Estimate Missing Climate Data

The EMCD tool is a Windows-based software application that efficiently estimates missing climate data. It simplifies the estimation process by modeling data from its historical records or leveraging information from other stations. The tool is known for its user-friendly interface and is designed to cater to researchers and analysts dealing with climate data. It offers an accessible solution for managing missing data in the domain.

The tool is organized into three distinct sections. In the initial section, the tool processes input data, identifies and addresses issues like incorrect or empty data, and locates missing datetime rows. The second section is dedicated to estimating missing data by drawing on information from other stations. The third section focuses on estimating missing data by analyzing the historical records of the current station.

1- Input Data and Find missing Data

This tool is designed for working with time series data. To get started, input an Excel file containing data from one or multiple stations. Ensure that your Excel file includes a Date column in the first column with the appropriate date format. Afterward, simply select your Excel file and the specific sheet

to load the data into the tool. If your data has a header in the first row, remember to check the "First Row is Header" checkbox.

Once the data loading is complete, proceed to set the time scale for your dataset. Choose from options such as yearly, monthly, daily, or hourly data, and specify the desired time step as needed.

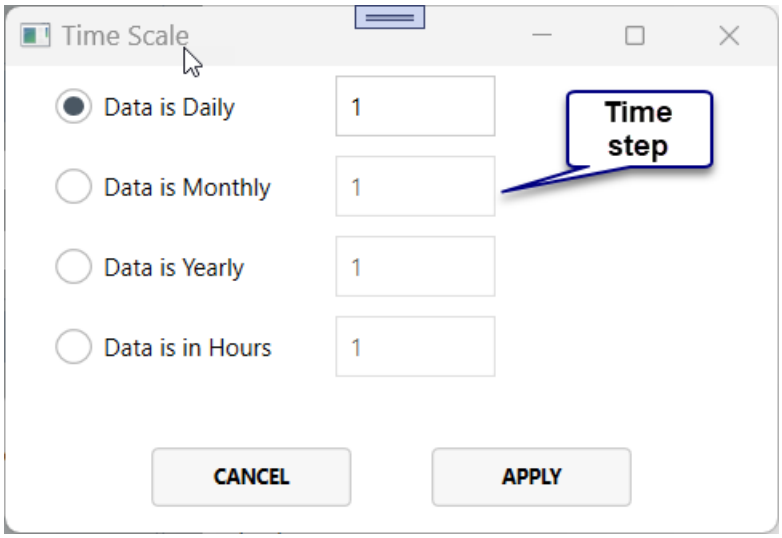

