The present work discusses the development of a nonlinear data-fitting technique based on genetic algorithm (GA) for the prediction of routine weather parameters using observations from Agro-Met Stations (AMS). The algorithm produces the equations that best describe the temporal evolutions of daily minimum and maximum near-surface (at 2.5-meter height) air temperature and relative humidity and daily averaged wind speed (at 10-meter height) at selected AMS locations. These enable the forecasts of these weather parameters, which could have possible use in crop forecast models. The forecast equations developed in the present study use only the past observations of the above-mentioned parameters. This approach, unlike other prediction methods, provides explicit analytical forecast equation for each parameter. The predictions up to 3 days in advance have been validated using independent datasets, unknown to the training algorithm, with impressive results. The power of the algorithm has also been demonstrated by its superiority over persistence forecast used as a benchmark.

1. Introduction

In the recent past, nonlinear data-adaptive approach of genetic algorithm [1, 2] for the forecast of a particular meteorological parameter has gained considerable ground. This is particularly advantageous when a location-specific forecast of a specific parameter is required and when a long time series of observations (either in situ or remotely sensed) of the relevant parameter exists. The algorithm is computationally extremely cheap and produces good quality forecasts. An additional advantage of genetic algorithm (GA) is that an explicit analytical forecast equation is obtained as an outcome of the algorithm [3, 4]. Also, the algorithm can work with only a small number of data points, typically of the order of a few hundred. Predictive skill of GA has been amply demonstrated in the cases of sea surface temperature (SST) in the Alboran Sea [5] and Arabian Sea [6], summer monsoon rainfall over India [7], SST and sea level anomaly in the Ligurian Sea [8], wave heights in the north Indian Ocean [3] and more recently, for basin-scale predictions of ocean surface wind in the north Indian Ocean [9] and wave heights in the Bay of Bengal [10]. Even tidal currents from a few tidal levels have been predicted using GA [11]. At the outset it should be clarified that we are not attempting to provide an alternate method of weather prediction, which can be done only by a numerical weather prediction (NWP) model, since only such a model can provide multilevel forecasts of a large number of meteorological parameters over a wide area (encompassing sometimes even the entire globe) at regular time intervals. Rather, GA should be viewed as a useful method of forecast for those who need the forecast of a few meteorological parameters at just a few locations.

The focus of our study is the prediction of daily maximum and minimum near-surface air temperature and daily maximum and minimum near-surface humidity (at 2.5 m height) and daily averaged wind speed at 10 m height at isolated locations which are representatives of very small areas (of the order of 500 m × 500 m). Such forecasts are integral components of several crop forecast models [12]. Forecasts of these parameters from NWP models represent average over a large area (typically of the order of 100 km × 100 km), which does not serve the purpose of crop forecast models.
GA is, of course, not the only available technique for predicting a time series. There are other prediction techniques which, like GA, operate directly on a data time series, for example, techniques of polynomial fitting, nearest neighbors, and artificial neural networks [13]. Unlike GA, these techniques, however, use a priori fixed architecture. The first two methods are suitable primarily for systems obeying relatively simple nonchaotic dynamics [13]. Artificial neural networks (ANNs) learn the forecast rules directly from the data using input, hidden, and output layers. Although suitable for chaotic dynamics, the disadvantage is that the network architecture (the number of hidden layers, the number of neurons in these layers, etc.) is fixed a priori, often quite heuristically [13]. Thus the degree of freedom in this technique for searching the forecast rule is limited. GA, on the other hand, finds the architecture itself from the data using Darwinian evolution (to be explained later). The most important feature of GA is that it allows an effective search in a high-dimensional phase space. This flexibility is the reason for its success for systems with complex chaotic dynamics like natural geophysical phenomena. Another noteworthy advantage is that, unlike ANNs, it provides explicit, easy to use in practice, analytical forecast equations, which has already been mentioned by us.

