Journal of Data Analysis and Information Processing, 2016, 4, 132-146
Published Online August 2016 in SciRes. http://www.scirp.org/journal/jdaip
http://dx.doi.org/10.4236/jdaip.2016.43012

Missing Data Imputations for Upper Air Temperature at 24 Standard Pressure Levels over Pakistan Collected from Aqua Satellite

Muhammad Usman Saleem[1,2], Sajid Rashid Ahmed[2,3]

[1]Institute of Geology, University of the Punjab, Lahore, Pakistan
[2]Collage of Earth and Environmental Sciences, University of the Punjab, Lahore, Pakistan
[3]Center for Geographic Information System, Punjab University Collage of Information Technology, Lahore, Pakistan
Email: osman.geomatics@gmail.com

Received 18 May 2016; accepted 28 August 2016; published 31 August 2016

Abstract

This research was an effort to select best imputation method for missing upper air temperature data over 24 standard pressure levels. We have implemented four imputation techniques like inverse distance weighting, Bilinear, Natural and Nearest interpolation for missing data imputations. Performance indicators for these techniques were the root mean square error (RMSE), absolute mean error (AME), correlation coefficient and coefficient of determination (R^2) adopted in this research. We randomly make 30% of total samples (total samples was 324) predictable from 70% remaining data. Although four interpolation methods seem good (producing <1 RMSE, AME) for imputations of air temperature data, but bilinear method was the most accurate with least errors for missing data imputations. RMSE for bilinear method remains <0.01 on all pressure levels except 1000 hPa where this value was 0.6. The low value of AME (<0.1) came at all pressure levels through bilinear imputations. Very strong correlation (>0.99) found between actual and predicted air temperature data through this method. The high value of the coefficient of determination (0.99) through bilinear interpolation method, tells us best fit to the surface. We have also found similar results for imputation with natural interpolation method in this research, but after investigating scatter plots over each month, imputations with this method seem to little obtuse in certain months than bilinear method.

How to cite this paper: Saleem, M.U. and Ahmed, S.R. (2016) Missing Data Imputations for Upper Air Temperature at 24 Standard Pressure Levels over Pakistan Collected from Aqua Satellite. *Journal of Data Analysis and Information Processing*, **4**, 132-146. http://dx.doi.org/10.4236/jdaip.2016.43012

Keywords

Missing Data Imputations, Spatial Interpolation, AQUA Satellite, Upper Level Air Temperature, AIRX3STML

1. Introduction

Climate data gather through remote sensing satellites contains missing data *i.e.* incomplete data matrices. A major cause of this missing data may result from insufficient sampling, errors in measurement or faults in data gathering [1] [2]. Using of missing data may lead to wrong results and analysis in the research [3]-[6]. It is, therefore, important to select the best estimation of these missing data through available sample data points [1] [7]. In literature [1] [7]-[9] we found mean value imputation method to estimate this missing information, but this method may disturb the data integrity. There are a number of interpolation methods that have been proposed for imputation of missing dataset [10]-[15]. The accurate method depends on the missing data mechanism [2] and schemes for interpolation, with spatial variation in air temperature [16]. In literature, there were several extensive spatial interpolation methods that have been used for air temperature data [17]-[22]. Air temperature data at different pressure levels captured by AQUA satellite contain randomly missing data. We employ Inverse Distance Weighting, Bilinear, Nearest and Natural interpolation techniques for filling these missing information accurately. Air temperature data at 24 standard pressure levels have been filled with these interpolation methods. According to [1] [2], [7] [8], [23], [24], we chose Root Mean Square Error (RMSE), Absolute Mean Error (AME), Correlation coefficient (corrl) and Coefficient of determination (R^2) as performance indicators for these imputations. Firstly missing values have been interpolated through these techniques, then 30% of sample data (total of 324) have been predicted through 70% of known datasets over each month. We take an average of each month's performance parameters over pressure levels. Scatter plots for bilinear and natural interpolation method have been created and investigated for best imputations. There was not a single study for missing air data imputation over Pakistan. So we make a strategy to check the best interpolation method for temperature data set through available data and software.

In Sections 2 and 3, we will describe data set and meteorology of the study area. Then in Sections 4 and 5, we describe methods and results of this study. Discussions and conclusions will be in Sections 6 and 7 respectively.

2. Data Set Used

To conduct this research work, we have used Atmospheric Infrared Sounder (AIRS) STM Lite Version 6 level 3 product monthly mean air temperature (K) from September 2002 to December 2015. AIRS is one of the instrument mounted on AQUA satellite, which was launched in May 2002 by NASA [25]. Goddard Earth Science Data and Information Service Center (GESDISC) provide air temperature data set on 24 hPa atmospheric pressure levels (*i.e.* 1000, 925, 850, 700, 600, 500, 400, 300, 250, 200, 150, 100, 70, 50, 30, 20, 15, 10, 7, 5, 3, 2, 1.5, 1 hPa). We have averaged each monthly temperature data over the time period. Through intensive programming, this data set has been processed in Matlab software for each pressure level. AQUA satellite has 2378 bands in the thermal infrared region (3.7 um - 15.5 um) and 4 bands in the visible region (0.4 um - 1.0 um) of electromagnetic spectrum [25]. This spectral range of thermal and visible bands provides accuracy of 1˚C in air temperature for 1 km of air mass thickness [25]. The version 6 of AIRS provide very high-resolution data set of 1˚ × 1˚ grid cell and has an advantage of providing good temperature free from atmospheric biases than version 5 [26]. AQUA provides temperature data set ranging from −180 to +180 longitudes and −90 to +90 latitude [25]. Version 6 has some gaps in the data set which we have filled with IDW, bilinear, nearest, natural interpolations techniques.

3. Study Area

Pakistan (60˚E - 78˚E, 20˚N - 38˚N) geographically lies northeast to china, West to Afghanistan, Iran to the south-west, India to east and South to the Arabian-Sea. Its climate varies from arctic like condition on snow covered mountains in northern areas to arid like conditions in hot deserts in the south-west regions. It consists of four provinces, Punjab, Sindh, Khyber Pakhtunkhwa (KPK), and Baluchistan. Punjab being the largest province mostly consists on the fertile region. This province shared its border with Indian Punjab. Both Punjab has same

fertile lands due to river flooding. Sindh which bordered with Indian provinces Rajasthan and Gujarat contain the Thar Desert. Baluchistan border with Afghanistan is a drier region. The overall climate of Pakistan is drier [27]. Its coastal areas facing 0.6 C to 1.0 C rise in temperature since the 1900 s [28]. Average rainfall varies from area to area but overall has about 10 inches of annually. Desert regions of Baluchistan get less than 5 inches of annual rainfall while Punjab got 20 inches of annual rainfall [27]. Northern mountainous areas face average coldest temperature of 23.88˚C in summer while on Baluchistan, it is about 26.66˚C. On the Indus plain, a temperature range of high of 32.22˚C to 48.88˚C [27] in summer, to a low 12.77˚C in winters. Hottest months in a year are May and June where temperature can reach to 50˚C and coldest month is January with temperature of 5˚C. Pakistan face monsoon season in which seasonal reversal of lower tropospheric winds appears [29]. It has four well define seasons: Winter, from November to February; Pre-monsoon (Hot), from March to mid of June; Monsoon, from mid of June to mid of September; Post-monsoon, from mid of September to October. Summer season is extremely hot and the relative humidity ranges from 25% to 50% [30] (see **Figure 1**).

