

## RAIN [Rainfall Analysis and Interpolation]

RAIN is a suite of deterministic and stochastic spatial interpolation methods developed by Dr. Ramesh Teegavarapu, Assistant Professor, Florida Atlantic University, Boca Raton, Florida, 33431, USA. Dr. Teegavarapu is leader of the Hydrosystems Research Laboratory (HRL) located in the Department of Civil, Environmental and Geomatics Engineering, FAU. The software provided is useful for estimation of missing precipitation data at a location. K-fold cross validation is possible by station selection option. For help with the software: mail to [rteegava@fau.edu](mailto:rteegava@fau.edu) or [ramesh@civil.fau.edu](mailto:ramesh@civil.fau.edu). Details of some of the interpolations methods that are available in this software are provided in the manual. The manual will be updated as soon as enhancements are made to the software. Installation of the software requires specific steps to be followed. These are detailed in the "instructions-for-installation.pdf" document. Few screenshots of the RAIN software are provided below.

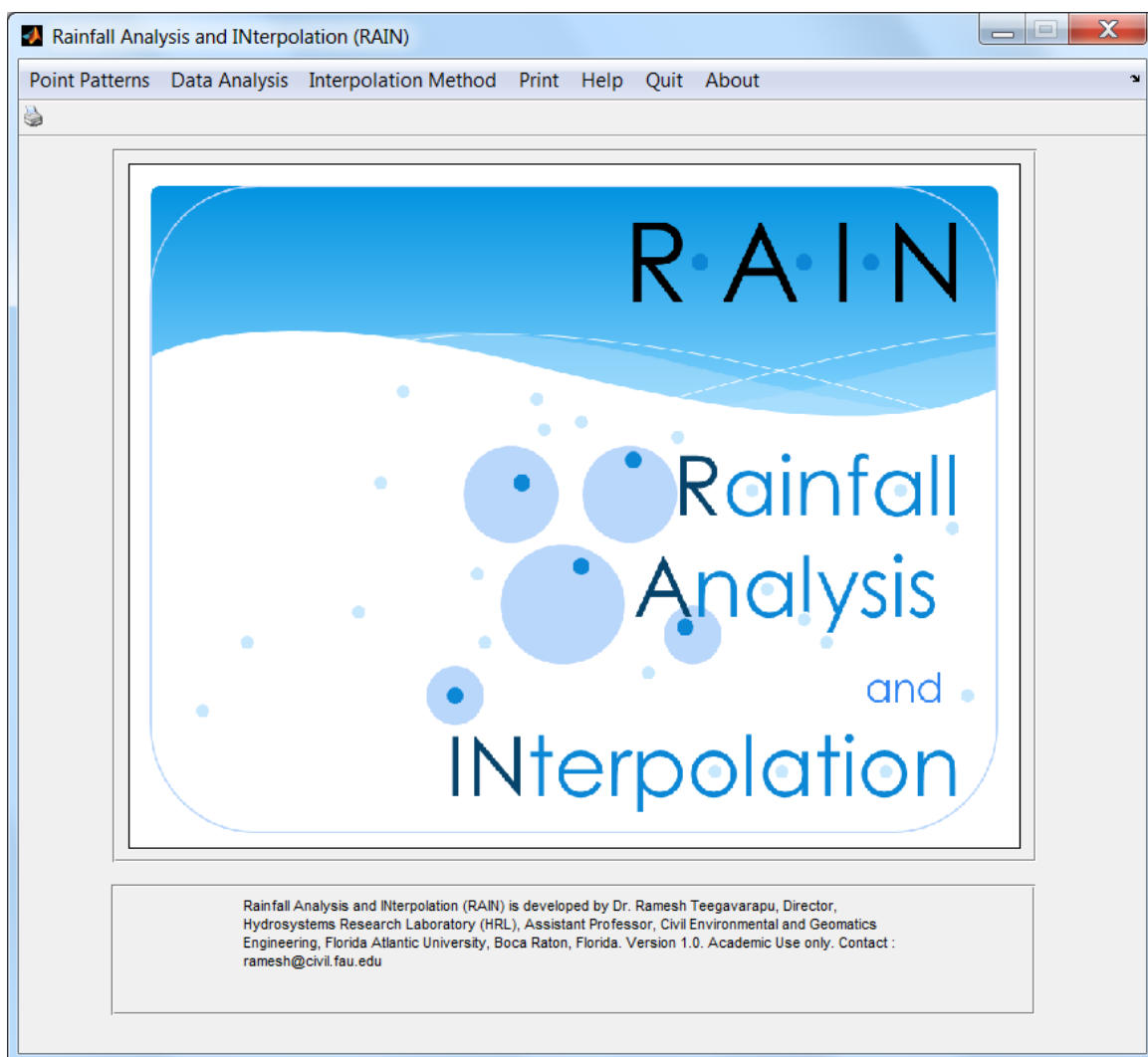


Figure A1. Main Screen Shot of RAIN software

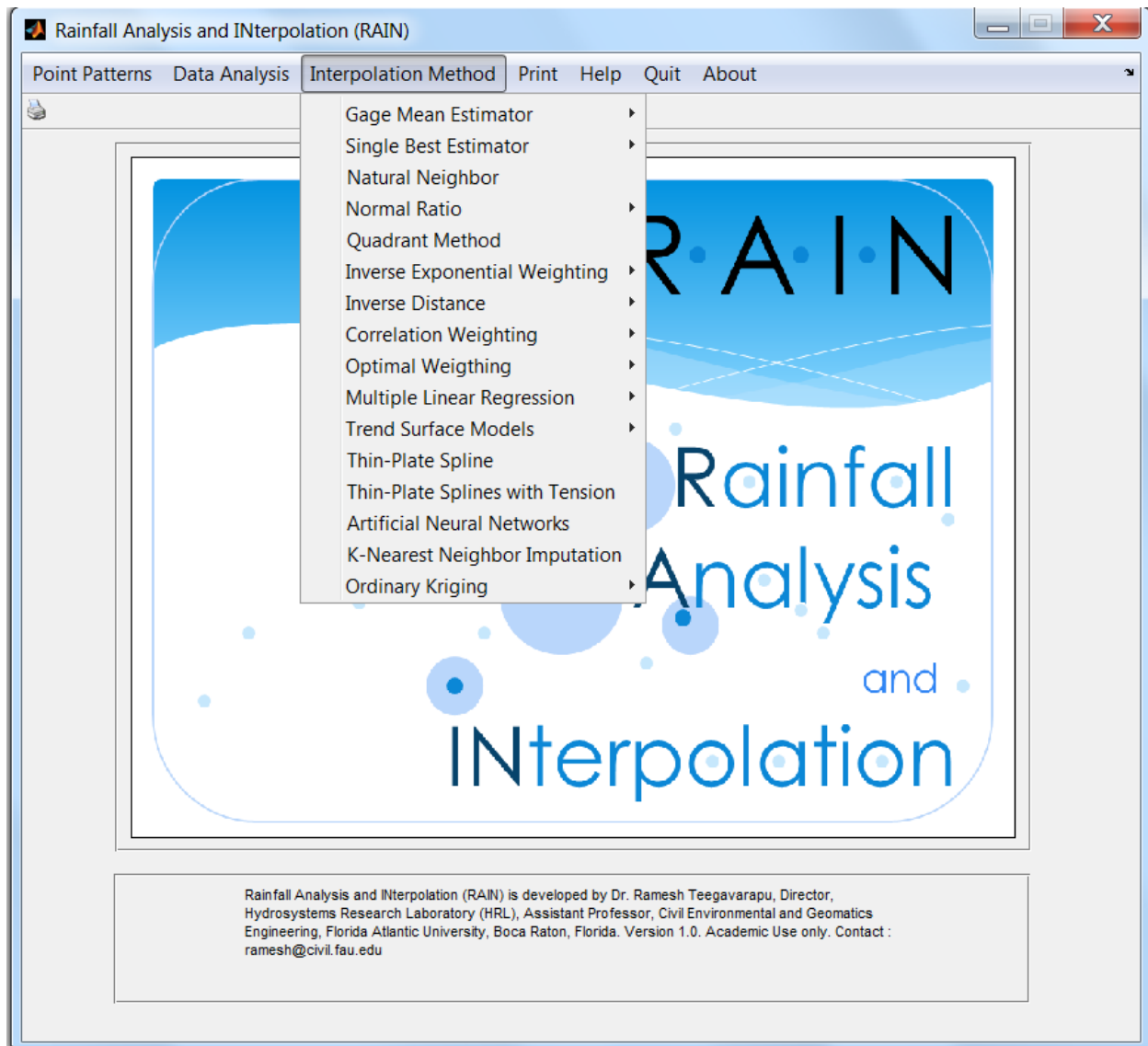


Figure A2. Interpolation methods available in RAIN software

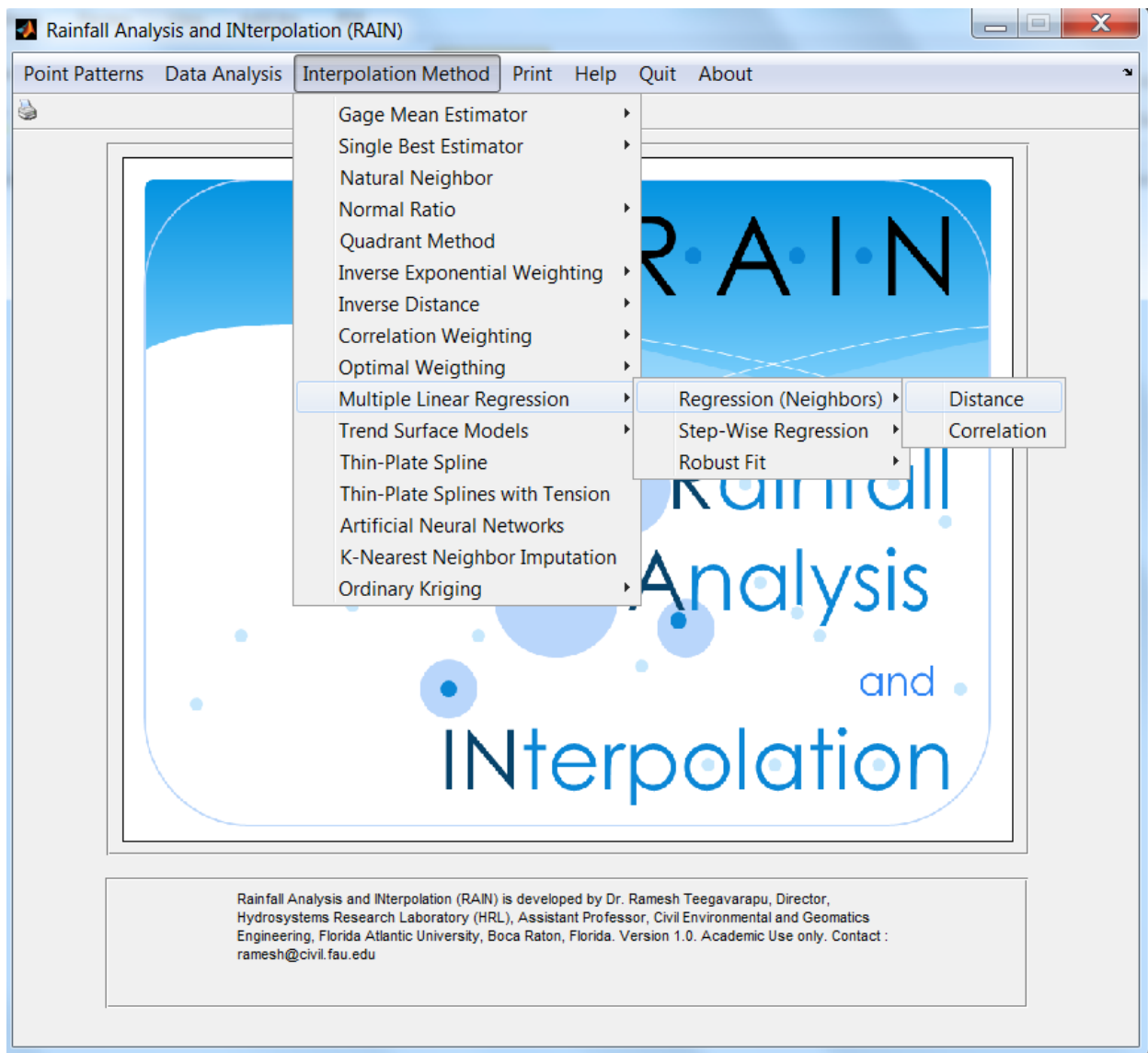


