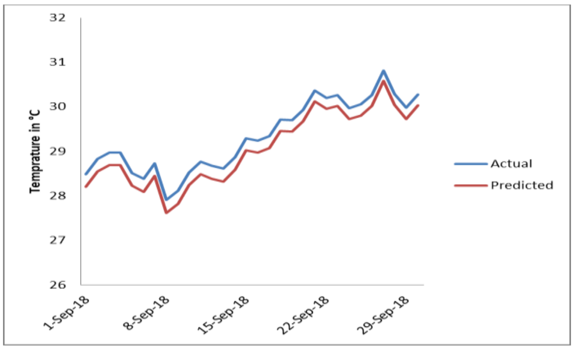


(c) *Induced error value in temperature data*

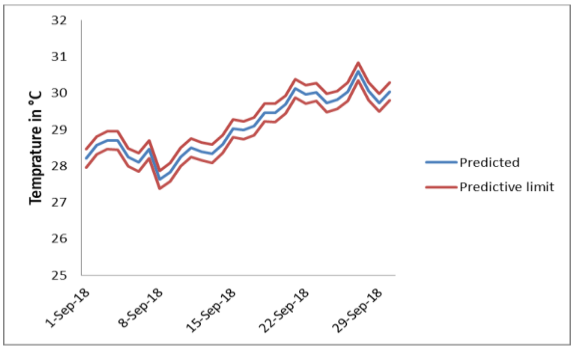


(d) *p-value of the error induced*

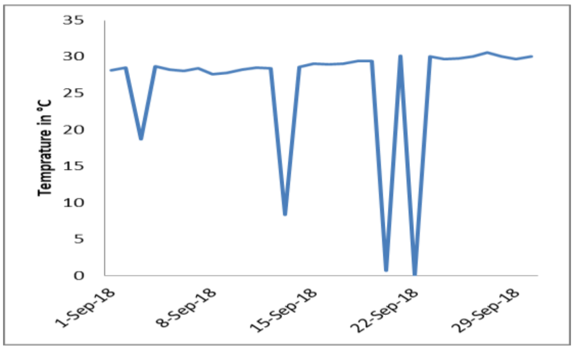
Figure 2. *LSTM Based DVC Model for Temperature Data*



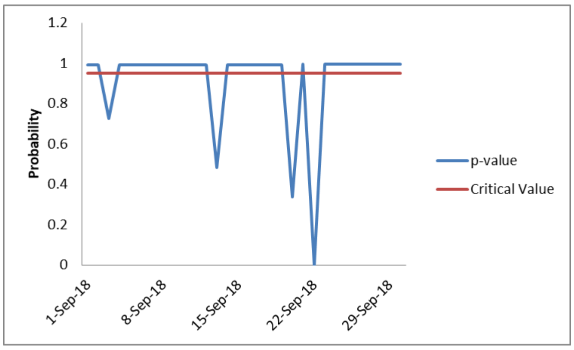
(a) *Predicted temperature by SARIMA model*