4. Research Method and Methodology

We have separated air temperature for each 24 atmospheric pressure levels using Matlab software. Through spatial interpolation techniques of Inverse distance weighting, bilinear, nearest, and natural we filled missing data in air temperature at each pressure level. In order to select best interpolation technique in our data set, we randomly make 30% of the sample data to predict from 70% remaining data set. Interpolated value of this 30% sample data set has been compared with its original ones and we calculated the root mean square error (RMSE) for each interpolation method. Besides other performance indicators like mean absolute error (AME), correlation coefficient (Corrl), the coefficient of determinations (R^2) and scatter plots have used in this research, to select best imputation method for upper air temperature data. Performance parameters were investigated after taking averaging over each month (Monthly results can be demanded by emailing the author).

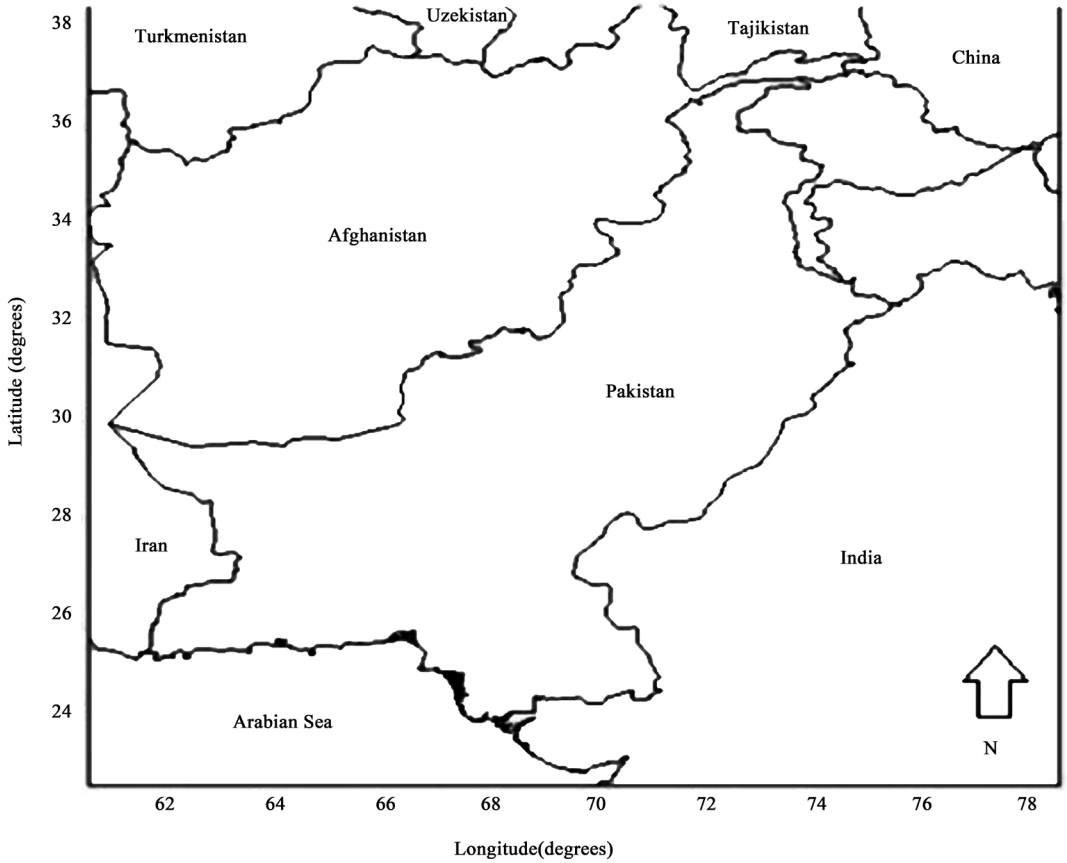


Figure 1. Geographical location of Pakistan as study area.

4.1. Inverse Distance Weighting

It is a weighted average deterministic approach, which assumes that things related more which are closer together. It applies weight to each observation in inverse proportion to its distance from the point where prediction is required [9]. More will be the distance of neighbors from a predicted point, less will be its weight in prediction. Here is mathematical formula is given by [23].

$$z\left(x_j\right)=\frac{\sum_{i=1}^{n} z\left(x_i\right) d_{ij}^{-r}}{\sum_{i=1}^{n} d_{ij}^{-r}} \tag{1}$$

where $z\left(x_j\right)$ is the interpolated value of missing data through IDW, d_{ij}^{-r} is the weighting factor which was calculated by the Euclidean distance between the neighbor $z\left(x_i\right)$ and point to be estimated $z\left(x_i\right)$. n is the total no of the sample used (total of 324) and r is the power of distance weighting. Setting search radius of 250 neighbor points and distance power of −2 in IDW algorithm created by [31] we filled the missing data with IDW technique in Matlab software. Extrapolation also has been done with IDW algorithm in data set imputation.

4.2. Nearest Neighbor Interpolation

In the nearest neighbor interpolation method (NNM) missing data imputed directly from the nearest neighbor of the data set [16] [23]. Extrapolation of temperature data was performed with NNM where it requires to extrapolate.

4.3. Bilinear Interpolation

Bilinear interpolation method (BIM) fits a straight line between starting and ending points of the missing data and then enable to interpolate missing values straightforwardly employ this linear equation. Reference [1] provides this formula for linear interpolation.

$$y=y_1+k\left(x+x_1\right) \tag{2}$$

$$k=\frac{y_2-y_1}{x_2-x_1} \tag{3}$$

$$x_1<x<x_2 \text{ and } y_1<y<y_2$$

Equation (2) is a linear equation of the straight line with x_1, y_1 are the coordinates of starting missing data and x_2, y_2 ending coordinates of the missing values, k being the slope or gradient of line calculated with (3). We have adopted the same scheme of 30% of the sample points as missing to interpolate with bilinear interpolation method. Extrapolation of the missing data has been done with linear interpolation method where it was required.

4.4. Natural Interpolation

The natural interpolation method is based natural neighbor around the missing data. In non-uniformly missing data, selection of natural neighbor is based on Delaunay triangulations method [32]. This interpolation smoothen the surface of the data set.

4.5. Performance Parameters

Several performance indicators were calculated for the best method of imputation. After reading literature, [1], [2], [7], [9], [16], [23], [24], [33], we select four performance parameters, root mean square error (RMSE), absolute mean error (AME), correction coefficient (Corrl) and coefficient of determination (R^2), as indicators, to select the best method for estimating missing air temperature data.

4.6. Root Mean Square Error (RMSE)

Root mean square error was commutated with Formula (4) given by [7].

$$\text{RMSE} = \left(\frac{1}{n} \sum_{i=1}^{n} \left[O_i - P_i \right]^2 \right)^{\frac{1}{2}} \tag{4}$$

where n is the total no of imputation points [1], P_i is the imputed data point and O_i is the observed data point. RMSE tells us about a total difference between actual and predicted air temperature data. Lower will be the RMSE, accurate will be imputations [33].

4.7. Absolute Mean Error (AME)

More residual error in comparison with RMSE can be investigated from absolute mean error (AME). It can be calculated from Equation (5) given by [1] and [7].

$$\text{AME} = \frac{1}{n} \sum_{i=1}^{n} \left| P_i - O_i \right| \tag{5}$$

where n is the total no of imputation points [1], P_i is the imputed data point and O_i is the observed data point. Its value ranging from 0 to ∞. 0 value indicates a perfect estimation of missing values.

4.8. Correlation Coefficient (Corrl)

This coefficient, correlate actual and predicted value of data set. Correlation of value 1, indicates a good estimation of predicted points with actual ones. Its value near to 0 indicates no or low correlation between the actual and predicted data set, concluding to bad imputation for data set.