Figure A3. Interpolation methods in RAIN software

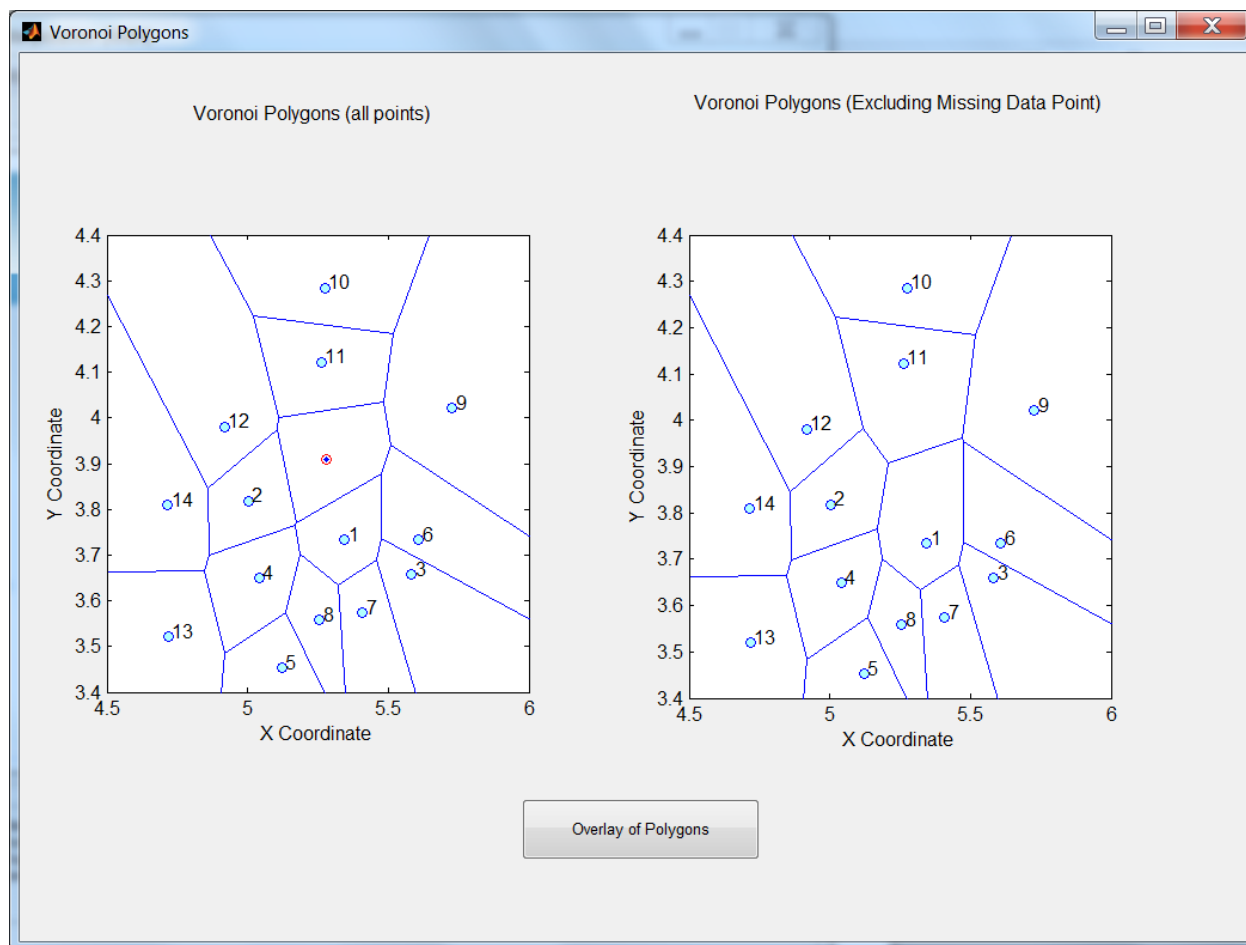


Figure A4. Voronoi polygons created using RAIN software

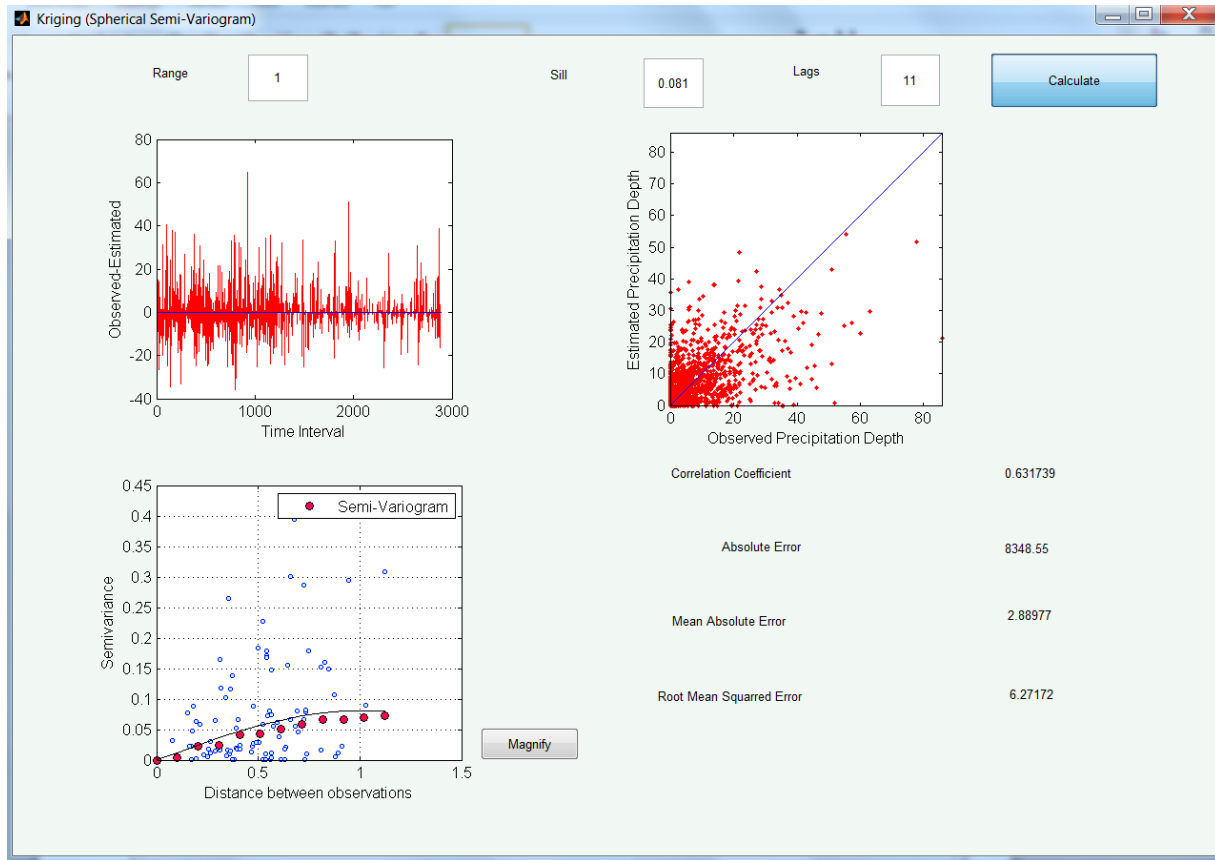


Figure A5. Results screen of RAIN software

## DATA Preparation for RAIN

The RAIN software requires few basic files to be provided by user. All files are ASCII text files. All these files are tab or space separated

Files required are:

File Name	Type	Details	Number of rows	Number of Columns
"allxy.txt"	ASCII – space or tab separated.	Two columns with X and Y coordinates	Number of stations	Two
