The following material is based on the book by Mujumdar & Nagesh Kumar, Floods in a Changing Climate: Hydrologic Modeling.

DATA PRE-PROCESSING

Large scale data requires pre-processing before it is used in downscaling models. Data pre-processing involves making changes to the data, without altering its properties, in order to reduce the errors caused by the unprocessed data.

Pre-processing includes:

1. Interpolation
2. Bias Removal
3. Dimensionality reduction

1. Interpolation:

Interpolation is the process of estimating new data points, using the known/existing data points. Interpolation is necessary before downscaling as the location of grid points of reanalysis data used as GCM inputs and the grid points of GCM outputs does not match. Figure DP 1 shows the mismatch between the grids of NCEP reanalysis data and MIROC3.2 GCM from the Centre of Climate System Research (CCSR, Japan) output variables.

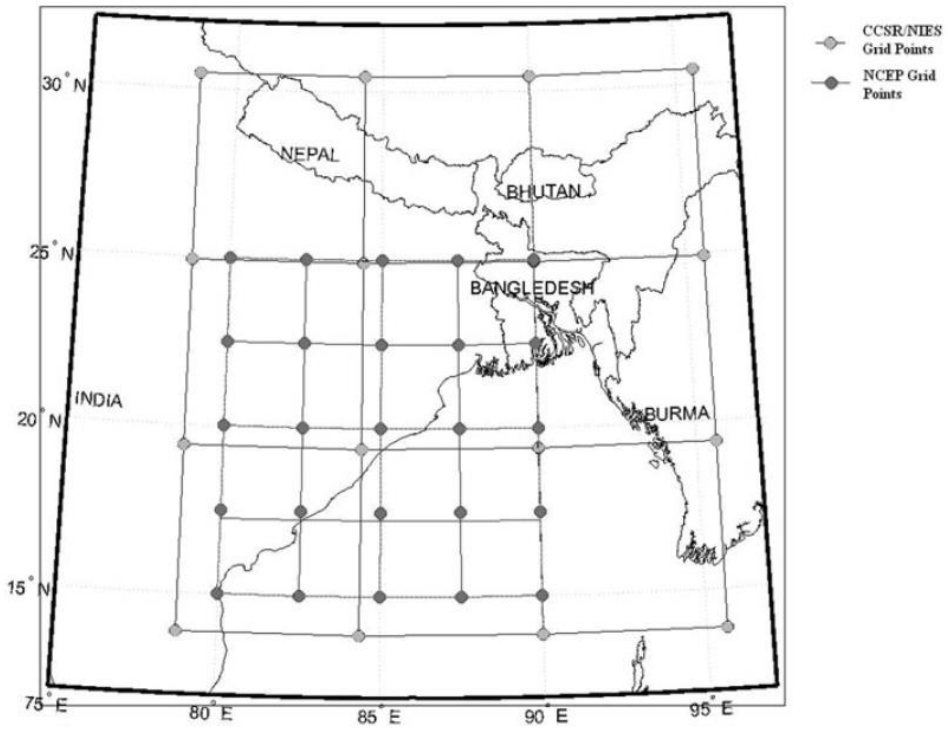

Figure DP 1. Disparity in the grids of reanalysis data and GCM grid points

The data on the NCEP grid has to be interpolated over the other, and this transfer of data from one geo-spatial point to other is called 'Change of Support Problems' (COSP). While choosing an interpolation method, one has to be extremely careful as it will effect/change the scale and support, i.e., the data available. In downscaling studies, linear, spherical, planar, and other interpolation methods are used, spherical being there most popular projection method. For large areas of earth, interpolating on a sphere, ellipsoid or geoid is the most accurate and geometrically consistent than planar interpolation method as planar interpolation methods does not take the shape of earth into consideration. Planar interpolation produces errors as large as 10 degrees as compared to other methods that take shape of the earth into consideration.

The National Centre for Geographic Information and Analysis (NCGIA) has developed a spherical interpolation kit, an open source facility, where several interpolation algorithms are provided which can adapt to the sphere, including inverse distance weighting, this plate splines, multi-quadrics, triangulation, and kriging in the toolkit.

2. Bias Removal:

The physical laws and process, used in GCM, are represented on large sized grid and integrated forward in time. Since GCMs work on large sized grid, the processes that operate on a small scale or which are difficult to understand/formulate are developed into modules and are called parameterizations. Parameterization causes a difference between the observed data and GCM simulated variables, for the present climate. This difference is called as bias.

