Meteorological Parameter Modeling with Different Soft Computing Techniques

Nilendu Das[1], Arpit Gupta[2], Rajarshi Bhattacharjee [1*], Shishir Gaur[1] and Anurag Ohri[1]

[1] Department of Civil Engineering, IIT(BHU), Varanasi-221005, India

[2] Department of Electronics Engineering, NIT, Raipur-492010, India

dasnilendu2016@gmail.com

arpit30nitrr@gmail.com

rajbhatt78645@gmail.com

shishirg.civ@iitbhu.ac.in

aohri.civ@iitbhu.ac.in

*Corresponding author: rajbhatt78645@gmail.com

Abstract-Meteorological measurements for weather forecasting and climatology have been carried out on a regular basis for centuries. There are several meteorological parameters like temperature, rainfall, humidity, speed of the wind, etc. By studying and observing these parameters, one can tell about the air pollution of an area or maybe the humidity present in the atmosphere. One can also predict cyclones or any natural calamities related to them. Numerous methodologies and strategies have been adopted for the analysis of these parameters. However, the data acquired can only be evaluated and interpreted after having statistically recorded medium-term and long-term atmospheric conditions. One of the most efficient tools for analysis is the soft computing techniques. These techniques have numerous advantages, as these techniques can be used for prediction studies and also for finding out any trends or patterns. In this paper, several soft computing techniques like linear regression, logistic regression, k-nearest neighbor, random forest regression (RFR), and support vector regression (SVR) are used for modeling of these meteorological parameters, and a comparative analysis has been shown. The linear regression technique is giving very poor results for the modeling of most of the parameters. RFR and SVR mostly showing high

accuracy rates for most of the meteorological parameters and these two techniques are quite efficient in comparison to other methods for showing the trend.

Keywords:* *Rainfall, Temperature, Random Forest Regression, Support Vector Regression, k-Nearest Neighbor