4.9. Coefficient of Determination (R^2)

A coefficient of determination provides us, the degree of correlation between actual and predicted data points [2]. It takes a value ranging between 0 and 1, with values closer to 1 indicate the best fit to the surface [2] and [7]. Reference [7] mentioned this formula for the coefficient of determination.

$$R^2 = \left[\frac{1}{n} \frac{\sum_{i=1}^{n} \left(P_i - A.I \right) \left(O_i - A.O \right)}{\partial_p \partial_o} \right] \tag{6}$$

In Equation (6), $A.I$ is the average of imputed points and $A.O$ is the average of observation points with $\partial_p \partial_o$ are the standard deviation of imputed points and observational data points respectively.

5. Results

These are the results of performance parameters derived from IDW, Bilinear, Natural and Nearest imputation methods. Performance parameters for each of interpolation method on upper air temperature have explained below over each pressure level.

5.1. Root Mean Square Error (RMSE)

Root mean square error through inverse distance weighting interpolation remains less than 0.2 on all levels except over 70, 100, 150, 250, 300, 400, 500, 600, 700, 850, 925, 1000 hPa where its value vary between 0.2 to 0.7 (see **Table 1**). RMSE in bilinear interpolation remains <0.03 over 1, 15, 20, 30, 50, 600, 700, 850, 925 hPa while this error little increase to 0.2 and 0.62 over 850, 925, 1000 hPa pressure levels (see **Table 2**). Same scenario for natural interpolation method, in which RMSE value remains 0.03 at all levels except than 850, 925, 1000 hPa levels (see **Table 3**). RMSE produce by nearest interpolation method looks somewhat like IDW technique (see **Table 4**). This RMSE going to increase from 0.1184 on 1 hPa level to 1.066 on 1000 hPa pressure level. This error was more than bilinear and natural interpolation methods on air temperature data (see **Tables 1-4**).

Table 1. Performance parameters for IDW method.

Pressure Level	RMSE	AME	Correlation	R^2
1 hPa	0.166469	0.072568	0.988667	0.971434
1.5 hPa	0.211402	0.092202	0.981786	0.958047
2 hPa	0.196899	0.086915	0.982793	0.959998
3 hPa	0.159178	0.069608	0.98045	0.955495
5 hPa	0.132347	0.055292	0.979928	0.954478
7 hPa	0.101749	0.043999	0.982191	0.958764
10 hPa	0.085341	0.0368	0.977734	0.950187
15 hPa	0.080505	0.034087	0.977044	0.948906
20 hPa	0.094717	0.038195	0.981358	0.957149
30 hPa	0.090669	0.04062	0.981759	0.957997
50 hPa	0.211602	0.090937	0.988755	0.971605
70 hPa	0.513806	0.232903	0.989783	0.973625
100 hPa	0.631246	0.287027	0.990918	0.975851
150 hPa	0.337199	0.153081	0.990553	0.975131
200 hPa	0.201157	0.085126	0.986582	0.967341
250 hPa	0.40177	0.178691	0.987243	0.968677
300 hPa	0.490221	0.222593	0.988619	0.971416
400 hPa	0.508421	0.22905	0.989708	0.973484
500 hPa	0.482349	0.206772	0.986345	0.966914
600 hPa	0.483208	0.198397	0.975668	0.946409
700 hPa	0.521899	0.210841	0.976866	0.948561
850 hPa	0.680485	0.268773	0.977312	0.949295
925 hPa	0.760635	0.291887	0.971972	0.93905
1000 hPa	0.550801	0.185222	0.973075	0.941161

5.2. Absolute Mean Error (AME)

Absolute mean error with IDW was 0.07256 over 1 hPa and going to increase 0.18522 till 1000 hPa (see **Table 1**). Although this AME was very less, but other imputation methods like bilinear, produce AME of 0.011 at 1hPa level and this error remains between 0.01 - 0.02 till 925 hPa level (see **Table 1**). Maximum error with bilinear interpolation method was 0.169 found over 1000 hPa level (see **Table 2**). Natural interpolation method also produces a very low absolute mean error, which remains from 0.010 to 0.08 till 850 hPa level. At 925, 1000 hPa levels, this error increase to 0.10 and 0.168 respectively (see **Table 3**). Imputations over 1 hPa to 200 hPa levels were good with very low AME of 0.053 to 0.056. But after 250 to 1000 hPa level this AME value going to increase to 0.263 showing the poor imputations over these pressure levels with the nearest interpolation (see **Table 4**).

5.3. Correlation Coefficient (Corrl)

Now correlation coefficient tells us what is the correlation between actual air temperature data and imputed data points. Its value close to 1 indicates a strong correlation between actual and predicted value with accurate imputation. Through IDW, this coefficient values was ranging 0.988 to 0.970 overall pressure levels (see **Table 1**). The correlation coefficient for bilinear interpolation methods remains within the range of 0.999 to 0.992, indicating a good relation between actual and predicted air temperature values (see **Table 2**). Imputations through natural interpolation methods produce same correlation results as that of bilinear interpolation (see **Table 3**). Imputations through nearest interpolation method produce worst correlation among all other techniques (see

Table 2. Performance parameters for bilinear interpolation method.

Pressure Level	RMSE	AME	Correlation	R^2
1 hPa	0.02679	0.011047	0.999488	0.992801
1.5 hPa	0.054627	0.022325	0.99788	0.989612
2 hPa	0.060734	0.023451	0.99656	0.987006
3 hPa	0.059314	0.024064	0.994355	0.982667
5 hPa	0.041878	0.017174	0.995857	0.985622
7 hPa	0.044549	0.017438	0.99466	0.983246
10 hPa	0.040881	0.016178	0.992926	0.979846
15 hPa	0.034114	0.013397	0.99408	0.982118
20 hPa	0.03006	0.01189	0.997128	0.988123
30 hPa	0.028061	0.010953	0.996916	0.987721
50 hPa	0.03166	0.0126	0.999639	0.9931
70 hPa	0.037411	0.015207	0.999893	0.993605
100 hPa	0.044706	0.017887	0.999928	0.993674
150 hPa	0.039325	0.014666	0.999839	0.993497
200 hPa	0.034996	0.01402	0.999436	0.992698
250 hPa	0.042946	0.016286	0.999582	0.992987
300 hPa	0.046799	0.018095	0.999772	0.993365
400 hPa	0.052031	0.020252	0.999841	0.993502
500 hPa	0.094673	0.029065	0.998563	0.990971
600 hPa	0.149939	0.044605	0.993462	0.980967
700 hPa	0.159515	0.051779	0.996594	0.987069
850 hPa	0.296113	0.104572	0.995724	0.985341
925 hPa	0.288459	0.105102	0.996338	0.986555
1000 hPa	0.625643	0.169008	0.992457	0.97894

Table 4). This coefficient values vary from 0.993 to 0.971 from 1 hPa to 1000 hPa levels indicating bad imputation of missing data through this method (see Table 4).

5.4. Coefficient of Determination (R^2)

The value of 0.971 to 0.93 through IDW interpolation indicating that it was not a good imputation method (see Table 1). In bilinear, its value 0.99 indicating a perfect fit to air temperature overall pressure levels except than 1000 hPa where it has a value of 0.978 (see Table 2). Even this value over 1000 hPa was more accurate then IDW method (see Table 2). The coefficient of determination values closer to 0.99 on all other pressure levels revealing accurate results with natural method (see Table 2). At 1000 hPa level, this coefficient has a value of 0.979. These results of natural interpolation look same as that from the bilinear method. Results through nearest interpolation method were same as that produce from IDW (see Table 4).

6. Discussions

Results indicate that bilinear and natural interpolation methods were best for upper air temperature imputation. Although IDW and nearest, results were also good, but we have to decide best method between bilinear and natural interpolation methods. We adopt the strategy of scatter plots for each of these interpolation methods over each pressure levels on each month of the year. Scatter plot created with plotting actual air temperature with

Table 3. Performance parameters for natural interpolation method.

