Online Short Course
MyStats 2022

Managing Uncertainties in Data Analytics

Prof Dr Fadhilah Yusof

4th October 2022

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

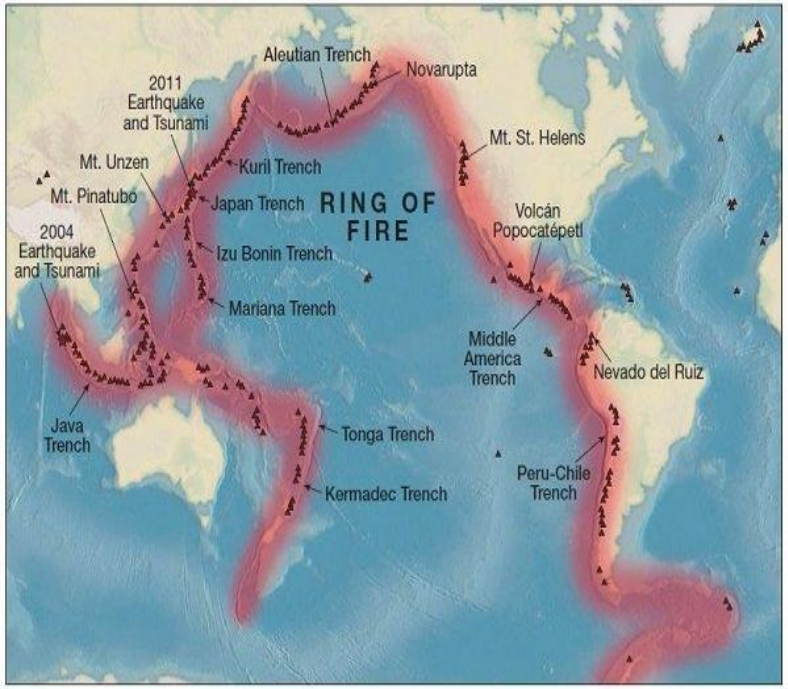
innovative • entrepreneurial • global

What is uncertainty?

You want to organize a party in your housing area in three months time. It will be a garden concept party in an open area. Your main concern is whether it will be raining or not during the event. There will be few canopies to provide some shades from morning sun but not from heavy rainfall. How do you make a decision on this matter?

This is an uncertainty situation. What will a researcher do to minimize this uncertainty?

What is uncertainty?



People living in the “ring of fire” region all their life which are prone to earthquakes and volcanoes eruption can still lead a normal life. Even though their livelihood are very uncertain, their countries’ economics and social well being are still moving forward in a positive way.

How they manage the uncertainty?

Uncertainty

- is a situation that involves imperfect or unknown information
- is essentially lack of information to formulate a decision
- Is when information is incomplete, inconsistent and uncertain. Hence, this information is unsuitable for solving a problem
- Is when there is lack of exact knowledge that enable us to reach to a perfectly reliable conclusion
- Requires reasoning along with possessing a lot of common sense
- Lack of sureness
- is defined as doubt
- Scientific uncertainty generally means that **there is a range of possible values within which the true value of the measurement lies**. Further research on a topic or theory may reduce the level of uncertainty or the range of possible values.

Uncertainty Versus Risk

- **Risk** refers to decision-making situations under which all potential outcomes and their likelihood of occurrences are known to the decision-maker,
- **Uncertainty** refers to situations under which either the outcomes and/or their probabilities of occurrences are unknown to the decision-maker.
- **Risk** can be controlled if proper measures are taken to control it. On the other hand, **uncertainty** is beyond the control of the person or enterprise, as the future is uncertain.

Categories of Uncertainty

- **EPISTEMIC UNCERTAINTY**
  - things we don't know because of lack of data or experience
  - things we could in principle know but don't in practice
  - insufficient measurement or modeling, missing data
  - a lack of knowledge about the system or phenomenon of interest.

- **REDUCIBLE**: can be alleviated by better models, more accurate measurement
- **Example**:
  - unknown equipment failure such as rain gauge blockage
  - unknown changes in the cross sectional velocity distribution of the channel.
- Hard to quantify

- **ALEATORIC UNCERTAINTY**
  - things that are simply unknown, like what number a die will show on the next roll
  - unknowns that differ on each run
  - arises from the inherent randomness of natural phenomena.
  - it is controlled by precision and accuracy of the data.
  - Aleatory uncertainty can be quantified in the form of probability distributions.

- **Irreducible**: cannot be eliminated through improvements in models or measurements
- Example:
  - Single measurement for rainfall or streamflow velocity due to limited precision of the sensor.
    - These uncertainties could be represented as normal distribution using a standard deviation estimate from the equipment manufacturer.

Sources of Uncertainty

- MEASUREMENT/EXPERIMENTAL UNCERTAINTY
  - input and output variables cannot be determined with absolute precision and accuracy. All measurements are prone to some imprecision.
  - Caused by imperfect instrument and sample disturbances during observation. Can be reduced by more information
  - experimental measurements have variability
- SAMPLING UNCERTAINTY
  - introduced when analyzing a random sample from a large population of interest. This random sample may capture effects that are spatially/temporally transient and overemphasize or miss effects. This variance is generally subsumed in an error term.
- Structural uncertainty
  - This can be regarded as a model's discrepancy or bias due to the fact that the model lacks exact knowledge about the underlying physics. It depends on the model's ability of representing real world process(es)
  - happens because any model is an approximation, or a best guess at what a true distribution might look like

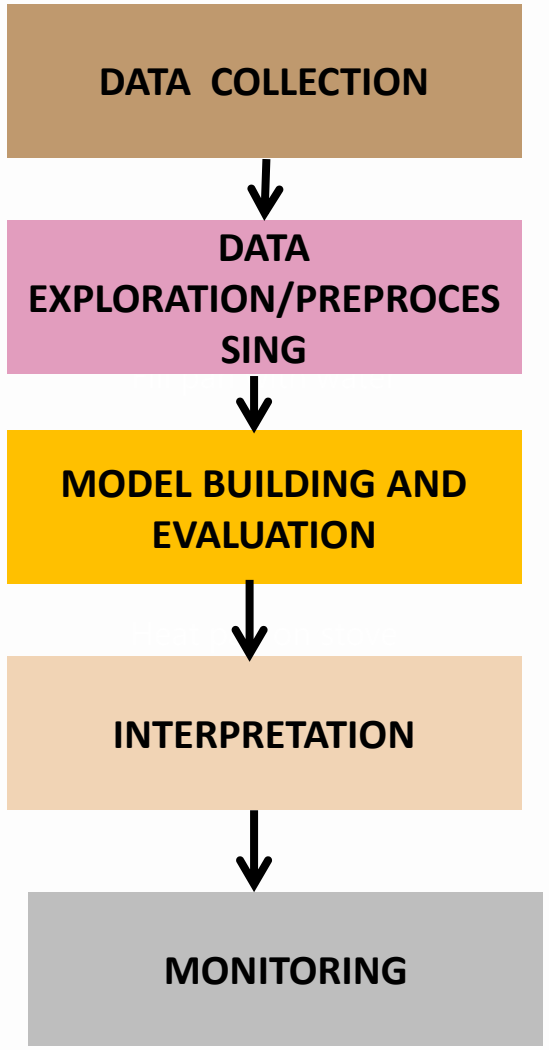
- INTERPOLATION UNCERTAINTY
    - lack of available data
    - interpolate/extrapolate for desired response
    - choice of interpolation method

- DATA PRE-PROCESSING UNCERTAINTY
    - uncertainty introduced by the decisions made in the selection of data as well as the definition, cleaning, and transformation of the input and output variables.
- ALGORITHMIC UNCERTAINTY
    - aka numerical uncertainty
    - numerical errors, approximations
    - translation of mathematical model to the computer
- STATISTICAL UNCERTAINTY
    - due to limited information such as limited number of observations. Can be reduced by obtaining more information.
- METHOD UNCERTAINTY
    - arises due to the choice of the implementation and computational method used to estimate parameters and/or generate predictions

What is data analytics?