As mentioned earlier, the focus is on the prediction of four maximum/minimum quantities and one average quantity on a daily basis. For this purpose, half-hourly observations collected by 5 micrometeorological towers (also known as Agro-Met Stations or AMS in brief) have been used. From these half-hourly observations, the quantities to be predicted have been computed and subsequently time series of these quantities have been subjected to forecast by GA, the details of which are described later in this paper.

2. Agro-Met Station (AMS) Data

The AMS are programmed to record air temperature, relative humidity, and winds at half-hourly interval at three vertical levels over different short vegetation cover types [14], apart from radiative and soil heat fluxes across different agroclimatic zones in India. The AMS data are available from the website (http://www.mosdac.gov.in/) maintained by Space Applications Centre, ISRO, India. These AMS stations are deployed during the year 2009. Good quality data from the next two years have been used in this study. Data from five AMS were used in the study (Table I). For each AMS, the last three months of data were used for validating the algorithm, while the remaining data were used for training the algorithm. Daily maximum and minimum values of air temperature and relative humidity were selected from half-hourly observations of these quantities, while daily average wind speed is simply the mean of all the 48 observations of wind speed at 10-meter height in a day.

3. Methodology

3.1. Genetic Algorithm (GA). The algorithm has been discussed in detail by several authors [1, 2, 9, 15]. However, for the benefit of the present readers, we describe below the essential details of the algorithm, borrowing heavily from the reference [15]. As mentioned above, the problem is to find a suitable prediction function $\beta$ such that the parameter concerned (say, $x$) at a particular time $t$ can be predicted from the past observations of the same parameter. In other words,

$$x(t) = \beta(x(t - \tau), x(t - 2\tau), \ldots, x(t - m\tau)), \quad (1)$$

where $m$ denotes the embedding dimension and $\tau$ denotes the time interval between successive observations. That such a function exists for a deterministic, albeit chaotic, time series is guaranteed by Takens’ theorem [16]. Unfortunately, the theorem is silent about the method to find the function. GA is one such method, which takes its cue from evolutionary biology. The starting point of the algorithm is an initial population of $N$ “equation strings” which are simply short equations connecting the past values of the parameter concerned. The connections are quite random, meaning two values of the parameter at arbitrary past times are connected randomly by an arithmetic operator (addition, subtraction, multiplication, or division). The coefficients of these connections are real numbers, taken randomly from a set, generally $[-10, +10]$. Every equation string is an approximation of the prediction function. The equation strings are ranked according to their strength, a measure of which is a well-defined mathematical expression. Before coming to this quantity, we note that (1) is valid only for one step ahead prediction. However, longer term predictability is also addressable by GA. This is simply the following obvious generalization:

$$x(t) = \gamma(x(t - \mu\tau), x(t - (\mu + 1)\tau), \ldots, x(t - (\mu + m - 1)\tau)), \quad (2)$$

valid for $\mu$ steps ahead prediction, $\mu = 1$ being the special case of one step ahead prediction defined by (1). Henceforth, we divide the time series into two parts, namely, the training set and the validation set. As the name signifies, the algorithm is trained using data in the training set, while the data in the validation set is withheld from the algorithm. Once the training is complete by virtue of satisfying a specific training criterion, an analytical forecast equation is obtained (which is simply the strongest equation set, or the fittest equation, following Darwinian terminology). This forecast equation is applied to the validation set, unknown to the algorithm, to see how the forecast performs. This is really a stringent test of the algorithm.

We now come to the actual training part. The $j$th equation string with mapping function $\gamma_j$ is used to compute estimates of all $x(t)$ in the training set, as a function of the previous values of the time series. The fitness for the equation string $\gamma_j$ can then be computed as

$$\Delta^2_j = \sum_{t-L+1}^{T} (x(t) - \gamma_j(x(t - \mu\tau), x(t - (\mu + 1)\tau), \ldots, x(t - (\mu + m - 1)\tau)))^2, \quad (3)$$
where \( L = m \tau \) and \( T \) is the total length of the training set.
One can further define a strength index for each individual in the following fashion:

\[
R_j = 1 - \frac{\Delta^2_j}{\delta^2},
\]

where

\[
\delta^2 = \sum_{t=L+1}^{T} (x(t) - \langle x \rangle)^2,
\]

where \( \langle x \rangle \) represents the mean value of the training data. \( R_j \) can be interpreted as the percentage of the training set’s total variance explained by the \( j \)th equation string. It is quite clear that the closer is the value of \( R_j \) toward unity, the stronger is the \( j \)th individual equation string.