"caldata.txt"	ASCII – space or tab separated.	Calibration data	Number of time intervals	Total number of stations
"valdata.txt"	ASCII – space or tab separated.	Validation data	Number of time intervals	Total number of stations
"pcor.txt"	ASCII – space or tab separated.	Single column data file with station numbers	Number of stations	One

Example files are provided with the software.

## METHODS

### 1. Inverse Distance (Reciprocal Distance) Method

The models developed in this study are compared with the reciprocal-distance (i.e., traditional inverse distance weighting) method that is most commonly used for estimation of missing precipitation data. The reciprocal distance method for estimation of missing value of an observation at base station,  $\phi_i^m$ , using the observed values at other stations is given by:

$$\hat{\theta}_m^n(x_m, y_m) = \sum_{j=1}^{ns-1} w_j(x, y) \theta_j^n(x_j, y_j) \quad \forall n \quad (A1)$$

$$w_j(x, y) = \frac{1}{d_{m,j}^{-f}} \left\{ \sum_{j=1}^{ns-1} d_{m,j}^{-f} \right\}^{-1} \quad \forall j \quad (A2)$$

where  $ns-1$  is the number of stations;  $\theta_j^j$  is the observation at station  $j$ ,  $d_{m,j}$  is the distance from the location of station  $j$  to the base station  $m$ ; and  $f$  is the exponent referred to as friction distance (Vieux, 2001) that ranges from 1.0 to 6.0.

