Recent developments and challenges ahead in hydrological modelling

A.W.Jayawardena

Department of Civil Engineering
The University of Hong Kong
Hong Kong

and

Research and Development Centre
Nippon Koei Co. Ltd (Consulting Engineers)
Tsukuba, Japan

November 12-14, 2014

Contents

- **Historical highlights of the pre-computer era**
- **Present (computer) era**
- **Highlights of the computer era**
  - **Conceptual models**
  - **Physics-based models**
  - **Data driven models**
- **Recent advances in data driven modelling**
  - **ANN, Support vector machines, Genetic algorithms and Genetic programming, Dynamical systems approach and chaos, Fuzzy logic, Neuro-fuzzy systems, ….**
- **Guiding principles and criteria for choosing a model**
- **Challenges in the choice of hydrologic models – Data issues; modelling issues; scale issues, parameter issues**

Historical highlights of the pre-computer era

- Egyptians learned to harness the waters of the NILE, measure the rise and fall of the river circa 3000 BC.
- Descriptive hydrology was conceived by Plato and Aristotle circa 400 BC.
- Romans constructed aqueducts without a complete knowledge of quantitative hydrology circa 300 BC.
- A reasonably clear idea of the hydrological cycle was put forward by Vitruvius circa 27-17 BC,
- Hydrological cycle in its present form is only after Leonardo Da Vinci (1452 - 1519).
- Perrault in 1674, and Mariotte in 1686 independently carried out experiments on runoff in France and concluded that the rainfall is sufficient to produce streamflow.
- Until then, the belief has been that streamflow was the cause and rainfall its effect.

Historical highlights of the pre-computer era….

- Hydraulics, Hydrology and Fluid Mechanics have no crisp boundaries
  - Chezy's equation (1769); Darcy's equation (1856); Manning's equation (1885) etc. are household names in hydrology too.
- Da Vinci (15^{th} century) – concept of hydrological cycle
- Rational method (First Hydrological Model) (Mulvany, 1850)
- Regression models - empirical
- Unit Hydrograph Theory (Sherman, 1932)
- Infiltration theories (Green & Ampt, 1911; Horton, 1933; Philip, 1954)
- Extreme value theory (Gumbel, 1941)
- Kinematic wave theory (Lighthill & Whitham, 1955)

Historical highlights of the pre-computer era…. - Analog models (now outdated)

- Analogy between water flow and electricity flow (late 1940's)
  - Pressure-voltage
  - Flow rate-current
  - Friction-resistance
  - Storage-capacitance
- Analogy between potential flow and magnetic field

Present (computer) era

- **Recent developments**
  - Due to advances in computing capabilities
  - Due to advances in methodologies (non-linearities, scaling etc.)
  - Remote sensing, GIS and Web
- **Dilemma for hydrologists**
  - Too many models and modelling techniques
    - Conceptual models
    - Physics-based models
    - Data-driven models
  - Which one to choose?

Highlights of the computer era

- **Conceptual models**
  - Tank model (Sugawara, 1956)
  - Stanford watershed model (Crawford and Linsley, 1966),
  - Xinanjiang model (Zhao, 1977, 1984; Zhao and Liu, 1995)
  - HEC series (USACE, 1960's onwards)
  - Linear cascade model (Nash, 1957, 1958, 1959, 1960)
  - Linear channel (Dooge, 1959)
  - Variable Infiltration Capacity (VIC) model (Wood et al, 1992)
  - ….
  - ….

Physics-based models

- Governing equations
  - Based on the laws of conservation of mass, momentum and energy (St.Venant's equations)
  - Assumptions and simplifications (Diffusion and kinematic wave approximations)
- Initial and boundary conditions
- Solution domain and discretization
- Solution scheme – Numerical methods
  - Finite difference methods
  - Finite element methods
  - Hybrid methods
- Parameter identification, calibration and validation

Physics-based models….

- Advantages
  - Potentially helps to understand the processes

- Disadvantages
  - Resources demanding

- Problems
  - Data resolution
  - Parameter estimation
  - Verification in a truly distributed sense is difficult if not impossible

- Example: **Système Hydrologique Europèen (SHE) Model)**
  - Evapo-transpiration component,
  - Unsaturated zone flow component described by 1-D Richards' equation,
  - Saturated zone flow component described by 3-D Boussinesq equation, and
  - Overland flow (2-D) and channel flow (1-D) components described by the diffusion wave approximation of the St. Venant equations.

Data Driven models

- Regression models
- Stochastic models – leading to time series analysis; Kalman filtering
- Artificial Neural Networks
  - Multi-layer perceptron, radial basis function networks, recurrent neural networks, product unit neural networks, wavelet neural networks etc.
- Support Vector Machines –classification and regression
- Genetic algorithms and Genetic Programming
- Fuzzy Logic models
- Neuro-fuzzy models
- Dynamical systems approach type models – chaos
- ....
- ....

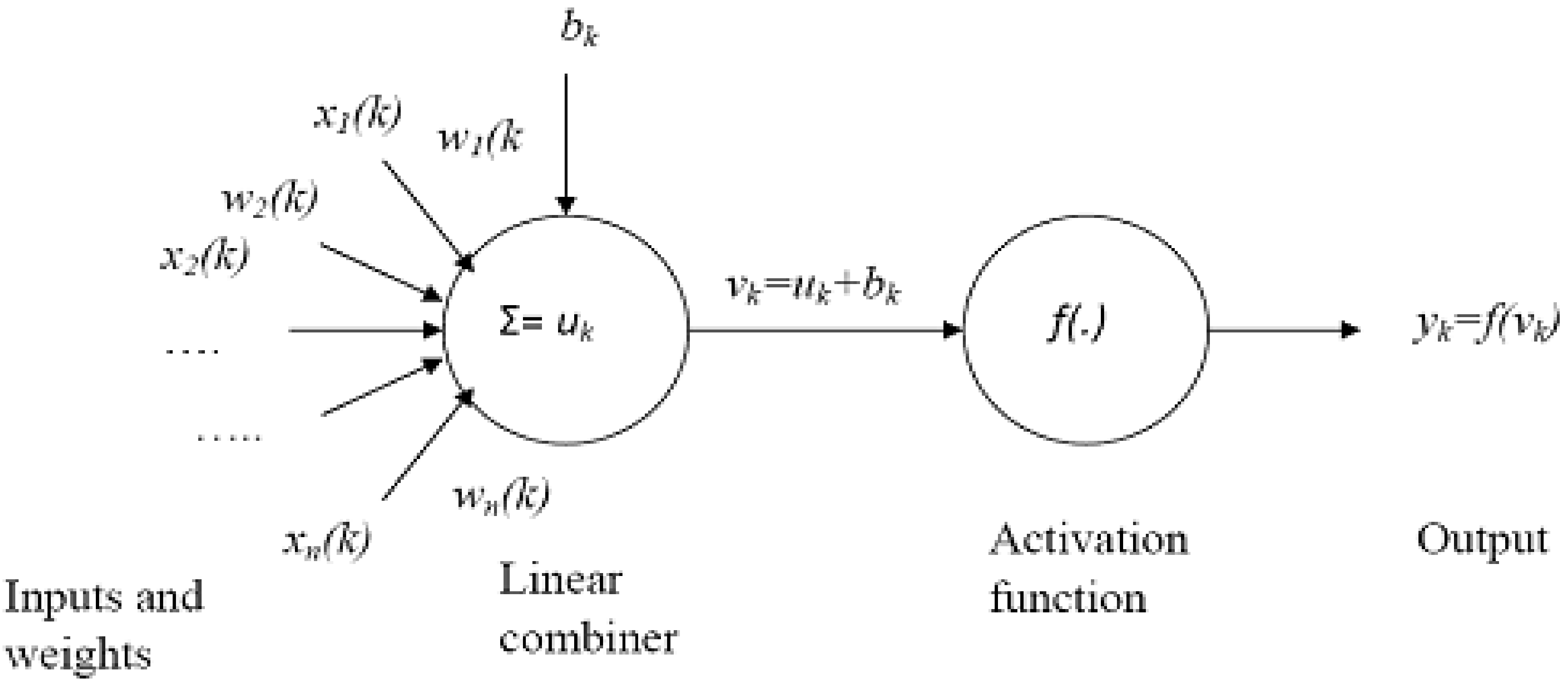
Why data driven models?