1 Introduction

Meteorology is a discipline of atmospheric science that deals mainly with weather forecasting issues. The study of weather forecasting is dated back up to several centuries, and with the due passage of time, new techniques have come into existence to enhance this forecasting in a much more efficient way. Meteorological processes are defined and quantified by several variables like temperature, air pressure, water vapour, etc. Several scales have been defined and used to predict the trends of weather on local, regional, and global levels (Hellmann, 2007). Due to the disturbances within the atmosphere, there can be some abnormal behavior in the meteorological parameters, but this abnormality is very short-lived. To know about the local climate of a place, we have to be well acquainted with the lower atmospheric phenomenon, and this phenomenon can be described by meteorological parameters. These parameters are very helpful for studying air pollution, avalanche warning, forestry, agriculture, water supply, town planning, etc. For example, if one has knowledge about solar radiation as well as air temperature and humidity, then that fellow can understand about chemical reactions of the pollutants present in the air. So a detailed study of air pollution can be done (Rösemann, 2011). Similarly, the understanding of cyclone genesis, its development, and features can be done with the help of meteorological parameters. This study is a very challenging subject, and it had been studied continuously for the last few decades till the present time (Henderson-Sellers, 1998). Without analyzing the observed atmospheric variables, it is impossible to do weather forecasting. This kind of analysis invariably involves the statistics in one form or the other. Statistical methods are used for decades for analysis of these parameters. A mainly concise introduction to the topic of statistical modeling of meteorology is given by a mathematical statistician Walter Freiberger, Brown University "*We believe that the behavior of the atmosphere is governed by certain dynamic equations, namely the Navier-Stokes equation, although we shall in practice certainly wish to use one of the several model atmospheres proposed for forecasting purposes. Statistics now come in through our ignorance or uncertainty of the initial weather field. This uncertainty will correspond to a probability superstructure imposed on the ensemble of possible solutions. And prediction would proceed from there.*" The data acquisition of the meteorological parameters can be evaluated and interpreted with various statistical and probabilistic methodologies. The variation in the various manifestation of the weather is very dynamic in nature because weather patterns can change very rapidly within a day also. In general, the weather phenomenon can vary day to day, month to month. The combination of the methods of getting strict dynamical solutions or what is also known as deterministic solutions, with the technique of describing quantitatively the initial and resulting in following uncertainties of meteorological variables is what we have known as the stochastic-dynamic prediction. The origin for stochastic-dynamic prediction is elementary; the zero or initial conditions in the time-dependent solution of the deterministic equations are mostly unknown. Therefore even if the model is made and available, the integration is intrinsically probabilistic in nature despite the fact that not overtly recognized as being so. By defining initial conditions in terms of probability density functions, the feature of the collection of possible initial conditions and adding equations were explaining the conservation of the probability, mainly involving the moments of probability distribution functions. So to put different data's which shows huge variation, in the order, we need to deal with the frequencies and averages. Most of the known tests of the hypothesis in the statistical domain and, in particular, the variance analysis are based on the assumption that the variables or parameters observed and recorded, more often than not, these variables are normally and homoscedastically distributed. Another assumption is that the expected effects of various methodologies are additive. The way workflow of the experiments has been designed, these two above-mentioned assumptions mostly dominate. At several points of the time, these assumptions are clearly contradicted by the observations, so at those moments, we conventionally make arduous efforts to approximate the analysis of variance situation by an appropriate transformation of the observable variables. In several cases, this conventional process is very successful. However, if we are working on the rainfall data, the rainfall amounts 'X' per experimental unit (day, storm, etc.), the hindrances are very considerable. The experimental study of weather science has two disadvantages in comparison to classical experimental studies like agriculture. First, in agricultural studies, a single experiment may test several parameters without any time loss. But in the case of weather studies, this sort of experiment is not possible. Second, in the case of a classical experiment, time duration maybe about a year or so, but for weather science experiments, the duration of experimental time increases significantly. Because of these two above-mentioned disadvantages, there is a difference in the relative importance of certain factors of the problem of the experimental set-up and in the problem of the suitable statistical test. For classical instances, there is the only trivial increase in cost. An experimenter with all his literature surveys and knowledge can set up a design that how long the experiment will be carried and what sort of results are expected. If somehow the mistake becomes larger than what is anticipated, that is repeatable, but the loss is not too havoc because the experimental setup can be carried out during the next year. But for weather studies, this kind of scenario may not exist because weather conditions may not be the same for two consecutive years for the same month. So the experimenter needs to be extra cautious while carrying out the experiment with weather studies (Freiberger and Grenander, 1965; Julian and Murphy, 1972; Neyman et al., 1969). In recent times, most analysis has been done with different numerical prediction methods. These models are the mechanism for combining the records recorded on an unequally spaced-net with the best guess of conditions existing at the analysis time. Such methods have practical advantages, a quantitative treatment of what is really done in interpolating from the unequally spaced observing net to the geometric grid of the numerical prediction model is almost impossible. An easy, purely statistical approach is to reduce the variation in between the calculated value of a scalar variable inside the grid point and the actual value recorded for that grid in the field with the help of some sensors. The development of the purely statistical theory was first formulated by Soviet scientists, mainly by *Gandin* in 1963, and it appears under the subject of *Optimum Interpolation Theory*. This theory tells us about the statistics calculated on the unequally-spaced observing net (covariance functions in space and possibly time) to interpolate to the geometric grid of the numerical model. To execute this theory, there is a mandatory assumption that the structure of the atmospheric scalar field is- I) isotropic II) homogeneous in time and space domain and III) homoscedastic i.e. space-Time plane of the scalar field, is independent of the absolute value of scalar itself. But unfortunately, these assumptions are not being fulfilled, so the validity of this theory is severely impeded. Now the scientific community started to do work for relaxing the severity of these assumptions. So it seems that the severity will only be reduced, and some progress will be made if- I) more computational power and II) more sophisticated theories are available. From a purely statistical least-square method, the “best analysis” is not considered, but instead of that “best analysis” is taken from those approaches that incorporate some dynamic constraints. Now with this change, a new class of analysis schemes comes into existence. In the meteorological science domain, the statistical computation of reliant variables has been performed as a function of two independent variables (space and time). The analysis is mostly concentrated upon second moments or the variances and co-variances. For the past few years, a great amount of work has been done for analyzing the meteorological variables in the Eulerian frame of reference. These variables have two different spectra i.e. temporal and spatial spectra. Earlier, the temporal spectrum of meteorological data used as an analytical analysis tool mostly by those people who calculate atmospheric turbulence. The interpretation of the Eulerian time spectra for meteorological variables is difficult because these variables show non-linear and dispersive nature in significant wave modes. To remove this non-linearity time series concept has been introduced. The advancement of Fast Fourier Transform (FFT) has made the analysis of spectrum radically more competent by diminishing the computational time and by giving full range of harmonic coefficients (Julian and Murphy, 1972). Time series analysis of meteorological data was also performed by some researchers (Cuervo et al., 2018).