### 2. Inverse Exponential Weighting Method (IEWM)

The IEWM uses a negative exponential function replacing the reciprocal-distance as weight in the traditional IDWM. IEWM is commonly used in the field of quantitative geography for surface generation. The weighting factor in IDWM,  $d_{m,j}^{-k}$ , is replaced by  $e^{-kd_{m,j}}$  in the equation 51. The most commonly used value for  $k$  is two, and usually several values are tested before arriving at a final acceptable value that improves the performance of method for estimation. In the current study a value of two is used.

$$\hat{\theta}_m^n = \frac{\sum_{j=1}^{ns-1} e^{-(2d_{m,j})} \theta_j}{\sum_{j=1}^{ns-1} e^{-(2d_{m,j})}} \quad (A3)$$

### 3. Single Best Estimator (SBE)

The single best estimator (SBE) is one of the simplest methods for estimating missing precipitation data. Data from the gauge "closest" to the gauge with missing data are used for estimation. This closest station can be selected by Euclidean distance or based on the strongest positive

correlation. The availability of historical data again is essential to use this estimator.

$$\hat{\theta}_m^n = \theta_s^n \quad s \in ns - 1, \forall n \quad (A4)$$

#### 4. Gauge Mean Estimator (GME)

Estimating missing precipitation values by this method is an arithmetic average of all the gauges reporting observed rainfall. The method is a special case of inverse distance weighting and correlation coefficient weighting where all the weights are raised to an exponent of zero. This method fails at estimating missing precipitation values in two situations: 1) when precipitation is measured at all or a few other stations but no precipitation actually occurred at the base station and 2) when precipitation occurred at the base station and no precipitation is measured or occurred at all the other stations. In case 1, the method provides a positive estimate, while in reality zero precipitation occurred at the base station. It is impossible to estimate non-zero precipitation values in the second case since the observations at all the other stations are zero. Data from other sources (e.g. radar-based precipitation estimates) could be used in this situation to estimate missing values.

$$\hat{\theta}_m^n = \frac{\sum_{j=1}^{ns-1} \theta_j}{ns - 1} \quad \forall n \quad (A5)$$

#### 5. Natural Neighbor Interpolation (NN)

Natural neighbor interpolation is referred to as area-stealing or Sibson interpolation. The natural neighbors of any point are those associated with neighboring Voronoi (Thiessen) polygons. These are constructed using two steps. In the first step, a Voronoi diagram is initially constructed based on all rain gauge stations excluding the station with missing precipitation data. In the second step, the location (point) of the rain gauge with missing precipitation data is then added to all other points (rain gauges) and a new Voronoi polygon is created. The two sets of polygons from two steps are overlaid on top of the each other. The proportions of overlaps among these new polygons and the initial polygons are then used as the weights.

$$w_{m,j} = \frac{A_{m,c}}{A_j} \quad \forall j \quad (A6)$$

$$\hat{\theta}_m^n = \left( \sum_{j=1}^{ns-1} w_{m,j} \theta_j^n \right) \quad \forall n \quad (A7)$$

#### 6. Normal Ratio Method

Traditional normal ratio method is also tested for estimation and for comparison purposes. The normal ratio method for estimating missing data at base station is given by:

$$\hat{\theta}_m^n = \frac{\theta_m^a}{ns - 1} \left( \sum_{j=1}^{ns-1} \frac{\theta_j^n}{\theta_j^a} \right) \quad \forall n \quad (A8)$$

where  $\hat{\theta}_m^n$  is the estimated value of the observation at the base station  $m$ ;  $ns-1$  is the number of stations;  $\theta_j$  is the observation at station  $j$ ,  $\theta_m^a$  and  $\theta_j^a$  are average annual precipitation values at the base station and at station  $j$ , respectively. The average annual precipitation values are obtained from long-term historical data.