- Data contain all the measurable information about the system
- Easier to formulate and interpret
- Only option when other approaches are infeasible
- No need for *a priori* understanding of the processes involved
- Particularly suited to theory weak data rich situations

Recent advances in data driven modelling

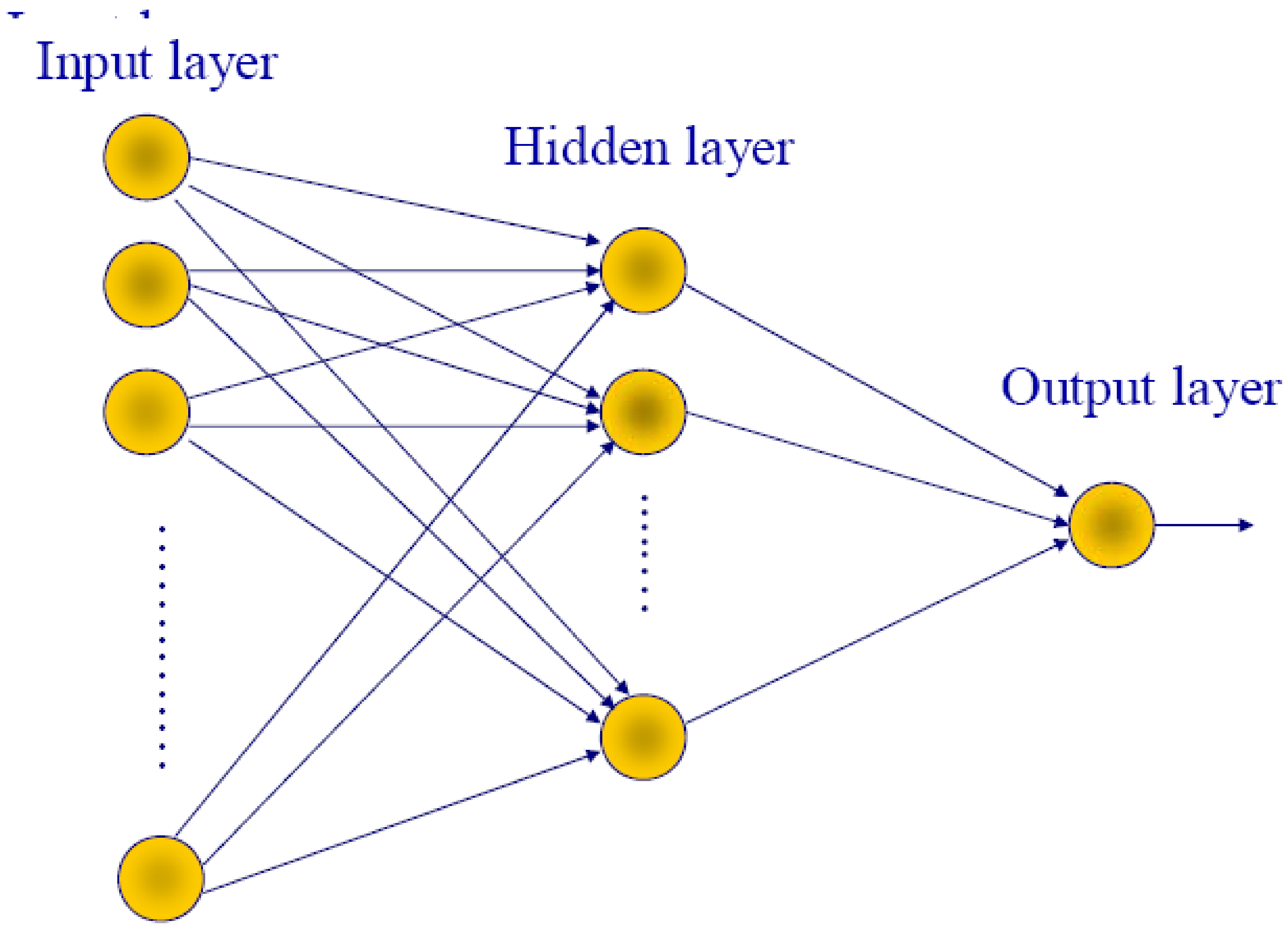
- Artificial Neural Networks
  - Multi-layer perceptron, radial basis function networks, recurrent neural networks, product unit neural networks, wavelet neural networks etc.
- Support Vector Machines –classification and regression
- Genetic algorithms and Genetic Programming
- Fuzzy Logic models
- Neuro-fuzzy models
- Dynamical systems approach type models – chaos
- ....
- ....

Artificial Neural Network

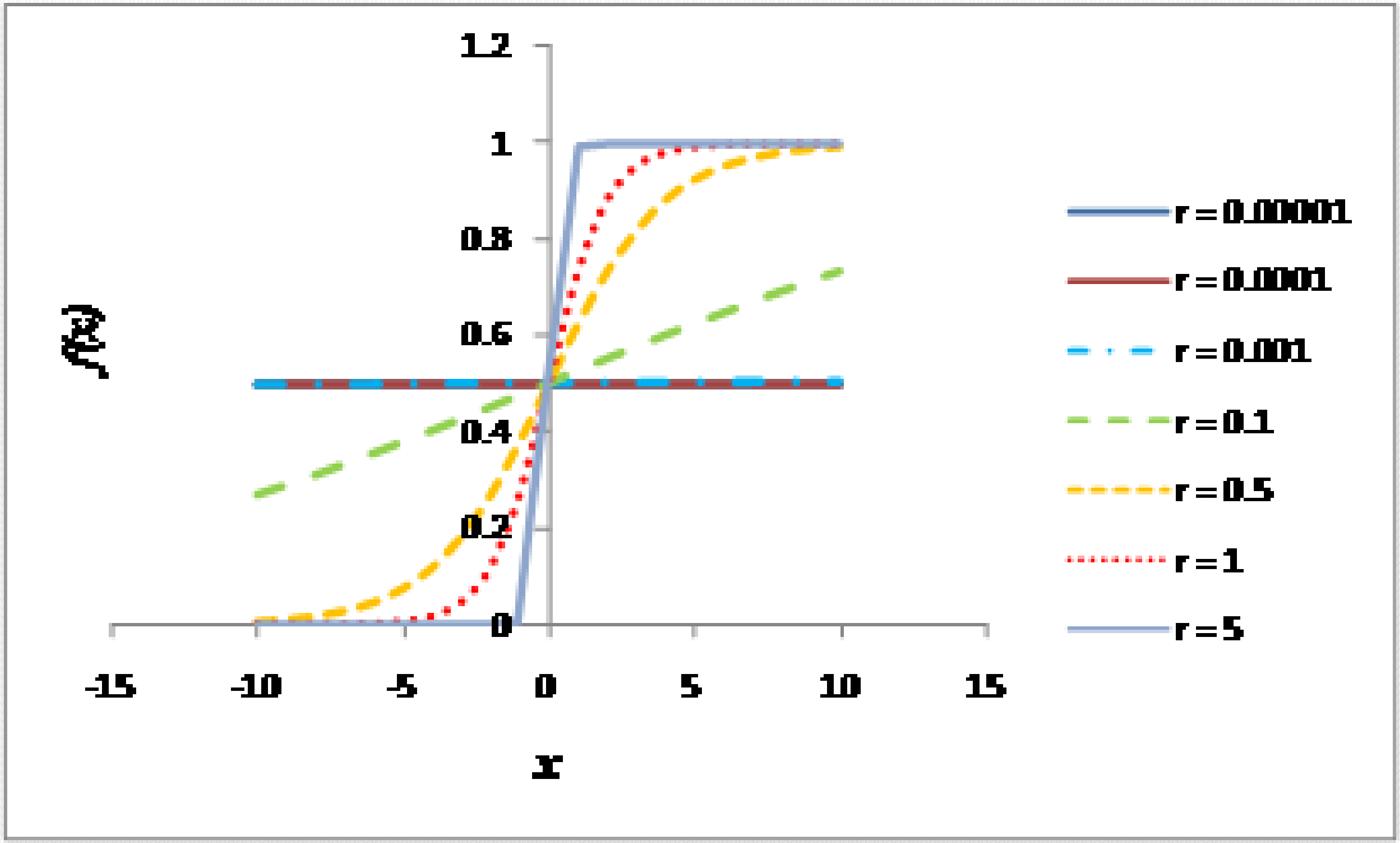


$$y_k = f(v_k) = f(u_k + b_k) = f\left(\sum_{j=1}^{N} w_{kj} x_j + b_k\right)$$

Structure of a Multi-layer perceptron (MLP) artificial neural network (ANN)