Pressure Level	RMSE	AME	Correlation	R^2
1 hPa	0.026192	0.01059	0.999505	0.992833
1.5 hPa	0.055985	0.022704	0.997788	0.989431
2 hPa	0.055027	0.022489	0.997227	0.988324
3 hPa	0.05697	0.023346	0.995403	0.98472
5 hPa	0.045424	0.01807	0.995226	0.984384
7 hPa	0.04693	0.018854	0.993896	0.98174
10 hPa	0.041709	0.017326	0.992901	0.979806
15 hPa	0.032621	0.013174	0.995344	0.984597
20 hPa	0.030227	0.012181	0.997049	0.987965
30 hPa	0.02517	0.009989	0.997308	0.988489
50 hPa	0.030252	0.012083	0.999664	0.99315
70 hPa	0.042121	0.017267	0.999876	0.993571
100 hPa	0.04164	0.01733	0.999938	0.993695
150 hPa	0.03327	0.013485	0.999886	0.993592
200 hPa	0.036728	0.014364	0.999415	0.992655
250 hPa	0.040868	0.015961	0.999691	0.993204
300 hPa	0.050005	0.018554	0.999829	0.993478
400 hPa	0.051152	0.019008	0.999831	0.993482
500 hPa	0.095453	0.030295	0.998929	0.991692
600 hPa	0.155535	0.043585	0.994595	0.983143
700 hPa	0.155907	0.051393	0.996526	0.986939
850 hPa	0.247188	0.087662	0.996911	0.98769
925 hPa	0.297521	0.111183	0.996102	0.986087
1000 hPa	0.582456	0.168721	0.992979	0.979983

imputed air temperature through these methods. Trend line which is the best-fit line in the cloud of scatter plot tells us how close the imputed value to its actual one. If clouds of points in scatter plot align perfectly around the trend line, we found it good imputation method.

Figures 2-4 were the scatter plots for each month of air temperature through bilinear interpolation method. After investigating each plot, we have found that over 1 hPa level, scatter plots were poorly fit in February, August, and May. August was the month in 1.5 hPa, where the scatter plots are spread around the trend line. At 3 hPa, in each month, trend line lies close to data points except in February, May, and July. Same spread of scatter plot has investigated in January, March over 7 hPa. Over 11, 20, 30, 200 hPa levels spread of scatter plots clouds was present in February and May. January and February scatter plots were scattered around the trend line over 400, 500, 600, 700, 850, 925 hPa pressure levels. Imputation through bilinear interpolation method of missing air temperature data found align with actual data values over 5, 50, 70, 80, 150, 250, 300, 1000 hPa pressure levels (see **Figures 2-4**).

Now **Figures 5-7** were the scatter plot investigation for natural interpolation method. Imputation through nat-

Table 4. Performance parameters for nearest interpolation method.

Pressure Level	RMSE	AME	Correlation	R^2
1 hPa	0.11849	0.053522	0.993688	0.981323
1.5 hPa	0.147994	0.063412	0.988543	0.971282
2 hPa	0.137701	0.059359	0.989902	0.973913
3 hPa	0.12883	0.055576	0.982947	0.96041
5 hPa	0.098991	0.04282	0.986068	0.966424
7 hPa	0.089466	0.037342	0.983107	0.960601
10 hPa	0.079866	0.032443	0.979374	0.953388
15 hPa	0.07194	0.028641	0.981414	0.957337
20 hPa	0.073197	0.029632	0.986802	0.96781
30 hPa	0.067964	0.029077	0.988025	0.970222
50 hPa	0.139988	0.061429	0.994903	0.983715
70 hPa	0.330818	0.144881	0.995352	0.984605
100 hPa	0.403023	0.180799	0.996283	0.986444
150 hPa	0.207653	0.094515	0.996273	0.986423
200 hPa	0.135824	0.057151	0.9931	0.980159
250 hPa	0.28262	0.125826	0.992193	0.978405
300 hPa	0.336921	0.148686	0.994391	0.982714
400 hPa	0.333406	0.148983	0.995103	0.984116
500 hPa	0.334415	0.143794	0.991464	0.977003
600 hPa	0.407548	0.150779	0.969235	0.935319
700 hPa	0.425033	0.159075	0.981251	0.957102
850 hPa	0.706973	0.23373	0.977612	0.949879
925 hPa	0.803679	0.268447	0.97417	0.943193
1000 hPa	1.066438	0.263442	0.97105	0.937176

ural method seems to be more disturb in certain months over each pressure level. February and May were the months at 1 hPa level, where the scatter plot has spread clouds around the trend line. At 1.5 hPa and 2 hPa levels, February, and May were the months for not providing good scatter plots (see **Figure 5**). January, February, May and October were the months of not good scatter plot over 3 hPa. 5 hPa seem to be less scatter of data points in April and October. Scatter plot of January, February, October, and February over 7 hPa level was obtuse (see **Figure 5**). 10 and 15 hPa levels found disturb in January, February, April and May. The large spread of clouds of the points observed in November and June at 20 hPa level (see **Figure 5**). Scatter plots in months of May, Jun and August found to disperse around trend line over 30 hPa level. 200, 250, 300, 400, 500, 600, 700, 850, 925 hPa were consistently spread around trend line in January, February, March and April (see **Figure 6** & **Figure 7**). Remaining pressure levels found to be best fitted with the trend line, indicating the results for good imputation of missing air temperature through natural interpolation method (see **Figures 5-7**).