Fig1 – Time Scale

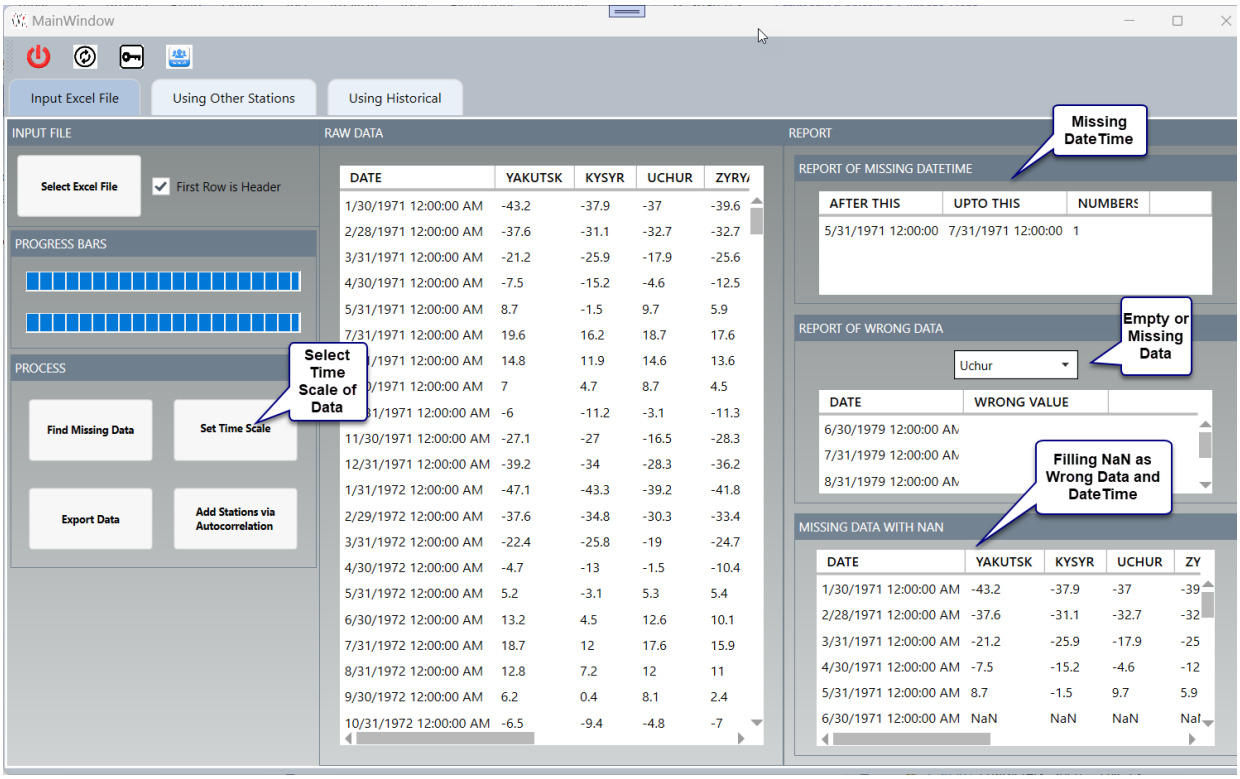

Fig2 – Find Missing Data

There is a button labeled "Add Station via Autocorrelation" that allows you to assess autocorrelation at various lags. Subsequently, you can apply the lag with the highest correlation to your data and save it as a new station. This method proves effective in modeling a station with missing data using its lagged data. For instance, in monthly data, the station often exhibits a strong correlation with its data from 12 months prior.

Now, by simply clicking the "Find Missing Data" button, the tool will identify missing datetime values and incorrect data, including empty cells and

poorly formatted cells. The results will be presented in three tables within the "Report" section on the right-hand side. Additionally, you have the option to export this information to Excel using the "Export Data" button.

2- Fill Missing Data Using Other Stations

On the second tab, choose a station with missing data as a dependent station. Additionally, select one or several stations that do not have any missing data as independent stations.

You can choose from four methods model the station based on one or several independent stations and then use it to fill in missing data. Two of these methods require additional information to run. The Inverse Distance Method (IDW) necessitates entering the latitude and longitude of stations to calculate distance, while the Neural Network method requires specific information as outlined below (For more description please check the Furmulas.pdf file):

