Expertise
and insight
for the future

Suraj Sharma

Spatial Interpolation in Water Runoff

Hydrology in Vietnam, Laos and Cambodia

Metropolia University of Applied Sciences

Bachelor of Engineering

Environmental Engineering

Thesis

22 December 2019

Author(s) Title Number of Pages Date	Suraj Sharma Spatial Interpolation in Water Runoff Hydrology in Vietnam, Laos, and Cambodia 35 pages + 8 appendices 22 December 2019
Degree	Bachelor of Engineering
Degree Programme	Environmental Engineering
Specialization option	GIS and Hydrology
Instructor(s)	Kaj Lindedahl, Senior Lecturer (Supervisor) Marko Kallio, Researcher, and Lecturer Sujan Dahal, GIS specialist Eija Koriseva, Lecturer

This thesis project was conducted on historical hydrological data. This thesis focuses on the water movement in the 3S basin. The 3S river basin is a transboundary river basin, contributing considerably to the geographical and economical activities for three countries, Cambodia, Lao PDR, and Vietnam. The main three rivers are Sekong, Sre Pok and Sesan, which affect the lots of surrounding livelihoods. The purpose was to use low-resolution runoff and observed streamflow data to determine a prediction of runoff in high resolution raster image with low-resolution runoff data and to compare high-resolution results with the standard error value obtained with the help of streamflow measurement.

This thesis presents different interpolation techniques and error methods. The presentation covers the method of inverse distance weighing, ordinary kriging, and topological kriging interpolation, providing a sound knowledge of how each of the methods works. The required procedure for the spatial data analysis was performed using R-studio and QGIS. Results obtained show, less error for high-resolution than standard low-resolution data. Among error propagation used, KGE value is mostly considered for a proper representation of the goodness of fit. On further analyzing, absolute difference, least-square difference, and ANOVA process were performed on the obtained error value. This suggests that the error value had no relation with each other according to the station while there was no significant difference between the method used. Finally, the absolute difference and least square difference between standard and methods, revealed that TK had the least deviation from standard than of other methods.

Keywords	Interpolation, 3S river basin, Hydrology

Acknowledgment

Writing this thesis would not have been possible without many helping hands. I was honored to get the privilege of writing this report from Marko Kallio (Researcher at Alto University). Without his help, it would not have been possible for me to complete this thesis project. Similarly, I would like to present my huge gratitude to Metropolia UAS which gave me suitable platform and environment for the completion of this thesis.

Above all, I am highly indebted to my supervisor Mr. Kaj Lindedahl who has stood by me during all the activities of this thesis. Similarly, I could not let myself go without huge thanking to Sujan Dahal (GIS specialist) and Eija Koriseva (Math teacher) who had helped me by giving their valuable time during difficult hours while doing this project.

Finally, I would also like to express my gratitude to all the people who have, directly and indirectly, helped me with the completion of this thesis.

Contents

List of Figure

List of Tables

List of Abbreviation

Abbreviations	*Definitions*
LIDAR	*Light Detection and Ranging*
MBS	*Multi Beam Solar*
LORA	*Linear Optimal Runoff Aggregate*
PDR	*People's Democratic Republic*
DEM	*Digital Elevation Model*
QGIS	*Quantum Geographic Information System*
HYMOS	*Hydro-Meteorological services*
HS grid	*Hydro Streamer grid*
*N/A**	*Not Available*
GoF	*goodness of fit*
GIS	*Geographic Information System*
IDW	*Inverse Distance Weight*
OK	*Ordinary Kriging*
TK	*Top Kriging*
B.L.U.E.	*Best Linear Unbiased Estimates*
ME	*Mean Error*
RMSE	*Root Mean Square Error*
Pbias	*Percentage Bias*
NSE	*Nash-Sutcliffe Efficiency*
KGE	*Kling Gupta Efficiency*
HydroSHEDS	*Hydrological data and maps based on Shuttle Elevation Derivatives at multiple Scales*
Avg	*Average*
idp	*inverse distance power*

1 Introduction

Understanding hydrology comprises understanding the nature and components of water. In other words, hydrology is the study done to understand the phenomenon of water circulation on the ground. Studying water, in the real world, helps us in many ways. Water is an essential element to survive in this environment. Most daily life activities are directly or indirectly influenced by the water. The study on the water leads us to its energy, its state, its atomic form, and many more features, by means of which can able to optimize its utilization. On the other hand, researchers are still working on many fields of hydrology to uncover the multiple riddles about hydrology. This thesis is a step taken in the hydrology to understand geographical water movement more deeply by converting low-resolution data to the high-resolution stream network data and comparing with the help of multiple error method. However, water is non-living material, and it shows its movement within its properties. The geographical flow direction of water (i.e. surface flow direction and underground flow direction), is determined by the elevation. This means water flows from higher elevation to the lower elevation, namely by gravity. Similarly, accumulation is the result of interruption to the flow of water in a certain location up to a certain level (elevation). The main aim of the thesis was to conduct an areal interpolation of runoff data from different interpolation methods and to detect different errors values for interpolation method used from different error formulas. Modern technology such as LIDAR (light detection and ranging) technology, MBS (multi beam solar) and aerial photogrammetry techniques were considered too expensive. Due to this, the field survey method is efficient for generating accurate bathymetric maps (Chia-Yu Wu, Joann Mossa, Liang Mao, Mohammad Almulla, 2019). When this type of survey is done repeatedly in the same location, it provides knowledge about the hydrological nature of water during various times and generates the time series data itself. This in return is very helpful for the better fit of the prediction curve for the study.

The 3S (Sesan, Sre Pok and Sekong) river basin was selected as the area to be studied in the thesis project. The 3S basin is sub-basin shared by three countries Laos, Vietnam, and Cambodia covering over 78,650 km^2 and serving more than 3.5 million people and richly comprise of natural resources. The river basin, on the other hand, contributing

most to national and regional development. In our study, this sub-basin contains 11 hydrological observation stations across the basin for the flow measurement in a different part. Our primary work is to interpolate the standard runoff value across the 3S basin with various techniques to find the runoff effect in streamflow followed by comparing the error results to the standard low-resolution data. This made it possible to detect the level of error deviation from the standard method, which in turn helped to select a suitable interpolation method among the various method. Further, the findings can help with the better prediction of flood risk, groundwater availability and irrigation needs. Additionally, our procedure aims to reveal the interpolated map and error values for the various method for better interpretation and conclusion.

2 Area of Interest

The 3S river basin is a transboundary river basin, also termed as the Sesan, the Sre Pok and the Sekong river basin. The river basin constitutes a significant part of the lower Mekong river basin with an area of 78,650 km^2. Figure 1 shows the location of the 3S basin.

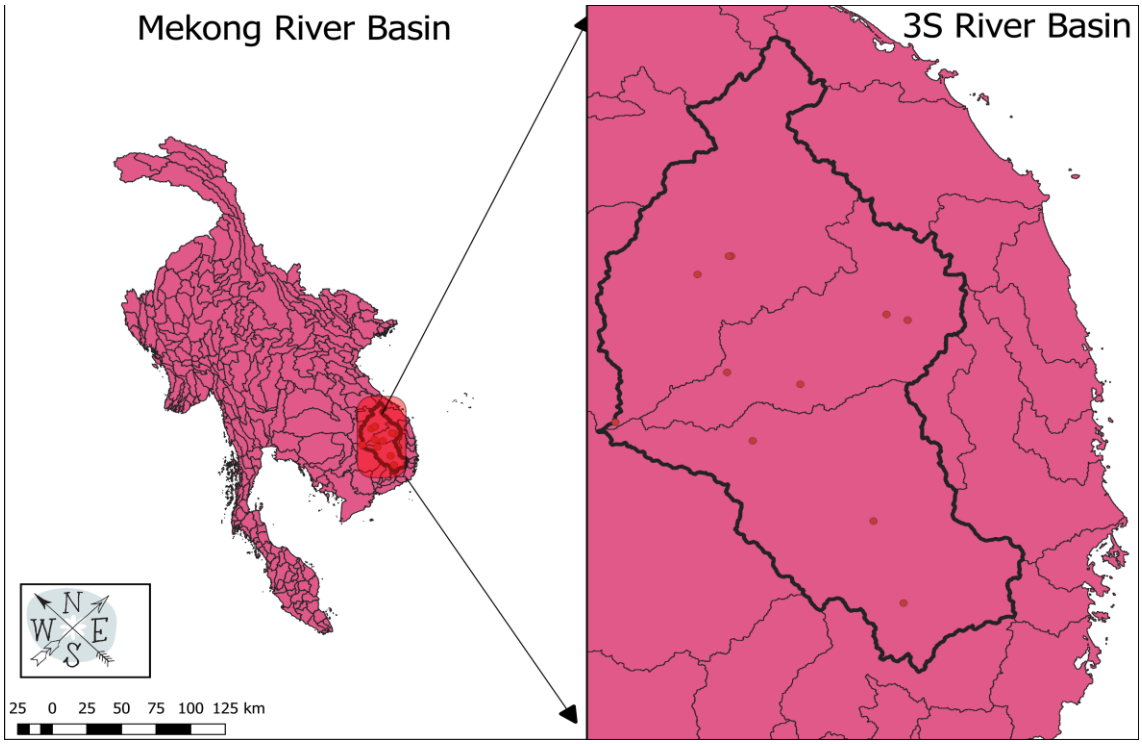

Figure 1: Location of 3S basin within the Mekong basin.

As much as 33% of the total area of the basin is in Cambodia, 38% in Vietnam and 29% in Lao PDR. The source of these three rivers is situated in the Central Highlands of Vietnam. Sekong flows through Lao PDR and Sesan and Sre Pok river flows through Vietnam to Cambodia before merging. These three rivers flow over 40 km to merge with the main stem of the Mekong River at Stung Treng in Cambodia. Figure 2 is the picture showing the daily activities of people in the Sesan river (ICEM Environmental Management, 2016).

Figure 2: People's livelihood in the Sesan river. Source: ICEM

3 Data and Product

3.1 Data source

With the proper interpretation of data, it is possible to extract reliable results with a concrete conclusion. To make this thesis project possible, the raw data was extracted from different sources and means. Some of the important sources for the data were used are listed below:

a. River network and basin outline for the 3S from online source Hydroshed organization.
b. Catchments for every individual river segment, delineated from a DEM.
c. HYMOS streamflow observation stations and streamflow measurements for all the stations are directly collected from Marko Kallio from his study.

d. River basin spatial polygon data frame from Hydroshed organization: an online source.

To compare the results, standard measurement for the runoff was taken from LORA – a raster format time series data available with the standard resolution of 0.5 degrees (Sanaa Hobeichi, Gab Abramowitz, Jason Evans, Hylke E. Beck, 2019). Sometimes it is possible to encounter certain errors during the collection of data or mistakes can randomly occur during the measuring process as well. For this reason, it is important to study data and procedurally omit such kind of data measurement. Data cleaning is a vital process in large data handling. Cleaning erroneous data values data from the raw data improves the reliability and validity of the results. This task is always a challenging task and improper cleaning could lead to wrong conclusions. Therefore, in order to make data cleaning effective following steps were taken into consideration.

a. Assembling data

It is not possible to withdraw information from data unless you have a clear vision about what kind of data it is. To understand data, scattered data was gathered in one place. To this end, QGIS (Quantum Geographic Information System) was utilized. Figure 3 demonstrates what the data looks like when it comes to one place.

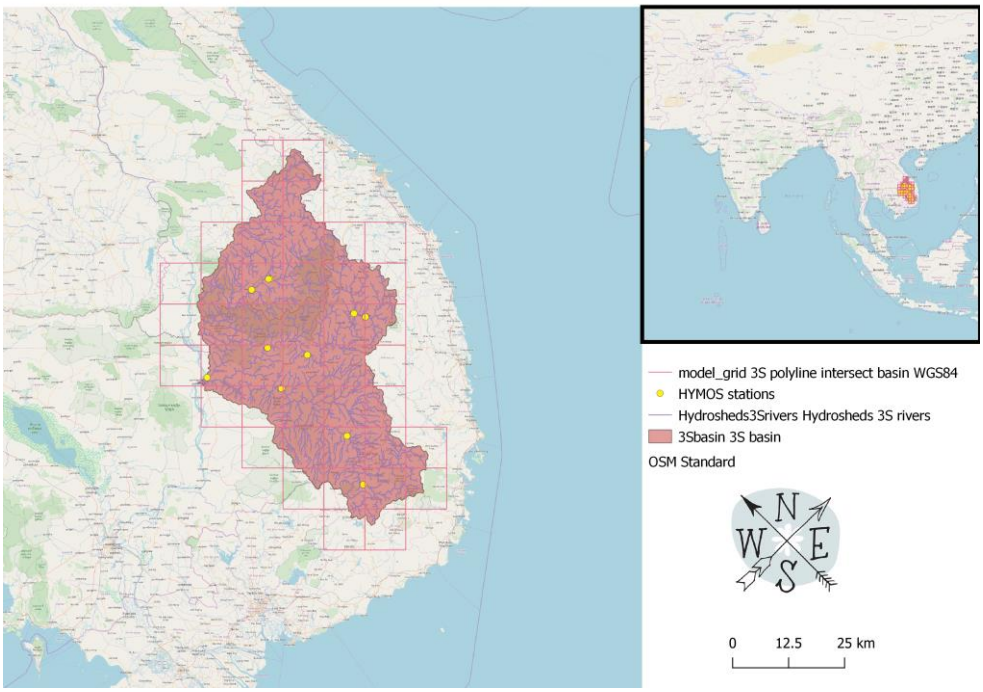

Figure 3: 3S basin in world map with measurement stations.

