International Journal of Advance Research in Science and Engineering
Vol. No.6, Issue No. 05, May 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

UNCERTAINTY ANALYSIS OF DEVELOPED ANN MODEL ALONGWITH FORWARD SELECTION TECHNIQUE IN PREDICTION OF DAILY ATMOSPHERIC BOUNDARY LAYER HEIGHT.

Mohit Mann[1], Kirti Soni[2], Sangeeta Kapoor[3]

[1]*Amity Institute of Applied Sciences, Amity University, Noida, UP (INDIA)*

[2] *CSIR-National Physical Laboratory, New Delhi, (INDIA)*

[3]*LNCT Bhopal (MP)*

ABSTRACT

This study aims to predict daily atmospheric boundary layer height in the atmosphere of Delhi by means of developed artificial neural network(ANN) model along with forward selection technique. Forward selection (FS) method is used for selecting input variables and developing hybrid model with ANN. From 8 input candidates, 4 input candidates are selected for FS. Evaluation of developed hybrid models and its comparison with ANN model fed with all input variables shows that FS method not only reduces the output error, but also computational cost due to less input. FS ANN model is selected as the best model after considering R^2, mean absolute error and also developed discrepancy ratio statistics; it is also shown that this model is superior in predicting daily atmospheric boundary layer height. Finally, uncertainty analysis based on this model is carried out for both simple and hybrid ANN model which shows that FS ANN model has less uncertainty; i.e. it is the best model which forecasts satisfactorily the trends in daily atmospheric boundary layer height.

Key Words :-Artificial Neural Network, Atmospheric boundary layer, Forecast, Forward Selection.

I. INTRODUCTION

Now days, Artificial Intelligence(AI) created methods are widely used and are considered to be one the best methods in comparison to customary statistical methods in most of the scientific fields. Various artificial intelligence models that address the nonlinearity such as Artificial Neural Network (ANN) and fuzzy inference system (FIS) techniques are successfully used for forecasting and air pollution modeling. ANN is a pattern matching technique that uses a pair of input and output sets. ANN has been generally applied to forecast stream flow, precipitation, and drought. FIS is data-driven model similar to ANN for representing linguistic fuzzy if-then rules that are difficult to formulate through a model with crisp parameters. In Fuzzy logic data was divided into training and testing phases. The model results were compared with measured data. The comparison depends on statistical characteristics, different error modes and contour map method. Selecting an input is one of the most vital step in ANN execution. This system is not planned to remove unnecessary inputs.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

For more number of input variables, noisy, terminated and unrelated variable may get involved in the set of data and which may instantaneously make some significant variables to get hidden. Hence there is a need to reduce the number of input variables. There are various approaches to reduce the quantity of input variables. One of the best methods that have been seen in the literature is Forward Selection (FS) method.

ANN model can predict uncertainty much easily as compared to any other statistical models. Though the forecast made are not that firm. This may result into uncertainty in the final outputs.

The present study aims to develop a reliable model for predication of Atmospheric Boundary Layer height. In this part, we will use Forward Selection method in ANN which we will call as hybrid ANN model and will compare the result with simple ANN model. Finally, uncertainty will be testified.

II. CASE STUDY AND DATA

The daily mean value of mixing height is calculated at CSIR-NPL, New Delhi for four days from January 9, 2014 to January 12, 2014 have been considered a sample site in the present study. So in total of 96 samples were collected and used for the prediction of daily Atmospheric Boundary Layer. Other metrological variables data and concentration of different pollutants was obtained from Central Pollution Control Board (CPCB), New Delhi. These four pollutants includes carbon monoxide (CO), nitrogen oxide (NO), sulfur dioxide (SO_2), and ozone (O_3). Other metrological parameters include temperature, humidity, solar radiation (SR) and wind speed (WS). Using all above variables prediction about daily atmospheric boundary layer in atmosphere can be done.

III. ARTIFICIAL NEURAL NETWORK (ANN)

Artificial Neural Network (ANN) is commanding non-linear modeling approach. Which is based upon human brain functioning. It is not probable to create an artificial brain, but then it is likely to have an easy artificial neurons. The ANNs methods can be created in several means and can attempt to simulate the brain in various alternative means. It is a suitable precise arrangement. ANNs are not intellectual; still they are better for identifying patterns and creating easy rules for tedious difficulties. They also have an outstanding training skill that is why they are frequently used in artificial intelligence research. It identifies and learns the correlated patterns between inputs values and objective values. It networks with nodes or neurons, which are interconnected to each other. A general three-layered neural network, consists of several elements nodes. These systems hold an input layer comprising of hubs speaking to diverse input variables, the hidden layer comprising of numerous concealed hubs and an output layer comprising of output variables. Numerous hypothetical plus investigational works have displayed that even a single hidden layer is enough for ANN to estimate several difficult non-linear function. One of the best part of ANN is that the output cells are not directly linked to the different transitional cells. Thus we observe a very minute variation in output and the learning by the system becomes slow.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

The back-propagation algorithm is generally used in between the layers of feed-forward ANNs which directly points that the artificial neurons are well-organized in layers, and provide their specific output in a forward direction, and after that the errors are circulated backwards. The network gets inputs with the help of neurons present in the input layer, and the output of the network is then provided by the neurons present on an output layer. There can be single or plenty of transitional hidden layers. The back propagation algorithm uses supervised learning, which means that we deliver the set of rules with examples as the inputs and outputs we want the network to calculate, and then the error (difference between real and predictable results) is calculated. The main idea behind the back propagation algorithm is to minimize any error which affects our final output, till the ANN completely picks up the training data. The training starts with arbitrary weights, and the objective is to regulate the data with the aim of minimizing the error.