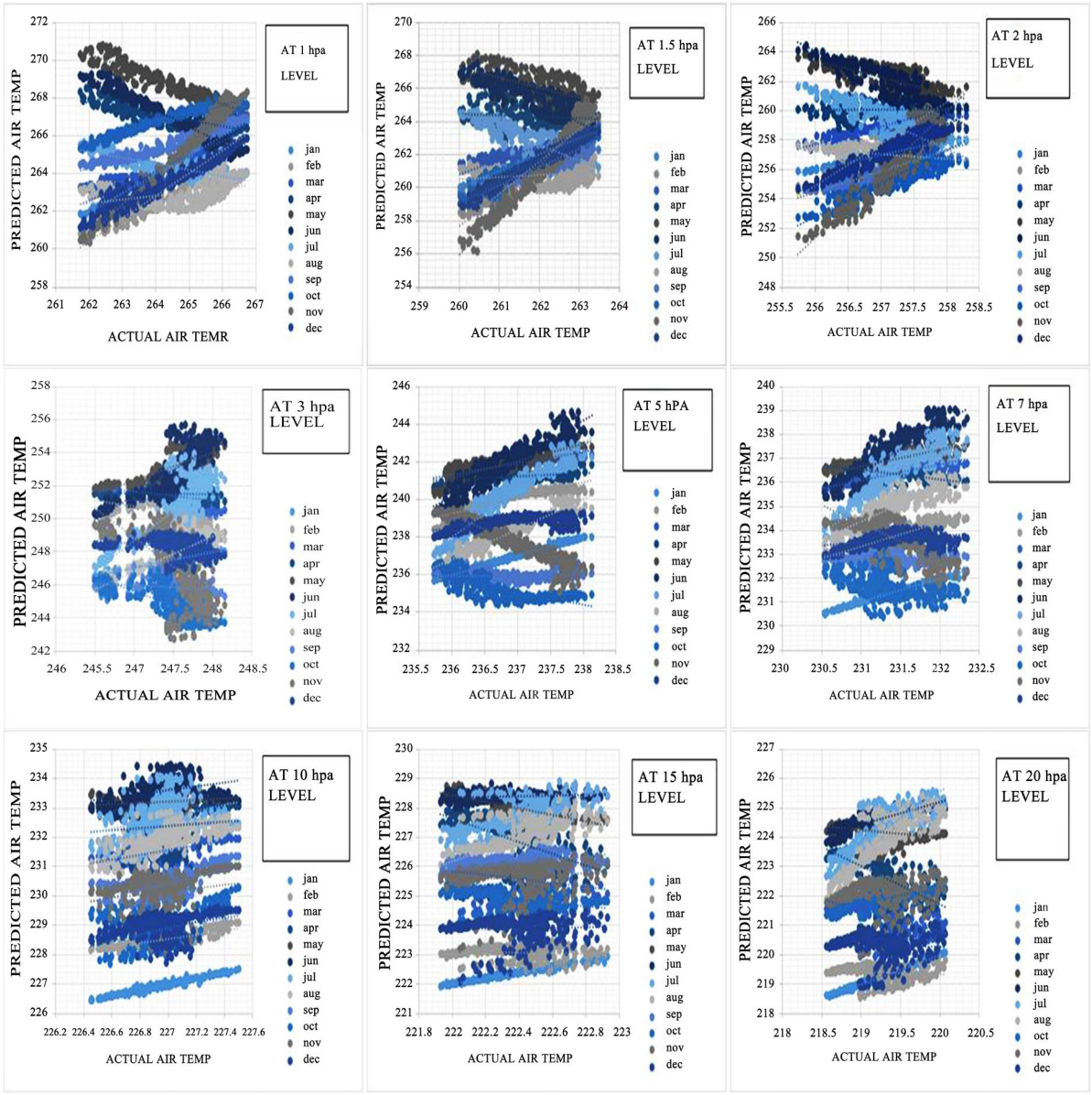
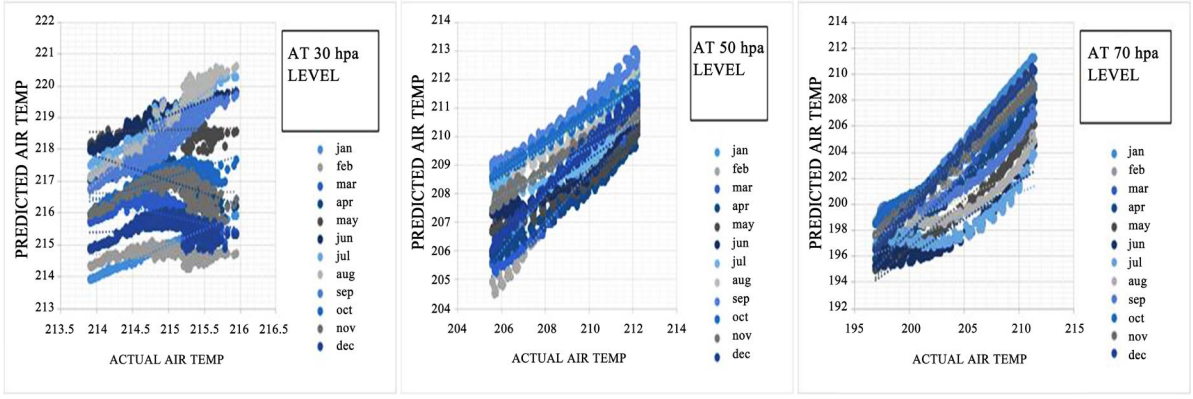


Figure 2. These scatter plots were created with bilinear interpolation method. Actual air temperature (K) with imputed air temperature (K). Top left plot is for 1 hPa level and bottom right is for 20 hPa level.



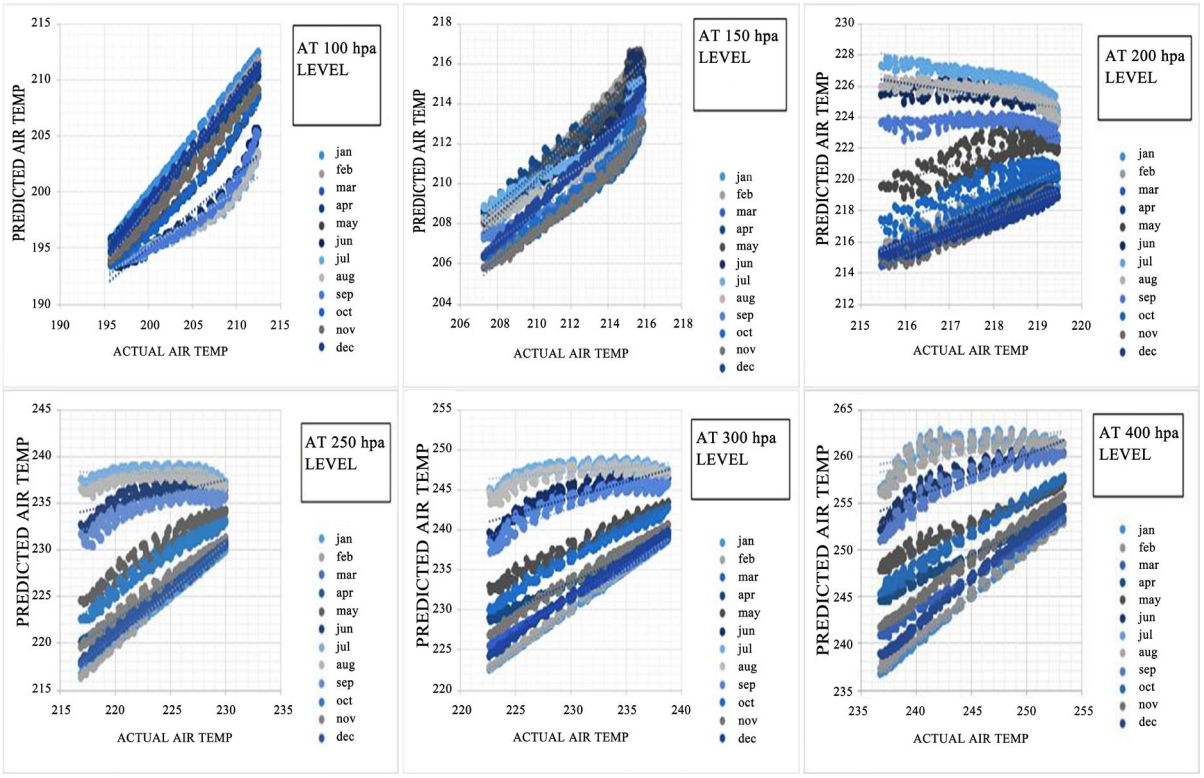


Figure 3. These scatter plots were created with bilinear interpolation method. Actual air temperature (K) with imputed air temperature (K). Top left plot is for 30 hPa level and bottom right is for 400 hPa level.