Some of the statistical and soft computing methods used in this work for the analysis of some meteorological parameters are described.

1.1 Linear Regression Model

In the statistical domain, linear regression is defined as the linear method of modeling the relation among scalar quantity (or dependent variable) with explanatory variable (or independent variable). In this method, the relationship is tuned using the *linear predictor function*, whose unidentified model variables are calculated from the data. This model has a focus on the conditional probability distribution of the output given the value of predictors rather than on the joint probability distribution of all of these variables. This analysis has extensive practical application. If the aim is prediction and forecasting, then the linear regression model is used as a fitting for a predictive model to an observed data set of values of the output and independent variables. This method uses the least square technique (Freedman, 2009; Seal, 1967; Yan and Su, 2009).

1.2 Logistic Regression Model

This model use probability for predicting the result. Each of the results detected is given a probabilistic value between 0 and 1and sum adding to one. This technique uses the logistic function to evaluate binary dependent variables, although several difficult extensions still exist. Mathematically, two possible values of the dependent variable are always associated with a binary logistic model, and these two values are 0 and 1. In this model, one or more independent variables are linearly combined, and their combining value is marked as 1. These independent variables are mostly of two types one is a binary variable, and another one is a continuous variable. The analogous probability of the value marked “1” can fluctuate between 0 (surely the value “0”) and 1 (surely the value “1”), hence the labeling. The function that converts the log-odds to the probability is the logistic function, hence the name. The probit function model can also be used instead of the logistic model, and this probit model has a different sigmoid function. The major feature of the logistic model is that if we increase one of the independent variables, it multiplicatively increases the odds of a given result at a steady rate, with each self-determining variable connected with its own parameter. If there are more than two levels of the dependent variable, then this binary logistic regression model can be extended to multinomial logistic regression. In this kind of logistic regression, categorical outputs with more than two values are modeled (Walker and Duncan, 1967).

1.3 k-Nearest Neighbor (k-NN) Algorithm

This algorithm is used for classification and regression. It is non-parametric in nature. For both classification and regression, the input has k nearest training examples in feature space. The output is reliant on if k-NN is used for either classification or for regression. In the case of classification, output comes as class membership. Classification of the object is done on the basis of plurality votes from its neighbors. The object is defined for that class that is most familiar among its k nearest neighbor. The value of k is typically small, and it is a positive integer. If the value of k is 1, then the object is allocated to the class of that single nearest neighbor. For the case of regression, the output is like the property value of the object. The average value of all the k nearest neighbor defined as property value of the object. In this algorithm there is a local approximation of the function under consideration, and all the calculation is postponed until the classification has been performed. This algorithm is a kind of instance-based learning (Altman, 1992; Garcia et al., 2011).

1.4 Random Forest Regression

Random forest is used for classification based on the collection of the learning algorithm. This technique can be used for the prediction of the continuous random variable. This regression and classification technique models a collection of decision trees to set a date. In each of the trees, the data are repetitively divided into more identical units, which are usually defined as nodes, for improving the prediction analysis of the response variable. Values of the prediction parameter are the basis of the split points. So the variables used for data splitting are considered as important descriptive variables—random forest models different decision trees to a preexisting number of bootstrapped data sets. In the random forest methodology, the output is the mode of the classes or average forecast of individual trees. *Tim Kam Ho* first proposed the algorithm for random forest regression using random subspace method (Barandiaran, 1998; Breiman, 2001; Everingham et al., 2016).

1.5 Support Vector Regression

Support vector algorithm is one of the tools of machine learning, and it works on feed-forward network mechanism. In this method, the learning algorithm is associated with supervised learning models that evaluate the data used for regression analysis and classification. Support vector regression is helpful for solving quadratic equations with linear constraints. The SVM is a linear machine of one output y(x), working in the high dimensional feature space formed by the nonlinear mapping of the N-dimensional input vector x into a K-dimensional feature space (K>N) through the use of the nonlinear function φ(x). The number of hidden units (K) is equal to the number of so-called support vectors that are the learning data points closest to the separating hyperplane (Cortes and Vapnik, 1995; Osowski and Garanty, 2007; Vapnik, 1998).

2 Materials and Methods

The meteorological data has been obtained from India Meteorological Department (IMD), Ozone Unit Banaras Hindu University. The data is of the Varanasi region. In this study, we have used data from 2014 to 2016 to train our models and data from 2017 to test our model. The meteorological parameters for consideration are rainfall, wind speed, relative humidity, and temperature. Daily data has been converted to monthly data by taking the average of a month. The value of the wind speed is given in *Knots*, but for calculation, we have converted this value into kilometer per hour by multiply the knot value by 2. Data pre-processing is not required in this case as there are no missing values. The numbers of observations for the meteorological parameters are given in **Table 1**. The flowchart of methodology has been given in **figure 1**.