IV. FORWARD SELECTION

For a small number of candidate covariates (N), a prediction model can be chosen by which certain criterion such as CVE (cross validation error), RMSE (Root Mean Square Error) can be computed for a given subsets of a predictors. But when N is very large, the computational load increases very rapidly. So, to overcome this load we need a well-defined step by step algorithms. One of the method is Forward Selection (FS) method. This method has been successfully used by many researchers to develop various accurate prediction models. FS is a linear regression model. Firstly, a dependent variable is chosen and a correlation with other explanatory variables is done. Then, they are arranged from the most correlated variable to a least related variable. The best correlated variable is taken as the first input. After that the second best correlated variable is chosen. It is observed that there is a significant change in the value of R^2 which is famously called as correlation coefficient. Coefficient of correlation is the grade of link among any two variable say 'x' and 'y'. The value of R lies in between -1 and 1. When the value of R is 1 it is said to be perfectly correlated. For the value -1, they are said to be perfectly opposites and for 0, it is said to have no correlation at all. Formula for R is shown below

$$r = \frac{n(\Sigma xy)-(\Sigma x)(\Sigma y)}{\sqrt{[n(\Sigma x^2)-(\Sigma x)^2]}\sqrt{[n(\Sigma y^2)-(\Sigma y)^2]}} \qquad (3.15)$$

Correlation can be rightfully explained for simple linear regression but in the case of multiple linear regression it is becomes tedious to explain each variable relation. Hence, we uses coefficient of determination (R^2) to fairly explain the relation. It is valuable as it provides the proportion of the variance of one variable to that of predictable value from the other variable. As it is the square of R value it range lies in between 0 and 1. R^2 can be used to explain both linear as well as multiple linear regressions.

FS method step is continued for N-1 times which help us in evaluating the effect that is caused by each variable to output of model. In the end, we obtain N number of subsets. The most favorable R^2 subset is taken as model input subset. After obtaining the most favorable R^2 for a set of variables there is no significant change in the value of R^2 when a new variable is added to it.

V. RESULT AND DISCUSSIONS

The numerical experiment for predicting the Atmospheric Boundary Layer height has been performed. The instrumental recorded data for ABL have been collected for every hour from 09 Jan. 2014 to 12 Jan. 2014. So,

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

in total 96 data is taken into consideration. The outputs are obtained by using ANN and FS ANN methods whose results have been assessed to calculate R^2, MAE, and descriptive statistics (d).

(A) For ANN

Figure 1 shows the graph between outputs of the network to the targeted value. The four plots represent training, validation, testing and all data. The dashed line in each plot represents the perfect result i.e. outputs is equal to targets. The solid line represents the best fit linear regression line between target and outputs. The value of R for training, validation and testing is found to be approximately 0.99 which shows that the output values are in perfect linear correlation with our target values. When correlation coefficient is multiplied with itself it gives us the value coefficient of determination. Hence R squared value obtained for calibration (which is the average of training and validation of R value and then its square) is 0.990 and for testing it comes out to be 0.995.