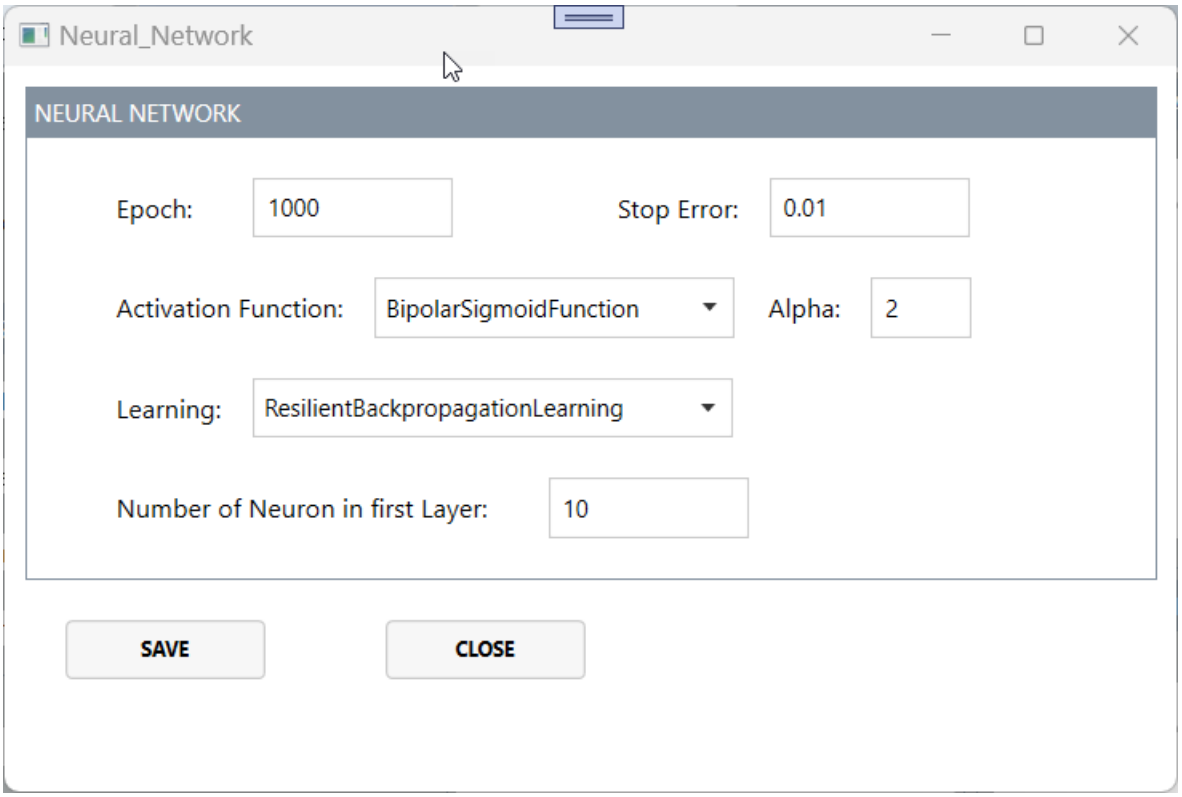

Fig3 – Neural Network

Once you've chosen the dependent and independent stations, simply click the "Fill Missing Data" button to model and fill the missing data.

After running the model, you'll obtain modeled data and the correlation table for all four methods. When using the Multiple Linear Regression method, you'll receive model results along with the modeling matrix. For the Artificial Neural Network, the results will include model results with RMSE and R-squared values.

