2023 International Conference on Machine Intelligence for GeoAnalytics and Remote Sensing (MIGARS) | 979-8-3503-4542-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/MIGARS57353.2023.10064528

A Comparative Study of Spatial Interpolation Methods for CMIP6 Monthly Historical and Future Hydro-climatic Datasets for Indian Region

Meghal Shah
Dept. of Computer Science & Engg., CSPIT, Charotar University of Science & Technology
India
mvshah1011@gmail.com

Amit Thakkar
Dept. of Computer Science & Engg., CSPIT, Charotar University of Science & Technology
India
amitthakkar.it@charusat.ac.in

Hiteshri Shastri
Dept. of Civil Engg., CSPIT, Charotar University of Science & Technology
India
hiteshrishastri.cv@charusat.ac.in

***Abstract*—The Global Climate Models (GCMs) provides one of the most credible datasets to understand the past and future of earth's climate. Regional Climate Models (RCMs) are applied to obtain climate information from GCMs to regional level. In practice, spatial interpolations are applied to re-grid coarser resolution GCM data to finer resolution before it is fed to RCM. Standard interpolation algorithms are proposed in literature, however mostly they are selected arbitrarily, especially for interpolation of GCM data. This study is envisaged to provide a comparative assessment of six interpolation methods for spatial interpolation of Sixth phase Coupled Model Intercomparison Project (CMIP6) datasets from six different GCMs. No existing study provides their comparative performance for CMIP6 variables specific to the Indian region. The performance evaluation of selected methods is carried out for seven important hydrological parameters at monthly time scale. The selected parameters are interpolated to 0.25 degree spatial resolution for the entire Indian region. Based on skill tests performed, the Bilinear and Bicubic interpolation methods are realized to be suitable choices for gridded data interpolation.**

Keywords— GCM, RCM, CMIP6, comparative, interpolation

I. INTRODUCTION

Interpolation is a simple mathematical method used to estimate an unknown value by using related known values. In a raster framework, this process is applied to create a surface based on values at isolated sample points where data on some phenomenon is collected. In general, this process may be described as mathematical estimation to values between these points. The major environmental variable data such as weather and soil properties etc. are collected at specific locations spaced apart uniformly or non-uniformly. The coarse resolution gridded GCM data resembles a uniformly spaced sampled location field. The finer spatial array of these data may provide a better accurate evaluation of the properties at unsampled locations than averaging between sampled locations. The measure of particular property in between these data points can be interpolated by fitting a suitable interpolation model that takes care for the expected variation. This problem of generating continuous surfaces with the help of irregularly distributed data points is a task for many disciplines[1]. Different methodologies are proposed in literature to perform this task. However, in most cases selecting an appropriate methodology to best reproduce the actual surface remains difficult.

The substitution of mean values for missing data disrupts the inherent structure of the data and leads to error in the matrix correlation[2]. The interpolation method performs better for all percentages of missing values. In comparison to methods of linear interpolation, the mean method results in very high inaccuracies.

The interpolation algorithms widely assists in forecast evaluation and image interpolation to predict unknown values for any geographic point data, such as chemical concentrations, elevation, noise levels, rainfall, etc. Different interpolation methods have their own advantages and limitations, depending on the characteristics of the point dataset. The methodology that is found suitable for one particular dataset may be non-suitable for another dataset, or for the same dataset obtained from a different location. Some of the climate variables for example rainfall depicts a higher spatial variability in both magnitude and frequency. A correct understanding of the spatial distribution of precipitation at different spatial scales makes key information for hydrological modeling, water resource management, agricultural activities, urban planning and many more. The selection of an appropriate spatial interpolation methodology is therefore highly important to provide reliable spatial distribution of precipitation in transforming from coarse to fine spatial resolution. Hence, it is important to provide a strong basis to select a suitable methodology for a particular point data set. At the same time it is equally important to specify objectives of interpolation, as the same evaluation criteria may not serve for different objectives.

This study provides comparative evaluation of six different methods namely : bilinear, bicubic, distance weighted average, nearest neighbor, first-order conservative and second-order conservative for re-gridding CMIP6 historical(1950-2014) and future data(2015-2044) for parameters namely precipitation(pr), maximum temperature(tasmax), minimum temperature(tasmin), wind(uas, vas), cloud cover(cl) and relative humidity(rh) to the India centric for three scenarios(SSP5-8.5, SSP2-4.5, SSP1-2.6) at monthly time-scales. The hydroclimatic variables are selected for their importance in hydrologic impact studies including rainfall-runoff prediction in different studies such as estimating rate of change of precipitation, humidity, evaporation, wind speed, temperature under changing climate.