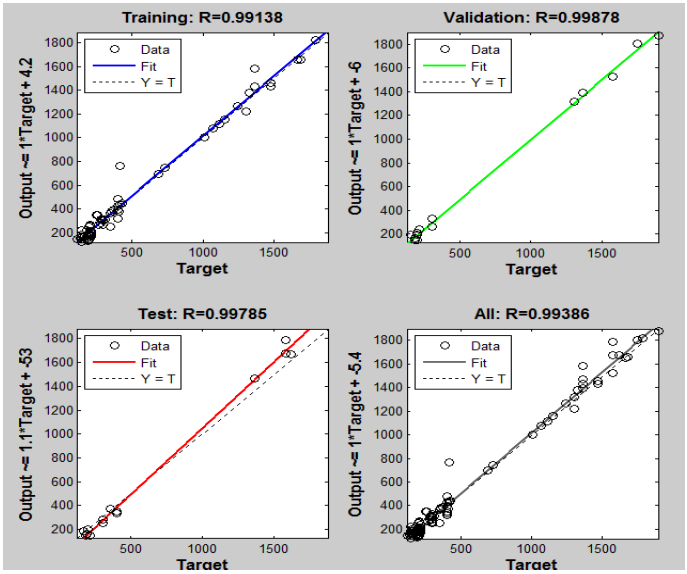


Figure 1: Plots of regression for training, validation and testing

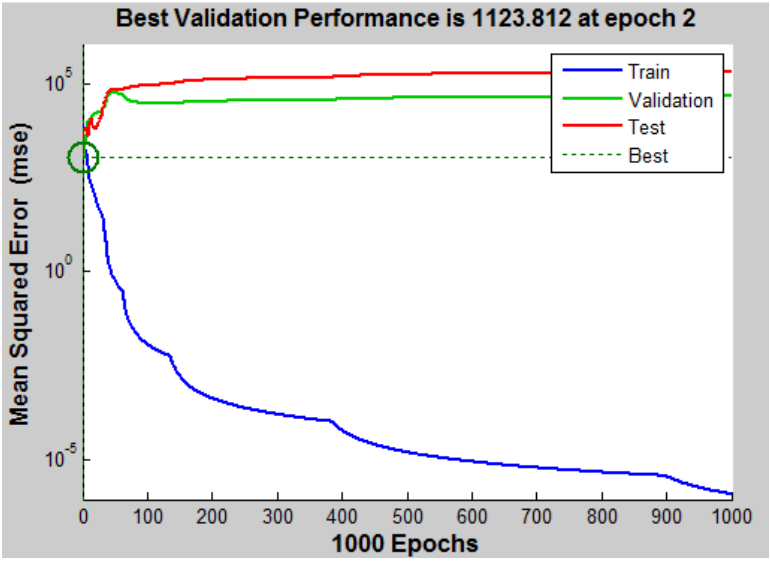
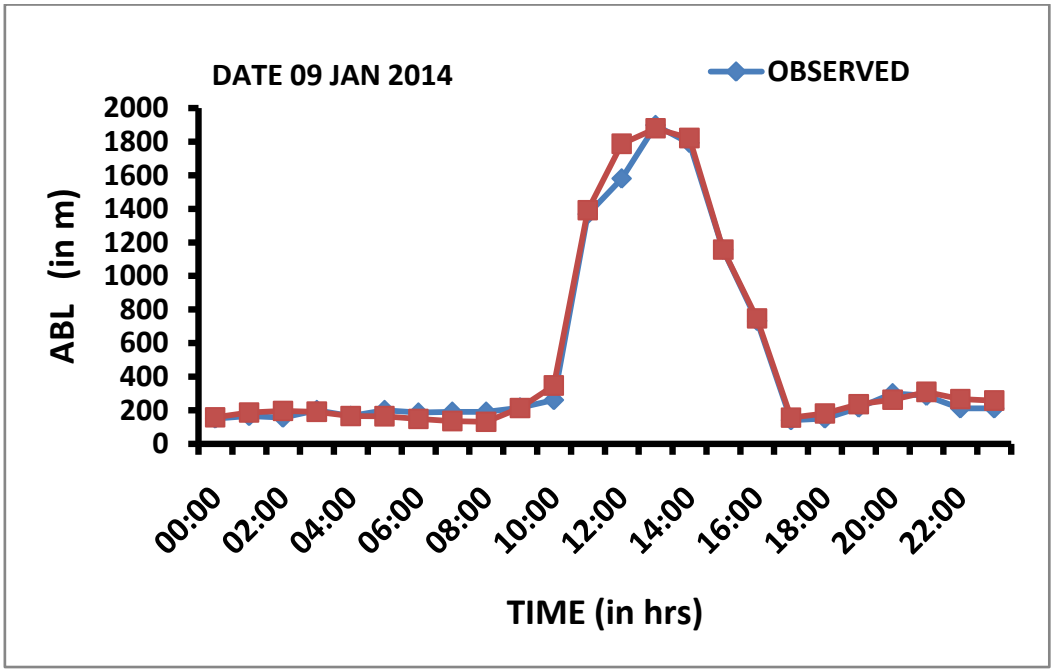


Figure 2: Plot of Performance for 1000 epochs

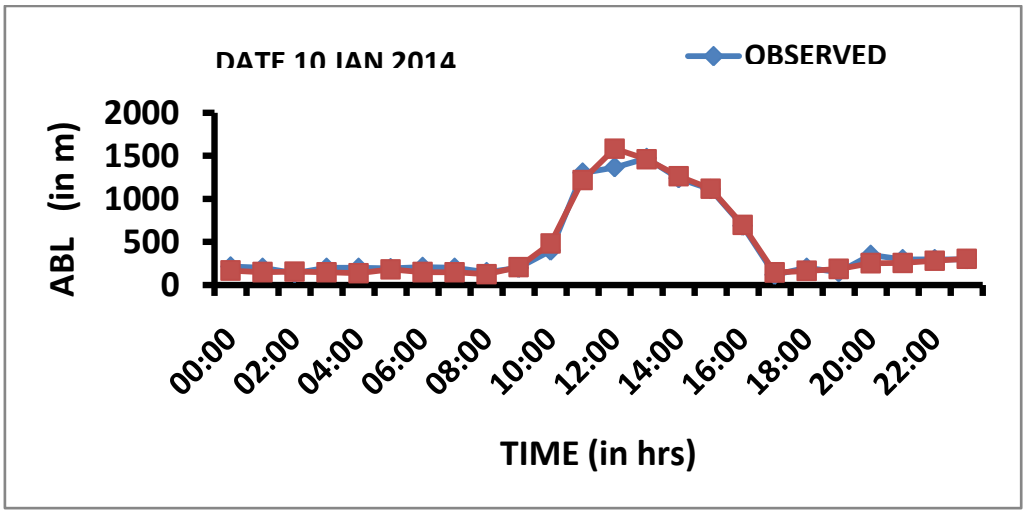
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

In the figure 2 relations between MSE and epochs is shown, where we notice that the test and validation lines are much similar but the training line decreases continuously indicating that the error in the output is decreased to much lower level.

The prediction is done by using ANN model and the outputs are finally obtained. Figure 3 shows the graphical representation of observed and predicted data for each hour of ABL of a day of using ANN model where blue line represents the observed values and maroon line shows predicted value. From the graphs itself it's clear that our modeling have shown a better result as expected. MAE value for ANN model is 39.79. Descriptive statistics (d) is 0.997 which is nearly equal to one which states that ANN model is highly accurate to determine the prediction about ABL height.