The major steps involved in the algorithm are (1) initialization, (2) computing the fitness, (3) ranking the agents, (4) choice of mates, (5) reproduction and crossover, and (6) mutation. These steps have been elaborately described in [15]. We choose not to dwell any further on them, since the method of forecast, operating on the principle that geophysical phenomena are generally persistent in nature and their change is generally not a rapid one. Any forecast is meaningless unless it is able to improve on persistence forecast defined by the equation

\[
x(t + n) = x(t),
\]

where \( n \) denotes the lead time of forecast. In other words, persistence assumes that the present conditions are not going to change at all for \( n \) forecast steps. The following improvement parameter normally indicates the degree of improvement:

\[
\eta = \left( \left[ \frac{1}{N} \sum_{i=1}^{N} (AMS_i - \text{PER}_i)^2 \right]^{1/2} \right. \\
\left. - \left[ \frac{1}{N} \sum_{i=1}^{N} (AMS_i - \text{GA}_i)^2 \right]^{1/2} \right) \times \left[ \left[ \frac{1}{N} \sum_{i=1}^{N} (AMS_i - \text{PER}_i)^2 \right]^{1/2} \right]^{-1} \times 100,
\]

where “PER” is the persistence forecast, “GA” is the forecast produced using genetic algorithm, “AMS” is the AMS-observed value, and \( N \) is the total number of forecasts. A positive (negative) value of this parameter signifies
Table 1: AMS observations used in study.

| Locations of AMS station            | Latitude   | Longitude | Training period | Validation period | Land type   | Climate type | Vegetation type |
|-------------------------------------|------------|-----------|-----------------|-------------------|-------------|--------------|-----------------|
| Bharatpur, Rajasthan                | 27°12′11.00″N | 77°27′09.00″E | May 2010–Feb. 2011 | Mar. 2011–May 2011 | Desert plain | Arid         | Agriculture     |
| Zonal Agricultural Research Station, Powarkhedha, Hoshangabad, Madhya Pradesh | 22°41′53.46″N | 77°44′55.06″E | Apr. 2010–Mar. 2011 | Apr. 2011–June 2011 | Central Plain | Subhumid      | Agriculture     |
| Range Office Research Centre, Dhordo, Kutch | 23°48′03.99″N | 69°30′05.97″E | May 2010–Jan. 2011 | Feb. 2011–Apr. 2011 | Gujarat plain | Semiarid      | Dry grass       |
| Naraingarh Seed Farm, Patiala, Punjab | 30°36′44.92″N | 76°19′17.75″E | July 2010–Oct. 2011 | Nov. 2011-Jan. 2012 | Trans-Gangetic plain | Subhumid | Agriculture |
| Chakdaha, Nadia, West Bengal        | 23°03′00.57″N | 88°32′00.61″E | Mar. 2010–Feb. 2011 | Mar. 2011–May 2011 | Lower Gangetic plain | Humid | Agriculture |

Table 2: Statistics of improvement parameter for different forecast lengths.

| Parameter | Improvement parameter (%) | 24 hr forecast | 48 hr forecast | 72 hr forecast |
|-----------|---------------------------|----------------|----------------|----------------|
| $T_{\text{min}}$ |                           | 5.3            | 8.5            | 8.8            |
| $T_{\text{max}}$ |                          | 4.1            | 6.2            | 2.5            |
| $\text{RH}_{\text{min}}$ |                        | 6.4            | 12.2           | 12.6           |
| $\text{RH}_{\text{max}}$ |                      | 4.6            | 2.0            | 7.4            |
| Mean WS10 |                           | 13.0           | 16.0           | 21.5           |