II. DATA USED

Global climate models(GCMs) consist of complex mathematical representation of interactions between the major climate system components namely oceans, atmosphere, land surface, and sea ice[10]. CMIP6 experiments provide global climate data for a wide range of climate variables generated from a number of GCMs. Monthly dataset from six GCMs listed in Table 1 are obtained from official CMIP6 data portal[3]. CMIP6 historical observations(1950-2014) and future observations(2015-2044) of climate variables pr, tasmax, tasmin, uas, vas, cl and rh to be obtained for selected six GCMs for three scenarios (SSP5-8.5, SSP2-4.5, SSP1-2.6) at monthly time-scales. To this end, seven surface variables are obtained for each model and ensemble run for the selected scenarios.

TABLE I: CMIP6 MODELS EMPLOYED WITH THEIR RESPECTIVE SPATIAL RESOLUTION

Model	Institution	Resolution (km)
BCC-ESM1	Beijing Climate Center Earth System Model	250×250
CanESM5	Canadian Center for Climate Modeling and Analysis	500×500
CNRM-CM6-1	National Centre for Meteorological Research	250×250
CNRM-ESM2-1	Centre National de Recherches Météorologiques	250×250
GISS-E2-1-G	Goddard Institute for Space Studies	250×250
MIROC6	Model for Interdisciplinary Research on Climate	500×500

Also, the National Centers for Environmental Prediction (NCEP) Reanalysis data set is a global level gridded data set providing representation of the state of the atmosphere of Earth by incorporating multiple observational datasets and outputs from numerical prediction models from 1948 to present. The NCEP determines data processing techniques, the best methods for processing it, and the best ways to deliver it to the users of meteorological, oceanographic, space weather, and hydrologic information. The data obtained for the chosen seven parameters is freely downloadable from its official website[4].

III. METHODOLOGY

The methodology includes three steps namely 1. Understanding data format and data extraction; 2. Understanding spatio-temporal characteristics of data; and 3. Applying interpolation algorithms. In the first step data for the acquired climate variables is extracted from the NETCDF compact data file format and stored in file format compatible with the Python programming. A smooth and efficient data extraction from multiple datasets of each GCM is efficiently achieved using Climate Data Operator (CDO)[5]. Statistical analysis is carried out to understand spatio-temporal characteristics of extracted data. The selected interpolation algorithms are implemented in Python programming language that provides powerful functionality and flexibility for this task as per need of the study. Execution of all the three methodological steps is carried out in the Param Shavak Supercomputer. The supercomputer facility is essentially used for ease in storage and management of the volumetric GCM data files and more importantly to avoid python container kernel crashing issues. At the same time, opting for the supercomputer, as it has multicores and high processing power, helps in achieving a great amount of computation in stipulated time periods. The following part of the methodology section provides a brief overview of mathematical functioning of selected interpolation algorithms.

A. Bilinear Interpolation (BiLIN):

BiLIN[6] is defined as linear interpolation on two directions or axes. The 2D designates two directions (x-y axes). For a set of data coordinates (x_k, y_k), where k = 1, 2 defines position of points Q_{11}, Q_{21}, Q_{12}, Q_{22}. For any given location between (x_k,y_k) BiLIN, helps to find value of data at point P. Here, R_1 and R_2 defined as:

$$R_1(x, y) = Q_{11} \cdot (x_2 - x) / (x_2 - x_1) + Q_{21} \cdot (x - x_1) / (x_2 - x_1) \quad (1)$$

$$R_2(x, y) = Q_{12} \cdot (x_2 - x) / (x_2 - x_1) + Q_{22} \cdot (x - x_1) / (x_2 - x_1) \quad (2)$$

Interpolated point P(x, y) is defined as:

$$P(x, y) = R_1 \cdot (y_2 - y) / (y_2 - y_1) + R_2 \cdot (y - y_1) / (y_2 - y_1) \quad (3)$$

B. Bicubic Interpolation(BiC):

BiC[7] is an augmentation of cubic interpolation, performed using either cubic splines, lagrange polynomials, or cubic convolution algorithms. BiC can be exemplified by functional value f and its derivatives fx, fy and fxy known at all four angles (0, 0), (1, 0), (0, 1), (1, 1) of units square, written as:

$$p(x, y) = \sum_{i=0}^{3} \sum_{i=0}^{3} a_{ij} x^i y^j \quad (4)$$

Where sixteen coefficients are calculated using sixteen equations formed using sixteen closest neighbors of point (x,y).