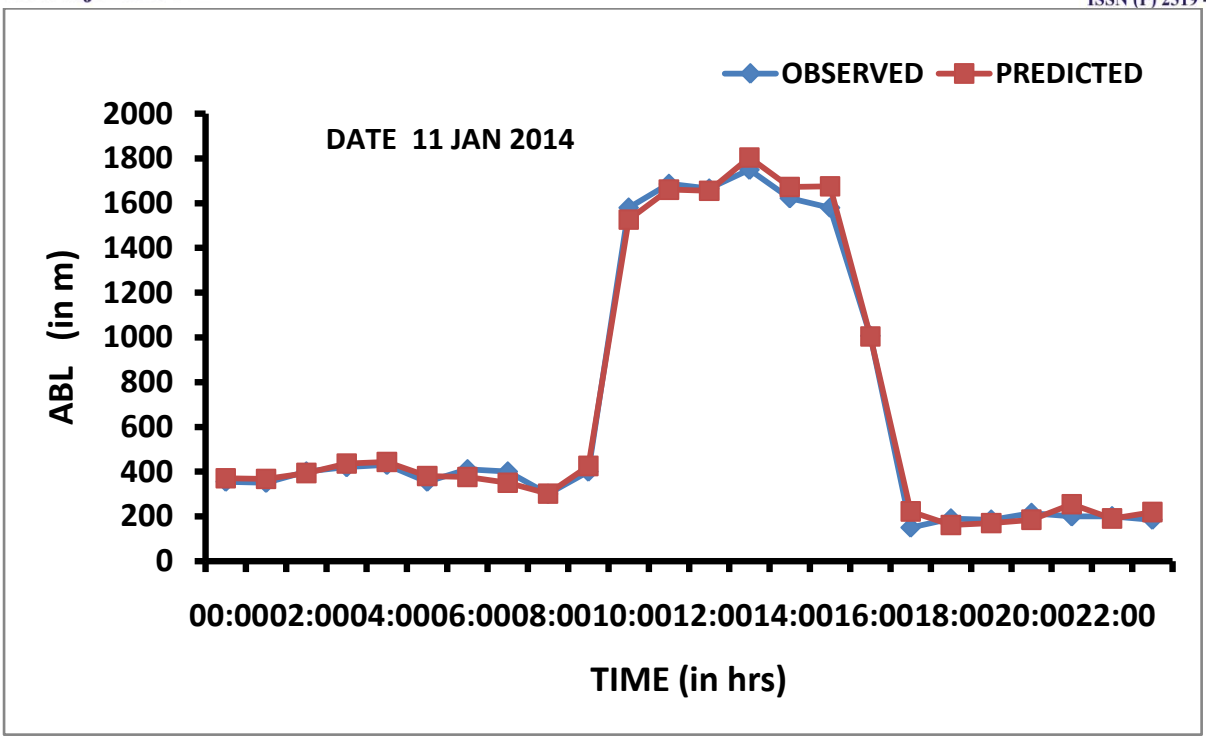


(a) 09 Jan 2014

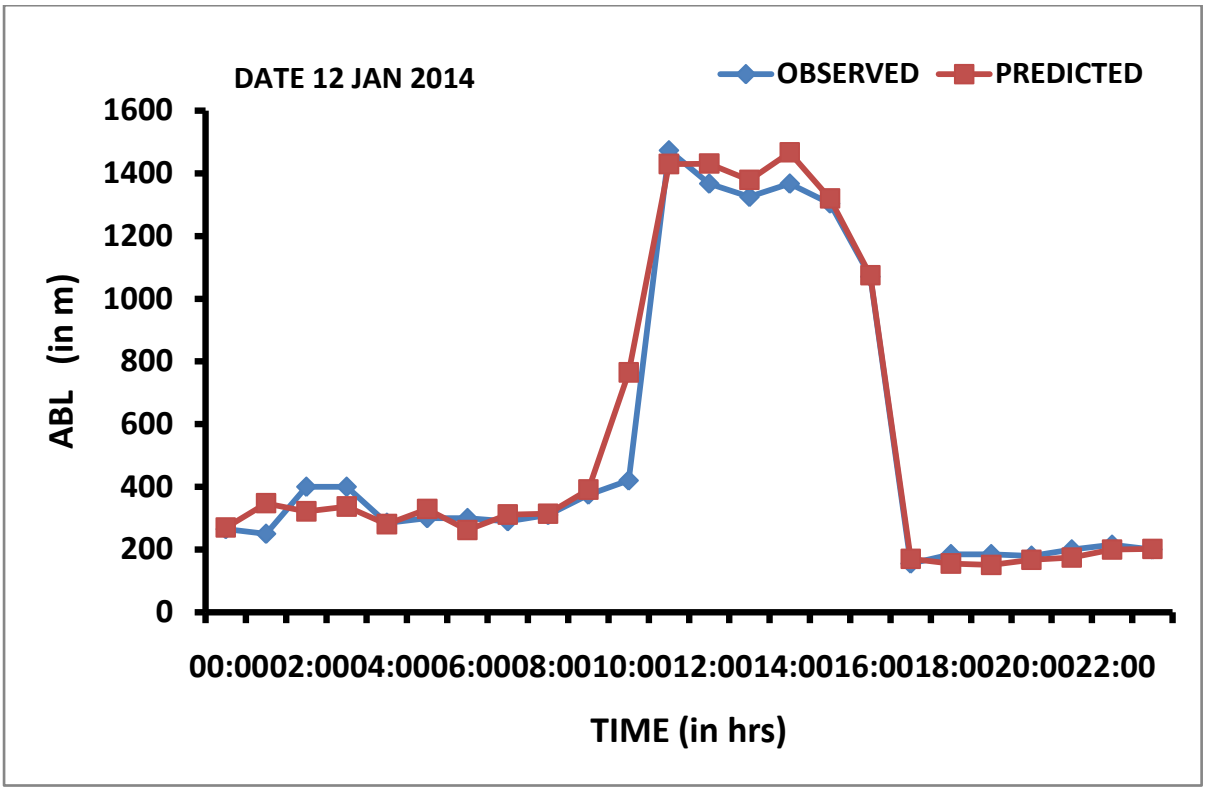


(b) 10 Jan 2014

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346



(c) 11 Jan 2014



(d) 12 Jan 2014

Figure 3: Observed and Predicted graphs of Atmospheric Boundary Layer Height per hour from 09 -12 Jan 2014 using ANN model.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