- Data analytics is the science of analyzing raw data to make conclusions about that information.
- Data analytics is a broad term that encompasses many diverse types of data analysis. Any type of information can be subjected to data analytics techniques to get insight that can be used to improve things.
- Data analysis is a sub-component of data analytics so the data analysis life cycle also comes into the analytics part, it consists of data gathering, data cleaning, analysis of data, and interpreting the data precisely so that you can understand what your data want to say.

General Data Analysis Framework



Process of collecting or acquiring data

Fix discrepancies, visualize and transform data into required form

Construct models and evaluate performance

Develop reasonable explanations for the obtained analysis

Based on the learned knowledge and domain expertise

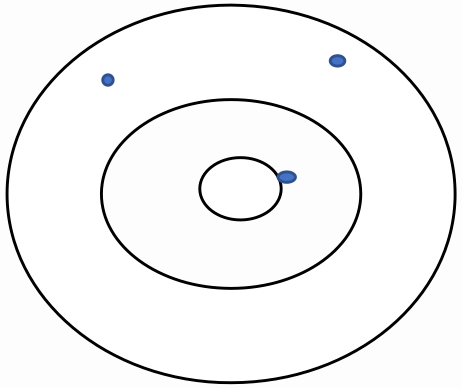
1. DATA COLLECTION

- Process of collecting or acquiring data and storing in a readily accessible form
- Requires thorough planning
  - Collection methodology – primary, secondary or tertiary (based on primary and secondary sources)
  - Selecting variable of interest
  - Determining sampling frequency
  - Assessing how much data to use

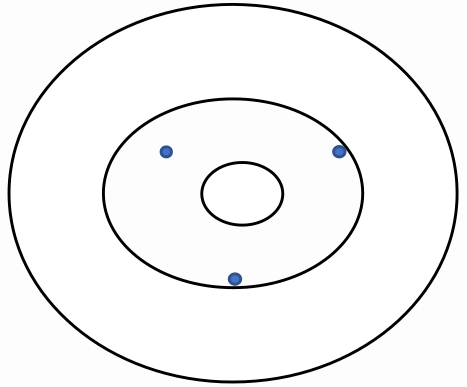
Uncertainties in data collection

- Measurement/Experimental uncertainty:
  - Systematic
    - due to faults in measuring instruments or in the techniques used
    - Decreases the accuracy of an experiment
    - Accuracy is how close a measurement is to the correct value for that m
    - **Low systematic** uncertainty is said to be **accurate.**
  - Random
    - associated with unpredicted variations in the experimental conditions or due to a deficiency in defining the quantity being measured
    - Random means “not precisely predictable or determinable”
    - Decreases the precision of an experiment.
    - Precision refers to how close the agreement is between the repeated measurement (repeated under same conditions)
    - **Low random** uncertainty is said to be **precise**

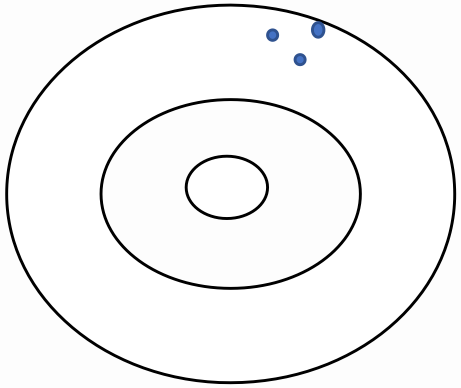
Precision and Accuracy



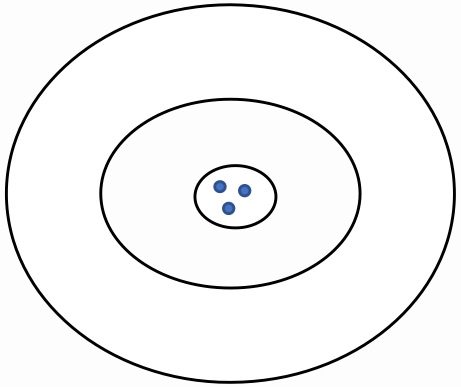
Not precise and not accurate



Not precise but accurate



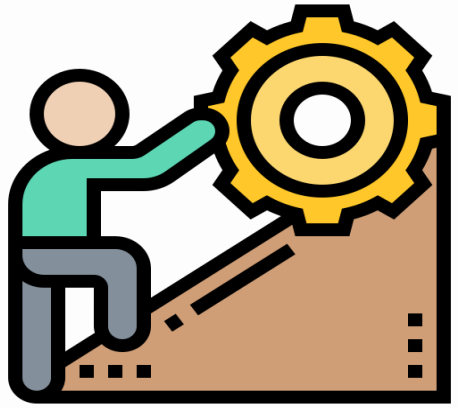
Precise but not accurate

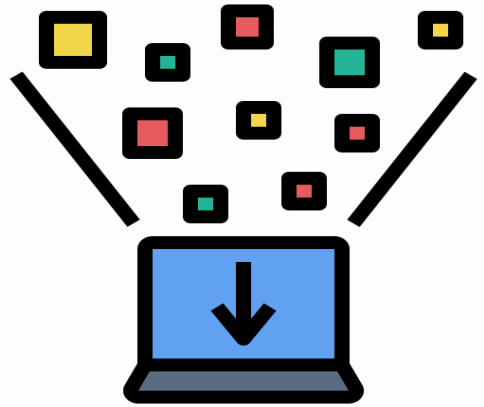


Precise and Accurate

Challenges in current data collection

- Inconsistent data collection standards.
- Context of data collection.
- Complexity.
- Lack of training in data collection.
- Lack of quality assurance processes.
- Changes to definitions and policies and maintaining data comparability.





2. DATA EXPLORATION AND PRE-PROCESSING

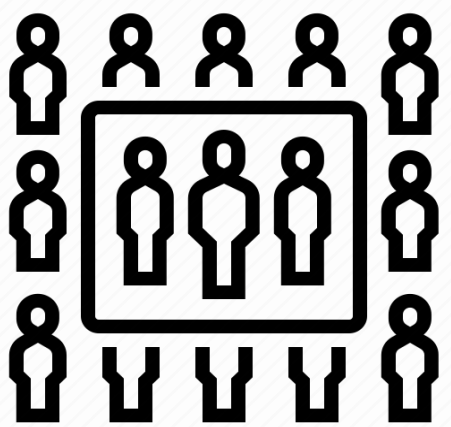
- To understand data characteristics
- During pre-processing key aspects that must be taken into consideration are:
  - Sample size- sampling method
  - Missing values – impute data or interpolation
  - Distributions
  - Initial patterns
  - Correlation
  - Variables that sensitive to the system of interest