Sigmoid activation function



$$f(x) = \frac{1}{1 + e^{-rx}}$$

Stopping criterion: Cross validation

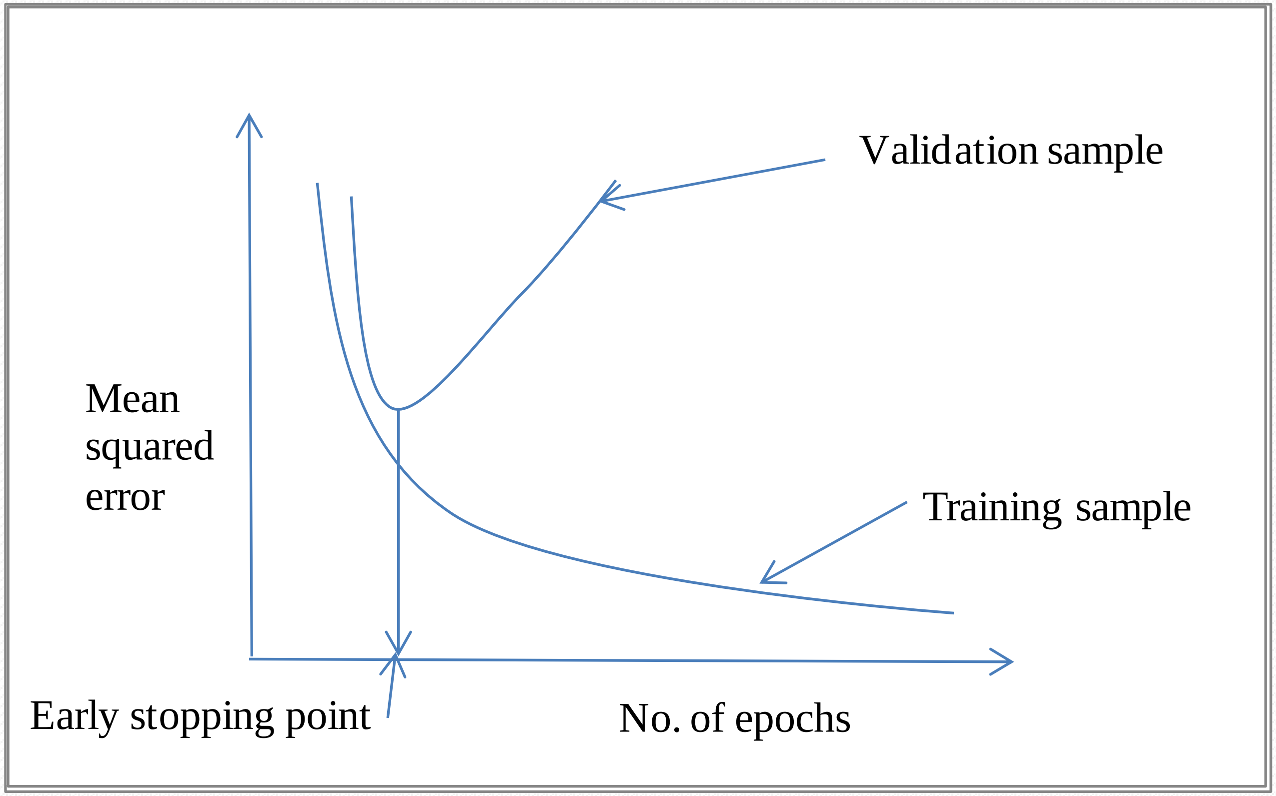
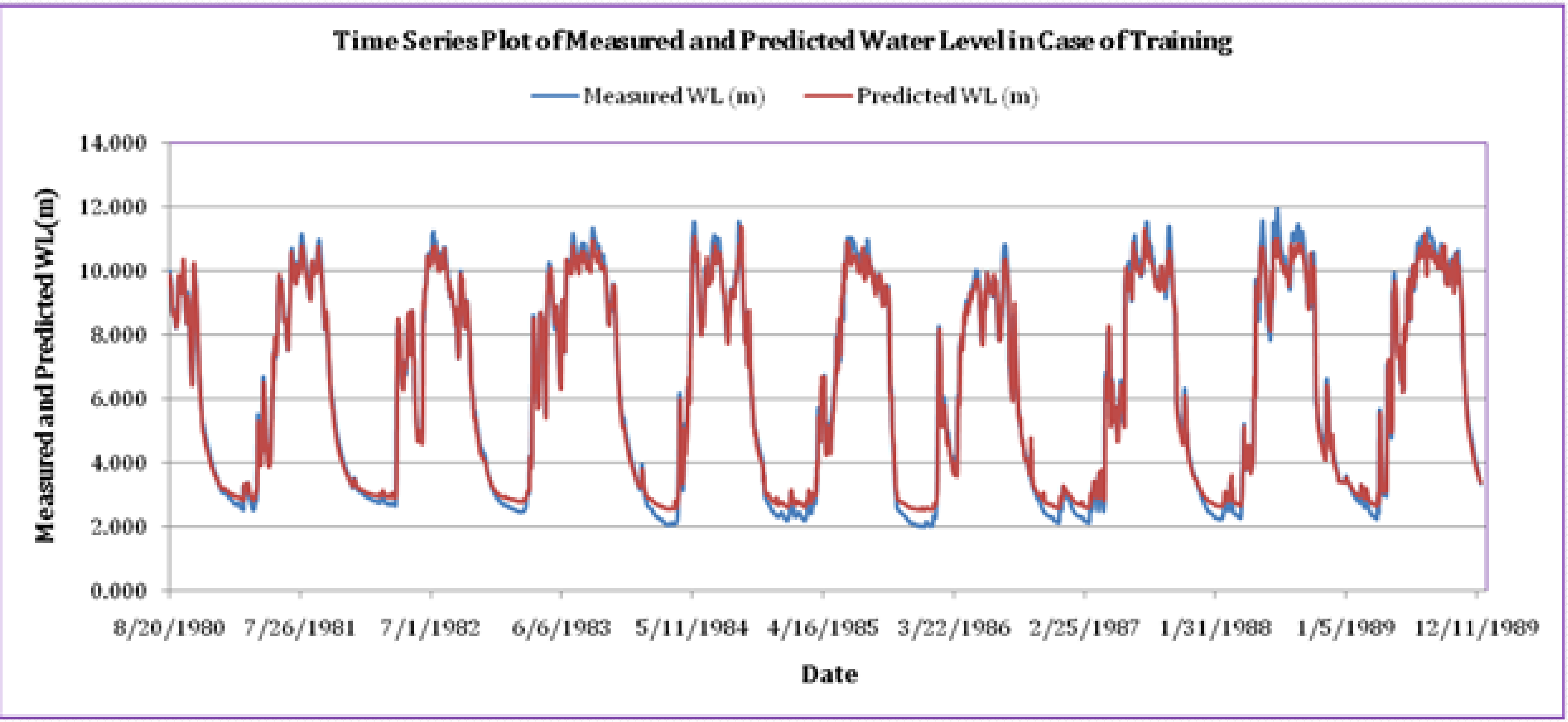