(B) FORWARD SELECTION

It can also be called as input selection method. In this study, we have a total of eight input candidates (CO, NO, SO_2 , Ozone, Temperature, Wind Speed, Humidity, and Solar Radiation). Forward Selection method is used as a linear input selection technique which determines the best suited subset from these eight variables. It can also be said that it is a linear model which is established using finest correlated subset on inputs. In the beginning a correlation among each input candidate is obtained and the output is calculated. Table 1 depicts the result of individual correlation of dependent variable (Atmospheric Boundary Layer Height) with other independent variables. In the very first step, the variable with the maximum correlation that is Solar Radiation with R^2 = 0.595, is carefully chosen as the first and the utmost vital input. This shows that solar radiation have a major impact on ABL height. Then the entire remaining variable according to the decreasing correlation is added to the model individually in step by step format. The modeling goodness is estimated by one of the parameter called as coefficient of determination or also called as R-squared (R^2). Secondly, it is temperature which plays a crucial role in variation of ABL height according to our data i.e. coefficient of determination for temperature is 0.465. Then we have wind speed at third number and the pattern goes as humidity, ozone, NO, CO, and finally in the end it is SO_2 . It is SO_2 which rarely affect ABL height since its R-squared value is less than 0.1.

Table 1: Individual correlation of dependent variable (ABL) with other independent variables

Correlation of ABL with	R^2
Solar Radiation (SR)	0.595
Temperature	0.465
Wind Speed (WS)	0.455
Humidity	0.432
Ozone	0.276
NO	0.149
CO	0.055
SO_2	0.015

After obtaining the individual values of R square for each single parameter, this process is repeated many times until that adding a new variable to input does not significantly advances the model output. It is observed that the value of R^2 increases as the new combination is made with a new variable such as value for individual solar radiation is 0.595 but as soon as temperature is added in combination with it the value increases to 0.68. This step is continued until the new variable is so selected that the increase in the value of correlation coefficient is over 5%. Lastly, input variables with utmost vital effect on output are selected and rest variables are removed. Result of Forward Selection procedure is depicted in Table 2 where four candidates according to their importance are selected as input variable. They are solar radiation, temperature, wind speed and humidity and the rest are removed. After these variables there is rarely any increase in in the value of determination coefficient so variation is almost negligible. Thus, inputs correlated to this value are carefully chosen.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Table 2: Result for forward selection procedure.

INPUT SUBSET	R^2
SR	0.595
SR, Temp	0.680
SR, Temp, WS	0.768
SR, Temp, WS, Hum	0.771[a]
SR, Temp, WS, Hum, O_3	0.773
SR, Temp, WS, Hum, O_3 , NO	0.773
SR, Temp, WS, Hum, O_3 , NO, CO	0.773
SR, Temp, WS, Hum, O_3 , NO, CO, SO_2	0.776

[a] After this value, variation of R^2 are negligible and thus, inputs related to this value are selected.

(C) For FS ANN

Figure 4 shows the graph between outputs of the network to the targeted value. The four plots represent training, validation, testing and all data. The dashed line in each plot represents the perfect result i.e. outputs is equal to targets. The solid line represents the best fit linear regression line between target and outputs. The value of R for training, validation and testing is found to be approximately 0.96, 0.99 and 0.99 which shows that the output values are in perfect linear correlation with our target values. When correlation coefficient is multiplied with itself it gives us the value coefficient of determination. Hence R squared value obtained for calibration (which is the average of training and validation of R value and then its square) is 0.961 and for testing it comes out to be 0.996.

In the figure 5 relations between MSE and epochs is shown, where we notice that the test and validation lines are greatly similar after 400 epochs but the training line decreases continuously indicating that the error in the output is decreased to much lower level.

The numerical experiment for predicting the Atmospheric Boundary Layer has been performed. The instrumental recorded data for ABL have been collected for every hour from 09 Jan. 2014 to 12 Jan. 2014. So, in total 96 data is taken into consideration. The outputs are obtained by using ANN whose results have been assessed to calculate R^2, MAE, and d.Figure 6 shows the graphical representation of observed and predicted data for each hour of ABL of a day using FS ANN where blue line represents the observed values and maroon line shows predicted value. From the graphs itself it is clear that our modeling have shown a better result as expected.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