(b) *Predicted temperature with predictive limit*

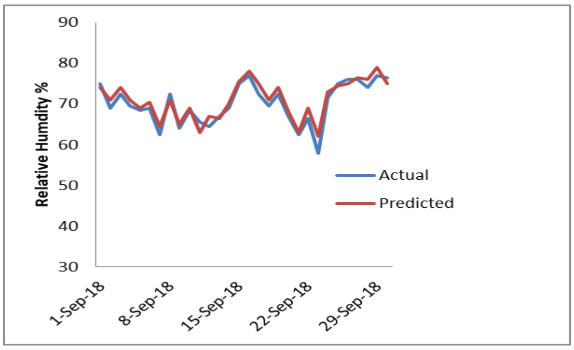


(c) *Induced error value in temperature data*

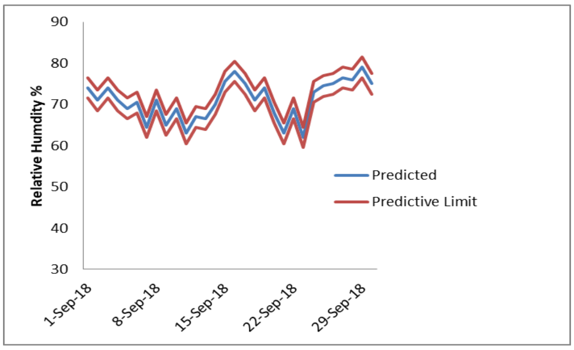


(d) *p-value of the error induced DVC SARIMA*

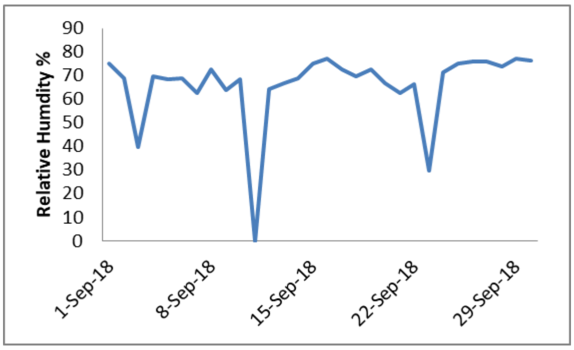
Figure 3. *SARIMA Based DVC Model for Temperature Data*



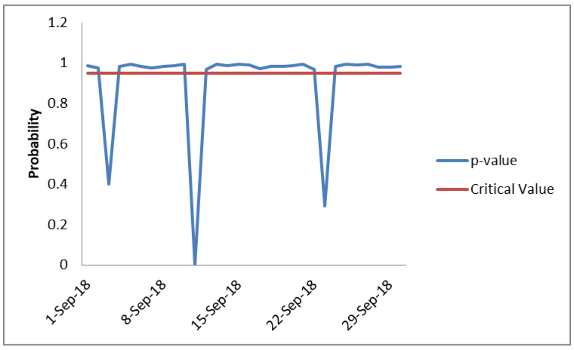
(a) *Predicted RH by LSTM model*



(b) *Predicted RH with predictive limit*

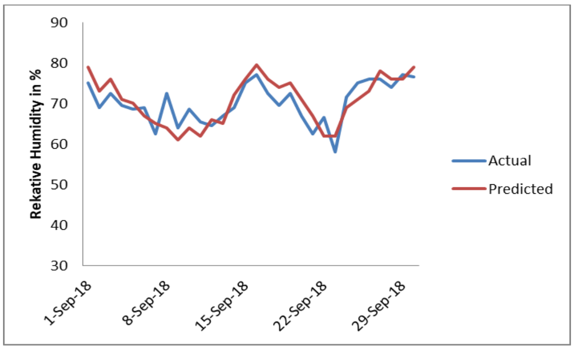


(c) *Induced error value in RH data*

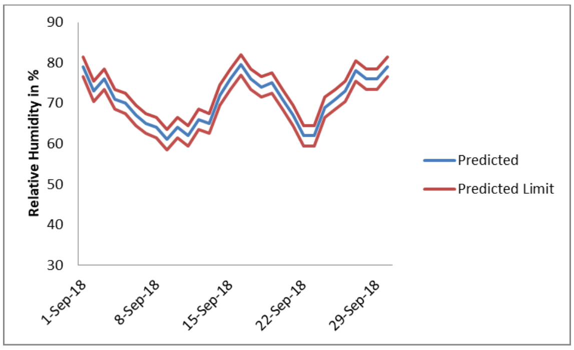


(d) *p-value of the error induced*

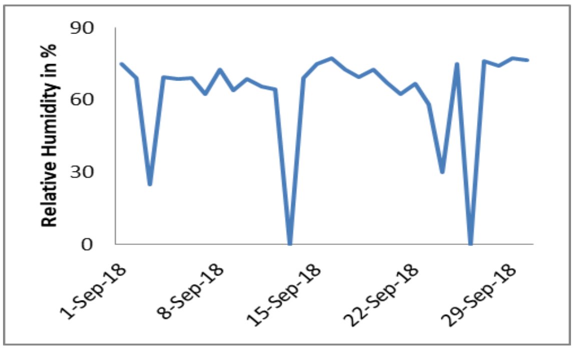
Figure 4. *LSTM Based DVC Model for Relative Humidity (RH) Data*