Fig. 7.6: Illustration of cross-validation

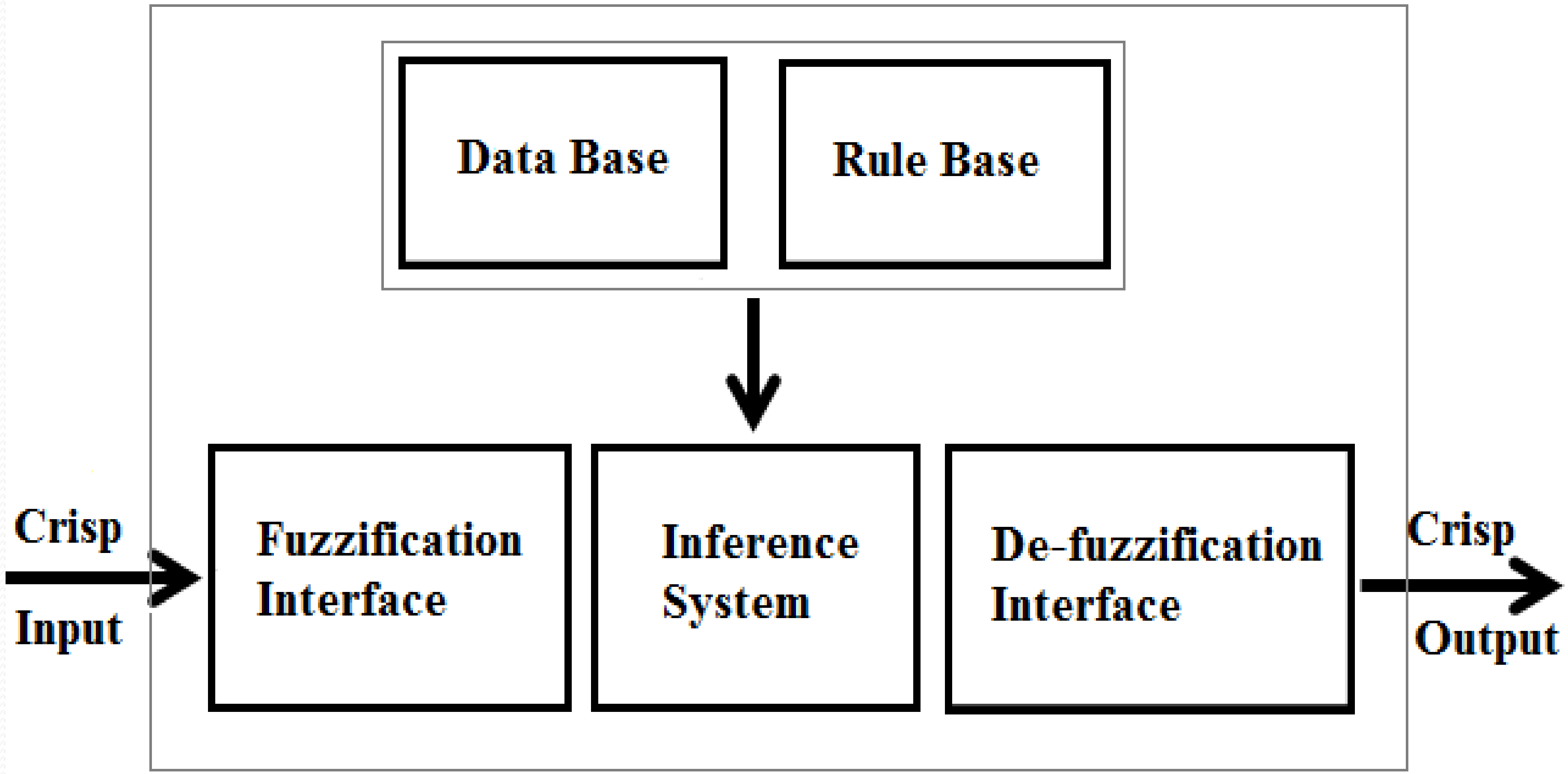
Application: Surma River, Bangladesh



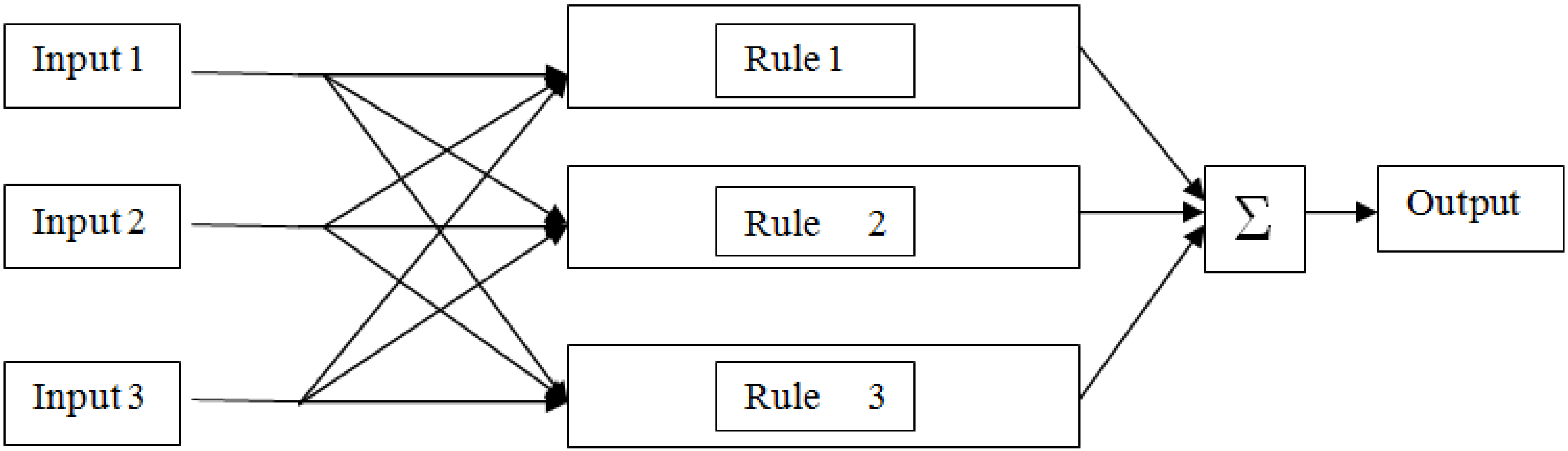
Fuzzy logic approach

- Traditional binary (crispy) logic vs. fuzzy logic
- Absolute truth vs. partial truth
- The key idea in fuzzy systems is to allow a partial truth to prevail which can be numerically described by a membership function that takes values between 0 and 1.
- Fuzzy logic enables embedding uncertain or imprecise reasoning in everyday life to computers which operate in exact deterministic ways
- Translate imprecise linguistic information sets into computer usable numerical language.

Structure of a fuzzy logic system



Information flow in a fuzzy system



Building a fuzzy system

- Steps in building a fuzzy system
  - Selection of the relevant input and output variables, choice of the possible term sets for each linguistic variable,
  - Choice of the type of membership functions,
  - Fuzzification of the crisp input and output variables,
  - Derivation of the rule set,
  - Fuzzy inference,
  - Aggregation of the outcomes of the rules and inferences and de-fuzzification.
- Choice of membership functions is rather subjective but is not due to randomness.
- Output of the aggregation process is a single fuzzy set for each output variable.

Crispy (binary) set

- In a crispy set, the membership function is of the form

$$f_A(x) = \begin{cases} 1 & x \in X \\ 0 & x \notin X \end{cases}$$

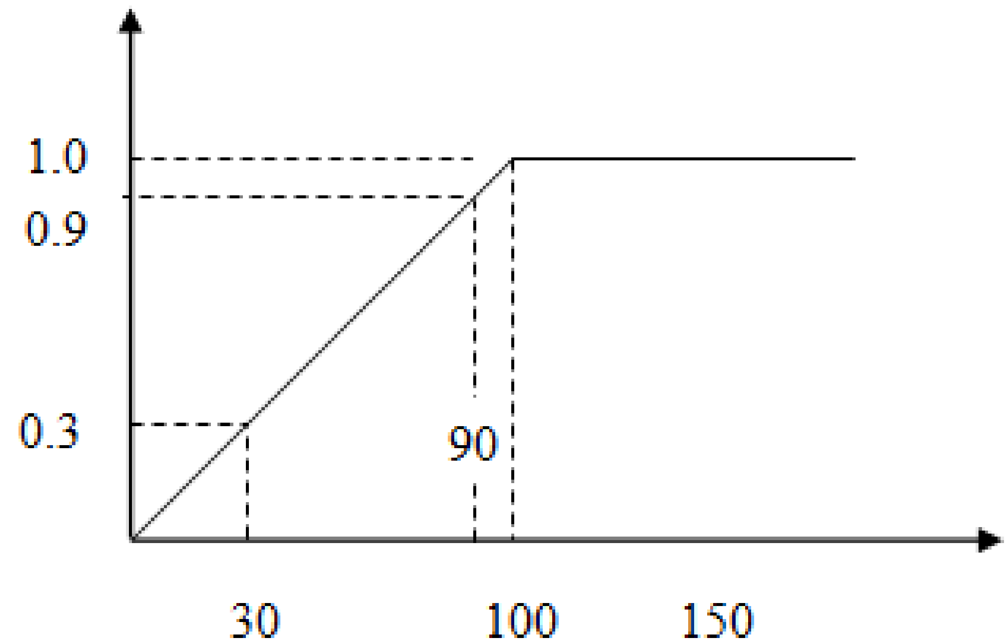
- For example, consider a rainfall of 100 mm/h. In a crisp system, the question 'Is the rainfall heavy?' has two possible answers: yes or no. If 100 mm/h is the mathematical boundary between 'heavy' and 'not heavy', the answer to any given rainfall intensity is unique, yes or no.

Fuzzy set

- A fuzzy set A in X can be defined as a set of ordered pairs A expressed as

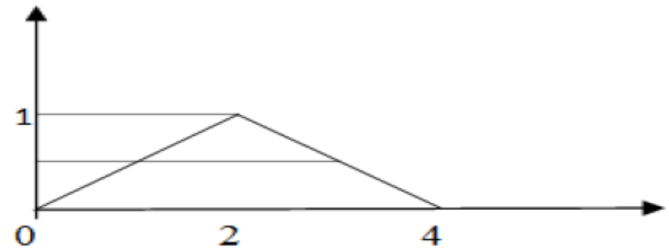
$$A = \{(x, f_A(x)) \mid x \in X\}$$

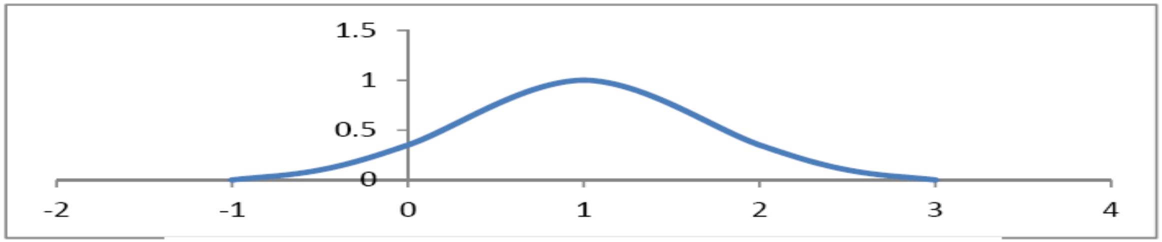
- Linear variation of rainfall from 0-100 mm/hr