Table 1 lists the measurement stations and their station codes, locations, and types. Types refer to the river types where normal refers to the ever-flowing river. and N/A* not available data.

Table 1: Showing the attributes of measurement stations

Station Code	Station Name	River	Country	Type
450701	Duc Xuyen	Krong Kno	Vietnam	Normal
451305	Ban Don	Ea Krong	Vietnam	Normal
450101	Lumphat	Sre Pok	Cambodia	Normal
440601	Trung Nghia	Krong Po Co	Vietnam	N/A*
440103	Andoung Meas	Sesan	Cambodia	N/A *
440102	Voeun Sai	Sesan	Cambodia	Normal
430106	Ban Veunkhen	Se Kong	LAO PDR	Normal
430105	--	Se Kong	LAO PDR	Normal
430103	Chantangoy	Se Kong	LAO PDR	Normal
430101	Ban Khmoun	Sekong	Cambodia	N/A *
440201	Kontum	Dak Kla	Vietnam	Normal

To measure goodness of fit (GoF), each station had time-series measurement flow data from the year 1985 to the year 2008 with unit m^3/s (cubic meter per second). The measurement error of the device was used to detect the flow rate of the stream is unknown. However, even in the absence of the measurement error details, a prediction can be made under the assumption that measurement error is very limited with a very minimal effect in the prediction method.

b. Sorting

In the general case, raw data is most likely to be processed for the generation of information. Moreover, the requirement of different nature of data (i.e. flow data, location data, time pattern on the source data, and length of data) at once for the analysis of data makes handling data more complicated during this thesis project. Unless sorting can be made as a prior procedure, it will be more difficult to withdraw the information from existing data. Sorting, in fact, is a process that involves data arrangement in a meaningful order and makes it possible to easily understand, analyze or visualize. Data can be sorted in many ways depending upon user needs or data nature. Data sorted in this thesis project was already in ascending order according to date. It is not wrong to assume that time series data are ordered by date. In this project, data were sorted on a daily basis. In this case, working on daily flowrate measurements may be more accurate but drawing conclusions from it may be misleading and difficult for the operational purpose. So, the data was divided on a monthly basis considering its central tendency as the mean value for each month.

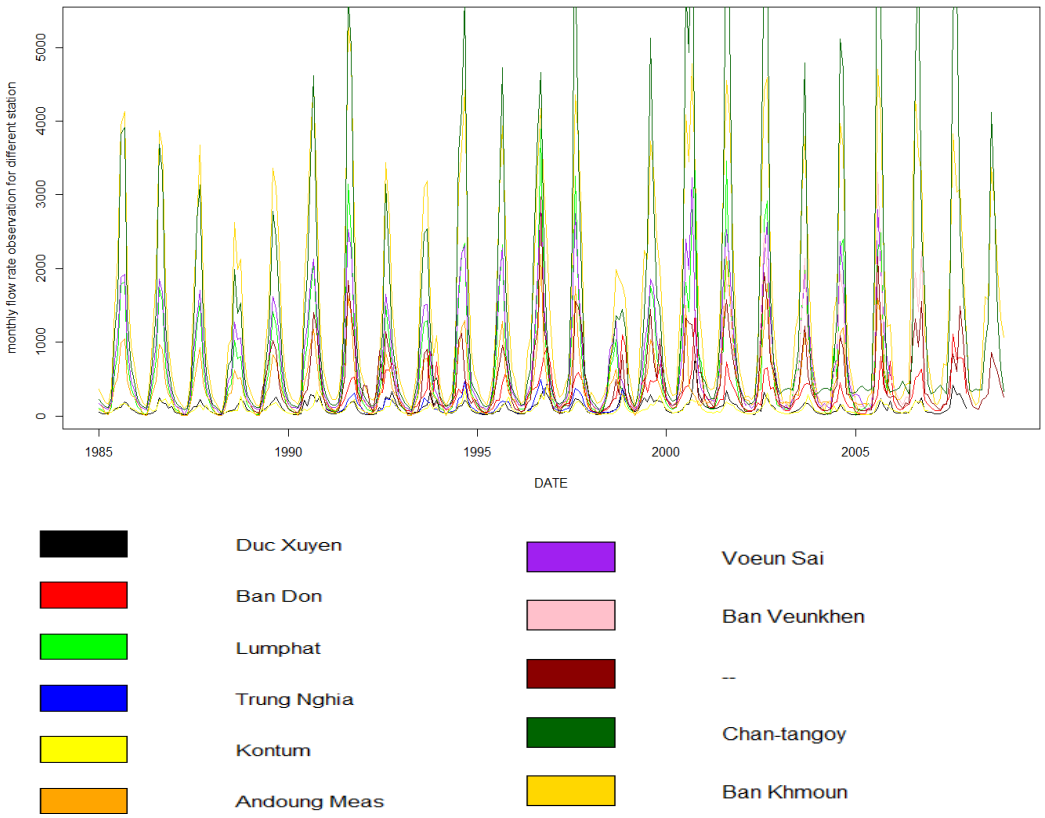

Figure 4: Average monthly flow-rate measurement for each station.

The graph in Figure 4 shows the mean monthly value of the flow rate in the data from the year 1985 to 2008 for the eleven different stations. Each station has a different color line for the ease of observation. It has been observed from the graph that the least flow rate was experienced in the year 1999 and the maximum flow rate was experienced in the year 2001. Additionally, station Chan-tangoy has the highest flowrate, while station Ban-Khmoun has the second-highest flowrate in the graph. Similarly, station Duc Xuyen has one of the smallest flowrate values in the graph. The following image gives the station locations in the 3S river basin:

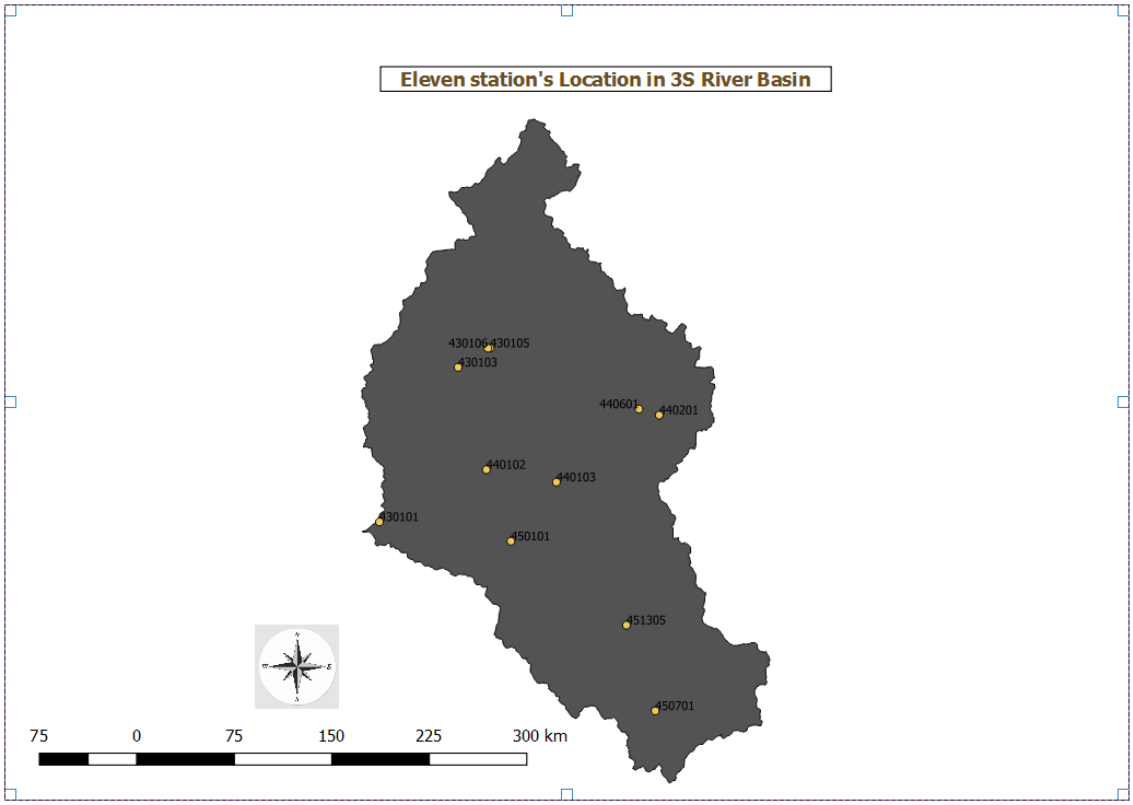

Figure 5: showing station's locations along with its HYMOS code.

Figure 5 shows the 11 stations with their respective locations in the 3S basin labeled by a unique HYMOS code. Observed flow direction was found to be from north-east direction to south-west direction, considering north-east as upstream zone and vice-versa.

c. Outliers

Outliers are the measured value which has a value far beyond the measurement pattern. Outliers are not necessarily an error. In other words, it is an event that can be described as unusual. In mathematical terms, the outlier is the point that falls more than 1.5 times the interquartile range below the first quartile and above the third quartile. A similar process is followed in R-studio while determining the outliers. In the equation, outliers can be expressed as follows:

$$\text{Outlier below} = Q1 - 1.5 * IQR \quad (1)$$

$$\text{Outlier above} = Q3 + 1.5 * IQR \quad (2)$$

For instance, Figure 5 shows the outlier as the point value for the Duc Xuyen station. According to the plot, the median lies between the values 80 and 100. However, the average monthly data included a value where the measured flow was as high as 800 cubic meters per second.

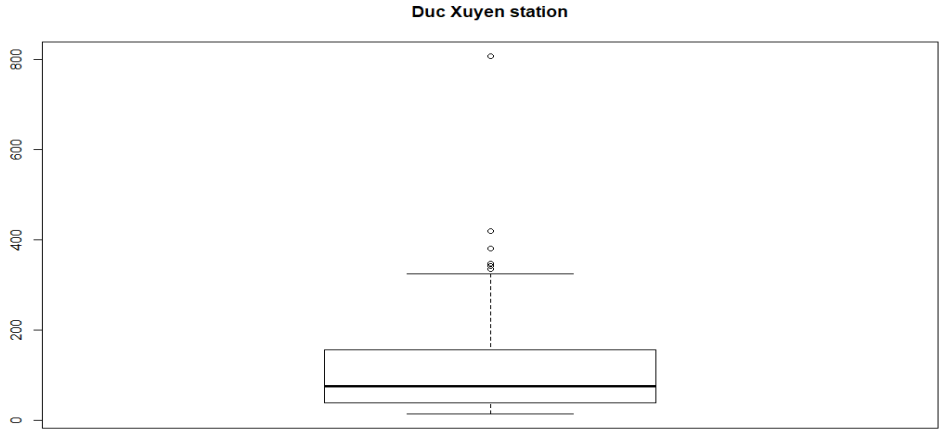

Figure 6: 1st quarter box plot for Duc Xuyen Station.

With Equation (1) and Equation (2) for outliers, we get a few value points, which seems to be unusual. Data processing was done on a monthly average, which in other words, is a summation of daily flow data for the whole month divided by total days in that month. If in that case the outlier from the data is omitted, then for that month all observation data will be zero value. Additionally, if we somehow are able to replace that zero value with overall average data, this will be predicting the possible values over observational values, which might not be acceptable for the work. Therefore, for this report, it will be prudent for the outliers to be as it is for processed of possible outcomes. Figure 6 indicates is the outcome of 1st and 2nd quarter data for the entire time-series data with its overall value for the 11 measurement stations.