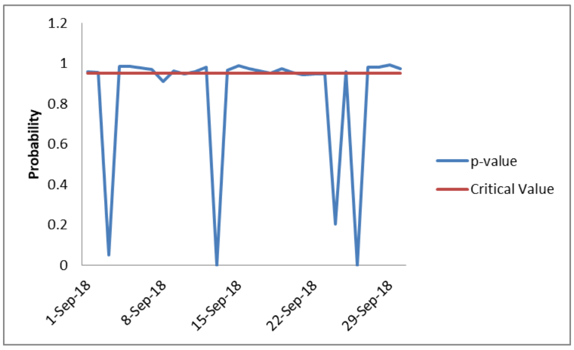
(a) *Predicted RH by SARIMA model*



(b) *Predicted RH with predictive limit*

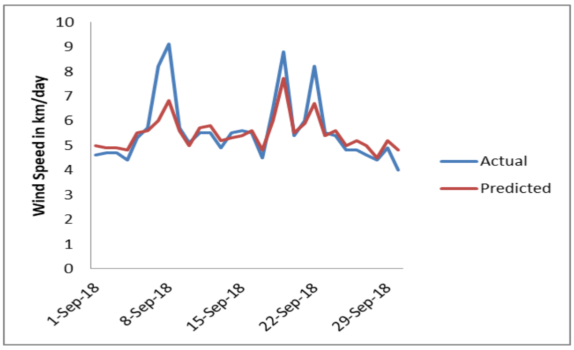


(c) *Induced error value in RH data*

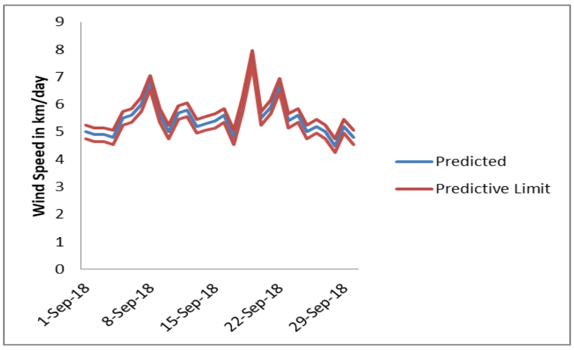


(d) *p-value of the error induced DVC SARIMA*

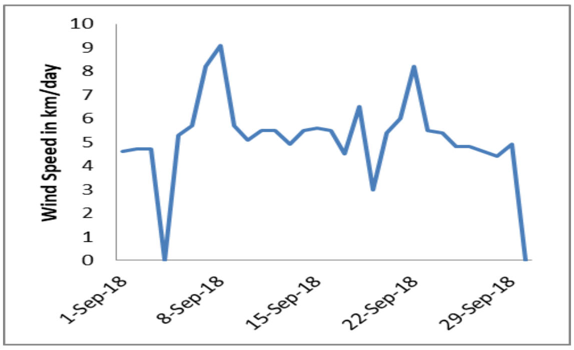
Figure 5. *SARIMA Based DVC Model for Relative Humidity (RH) Data*