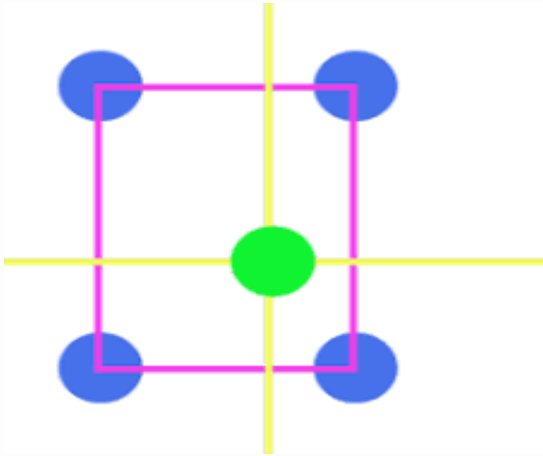
UNCERTAINTIES

- Since poor data quality may lead to inaccurate and misleading analyses, hence pre-processing is intended to improve model quality.
- There are four suggested steps: data cleaning, integration, reduction and transformation

Methods of Data exploration

- Use Histogram, box plot or scatter plot
- Interpolation and extrapolation
- Bias correction
- Missing data techniques
- Test on homogeneity
- Trend analysis
- Structural breakpoint
- Long memory or short memory
- Using plotting to visualize the distributions and trends
- Homogeneity of data
- Structural breakpoint and persistency of data processing
- Missing data techniques
- Clustering
- Dimension reduction





Missing Data

Wind Speed

7.1
8.7
8.4
7.4
9.0
6.4
7.2
12.9
9.1
-1.1
-1.1
-1.1
-1.1
-1.1
-1.1
-1.1
-1.1
5.7
15.7
6.4
7.0

260
290
280
230
230
170
170
330
240
-1
-1
-1
-1
-1
-1
-1
-1
210
260
340
240

-1.1
-1.1
-1.1
-1.1
-1.1
-1.1
-1.1
-1.1
-1.1
-1.1
11.5
8.8
10.6
9.0
8.9
10.4
9.9
12.3
4.9

Imputation Methods

Day	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	?	5.9	1.5	10.0	0.5	1.5	0.0	?	4.0	9.0	40.0	4.0
2	?	2.0	2.0	1.5	15.0	17.5	4.5	?	7.5	13.2	1.0	9.0
3	?	1.5	0.0	16.5	0.0	10.0	4.0	?	0.5	6.3	0.5	0.0
4	?	0.0	0.0	0.0	10.5	9.5	0.0	?	0.0	9.5	0.0	0.0
5	?	0.0	0.0	0.5	19.0	5.0	0.0	?	0.0	3.0	5.5	0.0
6	?	0.0	0.0	0.0	1.5	5.5	0.5	?	0.0	11.0	12.0	0.0
7	?	0.0	26.0	0.0	5.1	0.0	5.5	1.5	0.0	0.5	7.0	0.0
8	?	2.5	4.5	7.0	6.5	0.0	0.0	1.5	0.0	11.0	26.5	10.0
9	?	23.5	3.0	0.0	5.2	0.0	1.5	0.5	1.5	14.5	31.5	2.0
10	?	0.0	6.5	7.5	2.7	0.0	33.5	0.0	8.5	0.5	21.0	6.5
11	?	0.0	31.0	0.0	2.0	16.5	0.5	3.0	0.0	0.0	0.0	7.6
12	?	0.0	0.0	0.0	0.5	12.5	?	10.5	2.5	3.0	23.5	5.9
13	?	0.0	0.0	13.0	0.5	1.7	?	0.5	0.0	19.0	2.0	20.0
14	?	0.0	0.0	26.5	0.5	10.5	?	0.5	0.0	12.0	4.0	60.3
15	?	0.0	0.0	25.0	0.5	8.3	?	0.5	5.5	26.5	0.0	25.7
16	?	0.0	0.0	11.5	0.0	0.0	?	1.0	1.5	0.5	60.0	2.5
17	?	0.0	3.5	32.0	0.0	0.0	?	0.0	16.9	0.0	8.0	6.0
18	?	0.0	0.0	33.2	15.3	0.0	?	42.5	48.6	3.0	2.0	7.5

MISSING DATA

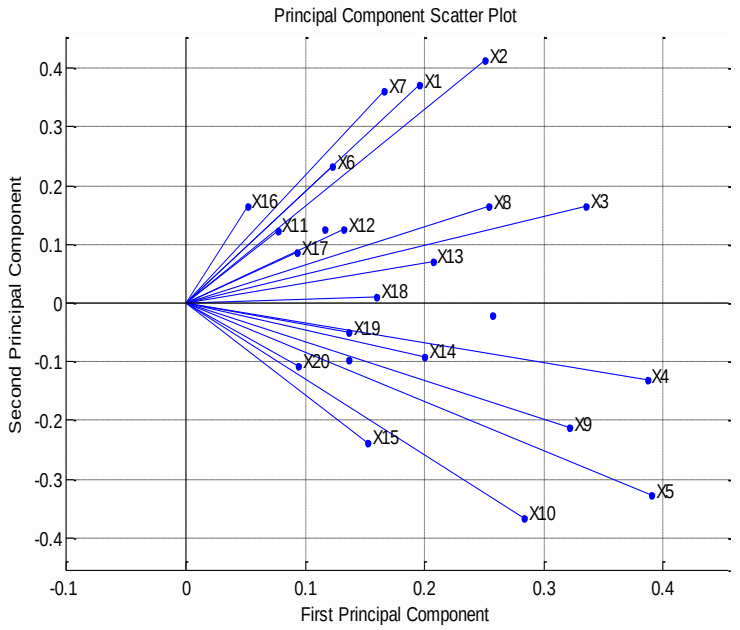
WHY DID IT HAPPEN?

HOW DO WE SOLVE IT ?

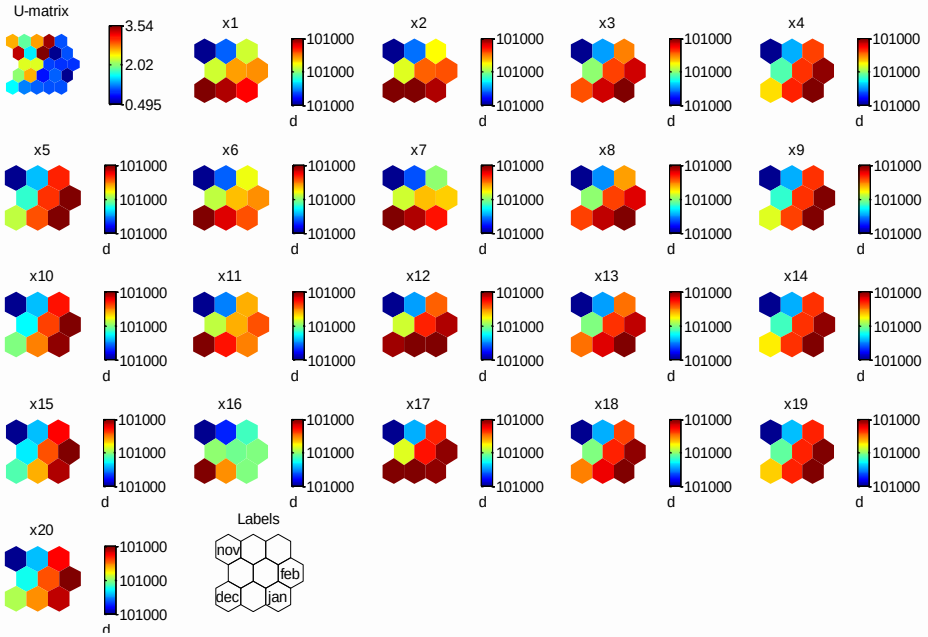
Nearest Neighbor Method?
Complete-case Analysis ?
Mean Imputation ?
Linear Regression ?
Expectation Maximization ?
Multiple Imputation ? OR
Artificial Neural Network ?