To office the future climatic conditions, this difference has to be removed. The systematic biases in the mean and variances of GCM predictors, corresponding to observed or reanalysis data, are removed using standardization. Standardization is there process of subtracting the mean from both NCEP and GCM outputs and dividing it by standard deviation. The baseline period selected, is the standard World Meteorological Organization baseline period, from 1960-1990.

The GCM data of a twentieth century experiment done using the baseline data was used to calculate the mean 'μ' and standard deviation 'σ' of all the variables at every single grid

point. The same was performed for NCEP variables for the baseline period. The kth predictor, at time t, after bias correction is given by:

$$v_{stan,t}(k) = \frac{v_t(k) - \mu_{v,1960-1990}(k)}{\sigma_{v,1960-1990}(k)}$$

Where, $v_t(k)$: Original value of *k*th predictor variable at time t

$\mu_{v,1960\text{-}1990}(k)$: mean value of the *k*th predictor variable for the period 1960-1990

$\sigma_{v,1960\text{-}1990}(k)$: standard deviation for the period 1960-1990

3. Dimensionality Reduction:

When a large number of predictors are involved in downscaling, it causes problems such as over fitting, extended time and memory for computation, and existence of significant correlation between predictors. In order to overcome these issues from occurring it is necessary to reduce the dimensionality of the data, while retaining the variability if the same. The dimensionality reduction is performed using Principal Component Analysis, an approach that transforms a large number of variables into smaller uncorrelated variables, called Principal Components. The eigen values and eigen vectors of the sample correlation/sample covariance matrix are used to find out the PC's. When the differences in magnitude of the variables are too high, sample correlation matrix is used to compute PCs. Alternatively standardization can be performed on variables, to make them all equally important as it creates variables such that the mean of each variable is 'zero' and the variance is 'one'. In this case PCs can be computed from sample covariance matrix. The eigen values of the covariance/correlation matrix are then found, called as orthogonal unit vectors, which are ordered in the decreasing order, and along this the original data are projected. The PCs are arranged such that the first principal component is in the direction along with which the data has most variance; the second PC is along the direction which has next highest variance, etc. Initially the first *k* PCs are chosen and the data is projected along these *k* PCs. The first few PCs of each variable, accounts for a large percentage variance and the PCs for which variance rises very slowly with number of PCs, are retained. These PCs that are retained are applied to reanalysis and GCM data to get projections in the principal directions. The below example is taken from Mujumdar & Nagesh Kumar, 2010, which demonstrates the calculation of Principal Components for a sample data set.

Example PC 1:

Sample data on which PCA is to be applied are shown in the Table 1. The data have three variables (or three dimensions), measured using five samples. The samples may be data measured at different times, in which case they would correspond to observations for time *t* = 1 to 5. Principal components of the data are to be identified, such that the maximum variance lies along the first PC, the next highest along the second PC, and so on.

Table 1: Sample data set

Data	**X**	**Y**	**Z**
Sample 1	2	1	3
Sample 2	4	2	3
Sample 3	4	1	0
Sample 4	2	3	3
Sample 5	5	1	9

Step 1: Standardize the data to zero mean and unit variance

The mean and variance of the sample data is computed as follows (Table 2):

Table 2: Sample statistics

	X	**Y**	**Z**
Mean	3.4	1.6	3.6
Variance	1.8	0.8	10.8

The data is then standardized to make the mean as ‘zero’ and variance as ‘one’. Standardization is done by subtracting the mean from the data and dividing it by standard deviation.

Mean (X):

$$\mathrm{EX} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Variance (X):

$$E(X - \bar{X})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X - \bar{X})^2$$

Standardized data:

$$X_s = \frac{X_i - \bar{X}}{std(X)}$$

The standardized data is given in Table 3.

Table 3: Standardized data

X	Y	Z
-1.043	-0.670	-0.182
0.447	0.447	-0.182
0.447	-0.670	-1.095
-1.043	1.565	-0.182
1.192	-0.670	1.643

Step 2: Compute the covariance matrix of standardized data

Note: If variables are not standardized, instead of covariance matrix , correlation matrix has to be used.

$$Cov(X,Y) = E[(X - EX)(Y - EY)]$$

$$Cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

The covariance matrix, $\psi = \frac{1}{n-1}XX^T$, where X is the $n*p$ zero mean matrix with p=3 variables and n=5 observations. The covariance matrix will be a 3*3 matrix as the number of variables in the data set is three.

$$X_s = \begin{matrix} -1.043 & -0.670 & -0.182 \\ 0.447 & 0.447 & -0.182 \\ 0.447 & -0.670 & -1.095 \\ -1.043 & 1.565 & -0.182 \\ 1.192 & -0.670 & 1.643 \end{matrix}$$