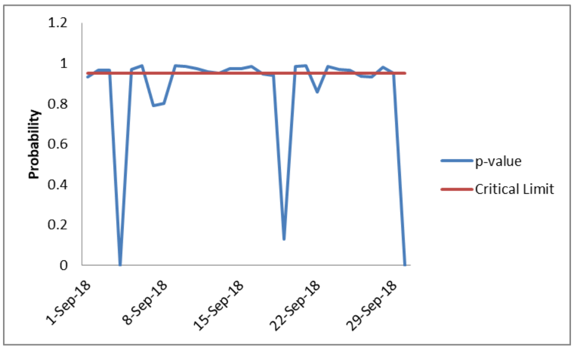
(a) Predicted WS by LSTM model



(b) Predicted WS with predictive limit

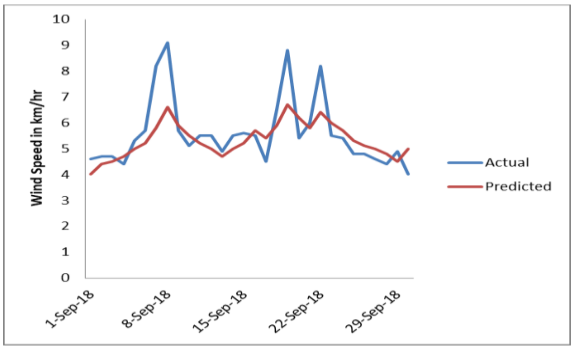


(c) Induced error value in WS data

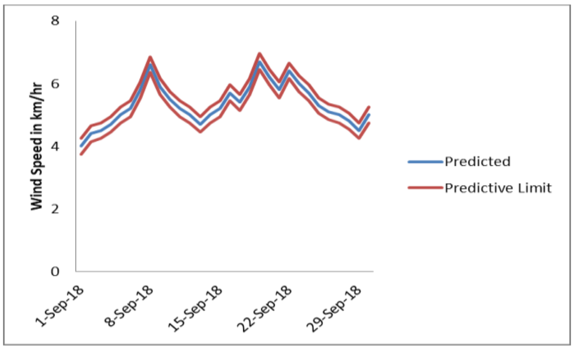


(d) p-value of the error induced

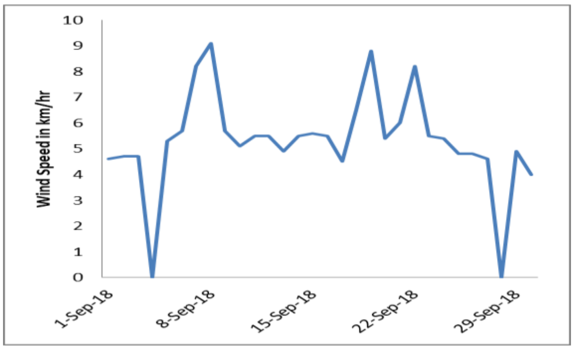
Figure 6. *LSTM Based DVC Model for Wind Speed (WS) Data*



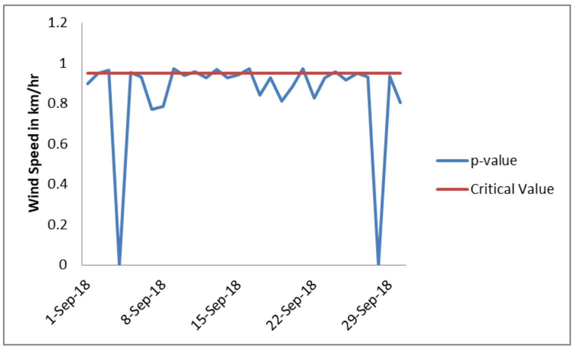
(a) Predicted WS by SARIMA model



(b) Predicted WS with predictive limit

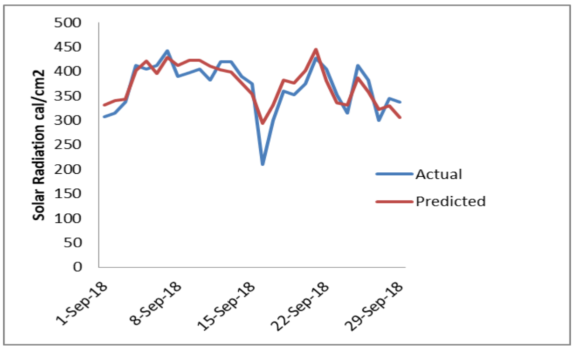


(c) Induced error value in WS data

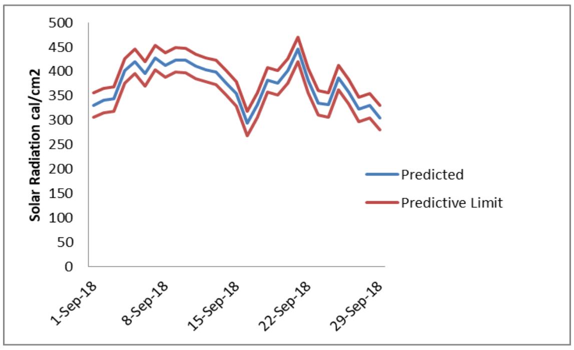


(d) p-value of the error induced DVC SARIMA

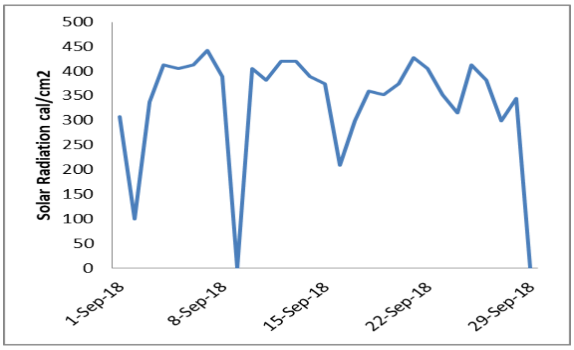
Figure 7. *SARIMA Based DVC Model for Wind Speed (WS) Data*