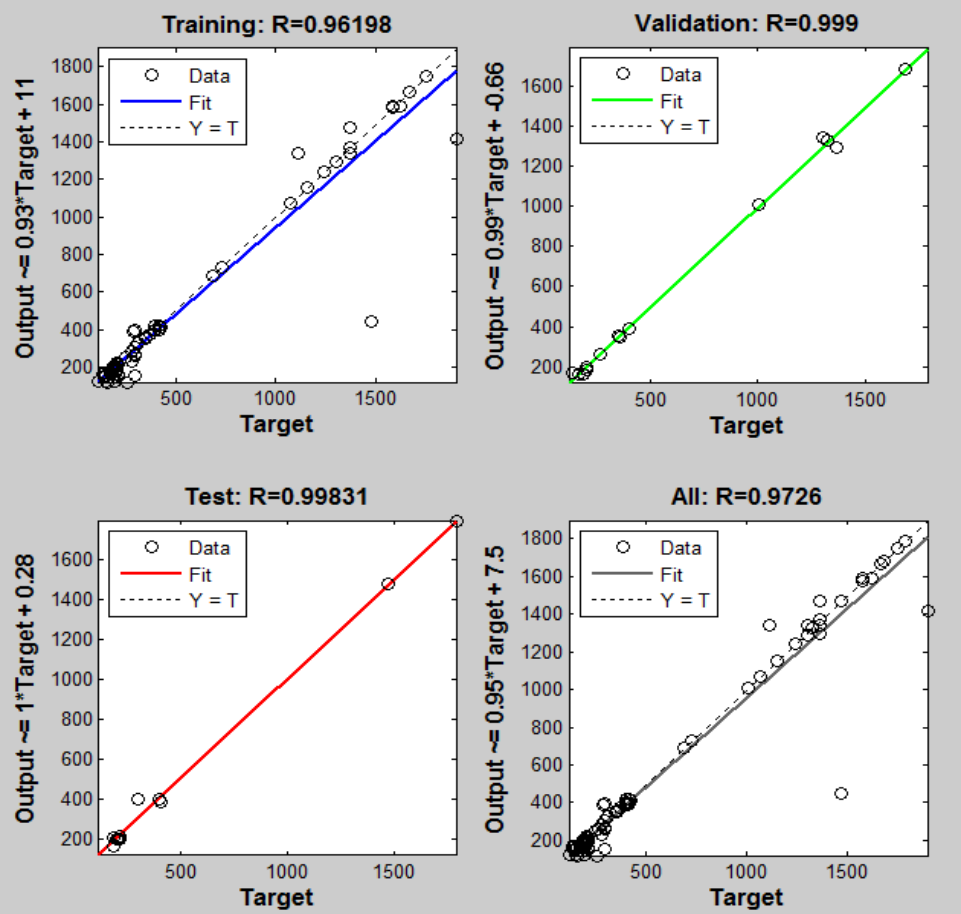


Figure 4: Plots of regression for training, validation and testing

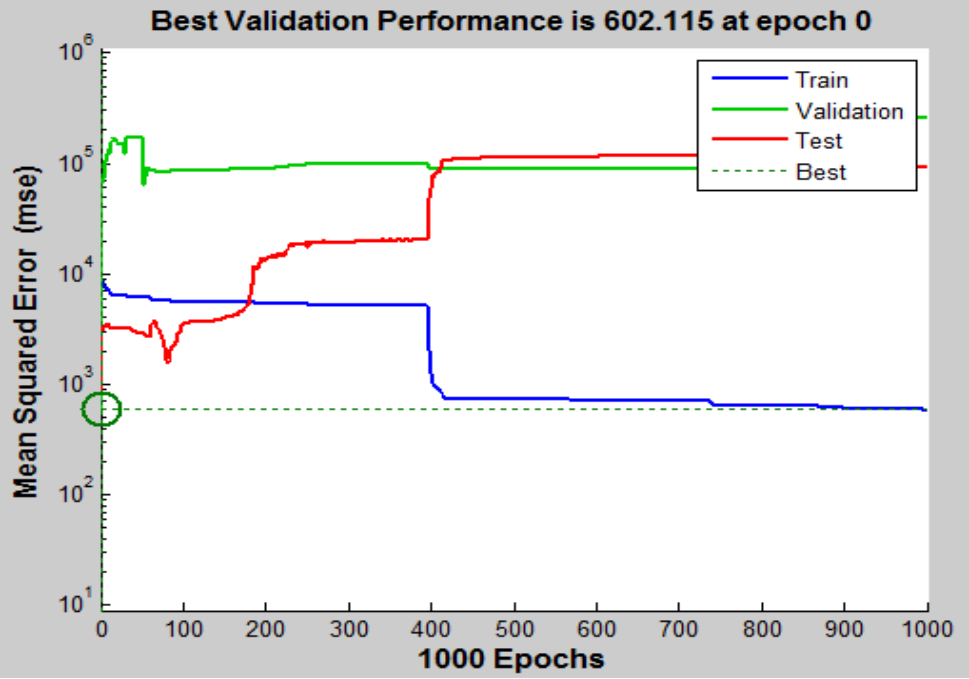


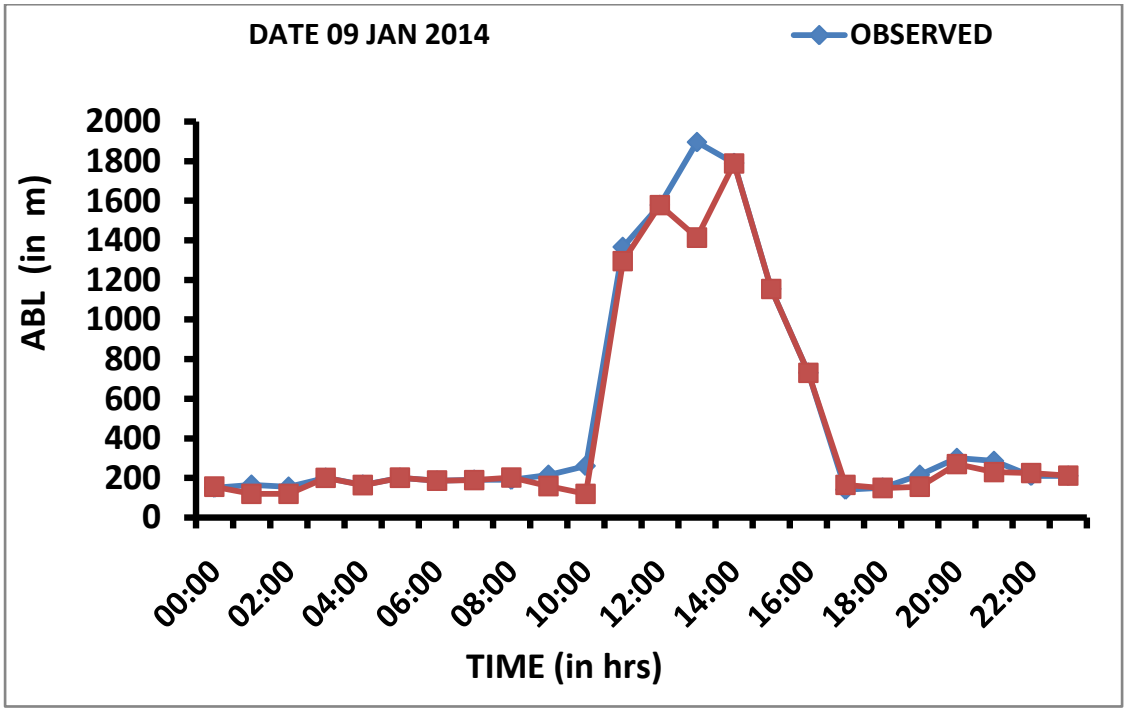
Figure 5: Plot of Performance for 1000 epochs