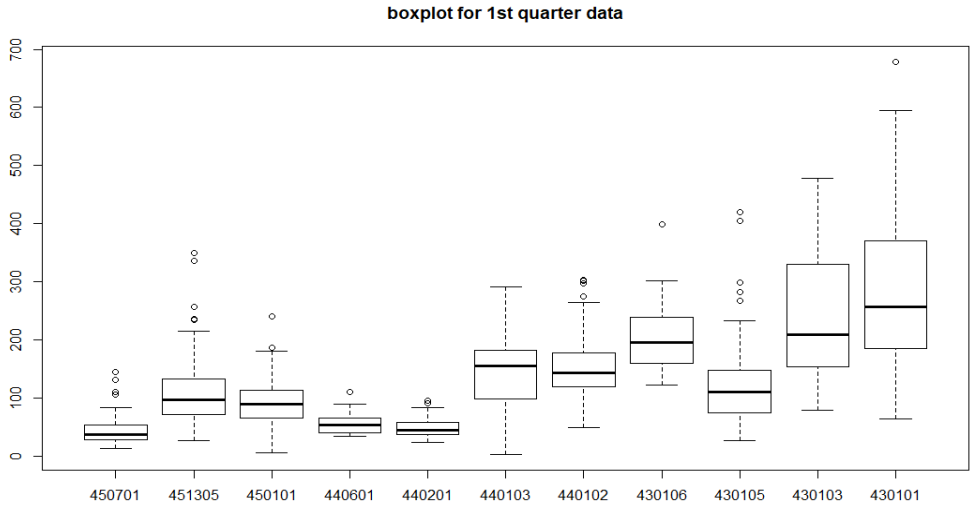

Figure 7: 1st quarter box plot for flow observation.

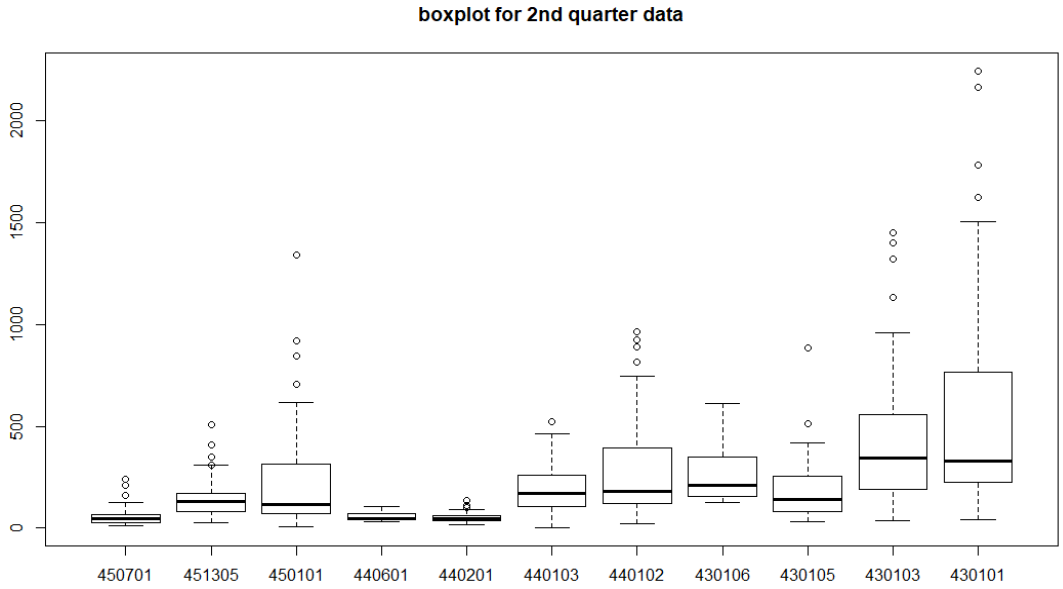

Figure 8: 2nd quarter box plot for flow observation.

Predicting that the flow of river water might vary seasonally, the first step to understand the data is to divide data into four parts according to the quarters of the year and to group flow measurement of all the stations into one box plot according to the quarter. This helped to organize the data in a few pictures and in a precise manner. Boxplots for 3rd and 4th quarter data can be seen in figures 8 and 9 below:

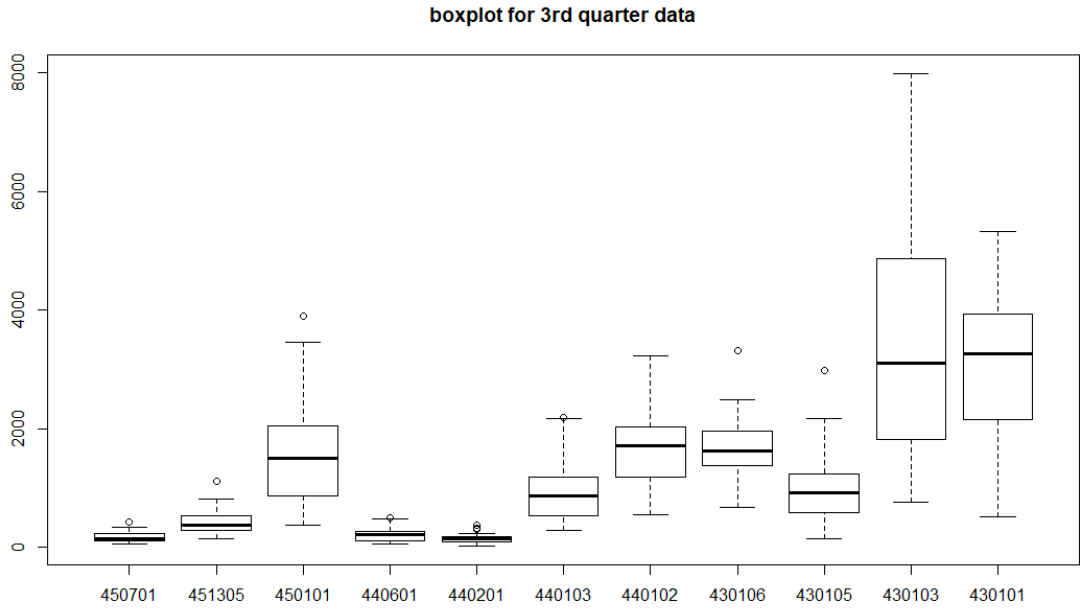

Figure 9: 3rd quarter box plot for flow observation.

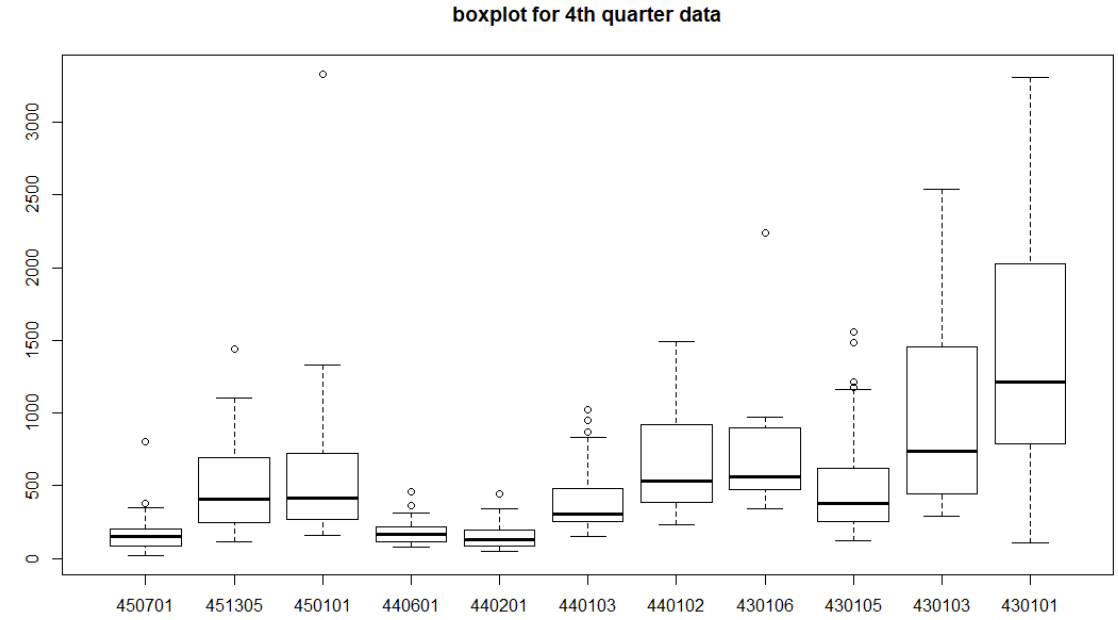

Figure 10: 4th quarter box plot for flow observation.

From the plot of Figures 8 and 9, it can be generalized that the average flow data of the 3rd quarter is higher than the 1st, 2nd and 4th quarters data. Similarly, the lowest average flow data were measured in the 1st quarter data (Figure 6). The flow values in the 2nd and 4th quarter data lie approximately in the same range. Besides, the mean flow of each station reveals that the maximum flow measurement value was measured at the stations Chantangoy(430103) and Ban Khmoun(430101), while the lowest value was observed at the stations Duc

Xuyen(450701), Trung Nghia(440601) and Kontum(440201). This observation also shows that flow value observed for each station was lower for the upstream zone than that of the low-stream zone.

3.2 Product

Hydrostreamer is a product package in the R-studio developed by Marko Kallio. It is capable of computing weight to the river stream according to the river length or area of the grid for the runoff calculation. It is also used for the interpolation and downscaling and runoff products to an explicit river network. Hydrostreamer is performing four main tasks for the calculation. First, it converts raster time-series to an HSgrid object of time-series of a certain time step. The time step can be either hour, day or month. Second, Hydrostreamer computes weight to the river segment according to each segment or grid. Weights are assigned from each polygon to the river segment within that polygon according to either the catchment area of the river segment or the river segment properties. Third, it downscales runoff, disaggregates the low-resolution runoff into each river segment, which is in other words, can be understood as assigning specific runoff to each river segment. Finally, Hydrostreamer applies river routing for the flow downstream by adding all runoff to every segment downstream (see the workflow depicted in figure 10 below).

Simple workflow direction for the product

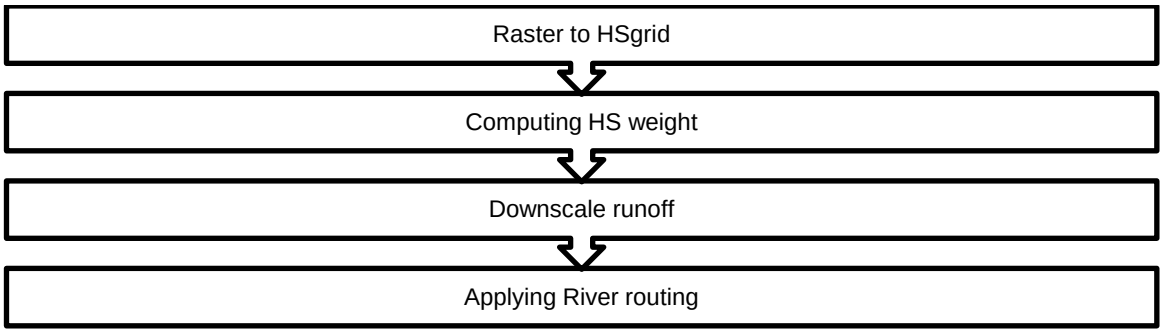

Figure 11: Working procedure of Hydrostreamer.

4 Literature review

Interpolation

Interpolation is the process of predicting an unknown value within the known standard values, while spatial interpolation deals with location, area and the predicted values from sample value of given location. The spatial interpolation in most of the cases is characterized by measured or digitized point data, which can be approximated by functions depending on location in a multidimensional space, vector or tensor field. A spatial prediction model is very complex. It is very hard to determine each and every variable's contributing level of changes during the process. This might result in the difference between the actual measurement and calculated measurement and can be also referred to as non-deterministic estimation.

Many prediction methods have been developed to approximate the value in spatial interpolation. Handling spatial interpolation process comes under the GIS and is normally based on the raster representation, which is the digital representation of heterogeneous datasets with different resolutions. This thesis project is influenced by the linear interpolation technique and, commonly understood as an areal-interpolation technique as well. This type of technique is widely used in science, especially when it involves spatial data and continuous phenomena. In this thesis project, checking error propagation has been done by performing a different method of interpolation i.e. Inverse Distance Weighing, Ordinary Kriging, and Topological Kriging.

4.1 Inverse Distance Weight Interpolation (IDW)

Inverse distance weighting interpolation is also termed as the basic interpolation technique, which states that all points are interdependent with each other based on their distance to each other given by the following formula:

$$H_p = \frac{\sum_{i=1}^{n} {h_i}/{d_i^2}}{\sum_{i=1}^{n} \left[{1}/{d_i^2}\right]}, \tag{3}$$

where H_p is the calculated value of point p in which interpolation is affected, h_i is the known value used to calculate the unknown value at point p, d_i is a distance from the

point *p*, and *n* is the number of points used in the interpolation procedure for estimating the values of point *p*.