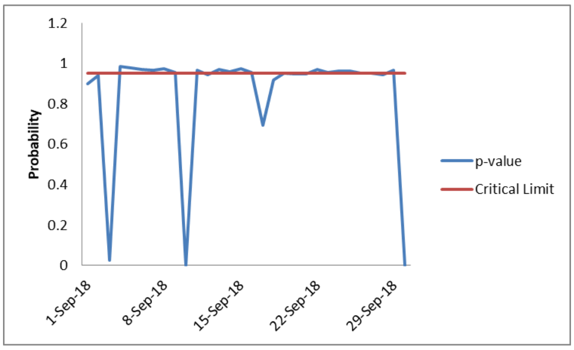
(a) *Predicted SR by LSTM model*



(b) *Predicted SR with predictive limit*

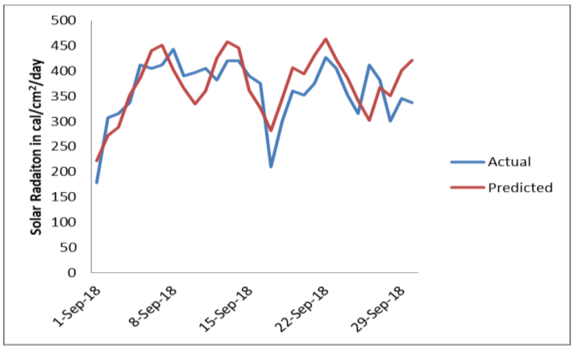


(c) *Induced error value in SR data*

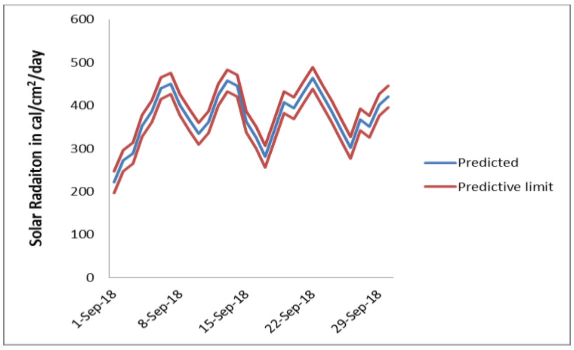


(d) *p-value of the error induced*

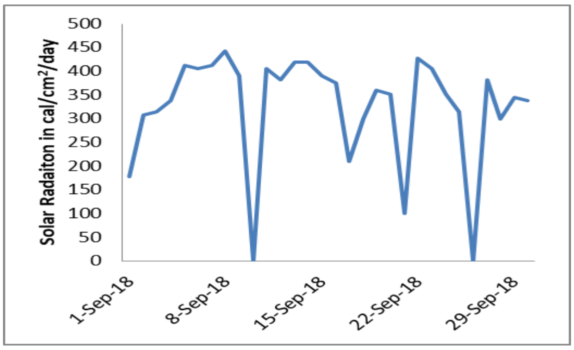
Figure 8. *LSTM Based DVC Model for Solar Radiation (SR) Data*



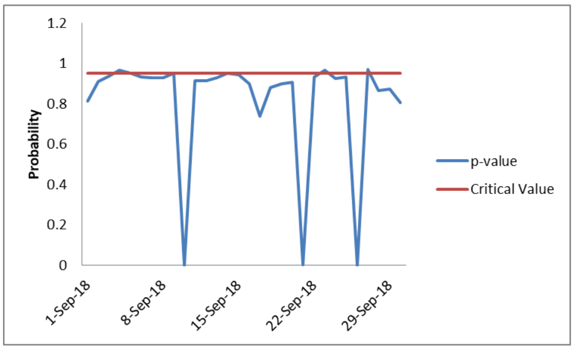
(a) *Predicted SR by SARIMA model*



(b) *Predicted SR with predictive limit*



(c) *Induced error value in SR data*



(d) *p-value of the error induced DVC SARIMA*

Figure 9. *SARIMA Based DVC Model for Solar Radiation (SR) Data*

Similar steps were followed for the other three parameters. From Table 2 the MAE, RMSE and MAPE for relative humidity, wind speed and solar radiation for DVC with LSTM shows better performace than DVC with SARIMA. Figure 4, 5, 6, 7, 8, and 9 shows performance of LSTM based DVC and SARIMA based DVC. It is clear that in all the four parameter LSTM based DVC performs better.