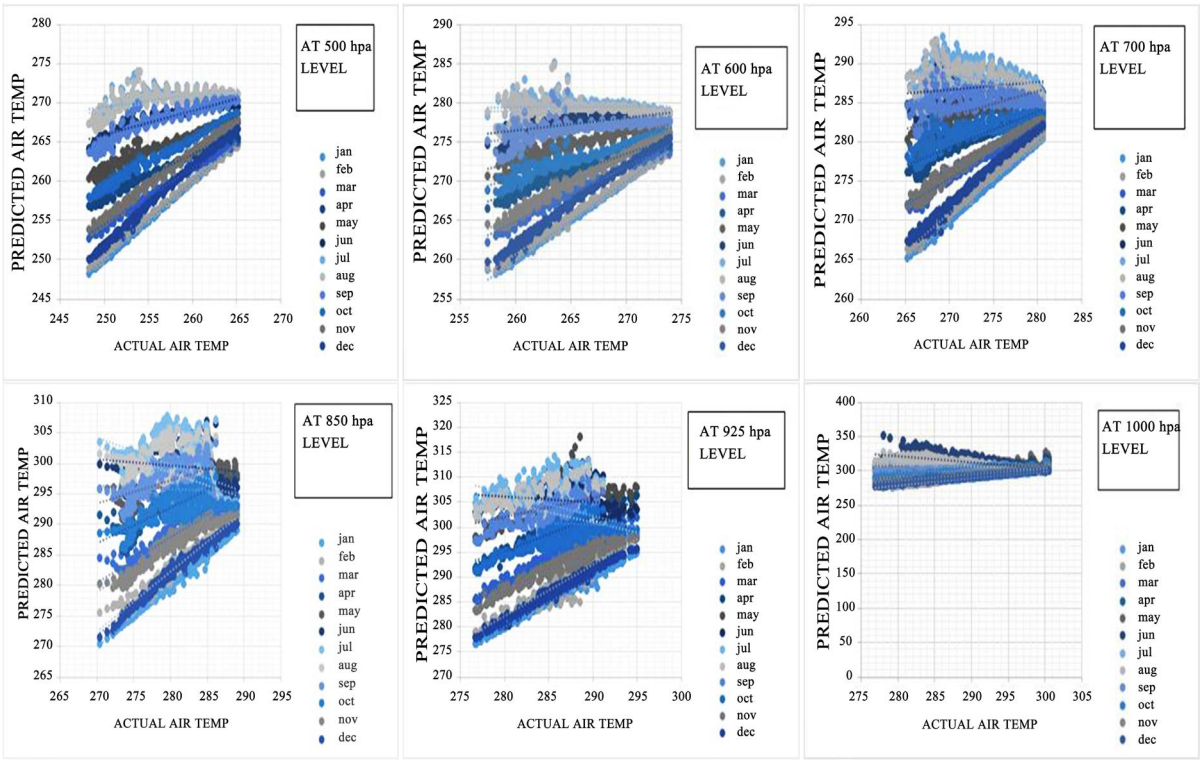


Figure 4. These scatter plots were created with bilinear interpolation method. Actual air temperature (K) with imputed air temperature (K). Top left plot is for 500 hPa level and bottom right is for 1000 hPa level.

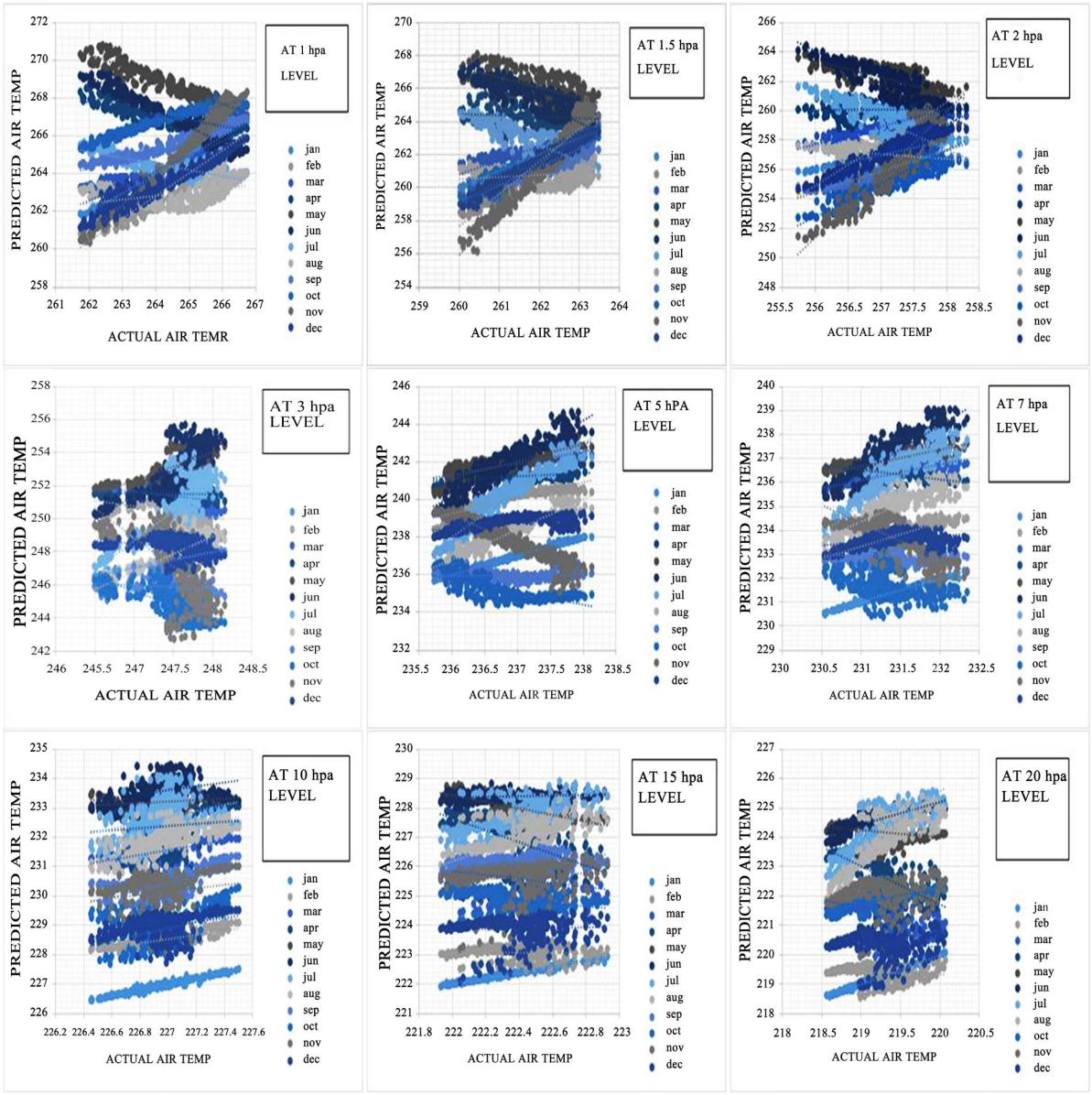
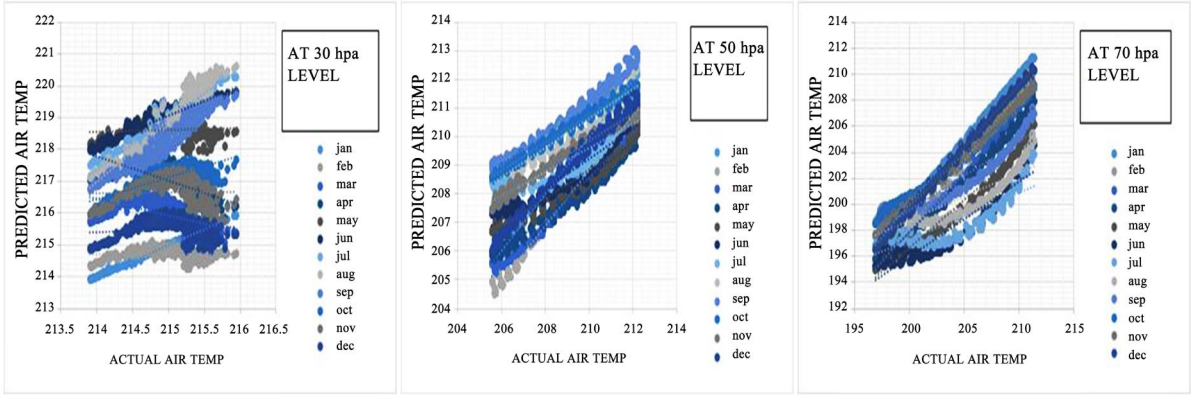


Figure 5. These scatter plots were created with natural interpolation method. Actual air temperature (K) with imputed air temperature (K). Top left plot is for 1 hPa level and bottom right is for 20 hPa level.