## 7. Multiple Linear Regression (MLR)

A multiple regression model can be developed using observations at stations (as predictors) and missing data to be estimated as dependent variable. The MLR model is given by:

$$\hat{\theta}_m^n = \alpha_o + \sum_{j=1}^{ns-1} \theta_j^n \alpha_j \quad \forall n \quad (\text{A9})$$

where,  $\alpha_j$  is the coefficient (or weight) for station  $j$  and  $\alpha_o$  is the constant term. There are few disadvantages of using MLR in the current context of estimating missing precipitation data without any stipulated conditions: (1) no constant ( $\alpha_o$ ) and non-negative coefficients ( $\alpha_j$ ). A constant positive value of precipitation is always realized when the MLR is used for estimation when the constant term is included and also the non-negative coefficients (or weights) will result in negative precipitation values.

## 8. Variants of Multiple Regression

Two variants of multiple linear regression methods that include; 1) step-wise regression and 2) robust regression are also evaluated in this study. These variants are generally not used earlier in the literature for estimation of missing data. Step-wise regression is a systematic method for adding and removing variables from a multiple linear regression model based on their statistical significance tested through F-statistic. The robust regression uses a method with iteratively reweighted least squares with a bi-square weighting function. The former regression helps in selecting the optimal number of variables in the multiple regression and the latter improves the regression by reducing the influence of the outliers than the original least-squares method.

## 9. Nonnegative least-squares (NLS)

Nonnegative constraints requirements to obtain positive weights can be enforced using the nonlinear least square constraint formulation defined by equation A10.

$$\text{Minimize} \quad \left\| \theta_j^n \alpha_j - \theta_m^n \right\|_2^2 \quad \forall n, j \quad (\text{A10})$$

**Subject to:**

$$\alpha_j \geq 0 \quad \forall j \quad (\text{A11})$$



The formulation minimizes the norm given by the equation A10 with constraint on the weights (inequality A11). This formulation provides nonnegative optimal coefficients when solved. The solution obtained from NLS is obviously better than that of multiple linear regression model as negative precipitation values are not possible using this model. However both MLR and NLS lack the conceptual superiority of models which allow the use of any objective function in any functional.

## 10. Trend Surface Models

Trend surface models use polynomial functions of different degrees to fit the surfaces to observations in space. Smooth and irregular surfaces may result depending on the nature of the polynomial or the degree of the polynomial adopted for the surface. Trend surface models using linear, quadratic and cubic functional forms are described by the equations A12, A13 and A14. The location of the observation points is specified by  $x$  and  $y$  coordinates (in Cartesian coordinate system) and parameters  $\{\varepsilon_i \mid i = 0 \dots 9\}$  of the trend surface models can be estimated using any non-linear least square regression optimization procedure.

$$\hat{\theta}_m^n = \varepsilon_o^n + \varepsilon_1^n(x) + \varepsilon_2^n(y) \quad \forall n$$

(A12)

$$\hat{\theta}_m^n = \varepsilon_o^n + \varepsilon_1^n(x) + \varepsilon_2^n(y) + \varepsilon_3^n(x^2) + \varepsilon_4^n(xy) + \varepsilon_5^n(y^2) \quad \forall n$$

(A13)

$$\hat{\theta}_m^n = \varepsilon_o^n + \varepsilon_1^n(x) + \varepsilon_2^n(y) + \varepsilon_3^n(x^2) + \varepsilon_4^n(xy) + \varepsilon_5^n(y^2) + \varepsilon_6^n(x^3) + \varepsilon_7^n(x^2y) + \varepsilon_8^n(xy^2) + \varepsilon_9^n(y^3) \quad \forall n$$

(A14)

The applicability of trend surface models for estimation of missing precipitation data for different time intervals requires generation of a surface for each time interval. A nonlinear regression approach is used for obtaining the coefficients in equations A12, A13 and A14.

## 11. Thin Plate Splines

Thin plate splines as exact spatial interpolators can be used to create surfaces that can help estimate values at a location in space. The expression for the thin-spline surface is given by:

$$\hat{\theta}_m^n = \varepsilon_o^n + \varepsilon_1^n(x) + \varepsilon_2^n(y) + \sum_{j=1}^{ns-1} \kappa_j d_j^2 \quad \forall n \quad (A15)$$

where  $d$  is the distance from the point where the estimate is required,  $\varepsilon_o$ ,  $\varepsilon_1$  and  $\varepsilon_2$ , and  $k_j$  are the parameters to be estimated based on data from the  $ns-1$  control points. Thin-plate splines have the same disadvantages as trend surface methods and for multi-time period estimation of missing precipitation data, the thin-splines need to be fitted for every time interval.