The Data Validation Correction (DVC) algorithm not only identifies the faulty data, it corrects the faulty ones with the predicted input data. It also can warn the user of system failure. With the help of the p-value and critical significance values this faulty data are identified. Once they are identified they are replaced by the predicted data or the substitute weather data. Continuous or prolonged detection of abnormalities that is, a consecutive occurrence of three values beyond the predictive limit will be classified as warning to the system. After identifing the error values with the help of p-value and critical value, error values are corrected by replacing them with forecasted values of that instant Figure 10 illustrates the replacement of error values with corrected values using the LSTM based DVC approach.

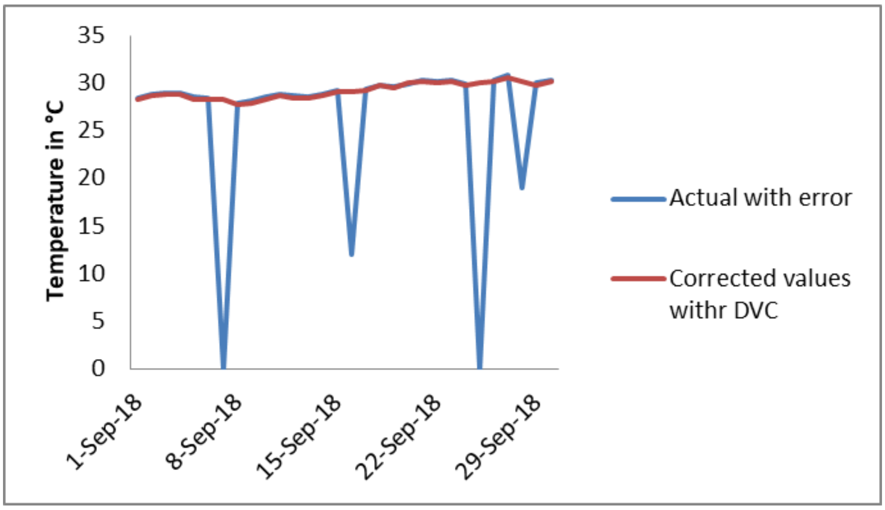


Figure 10. *Corrected Values with LSTM Based DVC Approach*

5. CONCLUSION

The contribution of ET_o in water balance calculation, irrigation planning and estimating yield prediction is immense it very important to accurately predict it. Therefore, accurate prediction is possible if the input weather data is fault free and of high quality. This lead to the development of data validation and correction algorithm to find faulty weather inputs from the sensor or any other source. To predict the input data or the weather data in the DVC approach, first SARIMA based prediction was used. SARIMA based approach though it boast of flexibility and less memory usage, low accuracy in predicting weather output data and difficulty in developing a single model to predicted multiple weather data stands against. Secondly the LSTM time series prediction model was used to improve the DVC algorithm. When using LSTM-RNN in DVC approach accuracy of the predicted weather data improved though at cost of increasing memory usage and longer computation time. The data fault detection was effectively done using statistical approach and then the faulty data was replaced by the predicted data. The prediction value of weather inputs like the temperature, relative humidity, solar radiation and wind speed yielded root mean square error (RMSE) value of 0.16, 1.66, 0.72 and 25.68 respectively. The proposed DVC approach works only with time series data and it’s basically a forecast based approach to detect and replace abnormalities in data. An approach for both time dependent and time independent data would increase the effectiveness of fault detection and correction.

CONFLICTS OF INTEREST

No conflict of interest was declared by the authors.

REFERENCES

Allen, R. G., Pereira, L. S., Raes, D., & Smith, M. (1998). Crop Evapotranspiration, guidelines for computing crop water requirements, FAO Irrigation and Drainage Paper No. 56. Food and Agriculture Organization of the United Nations, Rome, 300p.

Box, G. E., Jenkins, G. M., & Reinsel, G. C. (1994). Time Series Analysis, Forecasting and Control. Englewood Clifs, 598p.

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time series analysis: forecasting and control, 5th Edition. John Wiley & Sons, 712p.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys (CSUR), 41(3), Article No:15, 1-58.

Chatfield, C. (2000). Time-Series Forecasting. CRC Press, 280p.

Chauhan, S., & Vig, L. (2015). Anomaly detection in ECG time signals via deep long short-term memory networks. In: Proceedings of the IEEE International Conference on Data Science and Advanced Analytics (DSAA 2015), 1-7.

Cobaner, M. (2011). Evapotranspiration estimation by two different neuro-fuzzy inference systems. Journal of Hydrology, 398(3-4), 292-302.