Dimension Reduction Uncertainty

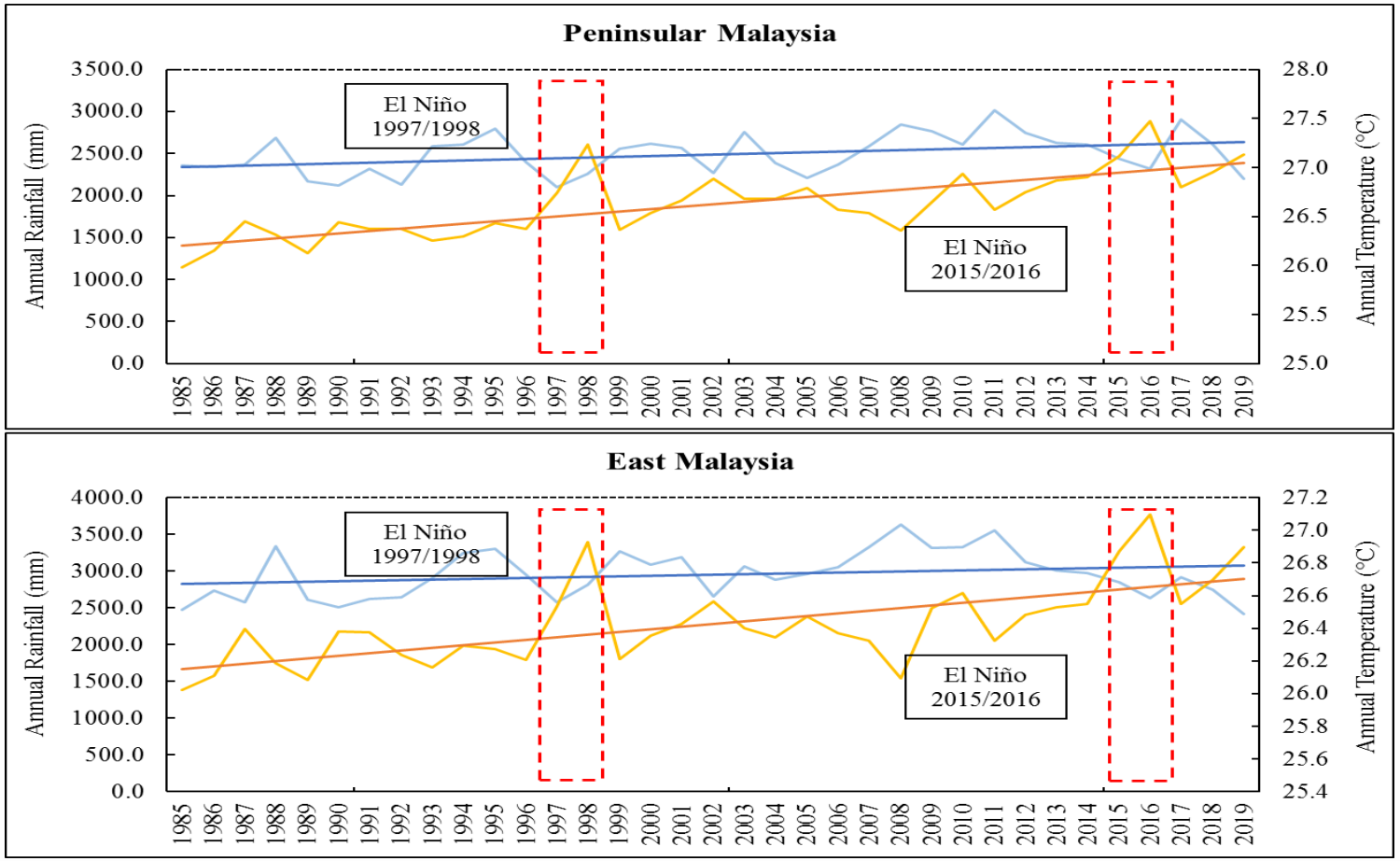
Principle Component Analysis



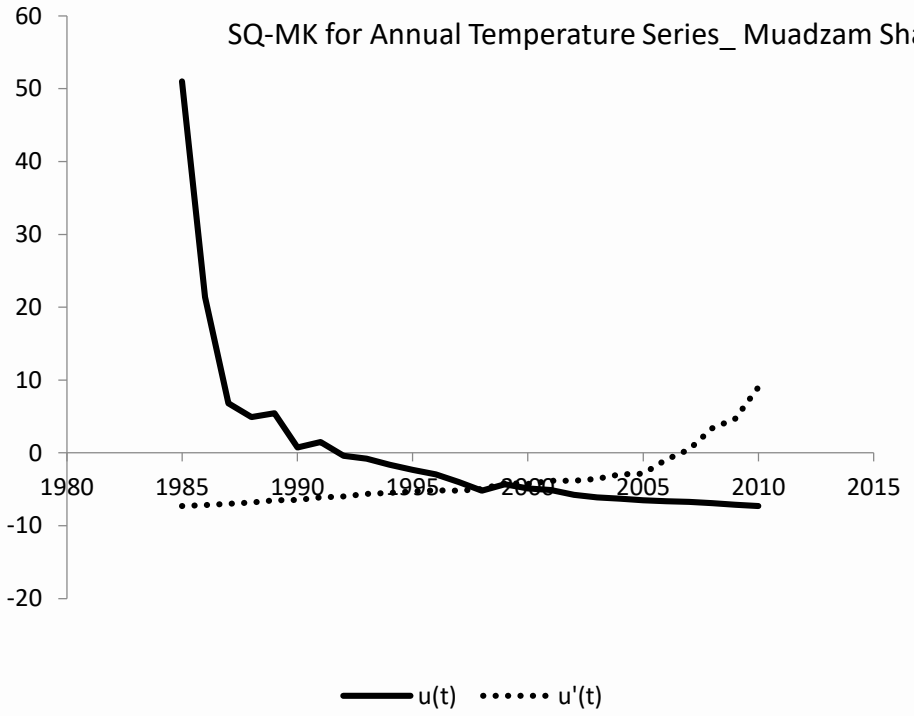
Self-Organizing Map

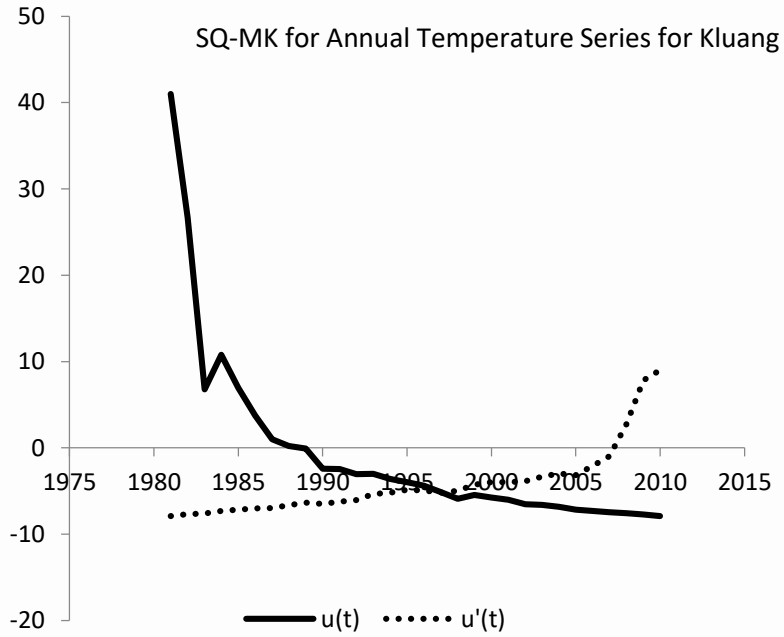


Trend Analysis

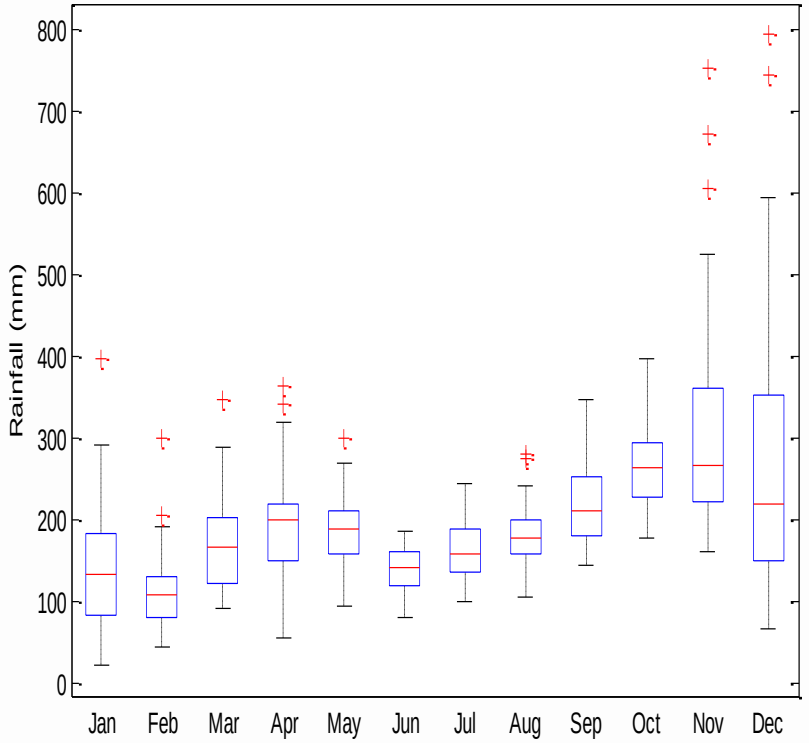


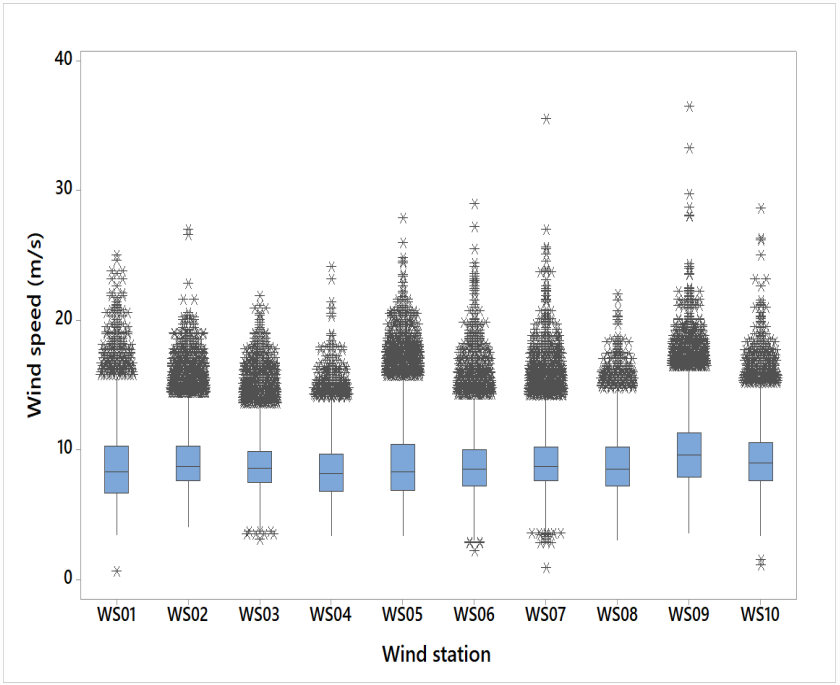
CHANGE POINT DETECTION & TREND ANALYSIS





Outliers in data set





3. MODEL BUILDING

- A **statistical model** is a mathematical model that embodies a set of statistical assumptions concerning the generation of sample data. A statistical model represents, often in considerably idealized form, the data-generating process
- A **mathematical model** is a description of a system using mathematical concepts and language.
- **Machine Learning** is the use of mathematical and or statistical models to obtain a general understanding of the data to make predictions

Uncertainties in modeling

- Model uncertainty is uncertainty due to imperfections and idealizations made in physical model formulations as well as in the choices of probability distribution types for the representation of uncertainties.
- Parameter uncertainty happens because **we don't know the exact, or "best" values in a population**—we can only take a good guess with sampling..
- The choice of distributions
- The assumption of independency between variables
- The parameter estimation techniques used
- The choice of model and the assumptions in the modelling
- Algorithmic uncertainty
- Statistical uncertainty

Example of Models

- **Supervised learning techniques** include regression models and classification models:
- **Regression model**: a type of predictive statistical model that analyzes the relationship between a dependent and an independent variable. Common regression models include logistic, polynomial, and linear regression models. Use cases include forecasting, time series modeling, and discovering the causal effect relationship between variables.