C. Distance Weighted Average Interpolation (DWA):

A specific case of weighted mean, DWA[8] is a measure of central tendency in which each data point's weighting coefficient is calculated as the inverse sum of distances between it and other data points. As a result, central observations in a dataset get the highest weights, whereas values in the tails get lower weight. Weighting coefficient for x_i is computed as inverse mean distance between x_i and other data points:

$$\underline{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \; where \; w_i = \frac{k}{\sum_{j=1}^{n} |x_i - x\vec{\square}|} \quad (5)$$

Here, coefficient k is any positive number used to prevent computational issues brought in by magnitude of distances between data points. It is beneficial to set k equals n (quantity of data points) or (n-1).

D. Nearest Neighbor Interpolation(NN):

The NN[9] does not take anisotropy into account and instead uses straightforward separation distance. NN lacks mathematical rigor. This technique extrapolates within the Z range of data and is an exact interpolator. Each grid node's Z value is just that of the original data point that is closest to it. Tied data points are sorted first on x, then y, and then z values if two or more points

tie for being nearest neighbor. The smallest value is selected as NN. Uncorrected NN Index is given as:

$$\text{NNI} = \frac{\bar{d}}{E(\bar{d})} \tag{6}$$

here, Average Nearest Neighbor Distance($\bar{d}$) and Expected Average Uncorrected Nearest Neighbor Distance(E$\bar{d}$) given as:

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} \text{ (3.4.2)} \quad \text{E}(\bar{d}) = 0.5\sqrt{\frac{A}{n}} \tag{7}$$

NN is best used for filling in holes or missing data points with regularly spaced data points.

E. First-Order Conservative Interpolation (FoC):

The objective of FoC[10] is to maintain the data field's integral over the course of interpolation from source to destination. Since each source cell data value is considered constant in FoC, its interpolation error is often higher than that of BiL. In this procedure, values of crossing source grids are combined to determine data values for a certain destination grid. The entire overlap between source and destination cell depends on the weight of the given source grid. The weight specifically represents the proportion of area of source and destination cell junction to area of total destination cell.

F. Second-Order Conservative Interpolation (SoC):

Similar to FoC, SoC[11] aims to maintain field integral by interpolation from source to destination. The SoC approach considers the source gradient, therefore it produces a smoother destination field that typically more closely resembles the source field. This is the difference between the first and SoC methods. Going from coarse source grid to finer destination grid makes distinction between FoC and SoC approaches very clear. Similar to FoC, SoC uses a mixture of values from overlapping source cells to determine values for a given destination cell. The weight of the source cell contribution to the sum is determined by how much of the destination cell it overlaps with. The field gradient across the source cell is taken into account in additional terms with SoC interpolation. The SoC approach outperforms FoC remapping by order of magnitude for all remappings from coarse to fine grid.

IV. RESULTS

The aim of this paper is to provide a basis for selecting the most suitable interpolation algorithm for the GCM hydroclimatic variables. The analysis carried out to understand the spatial data characteristics reveals that some of the selected hydroclimatic variables e.g. temperature and winds depict an overall uniform spatial field whereas precipitation and humidity depicts a nonuniform spatial field. Therefore it is essential to carefully select a suitable interpolation algorithm. Therefore it is important to define which interpolation method overall functions best, for different datasets. Here, the possibilities for having different viewpoints that can lead to different rankings of the methods is possible.

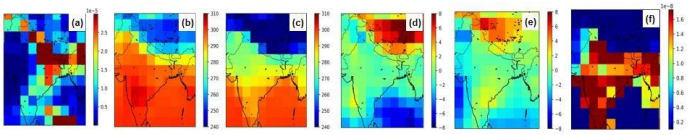

Fig. 1. Temporal mean values of raw GCM parameters (a)pr (b)tasmax (c)tasmin (d)uas (e)vas (f)rh for GCM CanESM5.

It is vital to link the quality of the interpolation to point data because the spatial distribution and variance of GCM data have a significant impact on the interpolation process. Here, it is important to mention that the selection of visual inspection methodology for validation is not applicable when the interpolated surfaces are expected to depict localized non uniformity. Further, it is to be noted that some indices are defined to give first idea about point data characteristics in terms of spatial homogeneity and roughness, essentially applied for spatially rough datasets. Theory at the same time suggests that the best interpolation is the one that minimizes the estimated error in an unknown point[12]. However, as the GCM data suffers from bias from the observed dataset, here we do not have knowledge of the true value in unsampled/resampled points[13]. Therefore, the procedure of cross validation is not applied in this study. The results for a spatially parameter, tasmax and a spatially nonuniform parameter, pr for two GCMs namely BCC-ESM1 and CanESM5 are presented herewith. Results of the other spatially uniform and nonuniform parameters are similar to the results presented for tasmax and pr herewith. The plots of selected parameters and GCMs are included because of the space constraint and also to avoid monotony of the paper.