Fischer, G., Nachtergaele, F. O., Prieler, S., Teixeira, E., Tóth, G., van Velthuizen, H., Verelst, L., & Wiberg, D. (2012). Global Agro-ecological Zones (GAEZ v3. 0) - Model Documentation. (IIASA: Laxenburg, Austria; FAO: Rome, Italy) pure.iiasa.ac.at/13290

Gardner, Jr, E. S. (2006). Exponential smoothing: The state of the art-Part II. International Journal of Forecasting, 22(4), 637-666.

Goh, J., Adepu, S., Tan, M., & Lee, Z. S. (2017). Anomaly Detection in Cyber Physical Systems Using Recurrent Neural Networks In: Proceedings of the 18th International Symposium on High Assurance Systems Engineering (HASE 2017), 140-145.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

Husein, M., & Chung, I. Y. (2019). Day-ahead solar irradiance forecasting for microgrids using a long short-term memory recurrent neural network: A deep learning approach. Energies, 12(10), 1856, 1-22.

Kim, S., Hong, S., Joh, M., & Song, S. K. (2017). DeepRain: ConvLSTM Network for Precipitation Prediction using Multichannel Radar Data. In: Proceedings of the 7th International Workshop on Climate Informatics. arxiv.org/abs/1711.02316

Lee, S., & Kim, H. K. (2018). ADSaS: Comprehensive Real-time Anomaly Detection System. 19th World International Conference on Information Security and Application (WISA 2018), Revised Selected Papers, 29-41, Springer.

Malhotra, P., Vig, L., Shroff, G., & Agarwal, P. (2015). Long Short Term Memory Networks for Anomaly Detection in Time Series. In: Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. (ESANN 2015), vol. 89, 89-94.

Martin, D. L., & Gilley, J. (1993). Irrigation Water Requirements, Chapter 2, Part 623 of the National Engineering Handbook, United States Department of Agriculture - Soil Conservation Service.

Mondal, P., Shit, L., & Goswami, S. (2014). Study of effectiveness of time series modeling (ARIMA) in forecasting stock prices. International Journal of Computer Science, Engineering and Applications, 4(2), 13-29.

O'Reilly, C., Gluhak, A., Imran, M. A., & Rajasegarar, S. (2014). Anomaly detection in wireless sensor networks in a non-stationary environment. IEEE Communications Surveys & Tutorials, 16(3), 1413-1432.

OECD (2008). Environmental outlook to 2030, The Organisation for Economic Co-operation and Development, 1-10.

Ponraj, A. S., & Vigneswaran, T. (2019). Daily evapotranspiration prediction using gradient boost regression model for irrigation planning. The Journal of Supercomputing, 76, 5732-5744.

Roesch, I., & Günther, T. (2019). Visualization of Neural Network Predictions for Weather Forecasting. Computer Graphics Forum, 38(1), 209-220.

Salman, A. G., Kanigoro, B., & Heryadi, Y. (2015). Weather forecasting using deep learning techniques. In: Proceedings of the International Conference on Advanced Computer Science and Information System (ICACSIS 2015), 281-285.

Sedgwick, P. (2014). Understanding statistical hypothesis testing. BMJ, 348.

Sharma, A. B., Golubchik, L., & Govindan, R. (2010). Sensor faults: Detection methods and prevalence in real-world datasets. ACM Transactions on Sensor Networks (TOSN), 6(3), Article No:23, 1-39.

Simmons, L. F. (1986). M-competition-A closer look at NAIVE2 and median APE: A note. International Journal of Forecasting, 2(4), 457-460.

Thiyagarajan, K., Kodagoda, S., Van Nguyen, L., & Ranasinghe, R. (2018). Sensor failure detection and faulty data accommodation approach for instrumented wastewater infrastructures. IEEE Access, 6, 56562-56574.

Tiwari, D., & Dinar, A. (2002). Balancing future food demand and water supply: The role of economic incentives in irrigated agriculture. Quarterly, Journal of International Agriculture, 41(1), 77-97.

Tomar, A. S. & Ranade, D. H. (2001). Pan coefficient determination for evapotranspiration at Indore, Madhya Pradesh. Indian J. Soil Conserv., 29, 173-175.

Watson, I., & Burnett A. D. (2017). Hydrology: An Environmental Approach, Routledge, 722p.

Venugopal, P., & Vigneswaran, T. (2019). State-of-Health Estimation of Li-ion Batteries in Electric Vehicle Using IndRNN under Variable Load Condition. Energies, 12(22), 4338, 1-29.