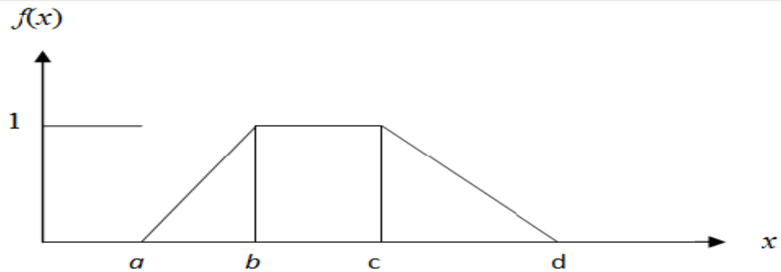


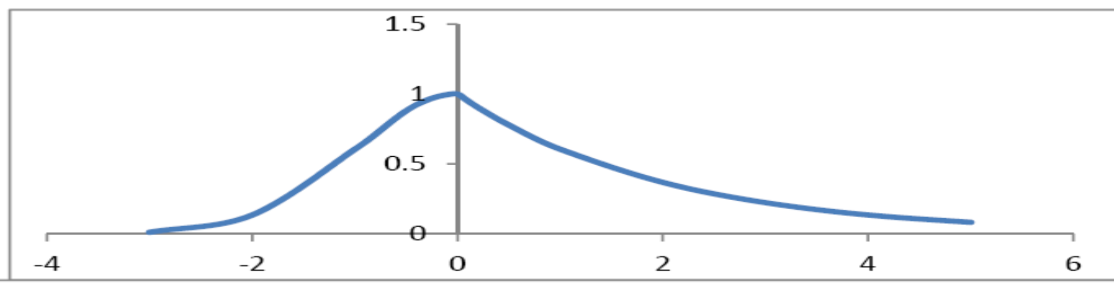
Linguistic variables

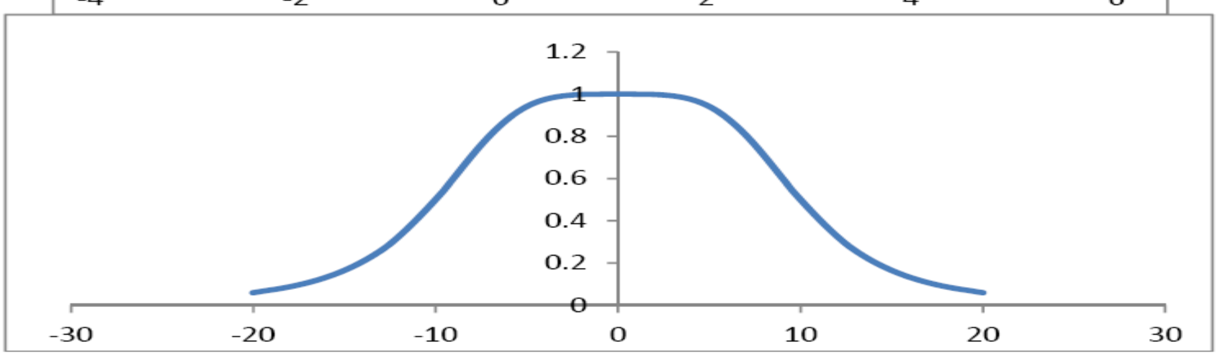
- An object can be described in many ways using modifiers (adjectives and/or adverbs) to describe a characteristic of the object or the object itself. Such descriptions that use natural languages convey imprecise meanings.
- For example a person may be described as 'very tall', 'tall', 'somewhat tall' or 'not tall' depending upon one's perception of 'tallness'. The modifiers used describe the degree of tallness in a linguistic way.
- For other variables, similar kind of linguistic variables can be assigned depending on the perception of its effect.
- Each linguistic variable can be assigned a membership value which will indicate its degree of effectiveness.
- Linking numerical values with the linguistic variables is subjective and problem specific.


1
0
2
4

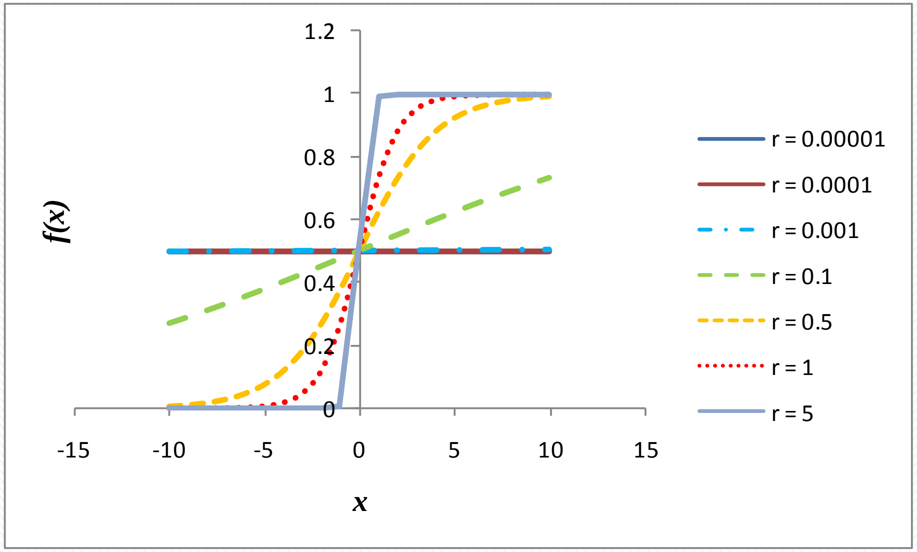

1.5
1
0.5
0
-2
-1
0
1
2
3
4

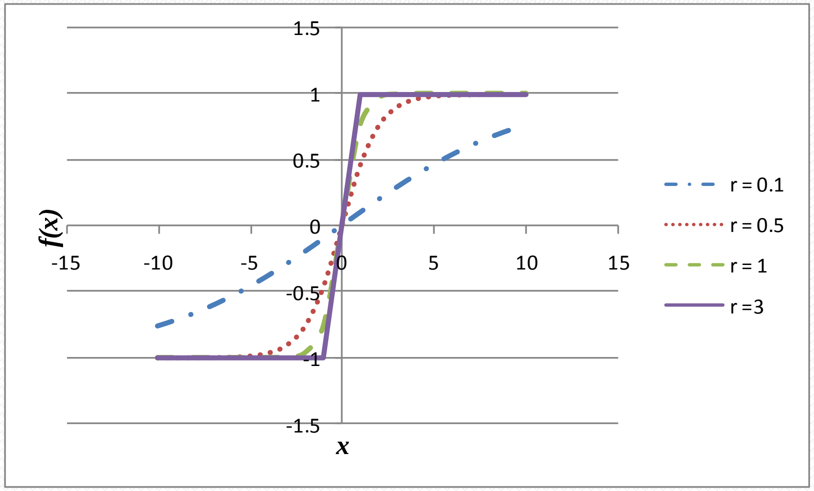

f(x)
1
a
b
c
d
x


1.5
1
0.5
0
-4
-2
0
2
4
6


1.2
1
0.8
0.6
0.4
0.2
0
-30
-20
-10
0
10
20
30

Sigmoid (logistic and tanh)





Fuzzy rules

- Fuzzy rules are based on common sense.
- Fuzzy rule base consists of rules that include all possible fuzzy relations between inputs and outputs.
- However it is to be noted that the number of rules increases exponentially with increasing number of inputs leading to what is known as the 'curse of dimensionality'.

Fuzzy rules....

- Rules are of the form IF.... THEN ......(part before THEN is the predicate (or antecedent) and the part after THEN is the consequent). For example,
- IF *rainfall is high,* THEN, *runoff is high*
- Sometimes rules are combined with logical operators 'AND', or, 'OR'.
- The activation of a rule is the deduction of the conclusion.
- For example,
  - IF *rainfall is high* AND *soil moisture is high* THEN *runoff is high*
  - *IF rainfall is high* OR *upstream discharge is high* THEN *runoff is high*

Fuzzy rule systems

- Two types of rule systems:
  - Mamdani (Mamdani and Assilian, 1975) type - fuzzy rule is also expressed in linguistic form.
  - Takagi-Sugeno-Kang (TSK) type (Takagi and Sugeno, 1974, 1985) - fuzzy rule is expressed as a mathematical function of the input variables which is more appropriate for neuro-fuzzy systems

Fuzzy inference system (FIS)

- Maps a given input to a corresponding output using fuzzy logic
- Combines the components such as membership functions, fuzzy logic operators and rules.
- Four well known inference mechanisms in fuzzy logic systems:
  - Mamdani,
  - Takagi-Sugeno-Kang (TSK),
  - Tsukamoto, and
  - Larsen.

Fuzzification of inputs

- Inputs and outputs are in most cases crisp numbers within the applicable range (universe of discourse)
- Prior to fuzzification, linguistic terms should be assigned to the crisp values of the variables within the universe of discourse.
- The number of linguistic terms to use is problem specific and rather subjective and not necessarily unique.
- The discretization should be sufficiently fine to describe the process variation adequately keeping the computer memory storage requirement as the limiting condition.
- Overlapping of membership functions is essential for smooth mapping.
- For example, to describe flow ranging from 0-50 m^3/s, linguistic terms such as dry weather flow, low flow, normal flow, high flow, flood flow may be used and assigned to equivalent crisp flow as follows:

Fuzzification of inputs….

- Dry weather flow

$$f_D(Q)=\begin{cases}1-\dfrac{Q}{10} & 0\le Q\le 10\\ 0 & Q\ge 10\end{cases}$$

- Low flow

$$f_L(Q)=\begin{cases}\dfrac{Q}{10} & 0\le Q\le 10\\ 2-\dfrac{Q}{10} & 10\le Q\le 20\end{cases}$$

- Normal flow

$$f_N(Q)=\begin{cases}\dfrac{Q}{10}-1 & 10\le Q\le 20\\ 3-\dfrac{Q}{10} & 20\le Q\le 30\end{cases}$$

- High flow

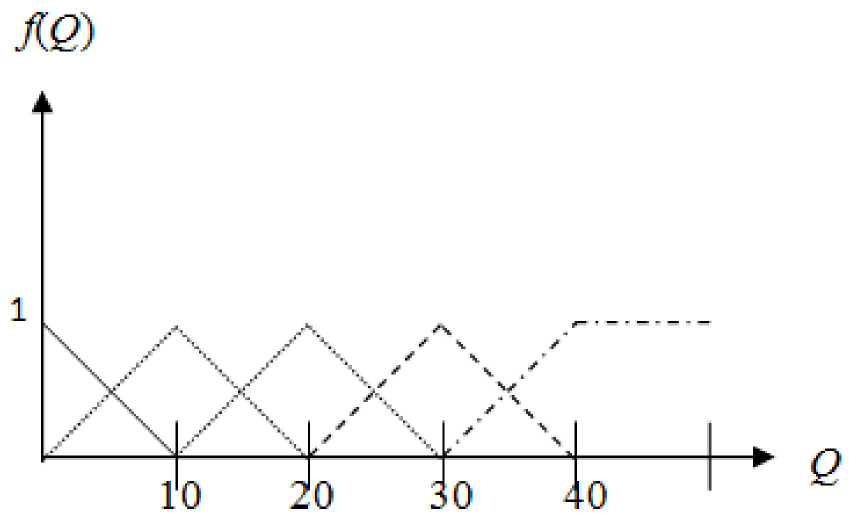
$$f_H(Q)=\begin{cases}\dfrac{Q}{10}-2 & 20\le Q\le 30\\ 4-\dfrac{Q}{10} & 30\le Q\le 40\end{cases}$$

- Flood flow

$$f_F(Q)=\begin{cases}\dfrac{Q}{10}-3 & 30\le Q\le 40\\ 1 & Q>40\end{cases}$$

Domain partitioning

- The domain partition can be represented by the 5 membership functions defined above.
- The output from the fuzzification process is a fuzzy number in the interval (0,1) that represents the degree of membership of the input variable.