Fig. 2 provides the result of temporal mean values of an interpolated GCM dataset for longitude 70°E to 96°E and latitude 6°N to 38°N covering the Indian region. The interpolated values between the grid points (6.9765, 70.3125); (9.7671, 70.3125); (6.9765, 73.1250); (9.7671, 73.1250) having data value 300.6K, 300.5K, 300.5K and 300.8K respectively are checked.

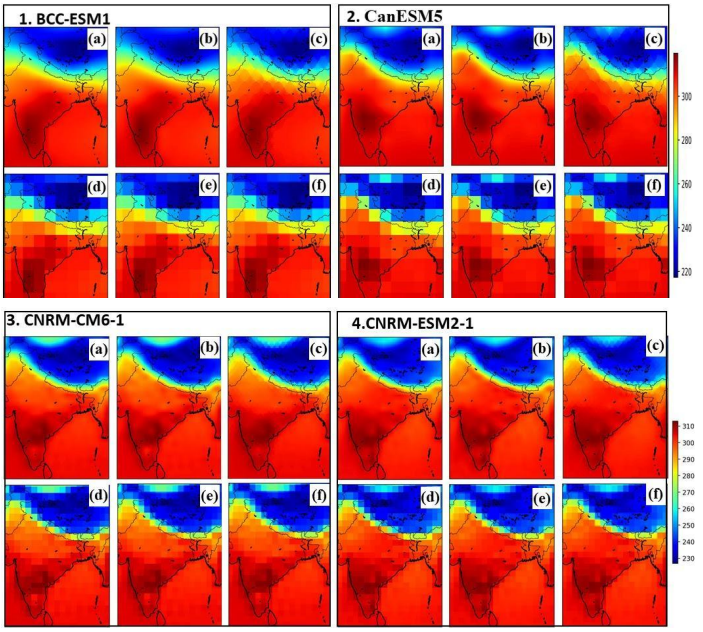

Fig. 2. Temporal mean values of interpolated GCM dataset for parameter maximum temperature (tasmax) for GCM 1. BCC-ESM1 2. CanESM5 3. CNRM-CM-1 and 4. CNRM-ESM-1 temporal mean of interpolated and re-gridded spatial mean by application of (a) BiL (b) BiC (c) DWA (d) NN (e) FoC (f) SoC.

Fig. 3 provides the result of temporal mean values of interpolated precipitation dataset over the Indian region.

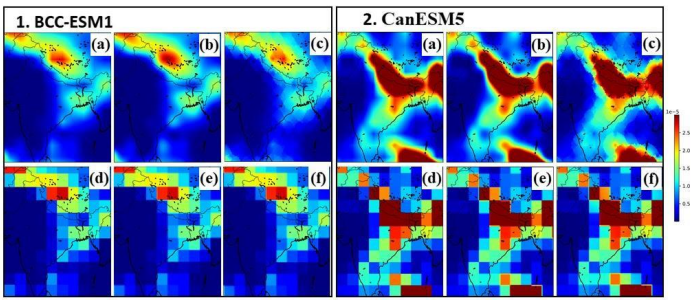

Fig. 3. Same as Fig. 2 but for parameter precipitation (pr)

In addition to the above results the overview of performance of selected interpolation methods for the variables uas and rh fields for CanESM5 is presented herewith.

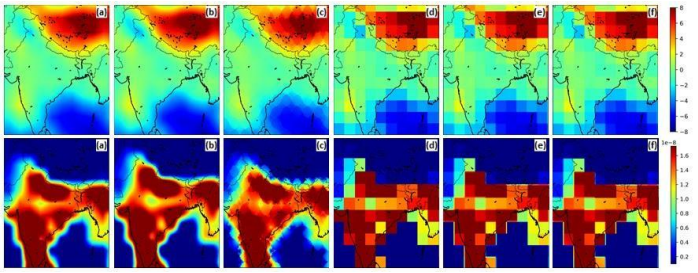

Fig. 4. Same as Fig. 1 but for uas - top panel (a) to (f) and rh- bottom panel (a) to (f) for CanESM5.