4. Results

4.1. Daily Maximum ($T_{\text{max}}$) and Minimum ($T_{\text{min}}$) Air Temperatures. As regards the minimum temperature ($T_{\text{min}}$), the forecasts for all the three lead times are near perfect, the correlation coefficient between predicted and observed $T_{\text{min}}$ being always more than 0.93. Compared to the observations, GA shows negative BIAS for all the forecast lead times. Further, it is observed that the absolute value of BIAS (Figure 1) is more as the forecast lead time increases. The RMSD displays increasing trend, while correlation coefficient displays decreasing trend, in conformity with our expectations. As mentioned previously, because of chaotic behaviour of weather parameters, this degradation is generic in nature and has been also observed in previous studies of similar nature. Similar to $T_{\text{min}}$, the maximum temperature ($T_{\text{max}}$) forecast is also degraded as the forecast horizon is increased from 24 hr to 72 hr period, as seen from Figure 1. The correlation coefficients greater than 0.94 are observed for all the cases of forecast which obviously demonstrates the power of GA forecast. Slightly less BIAS and RMSD are observed for $T_{\text{max}}$ forecast as compared to $T_{\text{min}}$ forecast. The same story of the degradation in forecast quality with increasing length of the forecast horizon is noticed here also. The positive value of the improvement parameter for both maximum and minimum temperatures (Table 2) demonstrates the superiority of GA over persistence. Although in Table 2 we show the improvement parameter by clubbing results of all the stations, it has been checked that the parameter is positive for each station. It can be also noted that the improvements for $T_{\text{min}}$ are better than those for $T_{\text{max}}$.

4.2. Daily Maximum ($\text{RH}_{\text{max}}$) and Minimum ($\text{RH}_{\text{min}}$) Relative Humidity. The comparison of 24 hr, 48 hr, and 72 hr forecasts of daily minimum ($\text{RH}_{\text{min}}$) and maximum ($\text{RH}_{\text{max}}$) relative humidity with corresponding AMS-observed humidities in the validation set is shown in Figure 2. Similar to temperature forecast, a systematic degradation in the quality of forecast is observed, as the length of the forecast horizon increases. RMSD of 8.6, 10.2, and 10.5% are noted in 24, 48, and 72 hr $\text{RH}_{\text{min}}$ forecasts, respectively. BIAS value of $\text{RH}_{\text{min}}$ is increasing and correlation coefficient is decreasing as the forecast length is increased from 24 hr to 72 hr. Of course, for the case of maximum humidity, the correlation and the RMSD of 72 hr forecast are marginally better than the corresponding quantities for 48 hr forecast, but the BIAS is much higher. Similar to temperature forecast, more percentage improvement is observed in $\text{RH}_{\text{min}}$ forecast compared to $\text{RH}_{\text{max}}$ forecast (Table 2).

4.3. Daily Average Wind Speed at 10 m Height (WS10). The comparison of daily average wind speed (WS10) forecast by GA with the AMS observations of the same quantity is shown in Figure 3. Continuous wind speed measurements are not available from two AMS (Naraingarh Seed Farm, Patiala, Punjab, and Chakdaha, Nadia, West Bengal) stations. The RMSD of 0.60 ms$^{-1}$, 0.68 ms$^{-1}$, and 0.73 ms$^{-1}$ are observed in 24 hr, 48 hr, and 72 hr wind speed forecast, respectively. A negative BIAS is observed in all the cases with correlation
Figure 1: Scatter plots of AMS-observed and GA predicted $T_{\text{min}}$ for (a) 24 hr, (b) 48 hr, and (c) 72 hr forecast and $T_{\text{max}}$ forecast for (d) 24 hr, (e) 48 hr, and (f) 72 hr.

coefficients of 0.68, 0.54, and 0.46 for 24 hr, 48 hr, and 72 hr forecast. The wind speed forecast also follows the general trend of forecast degradation with increasing lead time. Wind speed forecast using GA shows maximum percentage improvement (more than 12%) among all the five predicted parameters (Table 2).