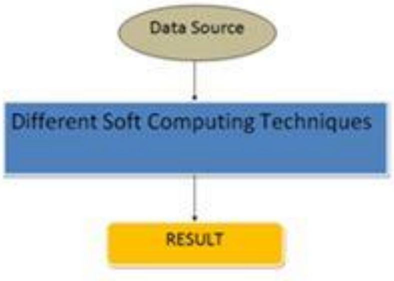

Figure1. Flowchart of the methodology

Table1. Parameters in the dataset

Parameter	Records	Units
Temperature	1461	°C
Relative Humidity	1461	%
Rainfall	1461	Millimeter (mm)
Wind Speed	1461	Knots (converted to Km/h)

3 Results and Discussions

The calculation and analysis have been done in a Python environment, and in this case jupyter notebook has been used. With the help of different soft computing techniques, we have predicted some meteorological parameters for the year 2017 for the Varanasi region.

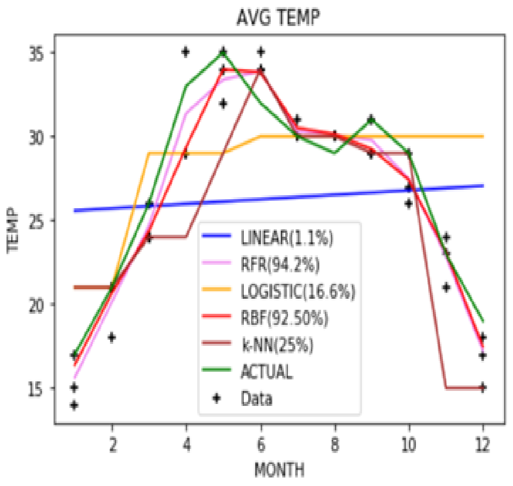

Figure2. Temperature prediction

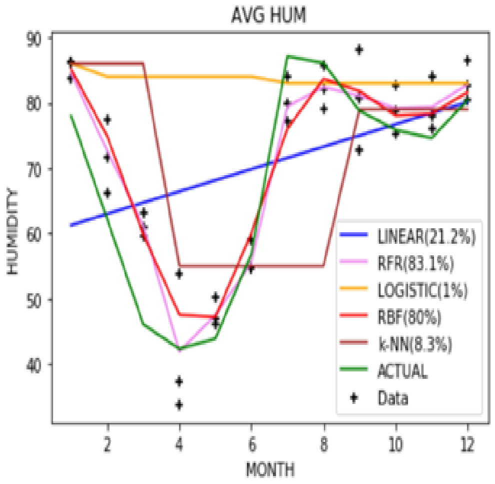

Figure3. Humidity prediction

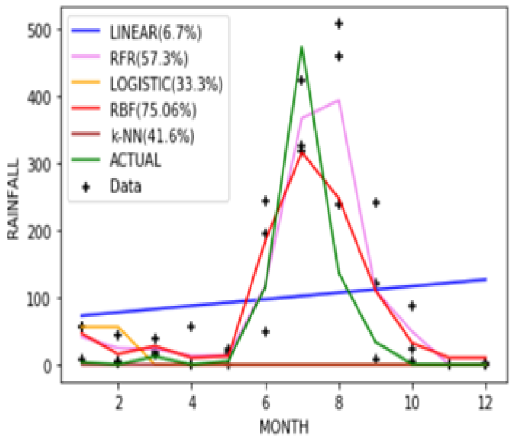

Figure4. Rainfall prediction

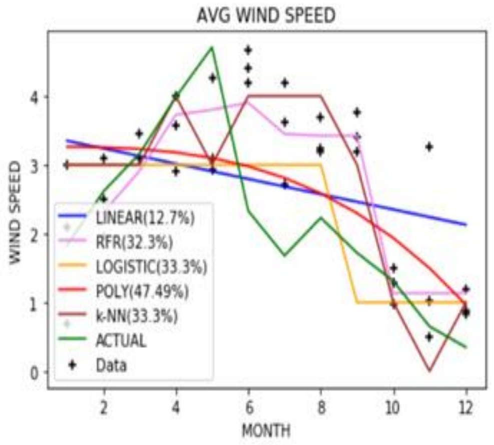

Figure5. Wind speed prediction

If one can analyze these comparative graphs (**figure 2, figure 3, figure 4, and figure 5**) for the parameters, then one can find that for temperature, and humidity prediction, the random forest regression method gives the highest accuracy. But for rainfall prediction, radial basis function (RBF) has the highest accuracy. RBF is a form of support vector machine (SVM). Lastly, for the wind speed forecast, the polynomial support vector machine (SVM) has been most accurate. It has also been seen from the above graphs that the accuracy percentage for wind speed data is the least for all the statistical methods applied to it. Even for the rainfall, the accuracy percentage is somewhat low. The advanced soft computing methods like random forest and support vector machines show a high accuracy rate in comparison to the linear regression technique. If more data can acquire then more improved results, one can expect.

4 Conclusions

Meteorology is very important because life cannot be sustained without air, especially oxygen. Weather forecasting plays a very important role in urban planning. Prediction of the weather for a long duration is also very vital for agricultural purposes. The study of the parameters related to meteorology also helpful to monitor global warming patterns. In short, meteorological parameters plays a very important role in our lives. Therefore the study and analysis of these parameters are also very essential. Soft computing techniques can be used very widely to analyze these parameters. These techniques can also be used for analyzing the trend and pattern. Several methods are applied, and that yields different results for different parameters.

References

Altman, N.S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression, *The American Statistician,***46,** no.3, 175-185.

Barandiaran, I. (1998). The random subspace method for constructing decision forests, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **20,** no.8, 832-844.