## 12. Thin Plate Splines with Tension

Thin plate splines with tension belong to the group of radial basis functions and are useful for fitting surface to values that vary smoothly in a spatial domain. They are variants of thin-plate splines that incorporate a tension parameter that allow controlling the shapes of membranes passing through control points. The equation for estimation of missing precipitation value ( $\hat{\theta}_m^n$ ) is given by equation A16:

$$\hat{\theta}_m^n = \vartheta_n + \sum_{j=1}^{n_s-1} A_{j,n} R(d_j)_n \quad \forall n \quad (\text{A16})$$

where  $\vartheta_n$  represents the trend function, and  $R(d_j)_n$  is the basis function. The basis function value is obtained by using equation A17.

$$R(d_j)_n = \frac{1}{2\pi\eta^2} \left[ \ln\left(\frac{d_j\eta}{2}\right) + c + k_0(d_j\eta) \right] \quad \forall n \quad (\text{A17})$$

The variables  $\vartheta$  and  $A_i$  need to be estimated,  $d_j$  is the distance between the station at which missing data are prevalent and any other station. The variable  $\eta$  is the tension (or weight) parameter,  $c$  is a constant (Euler's constant equal to 0.577215) and  $k_0$  is the modified Bessel function. When the tension parameter is set close to zero, the results from this method approximate to those from a thin plate spline method.

### 13. Geostatistical Spatial Interpolation

The methods developed in the current study are also compared with stochastic interpolation methods investigated in the previous studies. Ordinary kriging models using four different authorized semi-variograms are compared with all other methods developed in this study. Ordinary kriging is widely recognized as a stochastic interpolation method for surface interpolation based on scalar measurements at different locations. Kriging is an optimal surface interpolation method based on spatially dependent variance.

#### 13.1 Semi-Variogram Modeling

The degree of spatial dependence in the kriging method generally is expressed using a semi-variogram. The expression generally used to estimate the semi-variogram is given by

$$\gamma(d) = \frac{1}{2n(d)} \sum_{d_{ij}=d} (\theta_i - \theta_j)^2 \quad (\text{A18})$$

where  $\gamma(d)$  is the semi-variance which is defined over observations  $\theta_i$  and  $\theta_j$  lagged successively by distance  $d$ . Surface interpolation using kriging depends on the semi-variogram model that is selected which must be fitted with a theoretical form that can be used to estimate the semi-variogram values at arbitrary separation distance values. Depending on the shape of semi-variogram, several mathematical models are possible, including linear, spherical, circular, exponential and Gaussian formulations. A typical semi-variogram is shown in Figure A1 with the definition of sill and range.

### 13.2 Semi-Variogram Models

Several semi-variogram models generally are tested before selecting a particular one. The three most widely used semi-variogram models (spherical, exponential, Gaussian and circular) are given by equations A19, A20, A21 and A22.

$$\gamma(d)_1 = C_o + C_1 \left[ \frac{1.5d}{a} - 0.5 \left( \frac{d}{a} \right)^3 \right] \quad (A19)$$

$$\gamma(d)_2 = C_o + C_1 \left[ 1 - \exp \left( - \frac{3d}{a} \right) \right] \quad (A20)$$

$$\gamma(d)_3 = C_o + C_1 \left[ 1 - \exp \left( - \frac{(3d)^2}{a^2} \right) \right] \quad (A21)$$

$$\gamma(d)_4 = C_o + C_1 \left[ 1 - \frac{2}{\pi} \cos^{-1} \left( \frac{d}{a} \right) + \frac{2d}{\pi a} \sqrt{1 - \frac{d^2}{a^2}} \right] \quad (A22)$$

The parameters  $C_o$ ,  $d$ , and  $a$  are referred to as nugget, distance, and range. The weights obtained from kriging equations are used estimate the missing precipitation data at base station using equation A23.

$$\hat{\theta}_m^n = \sum_{j=1}^{ns-1} \lambda_j \theta_j^n \quad \forall n \quad (A23)$$

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE. **THIS SOFTWARE IS INTENDED FOR ACADEMIC USE ONLY** AND MAY NOT USED FOR INDUSTRY, CONSULTING OR FOR PROFIT PURPOSES.