5. Conclusion

A nonlinear data-fitting technique based on genetic algorithm has been applied for predicting daily maximum and minimum near-surface air temperatures, daily maximum and minimum near-surface relative humidities, and daily average wind speed at 10 m height at selected Agro-Met Stations. The forecast lead time varies from one to three days. The unique beauty of the method is that it is not computer-intensive, as huge data inputs are not required for carrying out the prediction. Neither does it require any auxiliary dataset (apart from the basic data consisting of the measurements of the parameter concerned at regular time intervals) and still manages to produce explicit analytical forecast equation. Thus it is ideal for those scientists, for example, scientists involved in crop forecast modeling, who only require forecasts of certain Agro-Met parameters at isolated locations, representative of a small area. The shortcoming is of course that it is not easy to generalize this algorithm for predicting weather parameters over the entire globe, or over large three-dimensional regions, although there are promising two-dimensional generalizations [5, 6, 8–10, 15].

The method is highly suitable for location-specific forecasts as demonstrated in the present study. The AMS forecast...
locations have been chosen carefully so as to reflect the diverse climates of the Indian landmass. Validation of the forecast was carried out with independent validation datasets, unknown to the training algorithm. It was found that the GA forecasts are of reasonably good quality. They are also improvements over corresponding persistence forecasts in each of the considered cases, as seen from the positive values of the improvement parameter. The improvement is more as length of the forecast horizon increases (barring isolated cases), as persistence begins to lose its predictability. We can possibly conclude by saying that the method advocated in the present paper is an extremely powerful technique for carrying out location-specific forecasts of Agro-Met parameters. It must have been noted that the study period consists of only two years. If we divide the dataset into seasons, there will be insufficient number of data points for a reliable GA forecast. More data points are required for studying seasonal variation of forecasts using GA method. GA forecast also depends on geographical locations of AMS station. Again, more data points are required to reach any conclusion.

Presently, India has 1100 INSAT-linked Automatic Weather Stations (AWS) maintained by ISRO (Indian Space Research Organization) as well as 1000 AWS maintained by India Meteorological Department (IMD). Apart from that, many private agencies install AWS for settling crop insurance claims by farmers based on weather derivatives.
Therefore, the benefits of application of this approach can be harnessed in economic terms if one applies this algorithm for the forecast of the AWS measured weather parameters.

Appendix

We provide below the representative equations obtained from GA for 24 hr forecast of three parameters at Bharatpur:

\[
T_{\text{max}}(t) = (T_{\text{max}}(t-1) - (T_{\text{max}}(t-6) - T_{\text{max}}(t-1))/((-1.79) + (((-4.25) + ((T_{\text{max}}(t-4))/((-9.76) + (T_{\text{max}}(t-1))))))) + (-2.35))}
\]

\[
R_{\text{Hmax}}(t) = (R_{\text{Hmax}}(t-1) - (R_{\text{Hmax}}(t-2) - R_{\text{Hmax}}(t-3))/((R_{\text{Hmax}}(t-2))/(9.80)/(R_{\text{Hmax}}(t-6)/(8.40) + \langle R_{\text{Hmax}}(t-3)/(R_{\text{Hmax}}(t-2)/R_{\text{Hmax}}(t-6)))))))
\]

\[
W_{\text{mean}}(t) = (W_{\text{mean}}(t-1) - ((W_{\text{mean}}(t-6) - W_{\text{mean}}(t-1)) - (W_{\text{mean}}(t-1) - (W_{\text{mean}}(t-1))/(W_{\text{mean}}(t-4)/(2.34)))))/(W_{\text{mean}}(t-4) + (-1.83) + (-6.82)))
\]

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors would like to thank Shri. A. S. Kiran Kumar, Director, SAC, and Dr. J. S. Parihar, Deputy Director, EPSA/SAC, for constant encouragement and guidance. The AMS data provided by http://www.mosdac.gov.in/ are acknowledged with sincere thanks. The authors would also like to acknowledge all the agencies such as state agricultural universities, centres of ICAR, and state forest departments involved in hosting AMS for obtaining seamless micrometeorological measurements.