Mamdani implication from antecedent to consequent – 'min' operator

- For two rules R_1 and R_2 given below, the Mamdani implication is

  R_1: IF x is A_1 AND y is B_1, THEN z is C_1
  R_2: IF x is A_2 AND y is B_2, THEN z is C_2
  (A, B and C are fuzzy numbers)
- The firing levels (or, the degrees of fulfilment) α_1 and α_2 for x_0 and y_0, are given as

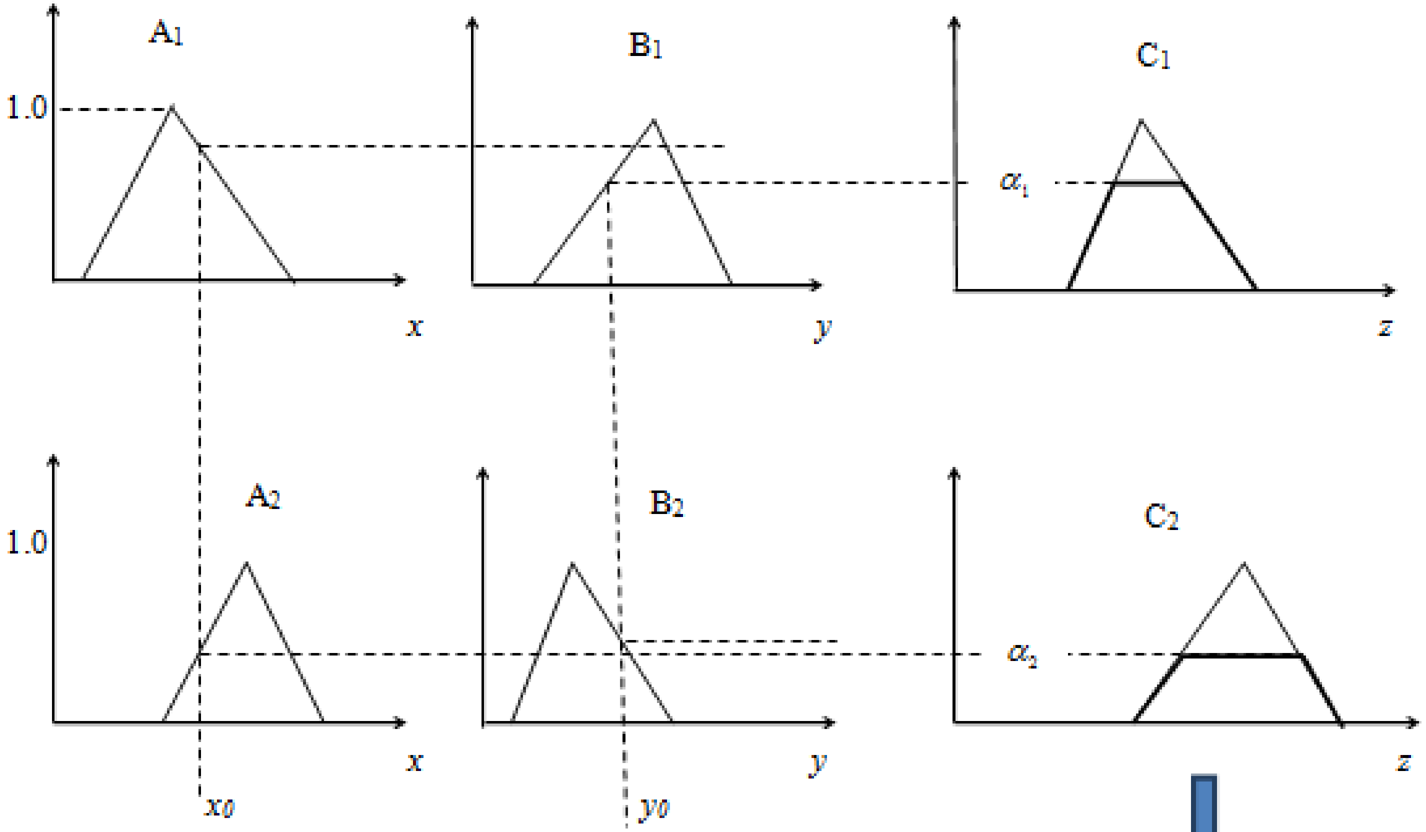
$$\alpha_1 = min(A_1(x_0), B_1(y_0)) \qquad \alpha_2 = min(A_2(x_0), B_2(y_0))$$

- The rule outputs respectively are

$$C_1'(z) = min(\alpha_1, C_1(z)) \qquad C_2'(z) = min(\alpha_2, C_2(z))$$

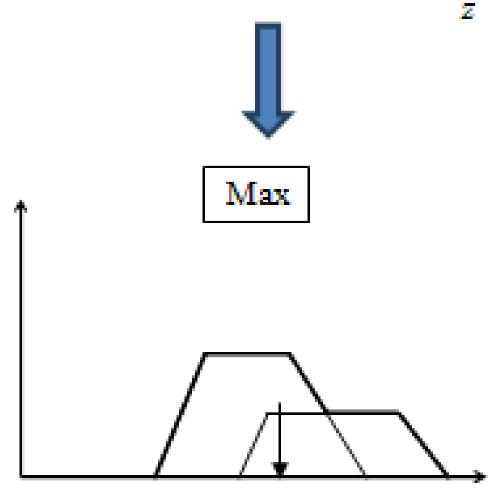
- The overall output is obtained as

$$C(z) = max(C_1'(z), C_2'(z)) = max\{min(\alpha_1, C_1(z)), min(\alpha_2, C_2(z))\}$$