The covariance matrix from the standardized data is obtained as:

$$\Psi = \begin{matrix} 1.000 & -0.458 & 0.442 \\ -0.458 & 1.000 & -0.153 \\ 0.442 & -0.153 & 1.000 \end{matrix}$$

Step 3: Compute the Eigen Values and Eigen Vectors of the covariance matrix

The equation, determinant of the covariance matrix gives the eigen values and eigen vectors of the matrix, by equating it to zero.

i.e., det($\psi - \lambda I$) = 0. The calculated eigen values and eigen vectors are:

$\lambda_1 = 0.435$ corresponding eigen vector is $v_1 = \begin{matrix} 0.748 \\ 0.483 \\ -0.454 \end{matrix}$

$\lambda_2 = 0.847$ corresponding eigen vector is $v_2 = \begin{matrix} -0.009 \\ 0.692 \\ 0.721 \end{matrix}$

$\lambda_3 = 1.718$ corresponding eigen vector is $v_3 = \begin{matrix} 0.663 \\ -0.535 \\ 0.522 \end{matrix}$

Verify that the sum of the main diagonal elements of the covariance matrix, ψ is equal to the sum of the eigen values, i.e., the largest eigen value is 3, accounting for 1.718/3 = 57.26% of the variance. The second largest eigen value is 2, accounting for 0.847/3 = 28.23% of the variance and eigen value 1 accounts for 0.435/3 = 14.5% of the variance. The eigenvectors are orthogonal and linearly independent of each other. Hence, the first PC will account for 57.26% of the variance of the original data and so on.

Step 4: Get the weights or loadings for the data

Weights or loading for the data are the eigenvectors, arranged in order from highest to lowest corresponding eigen values. Hence the components are:

$$w = \begin{matrix} 0.663 & -0.009 & 0.748 \\ -0.535 & 0.692 & 0.483 \\ 0.522 & 0.721 & -0.454 \end{matrix}$$

The weights of the first PC are: 0.663, -0.535, 0.522. These are the weights associated with the *X*, *Y*, and *Z* variables, respectively, to arrive at the projection in the direction of the first PC as below.

Step 5: Get the transformed data using projection

The transformed data is equivalent to the dot product obtained as:

$$H = XW$$

i.e.,

$$H = \begin{bmatrix} -0.428 & -0.586 & -1.022 \\ -0.037 & 0.173 & 0.633 \\ 0.082 & -1.259 & 0.507 \\ -1.625 & 0.962 & 0.059 \\ 2.009 & 0.708 & -0.179 \end{bmatrix}$$

Each component of H is a linear combination of components of X. Hence, for the first PC, H1 = 0.663*X − 0.535*Y +0.522*Z and so on.

$$H1 = (0.663)\begin{bmatrix} 2.0 \\ 4.0 \\ 4.0 \\ 2.0 \\ 5.0 \end{bmatrix} + (-0.535)\begin{bmatrix} 1.0 \\ 2.0 \\ 1.0 \\ 3.0 \\ 1.0 \end{bmatrix} + (0.522)\begin{bmatrix} 3.0 \\ 3.0 \\ 0.0 \\ 3.0 \\ 9.0 \end{bmatrix}$$ ……First PC

$$H2 = (-0.009)\begin{bmatrix} 2.0 \\ 4.0 \\ 4.0 \\ 2.0 \\ 5.0 \end{bmatrix} + (0.692)\begin{bmatrix} 1.0 \\ 2.0 \\ 1.0 \\ 3.0 \\ 1.0 \end{bmatrix} + (0.721)\begin{bmatrix} 3.0 \\ 3.0 \\ 0.0 \\ 3.0 \\ 9.0 \end{bmatrix}$$ ……Second PC

$$H3 = (0.748)\begin{bmatrix} 2.0 \\ 4.0 \\ 4.0 \\ 2.0 \\ 5.0 \end{bmatrix} + (0.483)\begin{bmatrix} 1.0 \\ 2.0 \\ 1.0 \\ 3.0 \\ 1.0 \end{bmatrix} + (-0.454)\begin{bmatrix} 3.0 \\ 3.0 \\ 0.0 \\ 3.0 \\ 9.0 \end{bmatrix}$$ ……Third PC

Step 6: Determine number of PCs (columns of *H*) to retain

The cumulative percentage of variance accounted for the PCs are found out, such that the first column accounts for 57.36% of the variance and so on.

REFERENCES:

Mujumdar, P. P., & Nagesh Kumar, D. (2010a). *Floods in a changing climate: Hydrologic modeling*. *Floods in a Changing Climate: Hydrologic Modeling*. Cambridge University Press. https://doi.org/10.1017/CBO9781139088428