In another words, this is also called distance-based interpolation. The closer to each other the unknown point and the sample point are, the higher the value is and vice versa. For instance, people closer to the source of sound can hear this better, and as the source goes further and further from the observation point, the intensity of sound reduced. A similar phenomenon can be experienced while interpolating with the process of IDW in which the value of the unknown point approaches more characteristics with the value of the nearest observational point according to a distance measurement. With IDW interpolation, it is more likely to get small peaks and pits around the sample data point in the raster image. These small peaks and pits are also termed as bull's eye phenomena (Burrough, 1986).

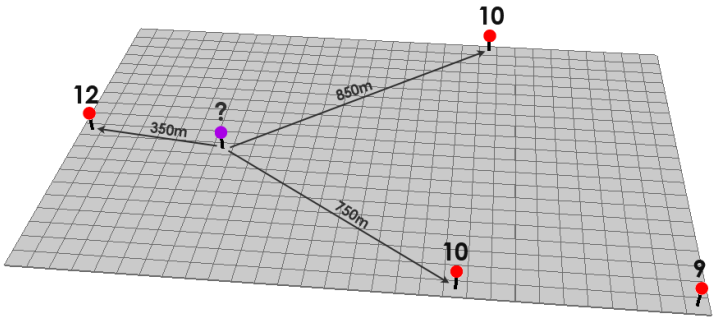

Figure 12: Illustration for IDW (GISGeography-online)

Figure 11 illustrates the known and unknown points. To find the value of unknown points from IDW we have distance from the known point to unknown point, which in spatial data can be retrieved from longitude and latitude. With the formula of IDW, it is possible to find the value of unknown points. By analyzing the behavior of the calculated value, the information about the calculated value can be obtained. The characteristics of the value will be according to the separation distance i.e. more characteristics obtained from a nearer distance and fewer characteristics from a farther distance. The interpolation results for IDW using general R-code is obtained as follows where inverse distance power (idp) is raised to power 2 as shown in the equation (3):

```
Library(gstat)
IDW = idw(formula, observation point, grid, idp=2)
```

4.2 Kriging

Kriging is an estimation approach that relies on linear weights and accounts for spatial continuity, data, closeness, and redundancy. In this approach, weights are unbiased and minimize the estimation variance. It is the procedure of obtaining the best linear unbiased estimates (B.L.U.E) of point values or of block averages (Armstrong, 1998). Best means the mean squared error is at its minimum, linear means the weighted mean is the estimate and unbiased means the mean expected error is zero. For instance, considering z as the given data and u_a as the location where a could be different integers for a different location. Then, $z(u_a)$ is the data values, $z^*(u_0)$ is an estimate, λ_a is the data weights and m_z is the global mean.

To determine the kriging estimation of an unknown point from the given data, $z(u_1)$, $z(u_2)$, and $z(u_3)$, we can use the following formula:

$$z^*(u_0) = \sum_{a=1}^{n} \lambda_a z(u_a) + (1 - \sum_{a=1}^{n} \lambda_a) m_z \ , \tag{4}$$

In Equation (4), the weighted value can be anything depending upon the nature of data. It can be distance or equal weight average of the data (i.e. $\lambda_a = 1/n$). In other words, it is simply a linear sum or weighted average of the data in its neighborhood. Those weights are allocated in such a way to minimize the estimation variance, which results in unbiased estimates.

4.2.1 Ordinary Kriging (OK)

Ordinary kriging method is the type of kriging interpolation method which considers the local variance of the data within the search parameters for the estimation. For the relevant weighting coefficient (λ_i), selected locations are assigned. This method assumes a constant unknown mean in the local neighborhood for each estimation point. In the ordinary kriging method, kriging variance is minimized using a linear external parameter called the Lagrange factor which helps to apply the condition that the sum of all weights is equal to 1. The equation of this method in matrix form can be illustrated as follows:

$$\begin{bmatrix} \gamma(Z_1 - Z_1) & \gamma(Z_1 - Z_2) & \cdots & \gamma(Z_1 - Z_n) & 1 \\ \gamma(Z_2 - Z_1) & \gamma(Z_2 - Z_2) & \cdots & \gamma(Z_2 - Z_n) & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \gamma(Z_n - Z_1) & \gamma(Z_n - Z_2) & \cdots & \gamma(Z_n - Z_n) & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix} * \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ \mu \end{bmatrix} = \begin{bmatrix} \gamma(X_1 - X) \\ \gamma(X_2 - X) \\ \vdots \\ \gamma(X_n - X) \\ 1 \end{bmatrix}, \tag{5}$$

Where γ is the variogram values, Z_1 to Z_n is the real values at the location 1 to n, X is a location where new value is estimated, and μ is the Lagrange factor.

According to Goovaerts (1997), key properties of OK variance is a dependency on two factors that is covariance model and data configuration and independent of data values. Dependency on two factors is an excellent feature while independent data values are considered to be a bad feature. It is because, with this feature, it is possible to get a high difference in prediction error for similarly valued data. Values for the OK interpolation is obtained by the following general code for R-studio:

```
Library(gstat)
OK = variogram (formula, location)
 OK = fit.variogram (formula, vgm())
 OK = gstat (formula, location, model)
 OK = predict (OK, grid)
```

4.2.2 Top Kriging

Top Kriging also usually termed as topological kriging is a type of kriging interpolation method in which the measurements are not point-values but are defined over the non-zero catchment area 'A'. This concept is built on the work of Sauquet (2000) with the aim to develop a prediction model more accurate without any further assumption. Unlike other interpolation, this interpolation technique considers the flow-nature of runoff. This means flow nature does not have more influence by the Euclidian distance over the up-stream and downstream catchment. In other words, downstream catchment must have to treat differently than that of the neighboring catchment. Furthermore, this method takes runoff generation into consideration as the point process which in turn takes indirectly consideration of different variables such as rainfall, soil characteristics, and evaporation. Besides these, routing of the stream network is also considered, which includes the accumulation of runoff along with the stream network. The name topological kriging is given because it takes into account the stream topology and nested catchment areas. Mathematically, Top kriging can be stated as follows:

$$\bar{z}(A) = \frac{1}{A}\int_A \omega(x).z(x).dx, \quad (6)$$

where $\hat{z}$ is a spatially averaged variable, $\omega(x)$ is a weighting function, and A is the spatial area. If A is accounted for non-zero catchment areas, then the method approach for this kriging system remains the same except for the variogram measurement i.e. γ- value.

This value between two measurements of catchment areas *A1* and *A2* can be obtained through the following equation:

$$\gamma_{12} = 0.5 * Var(z(A_1 - zA_2)), \quad (7)$$

Equation (7) suggests that Variogram values are found by integrating a point variogram over a large number of points in each of the catchments. Using gamma(variogram) value to the basic kriging equation matrix, weights 'λ_i' can be calculated in the normal way for the interpolation. It is to consider carefully in the top kriging that the integration is performed over the catchment area that drains to the outlet of the target catchment.

General code for Top-Kriging in R

Library(rtop)

Topkrige = createRtopObject(observation, predictionlocations, params)

Topkrige = rtopVariogram(Topkrige)

Topkrige = rtopFitvariogram(Topkrige)

Topkrige = rtopKrige(Topkrige)

4.2.3 Variogram

The variogram is a function characterizing the variability of samples along with an expectation of the random field $[Z(x) – Z(x+h)]^2$ (Journel and Huijbregts, 1978). It compares the sample in terms of distance and orientation and describes how the sample relates to one another in the space which in return helps to characterize the spatial continuity. Semi-variogram is half of the variogram. This summaries information, concerning the spatial distribution of a variable. Lag-distance (h) is related to the variogram. 'h' is the sampling distance at the position of the sample and starts from 0 because it is impossible to take two samples closer than no distance apart (Clark, 2001). While creating a variogram plot we can experience the direct relation to the variogram and Lag-distance(h). This means a variogram increase with the increase in Lag-distance. Construction of variogram includes both OK and Top kriging with different conditions and methods. Variogram can be expressed in mathematical term by the following equation:

$$2\gamma(h) = \frac{1}{N(h)} \times \sum_{n=1}^{N(h)} [Z_n - Z_{n+h}]^2, \quad (8)$$

where *N*(*h*) is the number of data pairs at the distance *h*. *Z*(*n*) is the value at location *n*, and *Z*(*n*+*h*) is the value at location *h* distance from *n*.

Theoretical variogram model

Variogram cloud is the point to represent the variability among the location with the distance. It is the prerequisite for the prediction of the kriging method. Fitting-variogram incorporates overall points and represents the value through the curve. The nature of the curve depends upon the model used. The model can be spherical, exponential or Gaussian (There are other models, but these are commonly used). Figure 12 is the picture of the theoretical variogram model which represent the overall variogram model in a similar way. The standard variogram model has nugget, range, and sill.

Nugget is an effect, experience in the variogram model where the semivariance curve intersects the y-axis. If the semivariance curve intersects y-axis at 2 then the nugget is 2. It is because, at lag distance zero, the value of the semivariogram is also zero. But, at an infinitely small separation distance, the value is slightly above than zero which in turn approximately observes as zero as shown in the picture.

Sill is the value from where the fitted curve flattens out.

Range in general terms is the distance between nugget distance and sill distance. Considering this, in the variogram model minimum value is the nugget and the maximum value is the sill.

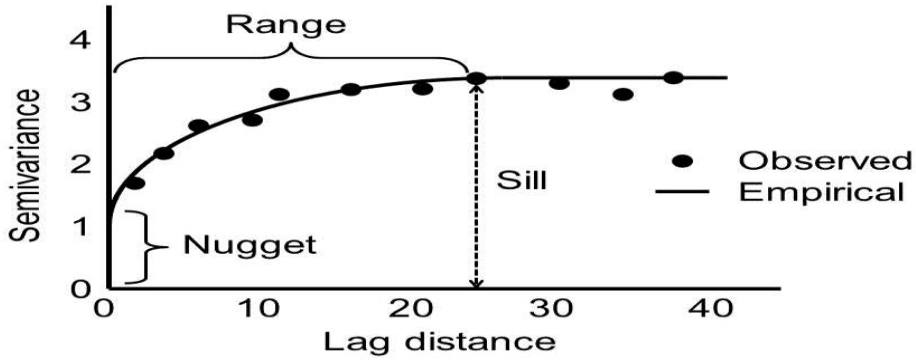

Figure 13: Illustration of the variogram model.

5 Workflow

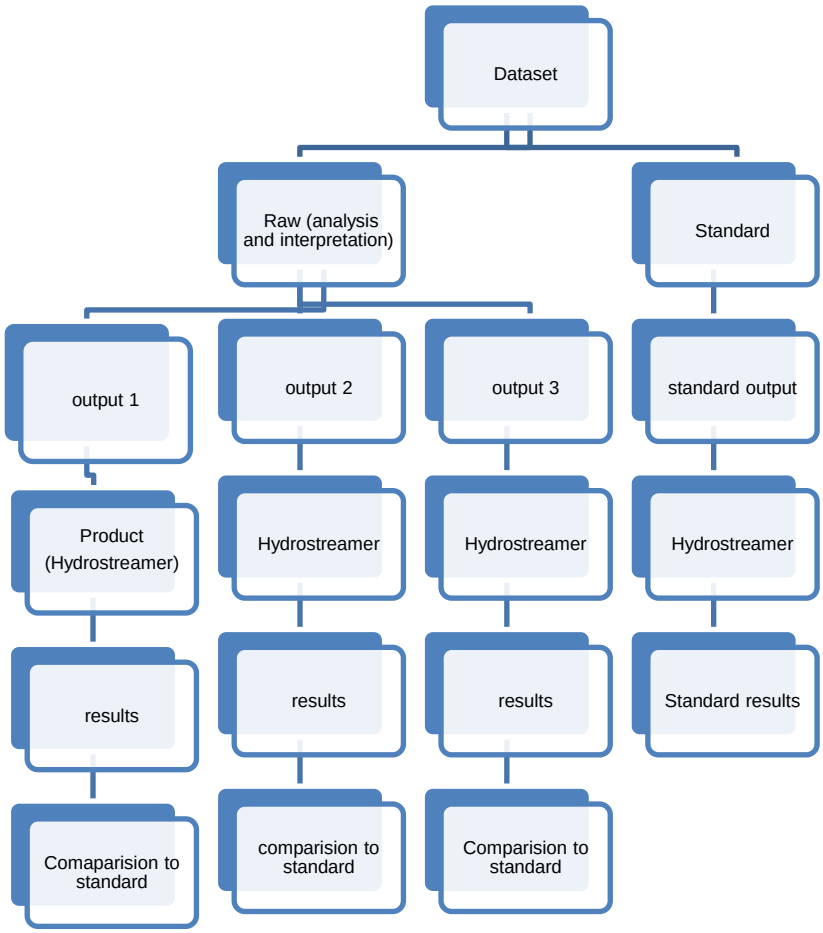

Figure 14: Workflow diagram

Figure 13 is a hierarchical work model that reveals the stepwise workflow during this project. The primary task was the collection of datasets from the various sources. Two types of datasets were collected. The first one is the standard data set from LORA and the second one is the raw data set which is flow measurements of sample locations in the 3S river basin.