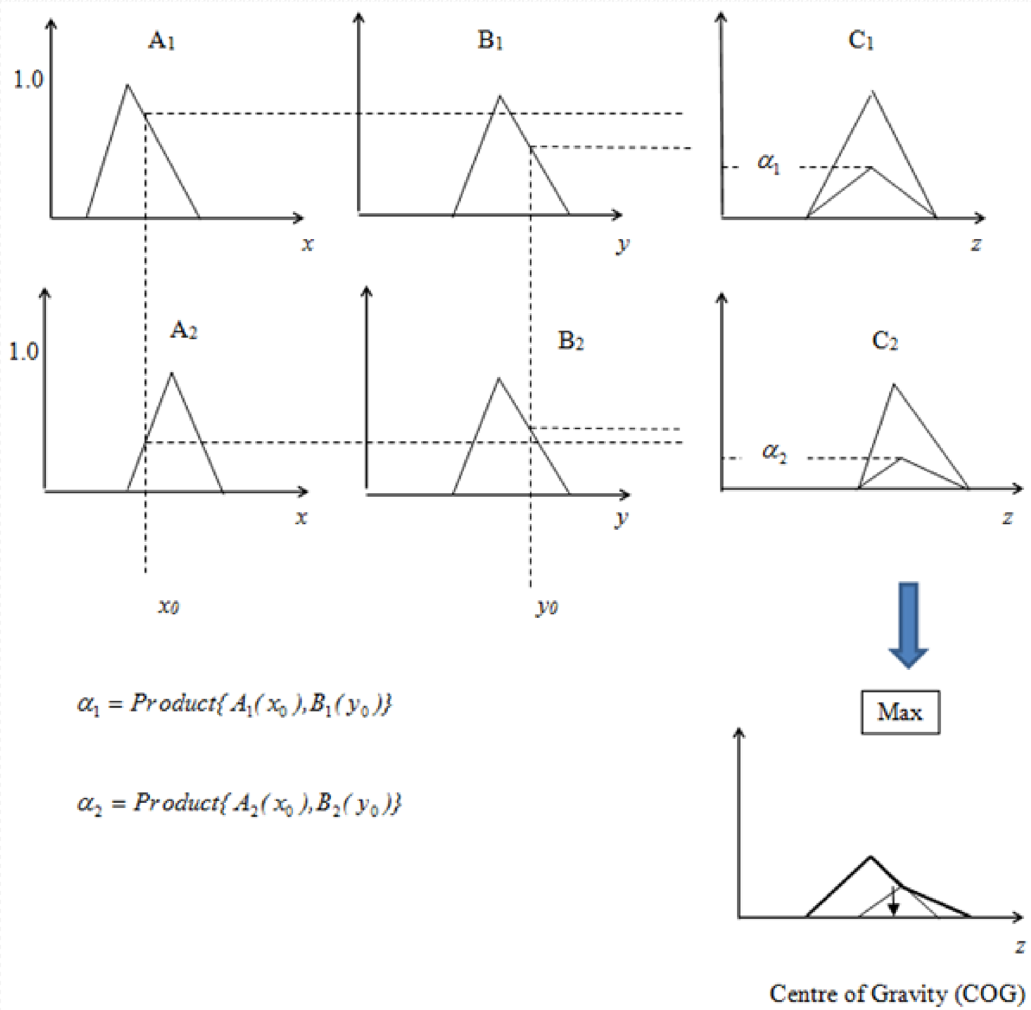
$$\alpha_1 = min(A_1(x_0), B_1(y_0))$$

$$\alpha_2 = min(A_2(x_0), B_2(y_0))$$



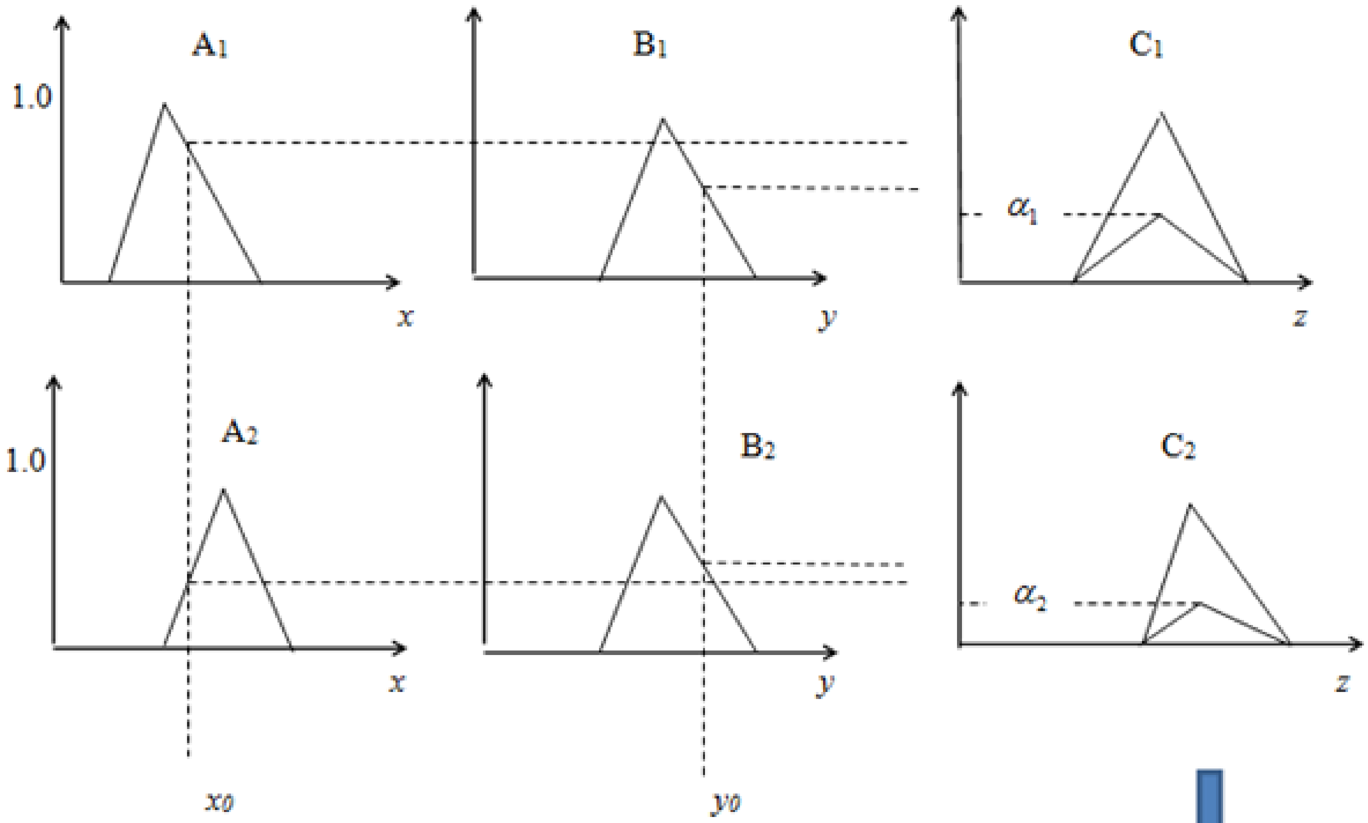
Centre of Gravity (COG)

Mamdani implication from antecedent to consequent – ‘product’ operator



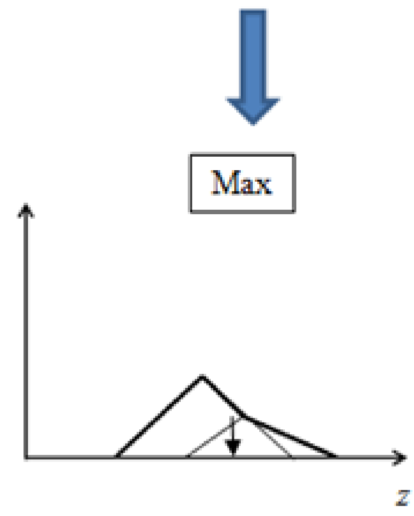
$$\alpha_1 = Product\{A_1(x_0), B_1(y_0)\}$$

$$\alpha_2 = Product\{A_2(x_0), B_2(y_0)\}$$



$$\alpha_1 = Product\{A_1(x_0), B_1(y_0)\}$$

$$\alpha_2 = Product\{A_2(x_0), B_2(y_0)\}$$



Centre of Gravity (COG)

De-fuzzification

- De-fuzzification is the last step in the fuzzy inference process since the final output of a fuzzy system has to be a crisp number.
- The input to the de-fuzzification process is the output fuzzy set from the aggregation process and the output is a single number.
- De-fuzzification is done according to the membership function of the output variable.
- There are several methods of which the centroid method is perhaps the most widely used. It returns the centre of area under the curve.
- Other methods such as the bisector of area, middle of maximum (average of the maximum value of the output set), largest maximum and smallest of the maximum can also be used.
- De-fuzzification causes some loss of information.

Takagi-Sugeno-Kang (TSK) Fuzzy Inference System

- For two rules R_1 and R_2, the TSK implication:

  R_1: IF x is A_1 AND y is B_1, THEN z_1 is $a_1x + b_1y + c_1$

  R_2: IF x is A_2 AND y is B_2, THEN z_2 is $a_2x + b_2y + c_2$

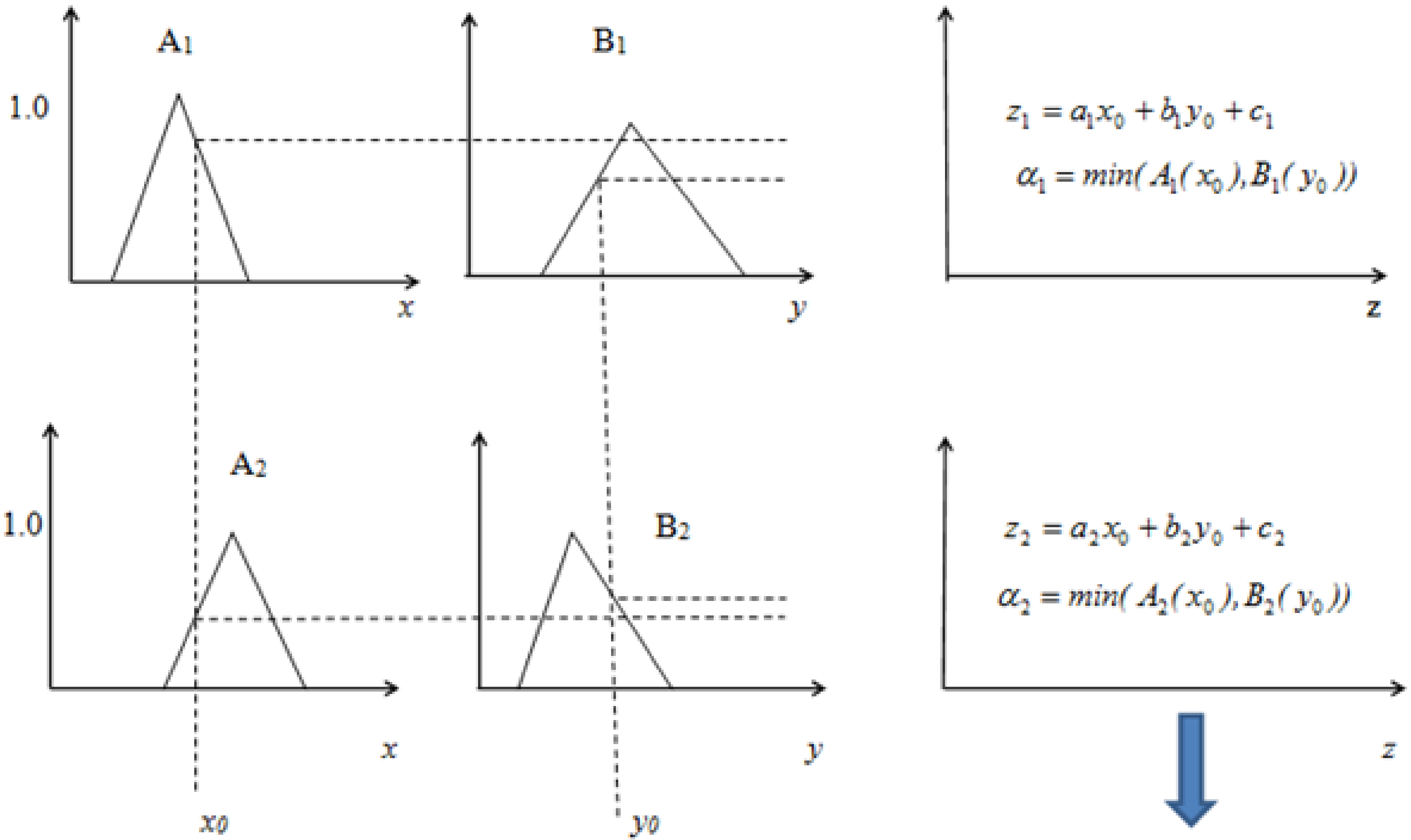
- Firing levels:

$$\alpha_1 = \min(A_1(x_0), B_1(y_0)) \qquad \alpha_2 = \min(A_2(x_0), B_2(y_0))$$