- **Classification model**: a type of machine learning in which an algorithm analyzes an existing, large and complex set of known data points as a means of understanding and then appropriately classifying the data; common models include models include decision trees, Naive Bayes, nearest neighbor, random forests, and neural networking models, which are typically used in Artificial Intelligence.

- **Unsupervised** learning techniques include clustering algorithms and association rules:
- **K-means clustering**: aggregates a specified number of data points into a specific number of groupings based on certain similarities.
- **Reinforcement learning:** an area of deep learning that concerns models iterating over many attempts, rewarding moves that produce favorable outcomes and penalizing steps that produce undesired outcomes, therefore training the algorithm to learn the optimal process.

Statistical Models

Three main types of statistical models

- Parametric: a family of probability distributions that has a finite number of parameters.
- Nonparametric: models in which the number and nature of the parameters are flexible and not fixed in advance.
- Semiparametric: the parameter has both a finite-dimensional component (parametric) and an infinite-dimensional component (nonparametric).

Uncertainties in building Statistical Model

- How to choose the best statistical model with the given variables?
- Is the purpose of the analysis to answer a very specific question, or solely to make predictions from a set of variables?
- How many explanatory and dependent variables are there?
- What is the form of the relationships between dependent and explanatory variables?
- How many parameters will be included in the model?
- Could the assumptions in the modeling be satisfied?

Challenges

- Normality of data
- Extreme and Heavy tailed data set
- Dependent or Independent Variables
- Data- too short
- Too many variables
- Assumption of stationarity but in real life – not stationary
- Not able to address extreme or high peak in the data series
- Sub scale data e.g in minutes or dailynot available, so use only monthly or yearly data.
- Linear or non-linear models?