Product (Hydrostreamer), capable of analyzing the raster stack form, is used for analyzing all output from the overall set of data. While routing runoff value to the stream, hydrostreamer did not consider all variables (e.g. soil absorption capacity, evaporation, and the direction of storm movement) Standard data is the time series raster data and can be used in the numerical format for analysis, which in return produces standard results.

On the other hand, the crucial task was to form time-series output from the interpolation method. Each of the interpolation methods comprises of time series raster output which

will go through the product and after each result obtained will then be compared with the standard result.

Working terminology and procedure

The report primarily deals with extracting information from data for the standard comparison. Without understanding data, conclusive working is not possible. Data were viewed in different format in prior and further process initialized. The process at first included the working grid with a similar projection for the required prediction map. Inverse distance weighting interpolation consumes less time among the other interpolation process. So, initiation of the process was done with IDW followed by OK and TK. The idea was to generate a monthly raster image from 1980 for each prediction method. IDW is only possible with the required number of points for interpolation, for which centroids from the grid are taken.

The process was initialized with the standard point value of runoff extracted from standard data LORA at the resolution of 0.5 degrees and substituting the runoff value to the centroids of the grid was then followed by methodological interpolation in high resolution, which in this report is 0.0391 degree. Doing so, It was possible to aggregate the runoff data to the stream and to analyze the variation of streamflow caused by the runoff aggregate. Standard data used in this report allowed seeing the results only in the low resolution, but the output built for the report provided the higher resolution data and enabled analysis of the data on high-resolution. Before analyzing error values, it was crucial to add the observation (flow data) for comparison, where the product in this project went through the identical month values for analyzing the process.

Figure 14 shows the raster images for an identical month, which shows how the runoff values are distributed according to the different prediction process. The first picture shows the distribution of runoff value through inverse distance interpolation. The second picture shows runoff distribution from ordinary kriging method, whereas the last picture shows runoff distribution through topological kriging method. All three images are shown with the contour lines with the same range of runoff values to understand the nature of runoff distribution by each method.

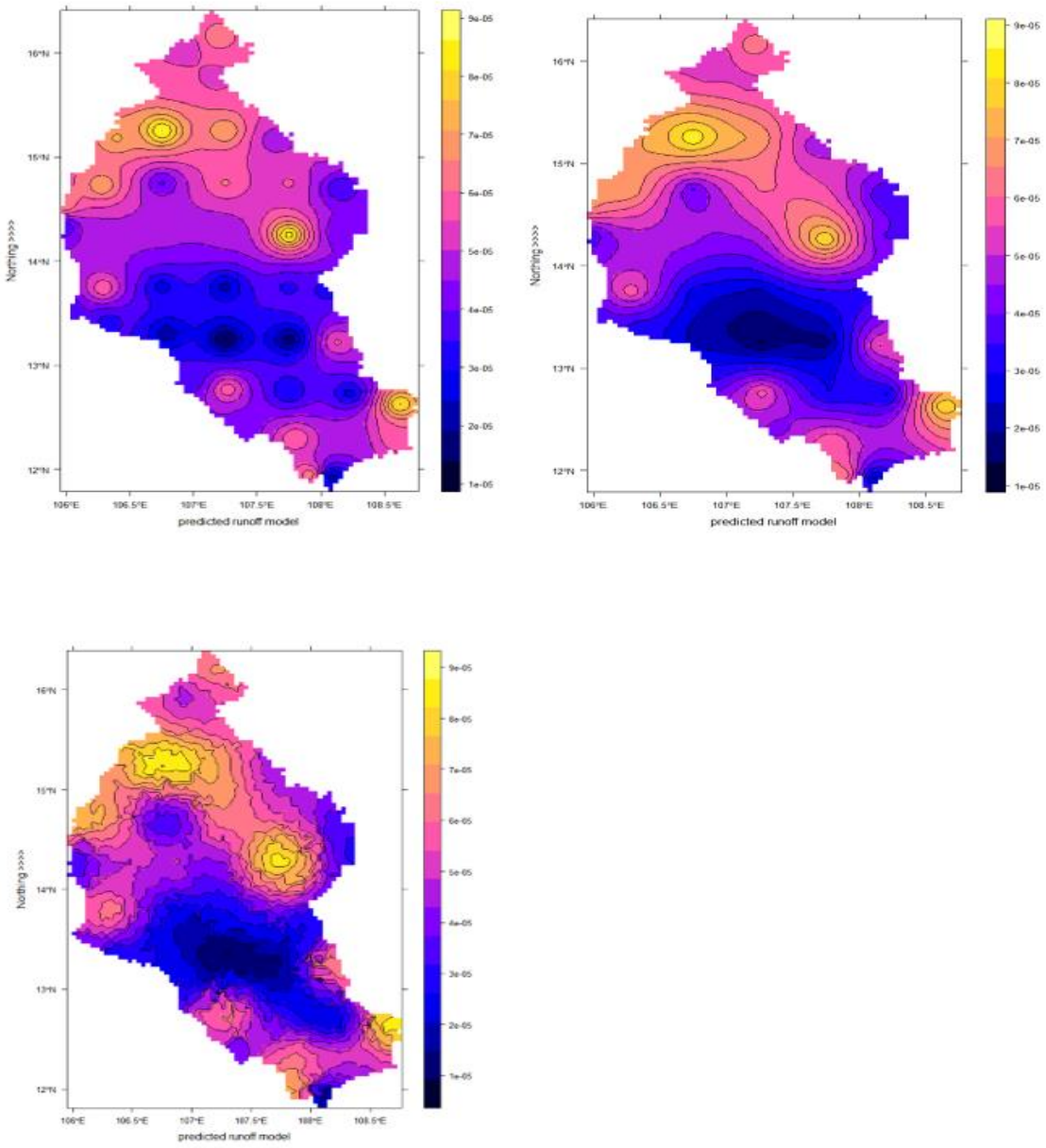

Figure 15: Interpolated raster image for IDW, OK and TK from left, right and down respectively

To understand the results, it is vital to understand the process of error measurement used for the results. Among various error methods, this project utilized six different error systems which are explained below:

a. Mean Error (ME)

   The mean error is an error measurement which is also called an average of all errors. For the single measurement, the mean error is the difference between the

actual value and the average value. For, multiple measurements it is mathematically written as follows:

$$ME = \frac{\sum(actual - avg\ value)}{N}\ , \tag{9}$$

where *N* is the number of measurements.

This error calculation method is often debated on because it is possible that the two different values, positive and negative cancel each other resulting in zero error. For example, two errors with -5 and 5 value each result mean error zero because they cancel each other. However, this method can still be utilized to gain a general view of data structure and information.

b. Root Mean Square Error (RMSE)

RMSE is the method selected by practitioners for frequent use to draw conclusions about forecasting methods, although it is unit free. It is the square root of a mean square error where the mean square value refers to the average squared difference between the estimated value and actual value. Mathematically, above can be shown as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2\ , \tag{10}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}\ , \tag{11}$$

Generalizing the above equation, it is found that RMSE is the square root of average squared error. According to the study made by J.Scott Armstrong and Fred Collopy (1992) about error measures and their comparisons among different error propagation, it has been stated that RMSE has given a low level of reliability among the error method used. This means the result from the RMSE method was more deviated from the actual results than the results of other methods.

c. PBIAS

Bias in a statistical term is the results in which the expected value of results differs from the true underlying quantitative parameter being estimated. For percentage bias, the results are as a percentage showing the tendency of simulated values to approach the observed ones. To get percentage error, the value of bias is divided by theoretical value followed by multiplication of 100.

$$Bias = Estimated\ value - observed\ value, \tag{12}$$

$$Pbias = \frac{Bias}{observed\ value} * 100\ , \tag{13}$$

In R-studio, Pbias can be calculated by Pbias function with the help of simulated values and observed values. Positive value and negative value are called as positive bias and negative bias. The higher these values are, the higher the bias is. Zero is considered to be the value where the non-bias condition prevails. The higher the difference from the optimal value 'zero', the more unacceptable the model will be.

d. Nash-Sutcliffe efficiency (NSE)

It is a coefficient used to evaluate the hydrological model's predictive power. Its value ranges from -∞ to 1. The value of 1 represents full efficiency in which modeled discharge perfectly matches observed data. 0 represents the prediction of models is approximately equal to the mean of observed data. The case of NSE value less than zero prevails if the observed mean is a better predictor than the model. The value of NSE can be derived as follows:

$$NSE = 1 - \frac{\sum_{t=1}^{T}(Q_m^t - Q_o^t)^2}{\sum_{t=1}^{T}(Q_o^t - \bar{Q}_o)^2}, \quad (14)$$

where Q_0 is the mean of observed discharge, $Q_0{}^t$ is observed discharge at time t and, Q_m is the modeled discharge.

In addition, the closer the value of NSE is to the 1, the more efficiency the model is. However, the obtained value is sensitive to the extreme values which are large outliers. If the data contains such extreme values, then results obtained through the process might be sub-optimal.

e. R-squared (R^2)

It is a statistical measure that reveals the closeness of data to the fitted regression line. The value of R-squared ranges from 0 to 1 where 0 represents the relation between the fitted line and data points is very less whereas 1 represents the strong relationship between data points and fitted line. It is also termed as a coefficient of determination. In a statistical model, its main objectives are predicting future outcomes or testing the hypothesis. For R-square to be one, the summation of deviation from fitted line value must be zero which in other cases can be understood as error must be zero.

Most common way of understanding R^2 is the following equation:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}, \quad (15)$$

$$where, SS_{res} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2, \quad (16)$$

$$SS_{tot} = \sum_i (y_i - \hat{y})^2 , \quad (17)$$

where $(y_i - \hat{y})^2$ is the proportional variance of the data and f_i is the fitted or predicted value. Moreover, the coefficient of determination gives information about the goodness of fit of the model and information about how well the regression coefficient approximate real data points.

f. Kling-Gupta Efficiency (KGE)

KGE is a goodness of fit measure which facilitates the analysis of correlation, bias, and variability in the context of hydrological modeling. This system was developed by Gupta at al. (2009) and was further revised by Kling et al. (2012) to ensure the bias and variability ratio not cross-correlated. Mathematically, the equation can be expressed as follows:

$$KGE = 1 - \sqrt{(CC - 1)^2 + (\frac{cd}{rd} - 1)^2 + (\frac{cm}{rm} - 1)^2}\ , \quad (18)$$

In the above expression, *CC* refers to Pearson coefficient value, the *cd* is a standard deviation and *rd* is a standard deviation of forecast values. Similarly, *rm* and *cm* are an average of observed and forecast values, respectively. KGE with a positive value is a good value for the goodness of fit measure, if it went to negative the fit is not considered good. Since KGE incorporates the bias and NSE factor as well, it is considered to be an important factor in determining the goodness of fit in this project.

6 Results

All our work reveals different errors value for different error methods for each interpolation method. From the standard data, we have error results which are shown in table 2 below.

Table 2: Error result between standard data set and flow observation

S.N.	Prediction name	Station Code	ME	RMSE	PBIAS %	NSE	R^2	KGE
1.	Ban Veunkhen	430106	285.25	564.23	159.6	-0.46	0.05	-0.83
2.	__	430105	-93.35	366.91	-25.1	0.35	0.48	0.27
3.	Chantangoy	430103	-777.25	1651.06	-61.2	-0.01	0.49	-0.07
4.	Trung Nghia	440601	36.29	88.54	83.5	-0.03	0.16	-0.14
5.	Kontum	440201	-24.19	61.67	-27.6	0.27	0.38	0.43
6.	Voeun Sai	440102	-179.10	537.02	-29.5	0.40	0.53	0.34
7.	Andoung Meas	440103	-37.84	303.61	-10.6	0.41	0.42	0.49
8.	Ban Khmoun	430101	-538.58	1077.07	-40.0	0.33	0.72	0.24
9.	Lumphat	450101	228.47	750.75	42.7	-0.06	0.17	0.24
10.	Ban Don	451305	107.83	215.32	52.2	0.24	0.45	0.36
11.	Duc Xuyen	450701	-15.68	79.98	-15.1	0.29	0.32	0.35

Table 2 represent different error calculated between flow discharge of each station and the standard runoff value to that station. Standard runoff value with the properties of the product used at the report, accumulated and routed to the river network associated to relative station. From the standard data (LORA.tif) file and observation, it is found that the strong correlation is 0.72 for the station Ban-Khmoun (430101) and the weak correlation coefficient is gained for station Ban-Veunkhen (430106) with the value of 0.05. KGE value associated with two relative stations is 0.24 and -0.05 respectively. Meanwhile, the highest value of KGE is for the station named Andoung Meas (440103) with a value of 0.49. Considering only KGE value, obtained three values are negative that is -0.83, -0.07, - 0.14 which are for the stations Ban Veunkhen, Chan tangoy, and Trung Nghia respectively. Considering the obtained fact from bias percentage, the highest positive bias percentage is 159.6 for station Ban Veunkhen and the highest negative bias percentage is -61.2 for the station Chantangoy. However, acceptable bias percentage is for station Andoung Meas with a value of -10.6 because of less deviation from optimal value zero.