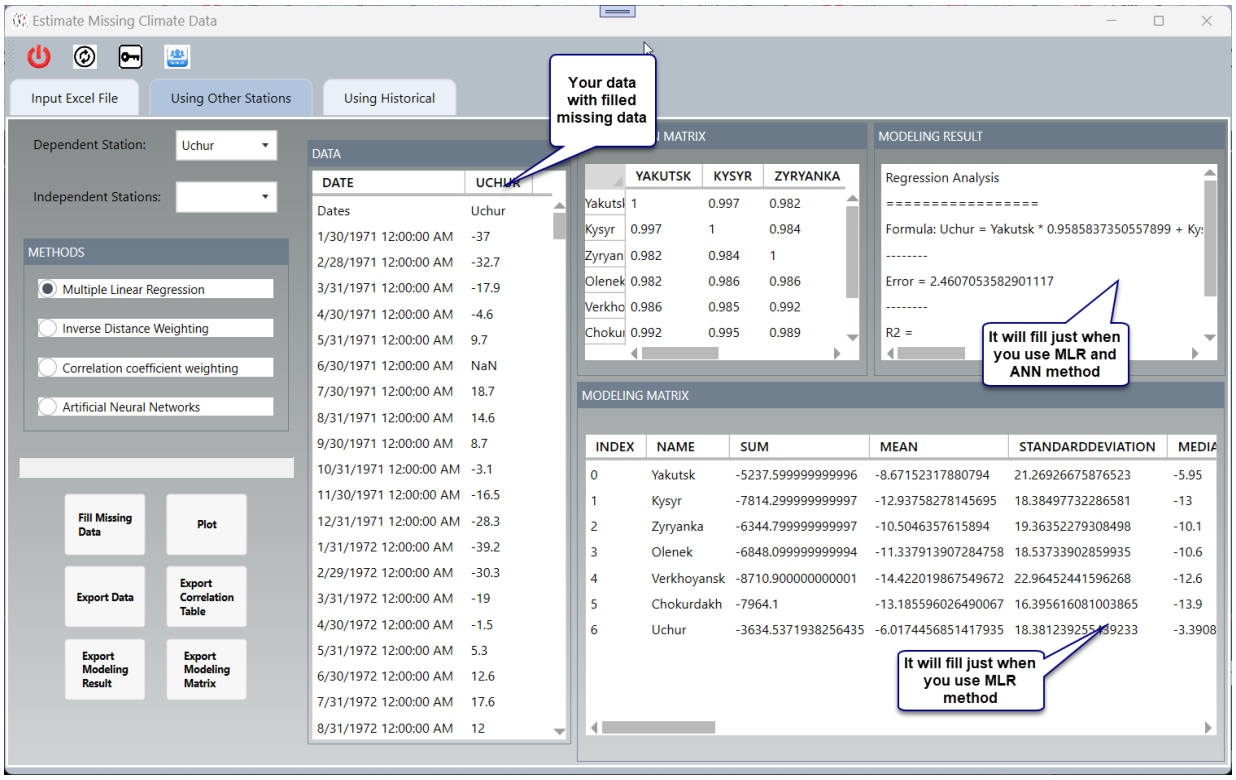

Fig4 – Fill missing data using neighbor stations

Select the "Plot" button to open a linear chart window displaying both modeled and raw data. You can export the chart to Excel for further editing. There are four buttons available to export all results from this tab to a file. Please be aware that if you have added a lagged station in the first tab, you can utilize it to model and fill missing data for a station. This is possible because the lagged station contains data in the missing cells of the station.

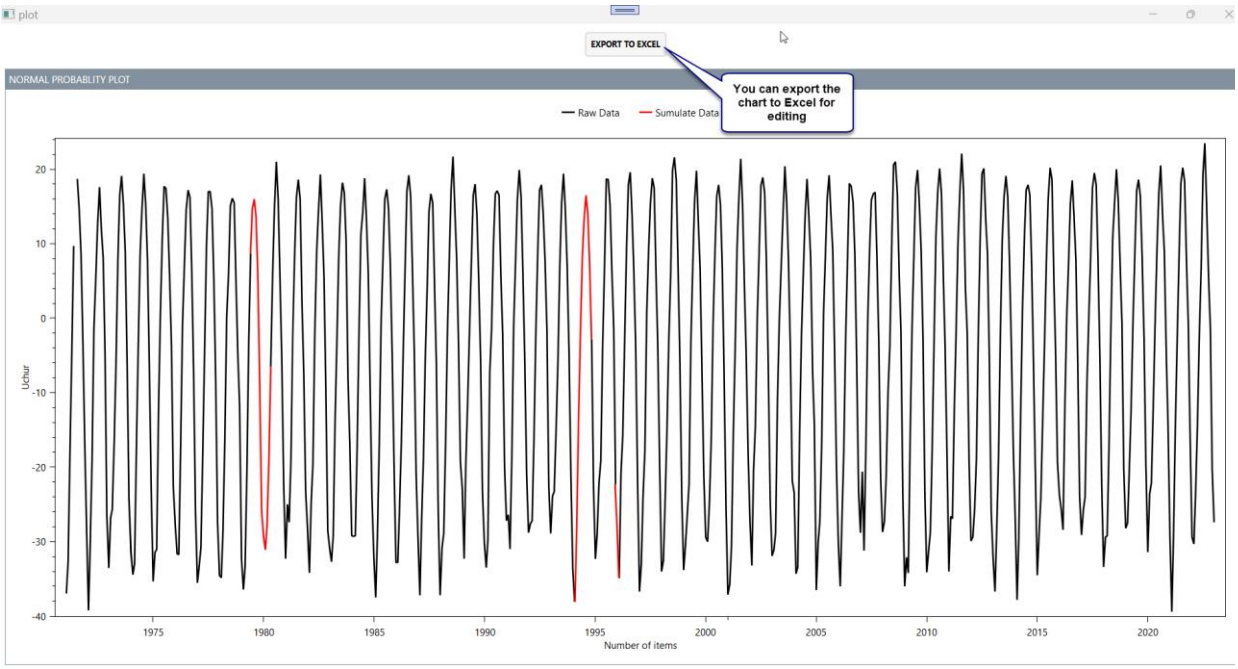

Fig5 – Plot missing data and raw data

3- Fill Missing Data Historical of Station

This tab is designated for filling in missing data of a station using its own data. Therefore, you need to choose a single station and select one of the four methods to fill in the missing data.