Best-fit distribution

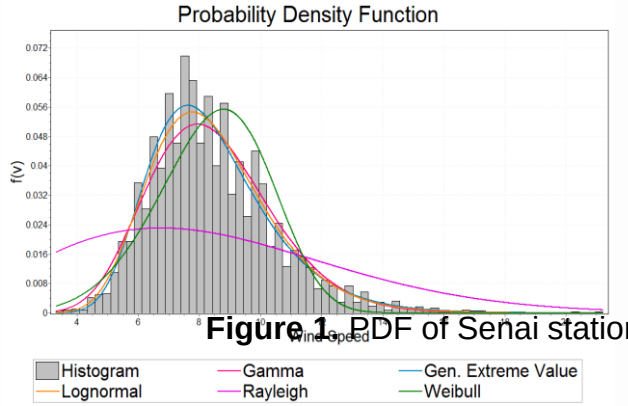
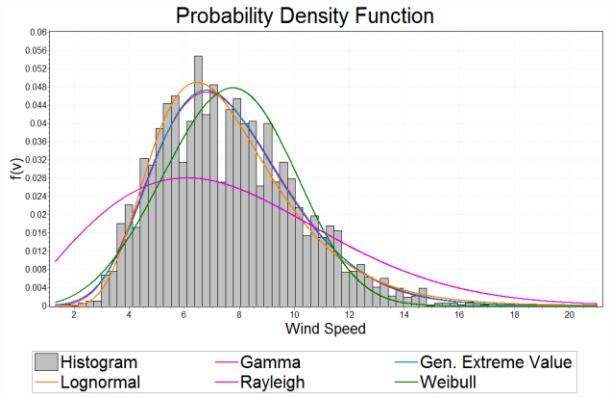


Figure 1. PDF of Senai station.



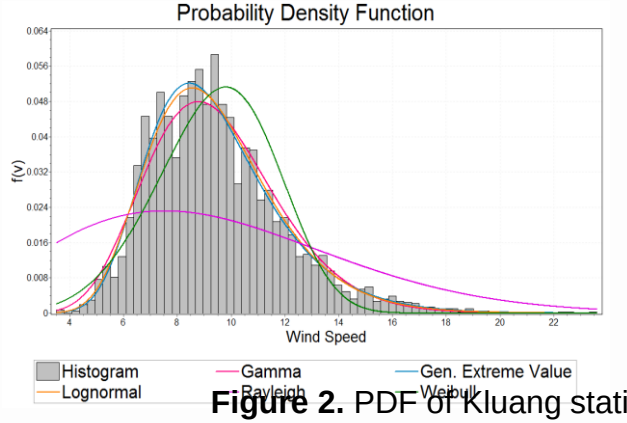
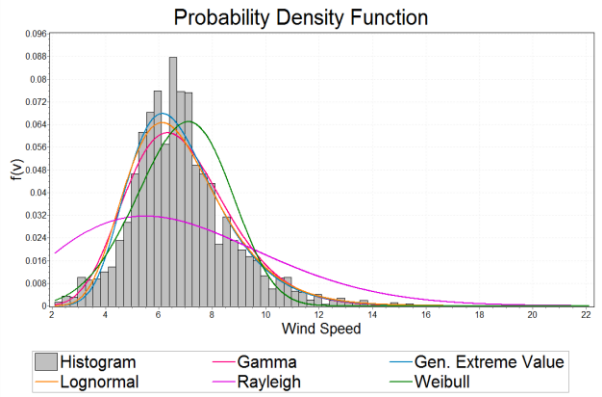


Figure 2. PDF of Kluang station.



4. MODEL EVALUATION

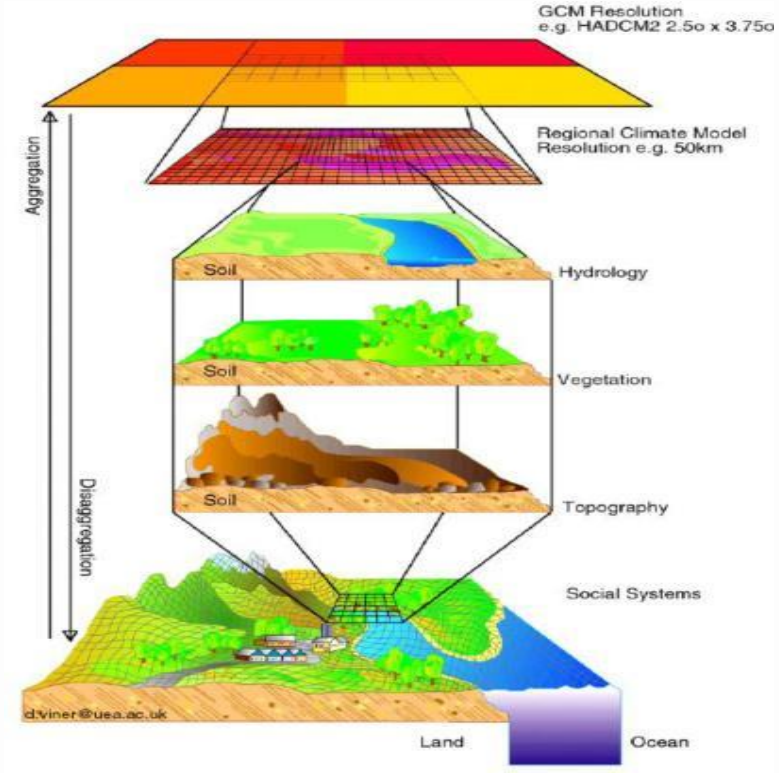
- To evaluate model robustness and accuracy
- Calibration and validation – 80% calibration 20% validation
- Model sensitivity analysis – to alter input variables and/or parameters of the model and study the subsequent changes in model output.
    - If small changes of output –output is robust to changes in parameter values, indicates that the uncertainty value is small. If large changes of output, means large uncertainty about the variable's value.
- Structural uncertainty can be evaluated by comparing model results with the observed data, but if data is not enough, expert assessment is needed.
- To compare performance of few models in getting the best model.
- Simulation – to verify the models established

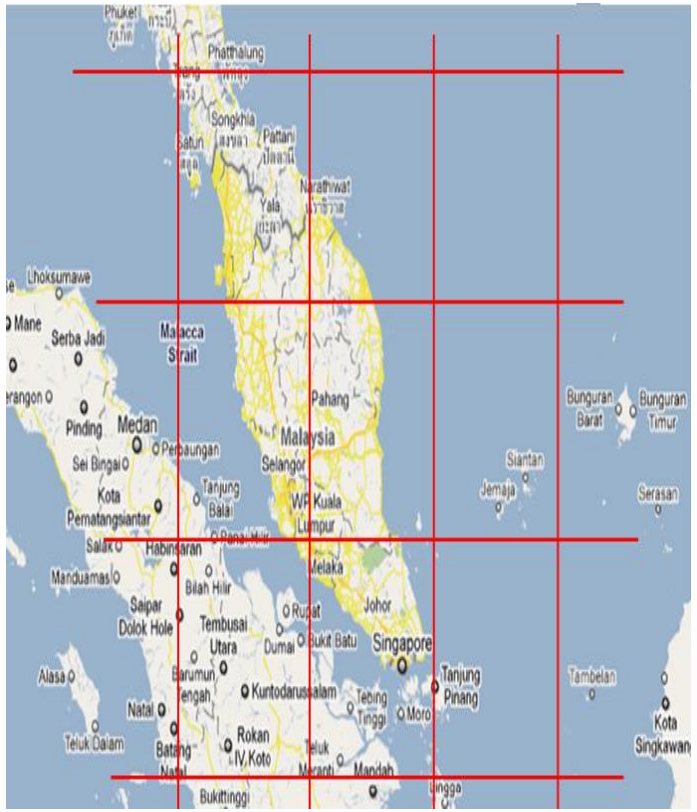
5. INTERPRETATION

- Develop reasonable explanations for the obtained analysis
- It could mean extracting important variables for model prediction.
- Literature and expert opinion are queried to make connections between model output and the phenomena for conclusions and decisions.
- If the inputs to a model are uncertain (which they inevitably are in many cases) than that there is an inherent variability (uncertainty) associated with the output of that model. Therefore it is very important that this should be communicated in the model predictions.
- The role of the uncertainty analysis is to assess the error in the model calculations. Uncertainty analysis **aims at quantifying the variability of the output that is due to the variability of the input**. The quantification is most often performed by estimating statistical quantities of interest such as mean, median, and population quantiles.
- Possibilities of making a wrong decision and prediction.