A similar procedure during the process has been applied to obtained error results for output but instead of Standard raster file, different raster files with high resolution have been constructed with three different methods (IDW, OK, and TK) for the assessment of

error value obtained for each method. Table 3 is the error calculation for the Inverse distance weighting method.

Table 3: Error results between IDW interpolation and observation flow measurement

S.N.	Prediction name	Station Code	ME	RMSE	PBIAS %	NSE	R^2	KGE
1.	Ban Veunkhen	430106	256.84	558.85	143.7	-0.43	0.04	-0.69
2.	__	430105	-112.55	374.70	-30.2	0.32	0.49	0.24
3.	Chantangoy	430103	-813.76	1634.89	-64.1	0.01	0.60	-0.06
4.	Trung Nghia	440601	41.11	94.35	94.6	-0.17	0.15	-0.17
5.	Kontum	440201	-16.14	51.37	-18.4	0.49	0.54	0.56
6.	Voeun Sai	440102	-195.32	587.37	-32.2	0.28	0.41	0.25
7.	Andoung Meas	440103	-50.50	317.66	-14.2	0.36	0.38	0.39
8.	Ban Khmoun	430101	-606.97	1102.47	-45.1	0.30	0.75	0.22
9.	Lumphat	450101	228.46	600.60	42.7	0.32	0.42	0.33
10.	Ban Don	451305	93.41	231.59	45.2	0.12	0.29	0.29
11.	Duc Xuyen	450701	-17.52	69.21	-16.8	0.47	0.54	0.44

In Table 3, the prediction name reveals the method used to create an interpolated raster data file. In this table, each error value can be experienced differently from the standard value table. Viewing PBIAS, NSE, $R^{2,}$ and KGE at once, high error value can be obtained from station Ban Veunkhen with the value 143.7, -0.43, 0.04, -0.69 respectively. This suggests that runoff model for this station is not much acceptable. Similarly, station Kontum has the lowest error for NSE and KGE with values 0.49 and 0.56 respectively. However, the highest correlation coefficient is 0.75 for station Ban-Khmoun and the least percentage bias value experienced for station Andoung Meas as -14.2. Additionally, the highest error value for the RMSE and ME is obtained for station Chantangoy with a value of 1634.89 and -813.76 respectively.

After IDW, the next step was a similar analysis through the ordinary kriging method which is shown in table 4.

Table 4: Error results between OK stations and flow observation measurement

S.N.	Prediction name	Station Code	ME	RMSE	PBIAS %	NSE	R^2	KGE
1.	Ban Veunkhen	430106	264.42	564.99	147.9	-0.46	0.04	-0.72
2.	__	430105	-106.96	371.29	-28.7	0.33	0.49	0.25
3.	Chantangoy	430103	-806.35	1625.57	-63.5	0.02	0.61	-0.05
4.	Trung Nghia	440601	40.44	95.08	93.1	-0.18	0.15	-0.15
5.	Kontum	440201	-18.38	52.60	-21.0	0.47	0.53	0.56
6.	Voeun Sai	440102	-201.85	595.89	-33.3	0.26	0.38	0.24
7.	Andoung Meas	440103	-55.33	317.93	-15.5	0.35	0.38	0.39
8.	Ban Khmoun	430101	-606.95	1096.52	-45.1	0.30	0.75	0.23
9.	Lumphat	450101	235.24	616.93	44.0	0.28	0.39	0.31
10.	Ban Don	451305	96.14	232.74	46.5	0.11	0.29	0.28
11.	Duc Xuyen	450701	-16.48	70.41	-15.8	0.45	0.52	0.42

With table number 4, different error propagation between standard runoff value and observed value for the stream method can be generalized ordinary kriging interpolation. From the table, different error values for the different station can be observed. Among the highest value of KGE and NSE method is 0.56 and 0.53 respectively for the station Kontum. Under this method, the correlation coefficient value is still highest for the station Ban Khmoun which is 0.75.

Topological kriging interpolation is performed with given standard runoff values and routed to the stream through as usual procedure performed for other two different methods and matched against the observed streamflow which results in the following table 5.

Table 5: Error results between TK interpolation and flow observation measurement

S.N.	Prediction name	Station Code	ME	RMSE	PBIAS %	NSE	R^2	KGE
1.	Ban Veunkhen	430106	265.43	566.18	148.5	-0.47	0.04	-0.73
2.	__	430105	-107.27	371.60	-28.8	0.33	0.49	0.25
3.	Chantangoy	430103	-804.13	1625.13	-63.4	0.02	0.60	-0.05
4.	Trung Nghia	440601	39.26	95.74	90.4	-0.20	0.14	-0.13
5.	Kontum	440201	-19.12	53.59	-21.8	0.45	0.52	0.54
6.	Voeun Sai	440102	-201.53	594.77	-33.2	0.27	0.38	0.25
7.	Andoung Meas	440103	-57.84	313.44	-16.2	0.37	0.40	0.40
8.	Ban Khmoun	430101	-599.13	1089.71	-44.5	0.31	0.74	0.23
9.	Lumphat	450101	243.94	631.72	45.6	0.25	0.36	0.28
10.	Ban Don	451305	95.90	233.01	46.4	0.11	0.28	0.27
11.	Duc Xuyen	450701	-15.88	70.51	-15.3	0.45	0.51	0.43

Table 5 is the topological kriging error value table among which for station code 440201 (Kontum), KGE value is 0.54 which means this has the best fit among the other stations. Similarly, the worst fit among the station's flowrate with the predicted runoff value is station 430106 (Ban Veunkhen) with KGE value negative 0.73. Also, according to the correlation coefficient, the highest correlation among runoff predicted and flowrate is with the station 430101 (Ban Khmoun) and the lowest correlation coefficient is the same station as of lowest KGE value.

7 Comparison

This section includes the difference between standard low-resolution data results and the results obtained through the interpolation technique that had been used for the report. Comparison is based on the amount of error results deviated from the standard error result to the output error value. The comparison might lead to an unacceptable negative value, for example, RMSE could be negative. This happens in the case where the standard error value is less than that of an interpolated error value. In order to interpret prudently and to avoid the above process, the system is directed by modulus sign. To understand the process, following mathematical notation will be helpful.

$$Comparison = |\, standard\ error\ value - interpolated\ error\ value \,| \quad (19)$$

Table 6 reveals the difference in error value between standard error value and IDW interpolated error value. The comparison results suggest that highest difference in KGE value is 0.14 for station 430106 and the lowest difference in KGE value is 0.01. Likewise, in correlation coefficient, station 450101 has the highest value with 0.25 and the lowest difference is 0.01 for station 440601 and 430106. The highest bias percentage difference is 15.9. The highest difference for ME and RMSE is 68.39 and 150.15 for station 430101 and 450101 respectively.

Table 6: Comparison table between Standard and IDW

S.N.	Prediction name	Station Code	ME	RMSE	PBIAS %	NSE	R^2	KGE
1.	Ban Veunkhen	430106	28.41	5.38	15.9	0.03	0.01	0.14
2.	__	430105	19.20	7.79	5.1	0.03	0.01	0.03
3.	Chantangoy	430103	36.51	16.17	2.9	0.02	0.11	0.01
4.	Trung Nghia	440601	4.82	5.81	11.1	0.14	0.01	0.03
5.	Kontum	440201	8.05	10.30	9.2	0.22	0.16	0.13
6.	Voeun Sai	440102	16.22	50.35	2.7	0.12	0.12	0.09
7.	Andoung Meas	440103	12.66	14.05	3.6	0.05	0.04	0.10
8.	Ban Khmoun	430101	68.39	25.40	5.1	0.03	0.03	0.02
9.	Lumphat	450101	0.01	150.15	0.0	0.38	0.25	0.09
10.	Ban Don	451305	14.42	16.27	7.0	0.12	0.16	0.07
11.	Duc Xuyen	450701	1.84	10.77	1.7	0.18	0.22	0.09

A comparison between standard error value and ordinary kriging interpolated error value can be seen from table 7. The highest difference in KGE value from standard was found to be 0.13 in the stations 440201 with the closest being 0.11 in the station 430106 whereas the lowest KGE, 0.01, was observed in station 440601. Moreover, the highest value of R^2, 0.22, was observed in station 450101 which was closely followed by station 450701 with the value of 0.20. Similarly, on the other extreme, stations 430101, 430105 and 440601 has value 0.0. followed by station 430101 with 0.03.

Table 7: Comparison table between Standard and OK

S.N.	Prediction name	Station Code	ME	RMSE	PBIAS %	NSE	R^2	KGE
1.	Ban Veunkhen	430106	20.83	0.76	11.7	0.00	0.01	0.11
2.	__	430105	13.61	4.38	3.6	0.02	0.01	0.02
3.	Chantangoy	430103	29.10	25.49	2.3	0.03	0.12	0.02
4.	Trung Nghia	440601	4.15	6.54	9.6	0.15	0.01	0.01
5.	Kontum	440201	5.81	9.07	6.6	0.20	0.15	0.13
6.	Voeun Sai	440102	22.75	58.87	3.8	0.14	0.15	0.10
7.	Andoung Meas	440103	17.49	14.32	4.9	0.06	0.04	0.10
8.	Ban Khmoun	430101	68.37	19.45	5.1	0.03	0.03	0.01
9.	Lumphat	450101	6.77	133.82	1.3	0.34	0.22	0.07
10.	Ban Don	451305	11.69	17.42	5.7	0.13	0.16	0.08
11.	Duc Xuyen	450701	0.80	9.57	0.7	0.16	0.20	0.07

On the edge, TK comparison to the standard has been done with the same procedure, reveals the following difference from the standard error values. The degree of deviation from the standard's data error is shown in table 8 below:

Table 8: Comparison between Standard and TK

S.N.	Prediction name	Station Code	ME	RMSE	PBIAS %	NSE	R^2	KGE
1.	Ban Veunkhen	430106	19.82	1.95	11.1	0.01	0.01	0.10
2.	__	430105	13.92	4.69	3.7	0.02	0.01	0.02
3.	Chantangoy	430103	26.88	25.93	2.2	0.03	0.11	0.02
4.	Trung Nghia	440601	2.97	7.20	6.9	0.17	0.02	0.01
5.	Kontum	440201	5.07	8.08	5.8	0.18	0.14	0.11
6.	Voeun Sai	440102	22.43	57.75	3.7	0.13	0.15	0.09
7.	Andoung Meas	440103	20.00	9.83	5.6	0.04	0.02	0.09
8.	Ban Khmoun	430101	60.55	12.64	4.5	0.02	0.02	0.01
9.	Lumphat	450101	15.47	119.03	2.9	0.31	0.19	0.04
10.	Ban Don	451305	11.93	17.69	5.8	0.13	0.17	0.09
11.	Duc Xuyen	450701	0.20	9.47	0.2	0.16	0.19	0.08

Table 8 shows the level of deviation from standard data error results to the TK error results. TK shows less deviation among the other two methods above. At station 440201 highest KGE difference was found to be 0.11 followed by station 440102, 440103 and 451305 with a KGE difference of 0.09 each. On the contrary, the least KGE difference was found to be 0.01 at stations 440601 and 430101. Similarly, the highest difference for R^2 was 0.19 at station 450101 and 450701. Meanwhile, 0.01 was found to be the least value difference for R^2 for two stations, 430106 and 430105. Largest percentage bias value difference was found to be 11.1 for station 430106 and the least value for Pbias was 0.2 for station 450701. Likewise, the highest value difference for ME and RMSE is 60.55 and 119.03 for station 430101 and 450101 respectively. At the other end of extreme value for ME and RMSE was 0.20 and 1.95 for the stations 450701 and 430106 respectively.