Comparing the output of the selected interpolation methods we see the results obtained with these are visually indistinguishable from one another. For example, the variable interpolated using NN algorithm shows a texture with edge jaggedness while that interpolated using the bilinear algorithm shows soft but blurred texture. At the same time the interpolation using the bicubic algorithm shows smoother but sharper texture and readily shows the spatial details. The same conclusions can be drawn for the different datasets and GCMs. Here, it is recognized that the inherently patchy nature of rainfall and humidity in space is well reduced by interpolation. At the same time differences introduced by different interpolation schemes are also evident.

CONCLUSION

The monthly data from six CMIP6 models are evaluated and the results are presented for parameter tasmax, pr, uas and rh. These models with 0.25° spatial resolutions after re-gridding and interpolation are chosen to assess the effect on skills of the various interpolation methods and compare them with the original GCM netCDF file. The results for the interpolation methods are not significantly different, but to some extent, some of them are slightly better than others[14].

Overall BiL and BiC shows comparatively better performance as compared to other methods like DWA, NN, FoC and SoC where we can still observe the larger uniform grids on plotting the dataset even after applying interpolation. It is recommended that interpolation algorithms' spatial capabilities be taken into account rather than competence as measured at specific spots. To evaluate climate variability at a local scale, especially for precipitation, maintaining the geographical distribution is more crucial. The outcomes in this case showed that conservative approaches, which preserve the spatial distribution of interpolated variables, would better suit spatial interpolation of variables.

This study importantly indicates that for maintaining the spatial distribution of the respective parameter depends on the choice of interpolation method. Furthermore, the gridded tasmax dataset from the GCMs found in this study is observed to be optimally interpolated using bilinear and bicubic interpolation methods. Similar research at different catchments or regions are advised for the cross-validation. This study is useful for re-gridding and interpolating the other likewise global climatic parameters from different GCM netCDF datasets. Also, from this study it is observed that few interpolating methods are not feasible to carry out re-gridding so a further approach would be to design an optimal algorithm by improvising those as stable skill tests.

The impacts of climate change in large part depend on local effects. Several algorithms are available to interpolate GCM data from a grid structure to interior points with a high degree of accuracy. However, before selecting an interpolation methodology users must weigh benefits and costs of different competing methods.

ACKNOWLEDGMENT

This work is supported by the establishment of Supercomputer facility at CHARUSAT Sponsored By Gujarat Council of Science And Technology (GUJCOST) in association with C-DAC.

REFERENCES

[1] Md Monowar Hossain, Nikhil Garg, A.H.M. Faisal Anwar, Mahesh Prakash, “Comparing Spatial Interpolation Methods For Cmip5 Monthly Precipitation At Catchment Scale”, J. Indian Water Resour. Soc., Apr., 2021.

[2] Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., Kolehmainen, M. (2004). “Methods for Imputation of Missing Values in Air Quality Data Sets” Journal of Atmospheric Environment, 38(18), 2895-2907.

[3] https://esgf-node.llnl.gov/search/cmip6/

[4] https://psl.noaa.gov/data/gridded/data.ncep.reanalysis.html

[5] CDO User’s Guide, Version 1.6.4, Jun., 2014

[6] D.J. DeRosier, P.B. Moore, “Reconstruction of three-dimensional images from electron micrographs of structures with helical symmetry”, Feb., 1970

[7] B.K. Bhattacharyya, “Bicubic Spline Interpolation As A Method For Treatment Of Potential Field Data”, Sept., 1969

[8] H.S. Shapiro, A.L. Shields, “On Some Interpolation Problems for Analytic Functions”, Jul., 1961

[9] P. Hart, “The condensed nearest neighbor rule”, May, 1968

[10] Bram van Leer, “Towards the ultimate conservative difference scheme. V. A second-order sequel to Godunov's method”, Sept., 1979

[11] Philip W. Jones, “First- and Second-Order Conservative Remapping Schemes for Grids in Spherical Coordinates”, Sept., 1999.

[12] C. Caruso, F. Quarta “Interpolation Methods Comparison”, Computers Math. Applic., Feb., 1997

[13] Shackley, S., Young, P., Parkinson, S., & Wynne, B. (1998). “Uncertainty, complexity and concepts of good science in climate change modelling: are GCMs the best tools?” Climatic change, 38(2), 159-205.

[14] Dianyuan Han, “Comparison of Commonly Used Image Interpolation Methods”, Mar., 2013