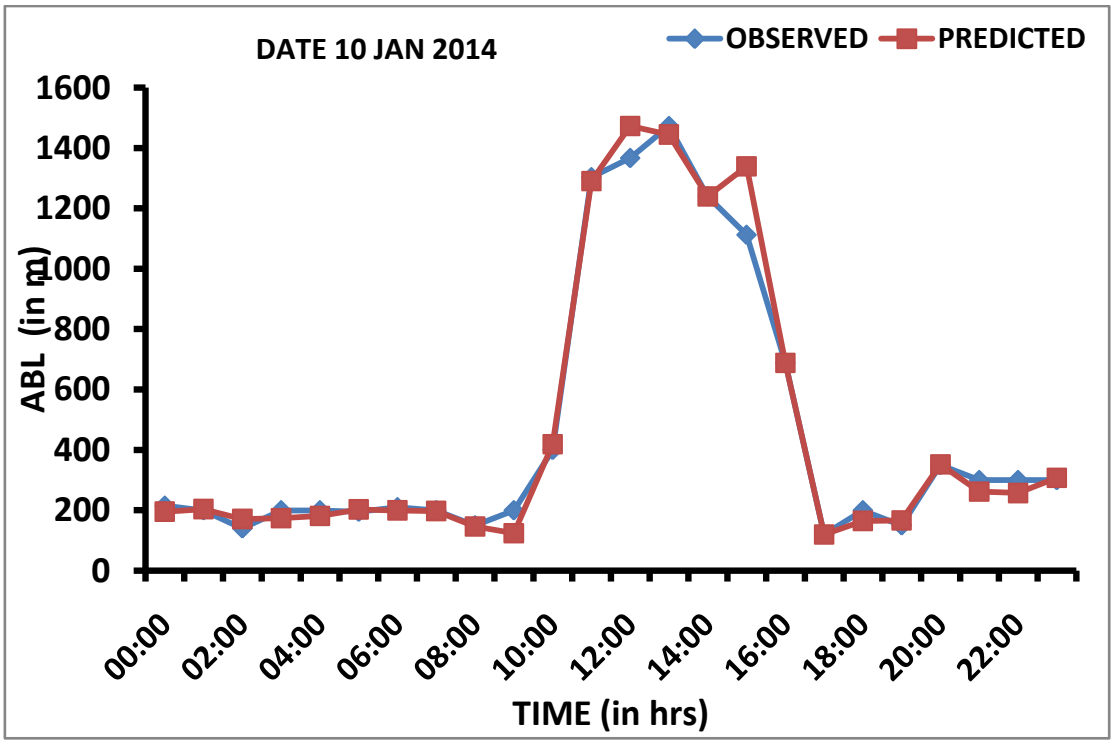
MAE value for ANN model is 27.08. Descriptive statistics (d) is 0.996 which is nearly equal to one which states that ANN model is highly accurate to determine the prediction about ABL height.

International Journal of Advance Research in Science and Engineering
Vol. No.6, Issue No. 05, May 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

(a) **09 Jan 2014**

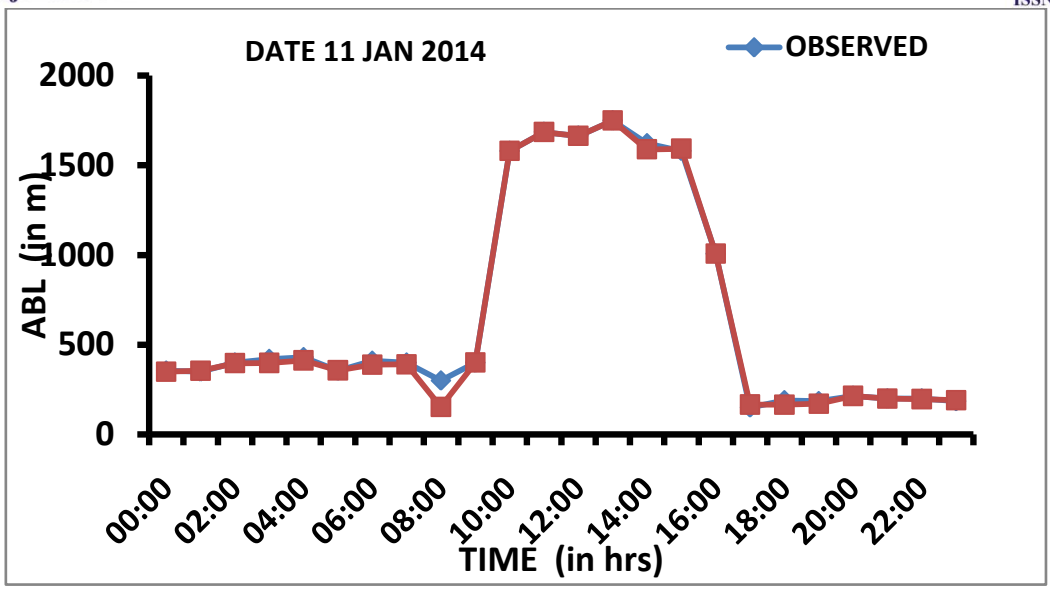
(b)