ANOVA and Least square deviation for KGE Value

Further analysis was performed for the KGE value obtained from all methods. Firstly, two-factor analysis of variance was done without replication considering modeled value as one factor and station value as other factors (Appendix). The obtained P-value among

the stations was $4.31*10^{-21}$ and among the methods was 0.941. This shows that there is no relation to the KGE value obtained among the stations and high similarity in the KGE value among the method used. Also, the summation of the least square deviation was done to ascertain the lowest value for the method (Appendix). The analysis shows that the lowest summation value was for TK which is 0.0554 followed by OK 0.0662 and IDW 0.078.

8 Conclusion

This chapter provides an overview of how prediction models vary along its error propagation to ascertain hydrological movement (runoff and streamflow). With the analysis of the results, it was found that high-resolution data has less error than that of low-resolution data. This means the observational value is more accurate in high resolution than in low-resolution format. It is surprising that despite the limitation of inverse distance weighting (IDW), error calculation for the IDW method shows less value for error results over the other two methods. Also, for the ordinary kriging (OK) method, the value obtained resembles some extent to the topological kriging (TK) method. On further analysis of error results in the prediction models, the error values do not differ significantly from each other. Similarly, the output of nonuniform error values for separate stations in each method is due to the difference in predicted runoff value due to different method for an individual stream.

Moreover, parameters for all three different methods were taken into consideration, where TK has the area as its important parameter for the interpolation method considering downstream and upstream zone. The other two methods use points system irrespective of the downstream and upstream zone. Additionally, from comparison tables, it has been revealed that TK was able to perform low-resolution data to high resolution with less deviation of errors from standard data than other methods. Finally, the analysis is done on the Kling Gupta Error (KGE) value for each method relative to the standard value. Because it is most relevant among other methods used for determining the goodness of fit as it comprises Bias and NSE within itself. This suggests that the least squared deviation from the standard error value is for the topological kriging method over inverse distance weighing and ordinary kriging method.

Despite the above facts and results, there is another type of interpolation technique as well. Also, the prediction model developed for this report is an areal interpolation. Similarly, actual geographical shape plays a major role in hydrology which in this report is not considered. Perhaps, considering all variables might show the fact differently, but this requires more deeper investigation and hard work.

References

Salas, Jose D, Rao S. Govindaraju, Anderson Michael, Arabi Mazdak, Frances Felix, Suarez Wilson, Waldo S. Lavado-Casimiro, Green, Timothy R., 2014. *Introduction to Hydrology*. New York: Springer Science+Business Media.

Wu Chia-Yu, Mossa Joann, Mao Liang, Almulla Mohammad, 2019. *Comparison of different spatial interpolation methods for historical hydrographic data of the lowermost Mississippi River*. Online: Research gate.

MacQuarrie Patrick R., Welling Rebecca, Rammont Lalita, Pangare Ganesh, 2013. *The 3S River Basin (Cambodia, Lao PDR, and Vietnam)*. Gland, Switzerland: IUCN.

ICEM Environmental Management, 2016. *3S River Basins Study set to begin.* Hanoi, Vietnam: ICEM.

Hobeichi Sanaa, Abramowitz Gab, Evans Jason, Beck Hylke E., 2019. *Linear Optimal Runoff Aggregate (LORA): a global gridded synthesis runoff product*. EU: Copernicus Publications.

Lehner B., Grill G., 2013. *Global river hydrography and network routing: baseline data and new approaches to study the world's large river system*, Hydrological Process. Online: HydroSHEDS.

Kallio M., 2018. *R package for downscaling off-the-shelf runoff products to explicit river network*. [online] Available at: https://github.com/mkkallio/hydrostreamer [Accessed 11 December 2019].

Kallio M., Guillaume J.H.A., Kummu M., Desalegn F., Virrantaus K., 2018. *Spatial allocation of low resolution runoff model outputs to a high resolution stream network*. [online] Available at: https://www.researchgate.net/publication/325066501_Spatial_allocation_of_low_resolution_runoff_model_outputs_to_high_resolution_stream_network?channel=doi&linkId=5af46d33a6fdcc0c030aed34&showFulltext=true [Accessed 11 December 2019]

Wahab Muhammad Abdul, 2017. *Interpolation and Extrapolation,* Topics in system Engineering. [online] Available at: https://www.researchgate.net/publication/313359295_Interpolation_and_Extrapolation [Accessed 18 December 2019]

Achilleous G.A., 2011. *The Inverse Distance Weighted interpolation method and error propagation mechanism – creating a DEM from an analogue topographical map,* Journal of Spatial Science, 56:2, 283-304. Greece: Taylor & Francis Online.

Burrough P.A., 1986. *Principles of Geographical Information Systems for Land Resources Assessment*. New York: Oxford University Press.

Armstrong Margaret, 1998. *Basic Linear Geostatistics.* New York: Springer-Verlag Berlin Heidelberg.

Goovaerts Pierre, 1997. *Geostatistics for Natural Resource Evaluation*. New York: Oxford University Press.

Skoien J.O., Merz R., Blöschl G., 2005. *Top-Kriging – geostatistics on stream networks*. Vienna: Copernicus GmbH.

Journel A.G., Huijbregts Ch.J., 1978. *Mining Geostatistics*. London and New York: Cambridge University Press.

Clark Isobel, Harper William V., 2001. *Practical Geostatistics 2000*. Ohio USA: Ecosse North America.

Sauquet Eric., Gottsghalk Lars, Leblois Etienne, 2000. *Mapping average annual runoff: a hierarchical approach applying a stochastic interpolation scheme,* Hydrological Sciences Journal, 45:6, 799-815. [online] Available at: https://www.tandfonline.com/doi/abs/10.1080/02626660009492385 [Accessed 18 December 2019]

Martin Bachmaier, 2008. *Variogram or semivariogram? Understanding the variances in a variogram,* Precision Agric. Germany: Springer Science+Business Media.

Chai T., Draxler R. R., 2014. *Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature*. EU: Copernicus Publications.

Armstrong J.S., Collopy F., 1992. *Error Measures for Generalizing About Forecasting Methods: Empirical Comparisons*. Pennsylvania: International Journal of Forecasting.

Moriasi D. N., Arnold J. G., Van Liew M. W., Bingner R. L., Harmel R. D., Veith T. L., 2007. *Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations*. Texas: American society of agricultural and biological engineers.

Pool Sandra, Vis Marc, Seibert Jan, 2018. *Evaluating model performance: towards a non parametric variant of the Kling-Gupta efficiency*. Hydrological Sciences Journal. [online] Available at: https://www.tandfonline.com/doi/full/10.1080/02626667.2018.1552002?src=recsys [Accessed 18 December 2019].

Appendices

Appendix 1. Codes written during the project

```
lora <- brick("lora combined.tif")
aoi <- read_sf("3Sbasin.gpkg")
aoi_sp <- as(aoi, "Spatial")
basin <- readOGR("HS_basins.gpkg")
HS <- raster_to_HSgrid(lora,
                       date = lubridate::dmy('1-1-1980') ,
                       "month",
                       aoi= aoi,
                       name = "LORA")

centroids <- st_centroid(HS)
class(centroids) <- c("sf", "data.frame")

grid <- as.data.frame(spsample(aoi_sp, "regular", n=4300))
names(grid) <- c("X", "Y")
coordinates(grid) <- c("X", "Y")
gridded(grid) <- TRUE
fullgrid(grid) <- TRUE
proj4string(grid) <- proj4string(aoi_sp)

n_tstep <- HS$runoff_ts[[1]] %>% nrow
r_idw <- list()
r_krige <- list()
# n_tstep <- 5
pb <- txtProgressBar(min = 0, max = n_tstep, style = 3)
for(tstep in 1:n_tstep) {
  centroids$runoff_ts <- unlist(lapply(HS$runoff_ts, function(x) x[tstep,2]))
  cent_sp <- as(centroids, "Spatial")

  runoff_idw <- idw(runoff_ts ~ 1, cent_sp, newdata=grid, idp=2)
  runoff_idw <- raster(runoff_idw)
  runoff_idw <- mask(runoff_idw, aoi_sp)


  runoff_krige <- variogram(runoff_ts ~1, locations = cent_sp)
  runoff_krige <- fit.variogram(runoff_krige, vgm(1,"Exp",300,0))
  plot(variogramLine(runoff_krige,250), ylim=c(0,0.0000000001), col='blue', lwd = 1)
  runoff_krige <- gstat(formula=runoff_ts ~1, location=cent_sp, model = runoff_krige)
  runoff_krige <- predict(runoff_krige,grid)
  runoff_krige<- raster(runoff_krige)
  runoff_krige <- mask(runoff_krige,aoi_sp)


  r_idw[[tstep]] <- runoff_idw
  r_krige[[tstep]] <- runoff_krige
  setTxtProgressBar(pb, tstep)
}
close(pb)
r_idw <- do.call("brick", r_idw)
r_krige <- do.call("brick", r_krige)
```

Figure 1.1: work code for IDW and OK

```
lora <- brick("lora combined.tif")
aoi <- read_sf("3Sbasin.gpkg")
aoi_sp <- as(aoi, "Spatial")
basin <- read_sf("HS_basins.gpkg")
basin_sp <- as(basin, "Spatial")
HS <- raster_to_HSgrid(lora,
                       date = lubridate::dmy('1-1-1980') ,
                       "month",
                       aoi= aoi,
                       name = "LORA")

centroids <- st_centroid(HS)
class(centroids) <- c("sf", "data.frame")

grid <- as.data.frame(spsample(aoi_sp, "regular", n=4300))
names(grid) <- c("X", "Y")
coordinates(grid) <- c("X", "Y")
gridded(grid) <- TRUE
fullgrid(grid) <- TRUE
proj4string(grid) <- proj4string(aoi_sp)

n_tstep <- HS$runoff_ts[[1]] %>% nrow

r_topkrige <- list()
# n_tstep <- 5
pb <- txtProgressBar(min = 0, max = n_tstep, style = 3)

HS_sp <- HS
class(HS_sp) <- c("sf", "data.frame")
r <- raster(aoi_sp)
res(r) <- 0.03913862
for(tstep in 1:n_tstep) {
  HS_sp$runoff_ts <- unlist(lapply(HS$runoff_ts, function(x) x[tstep,2]))
  HS_topkrige <- as(HS_sp, "Spatial")
  params <- getRtopParams()
  params$rresol <- 25
  params$maxdist <- 3
  params$gdist <- TRUE
  runoff_tk <- createRtopObject(HS_topkrige, basin_sp, runoff_ts~1, params= params)
  runoff_tk <- rtopVariogram(runoff_tk)
  runoff_tk <- rtopFitVariogram(runoff_tk)
  runoff_tk <- rtopKrige(runoff_tk)
  polys <- runoff_tk$predictions[,3]
  class(polys)
  r.polys <- rasterize(polys, r, field=polys@data)
  r_topkrige [[tstep]] <- r.polys
  setTxtProgressBar(pb, tstep)
}
close(pb)
r_topkrige <- do.call("brick", r_topkrige)
```

Figure 1.2: work code for TK

```
td <- readr::read_csv('flowdata_3s.csv')
td[1] <- as.Date(td$DATE, format = '%d/%m/%Y')
td[2:12] <- lapply(td[2:12], as.numeric)
monthly_td <- td %>%
  mutate(year = year(DATE),
         month = month(DATE)) %>%
  group_by(year,month) %>%
  summarise_all(mean, na.rm=TRUE) %>%
  mutate(date = ymd(paste(year,month,"01", sep="-"))) %>%
  ungroup %>%
  select(date, everything(),-year, -month)
river <- read_sf("Hydrosheds3Srivers.gpkg")
station <- read_sf("HYMOS stations.gpkg")
basins <- read_sf("HS_basins.gpkg") %>%
  rename(ARCID = X3S_drdir)
aoi <- read_sf("3Sbasin.gpkg")
colnames(monthly_td)[1] <- "Date"


HS1 <- raster_to_HSgrid(lora,
                        date = lubridate::dmy('1-1-1980'),
                        timestep = "month",
                        aoi= aoi,
                        name = "LORA",
                        verbose=TRUE)
Hsweights.monthly <- compute_HSweights(river,
                                       HSgrid = HS1,
                                       weights="area",
                                       aoi=aoi,
                                       basins = basins,
                                       riverID = "ARCID")
HSrunoff.monthly <- downscale_runoff(Hsweights.monthly, verbose=TRUE)
HS_runoffmet <- accumulate_runoff(HSrunoff.monthly, verbose=TRUE)

station_order <- colnames(monthly_td)[-c(1:2)] %>% match(station$Code)
riverIDs <- station$ARCID[station_order]
names(riverIDs) <- as.character(station$Code)[station_order]

HS_runoffmet <- add_observations(HS_runoffmet, monthly_td[,-2],riverIDs)

gof <- flow_gof(HS_runoffmet) %>%
  select(Prediction, Station, ME, RMSE, `PBIAS %`, NSE, R2, KGE)
gof
```