References

[1] G. G. Szpiro, “Forecasting chaotic time series with genetic algorithms,” Physical Review E, vol. 55, no. 3, pp. 2557–2568, 1997.
[2] Álvaro, A. Orfila, and J. Tintoré, “DARWIN: an evolutionary program for nonlinear modeling of chaotic time series,” Computer Physics Communications, vol. 136, no. 3, pp. 334–349, 2001.
[3] S. Basu, A. Sarkar, K. Satheesan, and C. M. Kishtawal, “Predicting wave heights in the north Indian Ocean using genetic algorithm,” Geophysical Research Letters, vol. 32, no. 17, 2005.
[4] S. Basu, K. Satheesan, A. Sarkar, and C. M. Kishtawal, “Ocean surface wind prediction in the north Indian Ocean from in situ and satellite observations using genetic algorithm,” Geophysical Research Letters, vol. 32, no. 24, Article ID L24601, 2005.
[5] A. Álvarez, C. López, M. Riera, E. Hernández-García, and J. Tintoré, “Forecasting the SST space-time variability of the Alboran Sea with genetic algorithms,” Geophysical Research Letters, vol. 27, no. 17, pp. 2709–2712, 2000.
[6] R. Sharma, S. Basu, A. Sarkar, and P. K. Pal, “Data-adaptive prediction of sea-surface temperature in the Arabian Sea,” IEEE Geoscience and Remote Sensing Letters, vol. 8, no. 1, pp. 9–13, 2011.
[7] C. M. Kishtawal, S. Basu, F. Patadia, and P. K. Thapalia, “Forecasting summer rainfall over India using genetic algorithm,” Geophysical Research Letters, vol. 30, no. 23, pp. 1–9, 2003.
[8] A. Álvarez, A. Orfila, and J. Tintoré, “Real-time forecasting at weekly timescales of the SST and SLA of the Ligurian Sea with a satellite-based ocean forecasting (SOFT) system,” Journal of Geophysical Research: Oceans, vol. 109, no. 3, 2004.
[9] S. Basu, R. Sharma, N. Agarwal, R. Kumar, and A. Sarkar, “Forecasting of scatterometer-derived wind fields in the North Indian Ocean with a data-adaptive approach,” Journal of Geophysical Research: Oceans, vol. 114, no. C9, 2009.
[10] A. D. Rao, M. Sinha, and S. Basu, “Bay of Bengal wave forecast based on genetic algorithm: a comparison of univariate and multivariate approaches,” *Applied Mathematical Modelling*, vol. 37, no. 6, pp. 4232–4244, 2013.

[11] P. G. Remya, R. Kumar, and S. Basu, “Forecasting tidal currents from tidal levels using genetic algorithm,” *Ocean Engineering*, vol. 40, pp. 62–68, 2012.

[12] R. Agrawal and S. C. Mehta, “Weather based forecasting of crop yields, pests and diseases-IASRI models,” *Journal of the Indian Society of Agricultural Statistics*, vol. 61, no. 2, pp. 255–263, 2007.

[13] H. D. I. Abarbanel, R. Brown, J. J. Sidorowich, and L. S. Tsimring, “The analysis of observed chaotic data in physical systems,” *Reviews of Modern Physics*, vol. 65, no. 4, pp. 1331–1392, 1993.

[14] B. K. Bhattacharya, C. B. S. Dutt, and J. S. Parihar, “INSAT uplinked Agromet Station—a scientific tool with a network of automated micrometeorological measurements for soil-canopy-atmosphere feedback studies,” in *ISPRS Archives XXXVIII-8/W3 Workshop Proceedings: Impact of Climate Change on Agriculture*, 2009.

[15] R. Sharma, A. Sarkar, N. Agarwal, R. Kumar, and S. Basu, “A new technique for forecasting surface wind field from scatterometer observations: a case study for the Arabian Sea,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 3, pp. 613–619, 2007.

[16] F. Takens, “Detecting strange attractors in turbulence,” in *Dynamical Systems and Turbulence*, D. Rand and L. S. Young, Eds., vol. 898 of *Lecture Notes in Mathematics*, pp. 366–381, Springer, New York, NY, USA, 1981.