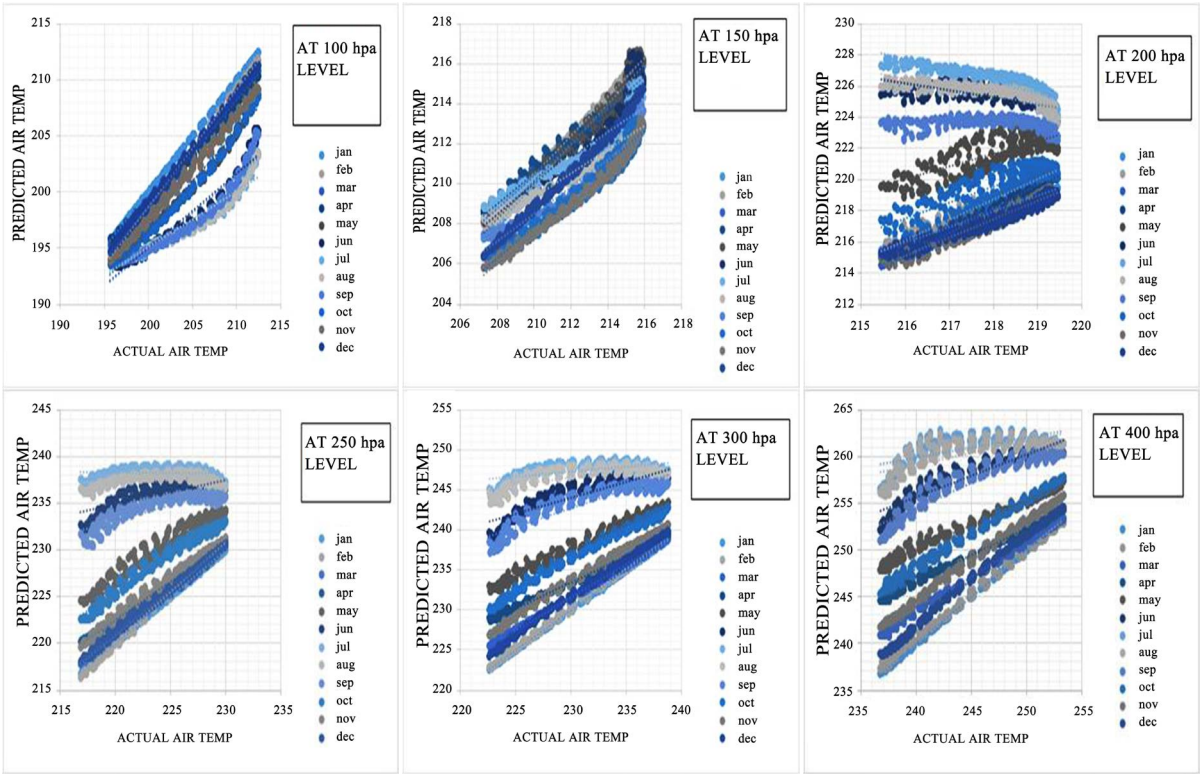


Figure 6. These scatter plots were created with natural interpolation method. Actual air temperature (K) with imputed air temperature (K). Top left plot is for 30 hPa level and bottom right is for 400 hPa level.

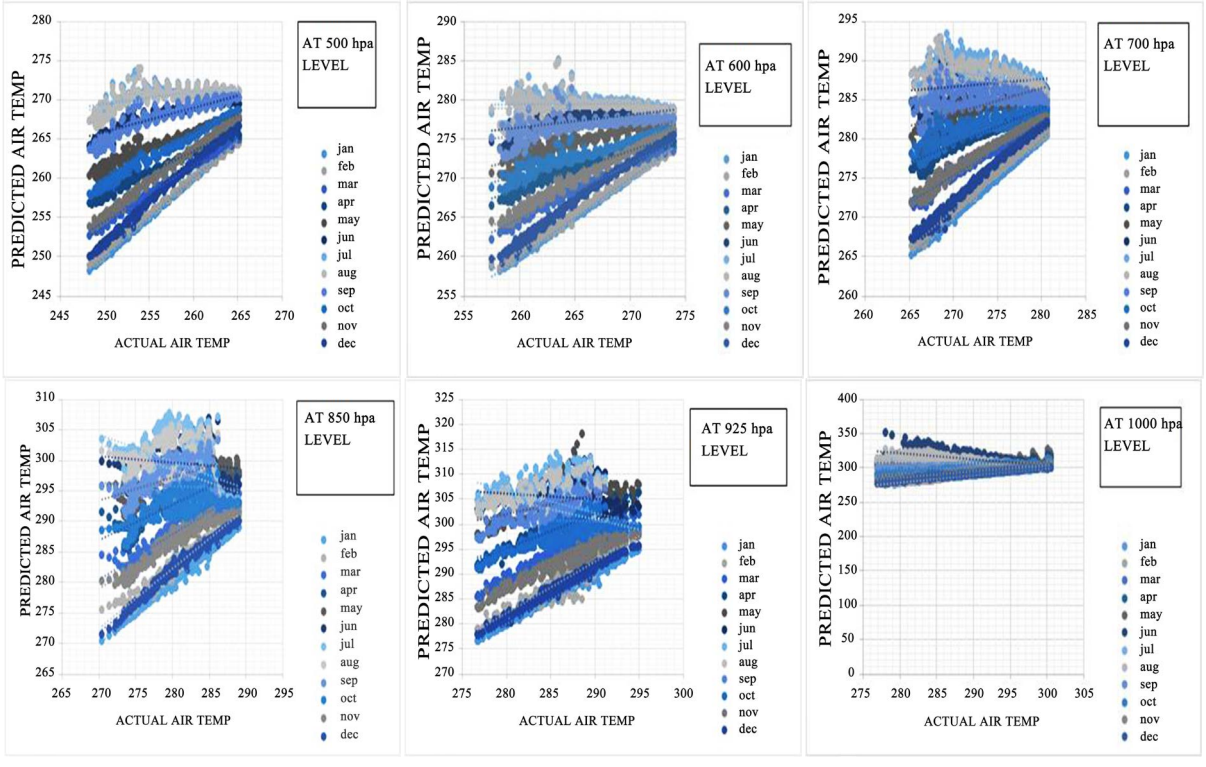


Figure 7. These scatter plots were created with natural interpolation method. Actual air temperature (K) with imputed air temperature (K). Top left plot is for 500 hPa level and bottom right is for 1000 hPa level.

7. Conclusion

Performance indicators make a room to reject inverse distance weighting and nearest interpolation methods for imputation (see **Table 1** and **Table 4**). Bilinear and natural interpolation methods were good in imputations over each pressure level. But after investigating scatter plots for each month, we conclude that bilinear interpolation method was the best for air temperature data imputations gathered from AIRS instrument of Aqua Satellite. This method produces an RMSE of only 0.99 in imputation over each pressure level. AME of value 0.01 in the majority of pressure level indicates the reliability of the RMSE, which is very accurate. The coefficient of determination remains >0.99 in all 24 pressure levels, concluding that only bilinear interpolation technique was the best (optimum) for imputation of air temperature data set. Due to very low errors in imputation with natural interpolation, it will be the second best imputation for air temperature data set.

Acknowledgements

I would like to acknowledge AIRS team for providing me air temperature data set. I would like to express my great regards to AIRS team members Edward T Olsen and Thomas Hearty, who guide me in every step to understand AIRS product. My deepest gratitude goes to Alessio Martino, University of the Rome, La Sapienza Italy for assisting me in this research. Also special thanks to Journal of Data Analysis and Information Processing for publication of this effort.