Figure 1.3: Work code for comparison through hydrostreamer

Appendix 2. Interpolation results

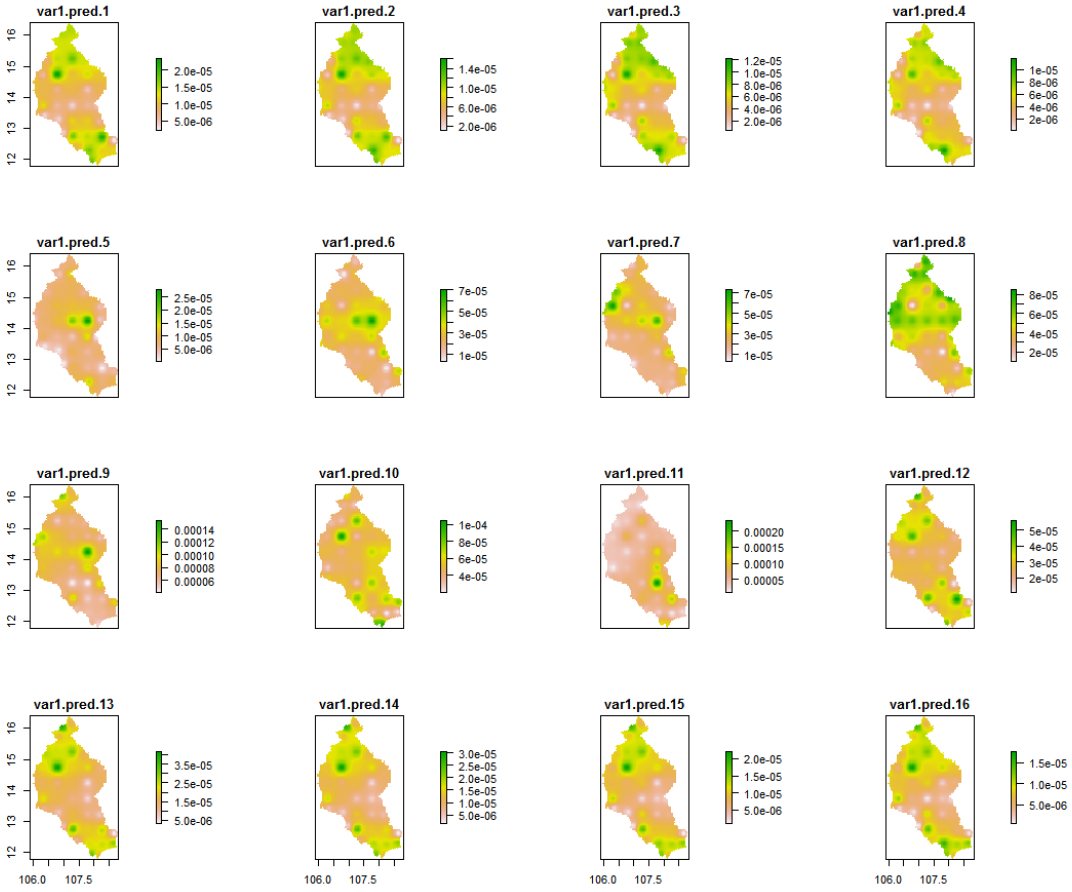

Figure 2.1: IDW interpolation for different layer

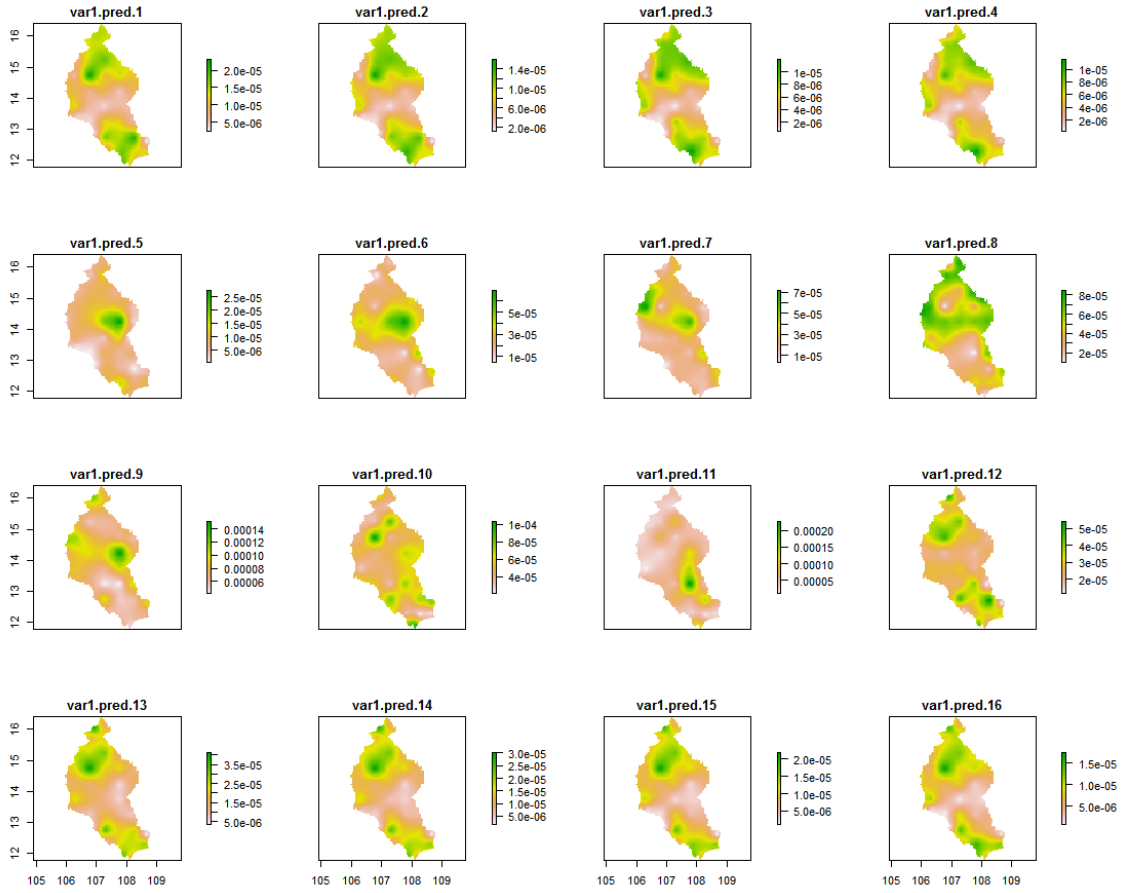

Figure 2.2: OK interpolation for different layer

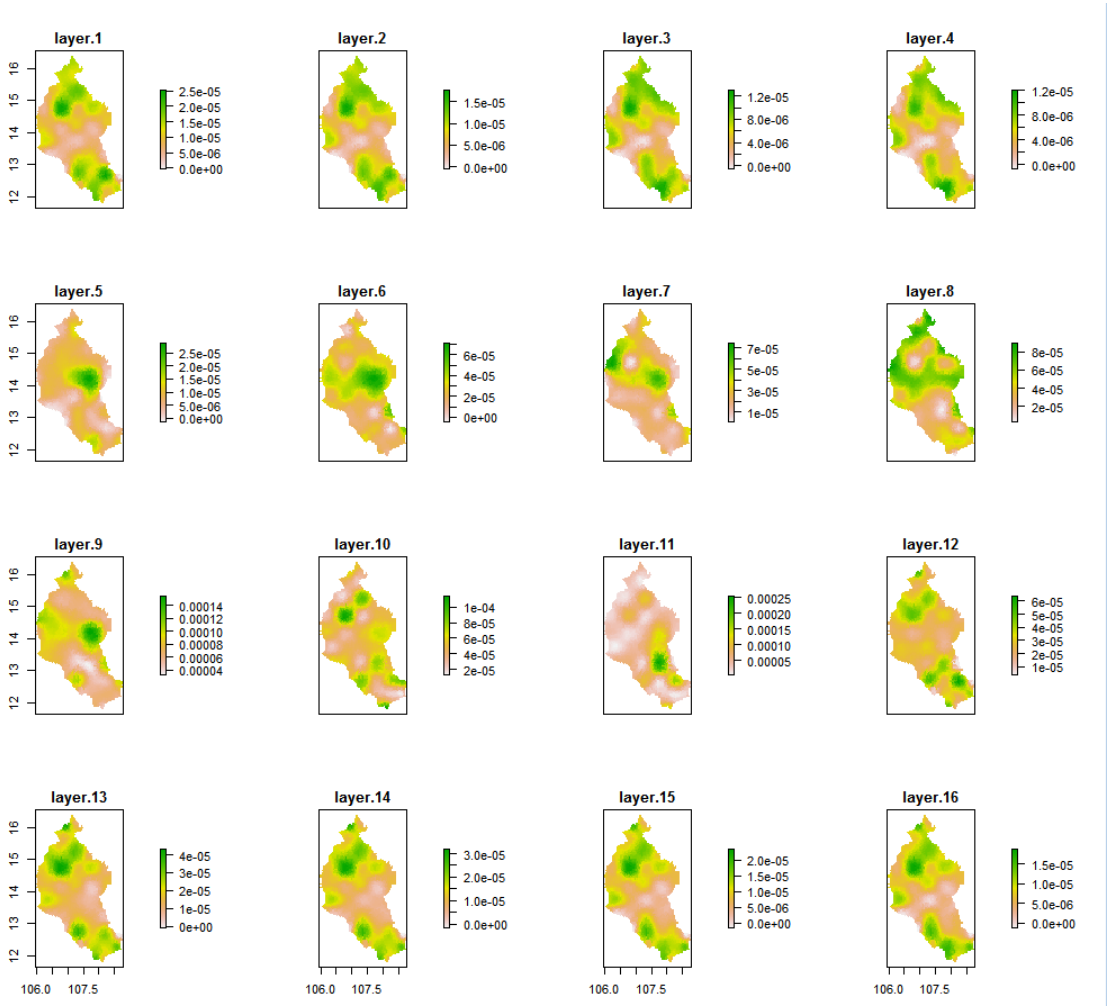

Figure 2.3: TK interpolation layer for different layer

Appendix 3. Least square deviation calculation

Station	TK	Standard	OK	IDW		(Standard - TK)^2		(Standard - OK)^2	(Standard - IDW)^2
S1	-0.73	-0.83	-0.72	-0.69		0.01	0	0.0121	0.0196
S2	0.25	0.27	0.25	0.24		0.0004	0	0.0004	0.0009
S3	-0.05	-0.07	-0.05	-0.06		0.0004	0	0.0004	0.0001
S4	-0.13	-0.14	-0.15	-0.17		0.0001	0	1E-04	0.0009
S5	0.54	0.43	0.56	0.56		0.0121	0	0.0169	0.0169
S6	0.25	0.34	0.24	0.25		0.0081	0	0.01	0.0081
S7	0.4	0.49	0.39	0.39		0.0081	0	0.01	0.01
S8	0.23	0.24	0.23	0.22		1E-04	0	1E-04	0.0004
S9	0.28	0.24	0.31	0.33		0.0016	0	0.0049	0.0081
S10	0.27	0.36	0.28	0.29		0.0081	0	0.0064	0.0049
S11	0.43	0.35	0.42	0.44		0.0064	0	0.0049	0.0081
					summation	**0.0554**	**0**	**0.0662**	**0.078**

Figure 3: showing summation squared difference of KGE value for each method in the excel sheet.

Appendix 4. Analysis of variance

Anova: Two-Factor Without Replication

SUMMARY	Count	Sum	Average	Variance
S1	4	-2.97	-0.7425	0.003692
S2	4	1.01	0.2525	0.000158
S3	4	-0.23	-0.0575	9.17E-05
S4	4	-0.59	-0.1475	0.000292
S5	4	2.09	0.5225	0.003892
S6	4	1.08	0.27	0.0022
S7	4	1.67	0.4175	0.002358
S8	4	0.92	0.23	6.67E-05
S9	4	1.16	0.29	0.001533
S10	4	1.2	0.3	0.001667
S11	4	1.64	0.41	0.001667
TK	11	1.74	0.158182	0.124276
Standard	11	1.68	0.152727	0.143762
OK	11	1.76	0.16	0.12514
IDW	11	1.8	0.163636	0.123485

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Rows	5.114468	10	0.511447	294.1142	4.31E-27	2.164579917
Columns	0.000682	3	0.000227	0.130696	0.941079	2.922277191
Error	0.052168	30	0.001739			
Total	5.167318	43				

Figure 4: ANOVA analysis performed for the Method used and Station to determine their relationship.