How to minimize uncertainty?

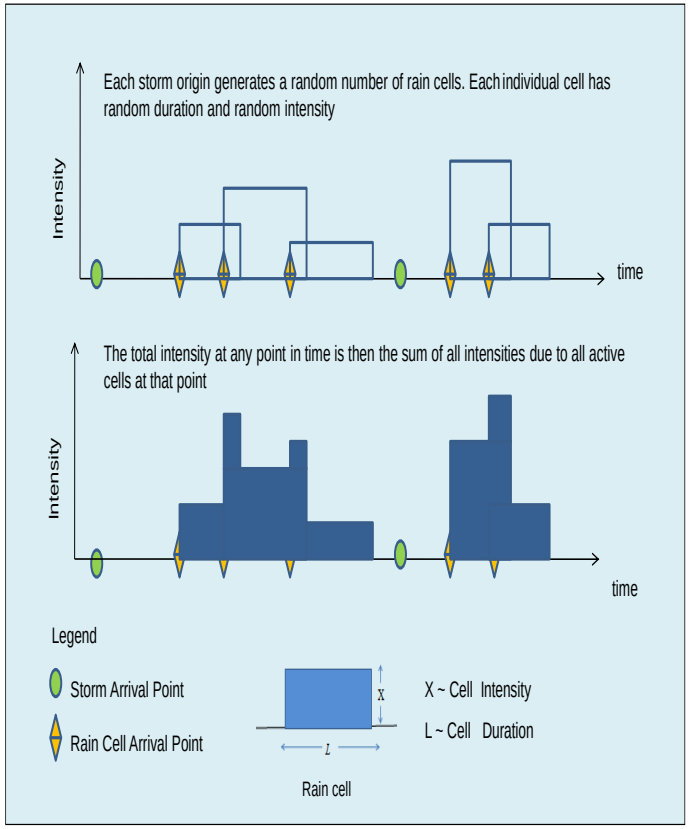
- Statistical Downscaling Study





1. GCM is able to make future climate projection but GCM runs on a relatively coarse resolution 150-300km x 150-300km.
2. Not able to be used for local impact studies because they require information at scales of 50 x50km(now 6x6km) .
3. Use Statistical Downscaling Method: to downscale the output of GCM predictors to the local predictands. The prediction of the rainfall trends can be done.
4. Methods used: Regression Based model.
5. Preprocessing : Dimension Reduction
6. Calibration and Validation- reduce uncertainties

Stochastic Generation Of Rainfall Process

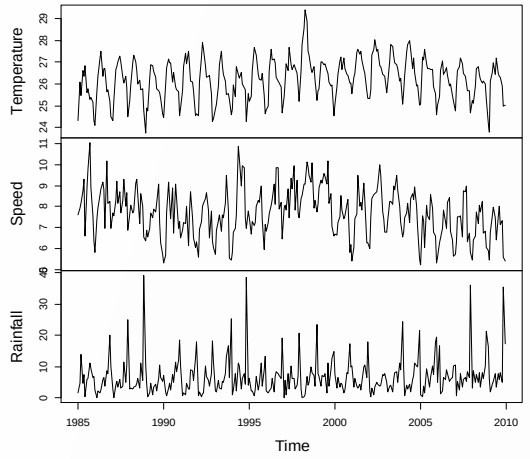


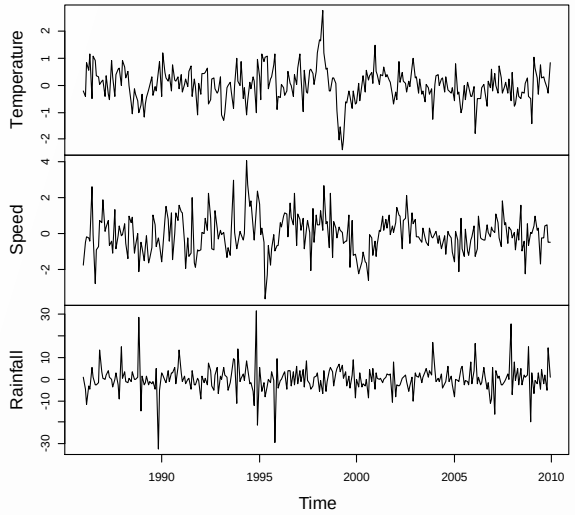
Parameter	Physical Process	Distribution
λ (Lambda)	Storm origin arrival rate	Poisson
β (Beta)	Waiting time for cell origins after the storm origin	Exponential
ν (Nu)	Mean number of cells	Poisson
η (Eta)	Duration of a cell	Exponential
θ (Theta)	Mean intensity of heavy cell	Mixed exponential
ξ (Xi)	Mean intensity of light cell	Mixed exponential
α (Alpha)	Mixing probability of cell intensity	Mixed exponential

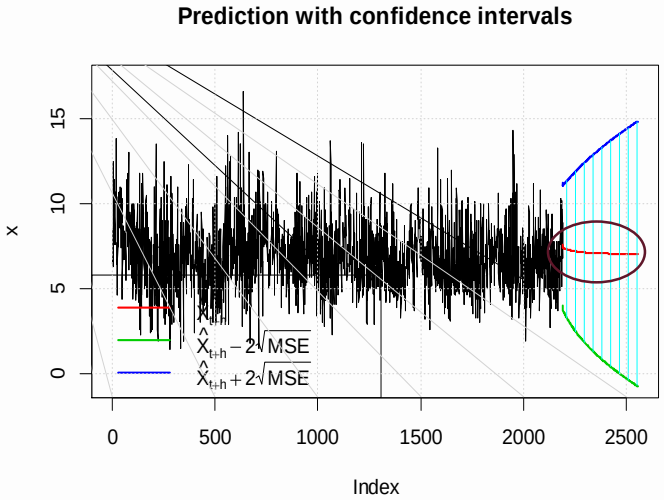
- Uncertainties
  - choice of distributions
  - Parameter estimations
  - Simulations
  - Comparing performances using physical and statistical properties
- Calibration and Validation process – to minimize uncertainty