Breiman, L. (2001). Random forests, *Machine Learning,***24,** no.2, 123-140.

Cortes, C. and Vapnik, V. (1995). Support-vector networks, *Machine Learning,* **20,** no.3, 273-297.

Cuervo, E.C., Achcar, J.A. and Andrade, M.G. (2018). Seasonal Hydrological and Meteorological Time Series, *Earth Sciences research Journal,* **22,** no.2, 83-90.

Everingham, Y., Sexton, J., Skocaj, D. and Inman-Bamber, G. (2016). Accurate prediction of sugarcane yield using a random forest algorithm, *Agronomy for Sustainable Development*, **36,** no.2, 27.

Freedman, D.A. (2009). *Statistical models: theory and practice,* Cambridge University Press, London, UK, 458pp.

Freiberger, W. and Grenander, U. (1965). On the formulation of statistical meteorology, *Review of the International Statistical Institute,***33,** no.1, 59-86.

Garcia, S., Derrac, J., Cano, J.R. and Herrera, F. (2011). Prototype selection for nearest neighbor classification: Taxonomy and empirical study, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **34,** no.3, 417-435.

Hellmann, G. (2007). The dawn of meteorology, *Quarterly Journal of the Royal Meteorological Society*. **34,** no.148, 221-232.

Henderson-Sellers, A., Zhang, H., Berz, G., Emanuel, K., Gray, W., Landsea, C., Holland, G., Lighthill, J., Shieh, S.L., Webster, P. and McGuffie, K. (1998). Tropical cyclones and global climate change: A post-IPCC assessment, *bulletin of the American Meteorological Society,***79,** no.1, 19-38.

Julian, P.R. and Murphy, A.H. (1972). Probability and statistics in meteorology: a review of some recent developments, *Bulletin of the American Meteorological Society,***53,** no.10, 957-965.

Neyman, J., Scott, E.L. and Wells, M.A. (1969). Statistics in meteorology, *Review of the International Statistical Institute,***37,** no.2, 119-148.

Osowski, S. and Garanty, K. (2007). Forecasting of the daily meteorological pollution using wavelets and support vector machine, *Engineering Applications of Artificial Intelligence*, **20,** no.6, 745-755.

Rösemann, R. (2011). *A Guide to Solar Radiation Measurement: From Sensor to Application : an Overview of the State of the Art : UV, Visible, Infrared*, Kipp & Zonen publication, Delft, Netherlands, 218pp.

Seal, H.L. (1967). Studies in the History of Probability and Statistics, *Biometrika*, **54,** no.1-2, 1-24.

Vapnik, V. (1998). *Statistical learning theory*, Wiley-Interscience Publication, New York ,USA, 768pp.

Walker, S.H. and Duncan, D.B. (1967). Estimation of the probability of an event as a function of several independent variables, *Biometrika*, **54,** no.1-2, 167-179.

Yan, X. and Su, X. (2009). *Linear regression analysis: theory and computing*, World Scientific, Singapore, 328pp.