References

[1] Junninen, H., Naska, H., Tuppurainen, K., Ruuskanen, J. and Kolehmainen, M. (2004) Methods for Imputation of Missing Values in Air Quality Data Set. *Atmospheric Environment*, **38**, 2895-2897, 2899-2900. http://dx.doi.org/10.1016/j.atmosenv.2004.02.026

[2] Rahman, M.G. and Islam, M.Z. (2011) A Decision Tree Based Missing Value Imputation Technique for Data Pre-Processing. *Proceedings of the Ninth Australasian Data Mining Conference*, Ballarat, 41-50.

[3] Han, J. and Kamber, M. (2006) Data Mining Concepts and Techniques. Morgan Kaufmann, San Meteo, California.

[4] Muraidjar, K., Parsa, R. and Sarathy, R. (1999) A General Additive Data Perturbation Method for Database Security. *Management Science*, **45**, 1399-1415. http://dx.doi.org/10.1287/mnsc.45.10.1399

[5] Muller, H., Naumann, F. and Freytag, J. (2003) In: *Data Quality in Genome Database*, Hummboldt Universitt Berlin, Institut fr informatik, Berlin.

[6] Abbas, Q.S. and Aggarwal, A. (2010) Development of a Structure Framework to Achieve Quality Data. *International Journal of Advance Engineering and Application*, 193-196.

[7] Mohammad Noor, N., Mustafa Al Bakri Abdullah, M., Shukri Yahaya, A. and Azam Ramli, N. (2007) Comparison of Linear Interpolation Method and Mean Method to Replace the Missing Values in Environmental Data Set. ICoSM, Penang, 85-87.

[8] Yozgatigil, C., Aslan, S., Lyigun, C. and Batmaz, I. (2013) Comparison of Missing Value Imputation Method in Time Series: The Case of Turkish Meteorological Data. *Theoretical and Applied Climatology*, **112**, 144-149.

[9] Robeson, M. (1994) Influence of Spatial Sampling and Interpolation on Estimates of Air Temperature Change. *Climate Research*, **4**, 120-124. http://dx.doi.org/10.3354/cr004119

[10] Tseng, S.M., Wang, K.H. and Lee, C.I. (2003) A Preprocessing Method to Deal with Missing Values by Integrating Clustering and Regression Techniques. *Applied Artificial Intelligence*, **17**, 535-544. http://dx.doi.org/10.1080/713827170

[11] Zhan, C., Qin, Y., Zhu, X., Zhang, J. and Zhang, S. (2006) Clustering Based Missing Value Imputation for Data Preprocessing. 2006 4*th IEEE International Conference on Industrial Informatics*, 16-18 August 2006, 1081-1086. http://dx.doi.org/10.1109/INDIN.2006.275767

[12] Junnunen, H., Niska, H., Tuppurainen, K., Ruuskanen, J. and Kolehminen, M. (2004) Methods for Imputation for Missing Values in Air Quality Data Sets. *Atmospheric Environment*, **38**, 2895-2807. http://dx.doi.org/10.1016/j.atmosenv.2004.02.026

[13] Schneider, T. (2001) Analysis of Incomplete Climate Data: Estimation of Mean Values and Covariance Matrices and Imputation of Missing Values. *Journal of Climate*, **14**, 853-871. http://dx.doi.org/10.1175/1520-0442(2001)014<0853:AOICDE>2.0.CO;2

[14] Little, R. and Rubin, D. (1987) Statistical Analysis with Missing Data. John Wiley and Sons Publishers, New York.

[15] Pyle, D. (1999) Data Preparation for Data Mining. Morgan Kaufmann Publishers, California.

[16] Stahl, K., Moore, R., Floyer, J., Asplin, M. and McKendry, I. (2006) Comparison of Approaches for Spatial Interpolation of Daily Air Temperature in a Large Region with Complex Topography and Highly Variable Station Density. *Agricultural and Forest Meteorology*, **139**, 224-236. http://dx.doi.org/10.1016/j.agrformet.2006.07.004

[17] Shumaker, L.L. (1976) Fitting Surfaces to Scattered Data. In: Lorentz, G.G., *et al*., Eds., *Approxmiation*, 2nd Edition, Academic Press, New York, 203-268.

[18] Gustavsson, N. (1981) A Review of Methods for Objective Analysis. In: Bengtsson, L., Ghil, M. and Kallen, E., Eds., *Dynamic Meterology*: *Data Assimilation Methods*, Springer, New York, 17-76. http://dx.doi.org/10.1007/978-1-4612-5970-1_2

[19] Franke, R. (1982) Scattered Data Interpolation: Tests of Some Methods. *Mathematics of Computation*, **38**, 181-200.

[20] Ngan Lam, S.N. (1983) Spatial Interpolation Methods: A Review. *The American Cartographer*, **10**, 129-149. http://dx.doi.org/10.1559/152304083783914958

[21] Burrough, P.A. (1986) Principles of Geographical Information Systems for Land Resources Assessment. Oxford University Press, Oxford.

[22] Thiebaux, H.J. and Pedder, M.A. (1987) Spatial Objective Analysis. Academic Press, New York.

[23] Ozaki, G. (2013) Missing Data Imputation of Climate Datasets: Implications to Modeling Extreme Drought Events. *Revista Brasileira de Meteorologia*, **29**, 23.

[24] Perry, M. and Hollis, D. (2005) The Generation of Monthly Gridded Datasets for a Range of Climatic Variables over the UK. *International Journal of Climatology*, **25**, 1045-1046. http://dx.doi.org/10.1002/joc.1161

[25] AIRS Laboratory Home (2016) http://airs.jpl.nasa.gov/mission_and_instrument/instrument

[26] Tian, B., Manning, E., Fetzer, E., Olsen, E. and Wong, S. (2014) AIRS Version 6 L3 User Guide: AIRS/AMSU/HSB Version 6 Level 3 Product User Guide. Pasadena, CA.

[27] Saleem, M. (2016) Statistical Investigation and Mapping of Monthly Modified Refractivity Gradient over Pakistan at 700 hecto Pascal Level. *Open Journal of Antennas and Propagation*, 3.

[28] Iqbal, M. and Quamar, J. (2011) Measuring Temperature Variability of Five Major Cities of Pakistan. *Arabian Journal of Geosciences*, **4**, 595. http://dx.doi.org/10.1007/s12517-010-0224-0

[29] Haider, K., Rasul, G. and Afzaal, M. (2008) A Study on Tropical Cyclones of the Arabian Sea in June 2007 and Their Connection with Sea Surface Temperature. *Pakistan Journal of Meteorology*, **4**, 38.

[30] Pakistan Meteorology Department (2009) Climate Change Indicators of Pakistan. Pakistan Meteorology Department.

[31] Langella, G. (2010) File Exchange-MATLAB Central. Mathworks.Com. http://www.mathworks.com/matlabcentral/fileexchange/27562-inverse-distance-weighted--idw--or-simple-moving-average--sma--interpolation/content/gIDW.m

[32] Boissonnat, J.-D. and Cazals, F. (2000) Smooth Surface Reconstruction via Natural Neighbour Interpolation of Distance Functions. *Proceedings of the* 16*th Annual Symposium on Computational Geometry*, Hong Kong, June 2000, 224. http://dx.doi.org/10.1145/336154.336208

[33] Price, D., McKenney, D., Nalder, I., Hutchinson, M. and Kesteven, J. (2000) A Comparision of Two Statistical Methods for Spatial Interpoaltion of Canadian Monthly Mean Climate Data. *Agricultural and Forest Meterology*, **101**, 84. http://dx.doi.org/10.1016/S0168-1923(99)00169-0