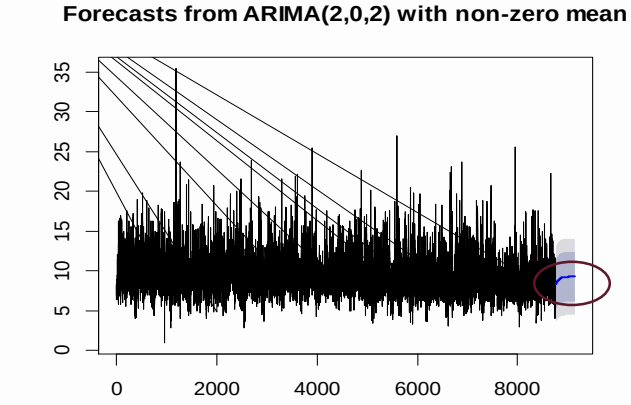
Time Series Analysis

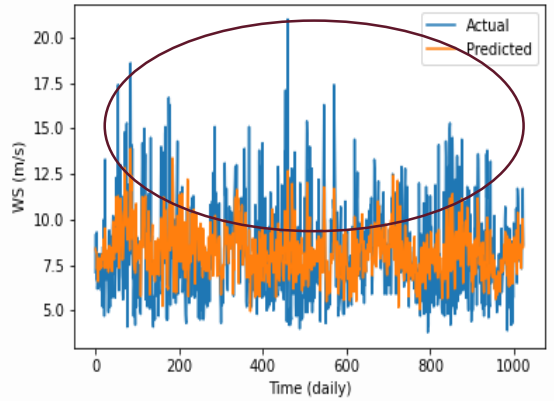
Models uncertainty

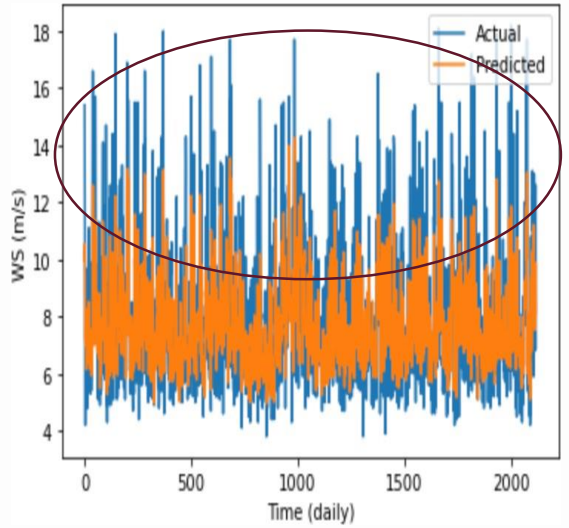






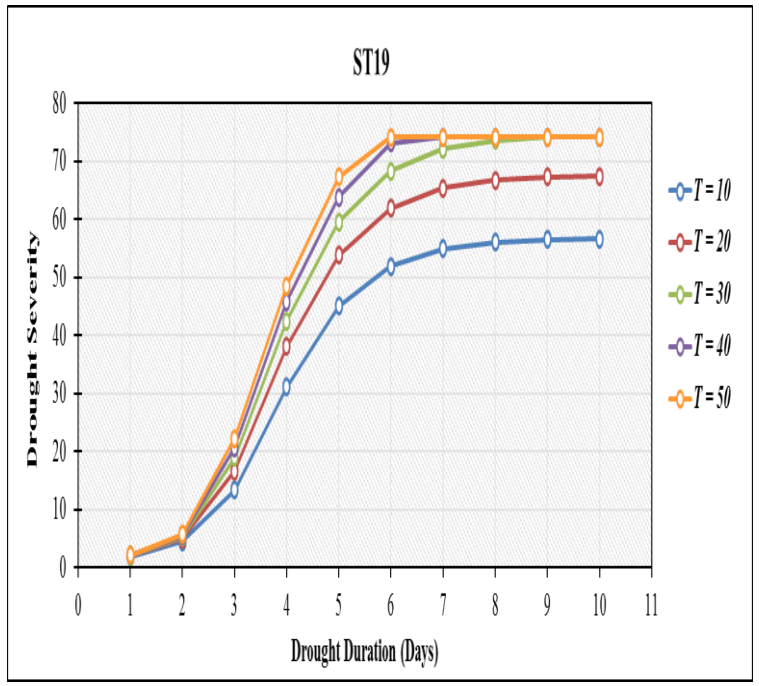




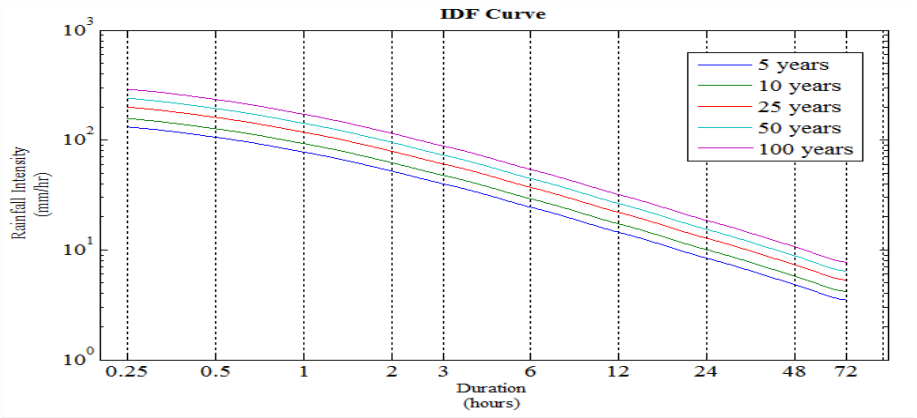


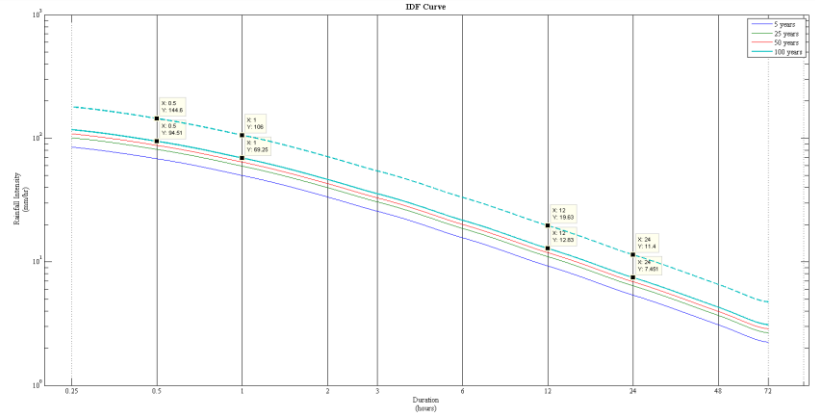
Drought and Flood Analysis

SEVERITY DURATION FREQUENCY CURVE



INTENSITY DURATION FREQUENCY CURVE





Historical data was used in this study . To minimize uncertainty, include climate change factor.

Discussion

What is uncertainty?

You want to organize a party in your housing area in three months time. It will be a garden concept party in an open area. Your main concern is whether it will be raining or not during the event. There will be few canopies to provide some shades from morning sun but not from heavy rainfall. How do you make a decision on this matter?

This is an uncertainty situation. What will a researcher do to minimize this uncertainty?

What is uncertainty?

People living in the "ring of fire" region all their life which are prone to earthquakes and volcanoes eruption can still lead a normal life. Even though their livelihood are very uncertain, their countries' economics and social well being are still moving forward in a positive way.

How they manage the uncertainty?

The future cannot be predicted with precision, but that doesn’t mean we can’t prepare for it. But rather than using uncertainty as an excuse for inaction, policymakers and others must embrace it, incorporate it into their decision-making and make the best plans now to protect vulnerable communities from future events. Two of the main ones are the **resilient** and **adaptive** approaches.

UNCERTAINTY QUOTES

We as human are very uncomfortable with uncertainty. God says in the Qur'an "Indeed, mankind was created anxious" ***(Qur'an, 70:19).***

As far as the law of mathematics refer to reality, they are not certain; and as far as they are certain, they do not refer to reality. ***Albert Einstein***

Ignorance more frequently begets confidence than do knowledge. ***Charles Darwin***

Uncertainty is an uncomfortable position. But certainty is an absurd one. ***Voltaire***

THANK YOU

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

In the Name of God for Mankind

www.utm.my