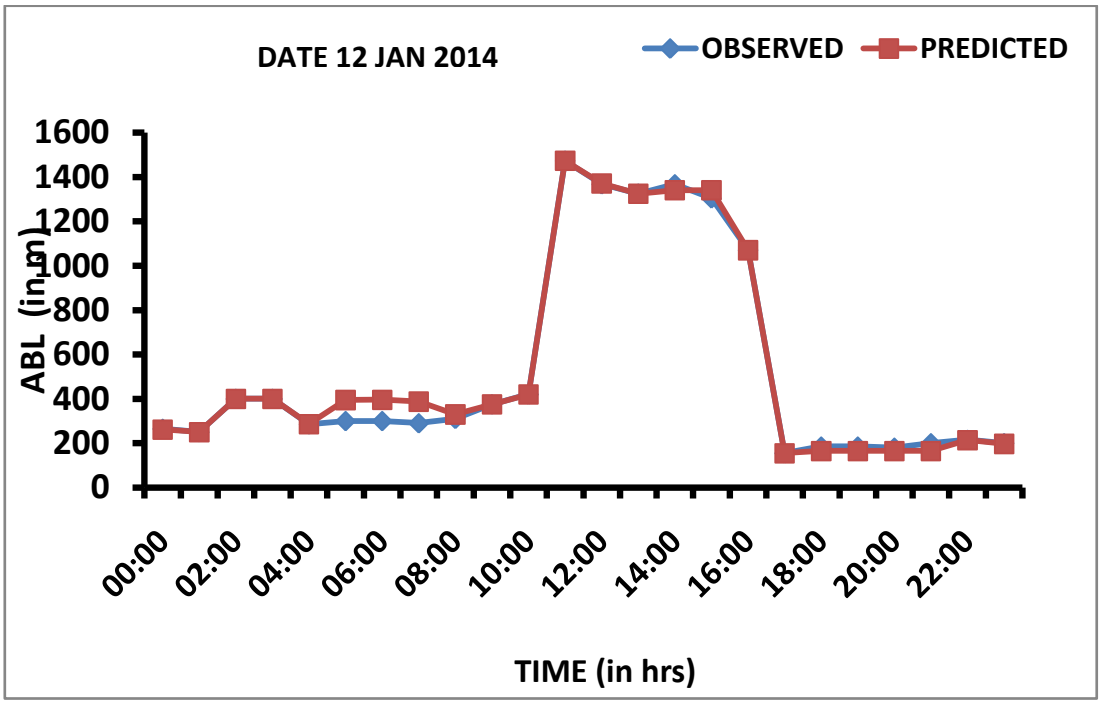


(c) 10 Jan 2014

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346



(d) 11 Jan 2014



(e) 12 Jan 2014

Figure 6: Observed and Predicted graphs of Atmospheric Boundary Layer Height per hour for 09 -12 Jan 2014 using FS-ANN model.

Both models predicted the value for daily atmospheric boundary layer height to their best. But there are some differences in both of them. For ANN model the number of inputs is more i.e. 8 where as for FS ANN model it less i.e. 4 which make the calculation much easier and faster in case of FS ANN

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

model. For R squared value, ANN model is more efficient than FS ANN model but testing results of FS are a bit more accurate.

From the Table 3 it is seen that Mean absolute error for ANN model is more which shows that uncertainty is high in case of this model whereas FS ANN model high much accurate prediction. Descriptive statistics for both the result is nearly equal to one which shows both models are preferable in the prediction statistics.

Table 3: Results of calibrating and testing of ANN and FS-ANN models

Model	Number of Input Variables	Calibrating R^2	Testing R^2	MAE	d
ANN	8	0.990	0.995	39.79	0.997
FS ANN	4	0.961	0.996	27.08	0.996

III. CONCLUSION

This study objective is to develop a suitable prediction model for determining atmospheric boundary layer using ANN model. Subsequently, input selection is one of the utmost important steps in modeling, Forward Selection (FS) scheme is used and the model is established. The goodness of respective model is estimated by means of d, R^2, and Mean Absolute Error (MAE) statistics. As a final point, uncertainty analysis of FS ANN is originated to be superior model. The subsequent conclusion can be drawn as of the current study:

1. Input selection advances prediction ability of ANN model. It moderates not simply the output error, but similarly on the other hand reduces the time period of calculation caused by having fewer numbers of input variables.
2. Four is the number of input variables selected for FS ANN model in comparison to eight candidates chosen for simple ANN model. FS established model is more appropriate than simple models.
3. As R^2, d and MAE data, FS ANN is initiated to be well improved model.
4. FS ANN model provides the accurate prediction of ABL for each day as compared to prediction done by ANN.

V. REFERENCES

[1.] Perez, P., Trier, A., Reyes, J., 2000. Prediction of PM2.5concentrations several hours inadvance using neural networks in Santiago, Chile. Atmospheric Environment34, 1189-1196

[2.] Noori,R.,Abdoli,M.A.,Jalili-Ghazizadeh,M.,Samifard,R.,2009 .ComparisonofANNandPCA based multivariate linear regression applied to predict the weekly municipalsolid waste generation in Tehran. Iranian Journal of Public Health 38, 74- 84.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

[3.] Noori, R., Karbassi, A., Farokhnia, A., Dehghani, M., 2009.Predicting the longitudinaldispersion coefficient using support vector machine and adaptive neuro-fuzzyinference system techniques. Environmental Engineering Science 26, 1503-1510.

[4.] Noori, R., Sabahi, M.S., Karbassi, A.R., 2009. Evaluation of PCA and gamma testtechniques on ANN operation for weekly solid waste predicting. Journal ofEnvironmental Management doi:10.1016/j.jenvman.2009.10.007