If you choose the Last Observation Carried Forward (LOCF) or Next Observation Carried Backward (NOCB) method, you have the flexibility to adjust the lag values. The default lag value for these methods is set to one (based on the original method), implying that the tool will fill the missing data using the value from the cell before or after. However, this might not

yield optimal results for many datasets. For instance, in monthly time series, it is advisable to set the lag equal to 12. You can assess the autocorrelation of the data in the first tab to identify the most suitable lag value.

Additionally, it's important to note that if you have missing data in the first n items (where n is the number of lag), LOCF won't have any preceding data to fill the missing values. Similarly, when using NOCB, be mindful of missing data in the last n items (where n is the number of lag).

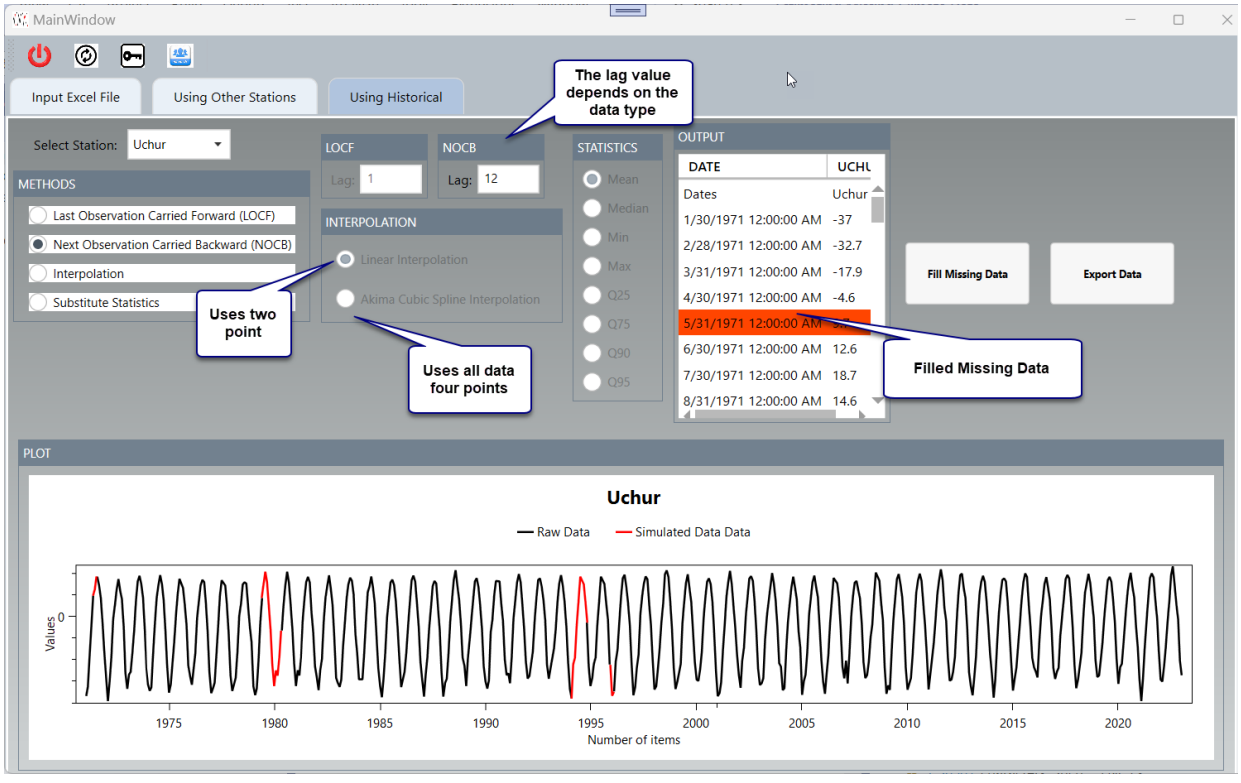

Fig6 – Fill missing data using its own data

The interpolation method consists of two sub-items: Linear Interpolation and Akima Cubic Spline Interpolation. Linear Interpolation is a straightforward method that determines the missing point using the data before and after the gap. On the other hand, the Akima Cubic Spline Interpolation method is more advanced, utilizing two points and the slopes before and after those two points in the vicinity of the missing data. Check the formulas in pdf file.

The Substitute Statistics method fills in missing data by employing one of the long-term statistics from the dataset. This includes options such as mean, median, minimum, maximum, or quantiles at specific percentiles such as 25, 75, 90, and 95.