- Rule outputs:

$$z_1 = a_1x_0 + b_1y_0 + c_1 \qquad z_2 = a_2x_0 + b_2y_0 + c_2$$

- Overall output: $$z_0 = \frac{\alpha_1 z_1 + \alpha_2 z_2}{\alpha_1 + \alpha_2}$$



Weighted Average

$$z = \frac{\sum_{i=1}^{k} \alpha_i z_i}{\sum_{i=1}^{k} \alpha_i}$$

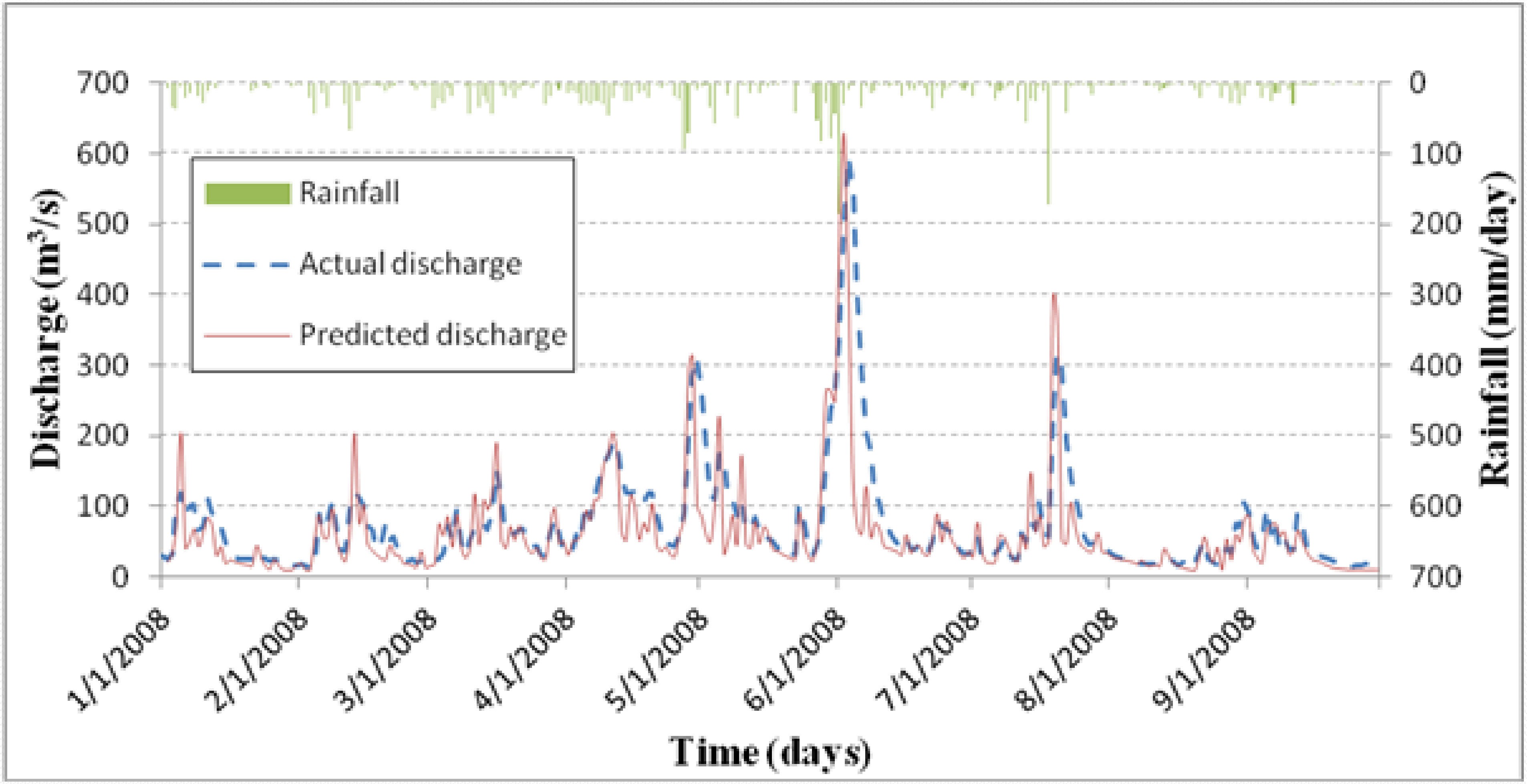
Antecedent → Consequent

$$Q_{com} = \sum_{i=1}^{N} \frac{w_i}{\sum_{i=1}^{N} w_i} \left(a_i q_1 + b_i q_2 + c_i \right)$$

Applications

- Many industrial applications (e.g. train braking system, washing machines, auto-focussing in cameras, air conditioners, photo copying machines etc.)
- Hydrological applications include
  - Water level prediction
  - Rainfall-runoff modelling
  - Infiltration modelling
  - River discharge prediction
  - Hydrological time series modelling
  - …….

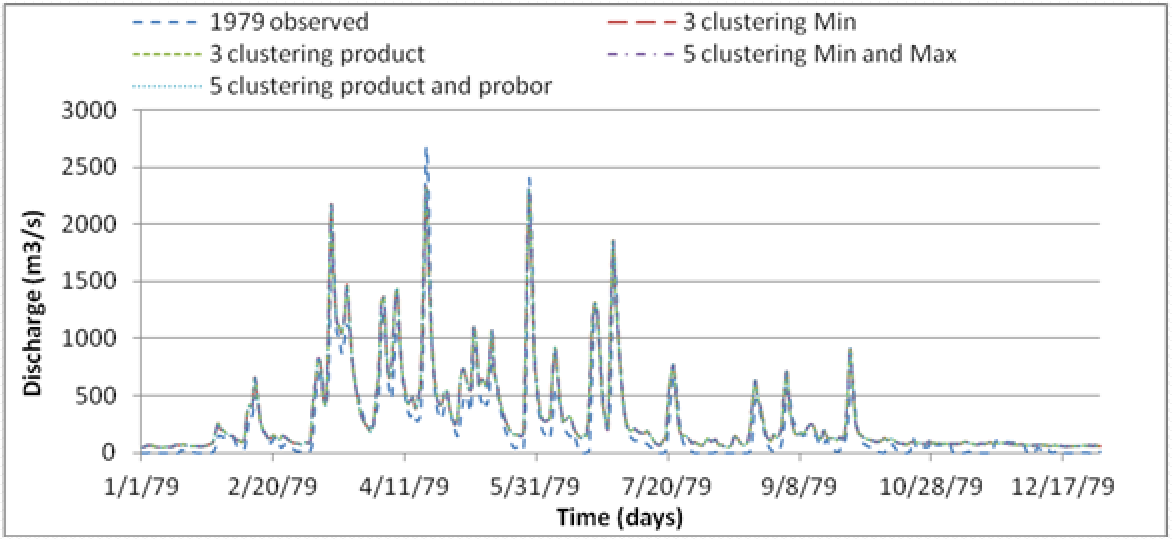
A typical result



Performance indicators

Performance Indicator	9 Rules		15 Rules		25 Rules	
	1- day lead-time	2- day lead-time	1- day lead-time	2- day lead-time	1- day lead-time	2- day lead-time
Mean Absolute Error (MAE)	35.344	34.005	27.010	29.452	24.326	25.600
Relative Root Mean Square Error(RRMSE)	0.688	0.656	0.616	0.572	0.544	0.564
Coefficient of Efficiency (EF)	0.566	0.606	0.652	0.700	0.729	0.708
Coefficient of Determination (CD)	0.808	0.806	0.904	0.903	1.211	1.122

A typical result



Performance indicators with 3 (top) and 5 (bottom) cluster centers

Performance indicator	Minimum		Product		Optimum value
	Calibration	Verification	Calibration	Verification	
MAE	84.176	84.495	84.882	84.793	0
RMSE	172.711	135.945	176.292	136.288	0
RRMSE	0.436	0.4222	0.445	0.423	0
EF	0.931	0.9468	0.928	0.947	1.0
CD	0.964	0.9602	0.957	0.968	1.0

Performance indicator	Min and Max		Product and probability		Optimum value
	Calibration	Verification	Calibration	Verification	
MAE	75.113	74.579	74.961	73.664	0
RMSE	148.517	133.085	148.028	128.143	0
RRMSE	0.375	0.413	0.374	0.398	0
EF	0.949	0.949	0.949	0.953	1.0
CD	1.054	1.032	1.053	1.031	1.0

Guiding principles and criteria for choosing a model

- Should be useful to solve or understand a particular problem under a given set conditions and constraints.
- A reasonable balance between the costs and benefits; Many models and modelling techniques add only a marginal value at an unjustifiable cost
- Resource-driven or needs-driven?
- Simple models or complex models?
- Whether the model is for a specific purpose to solve a problem or for an academic purpose for better understanding of the system
- Opinions are divided – whether it is the end result that matters or how it is obtained?

Challenges in the choice of hydrologic models – Data issues

- For simple hydrologic models, the basic input is the rainfall, which varies spatially and temporally. A reasonable spatial and temporal resolution is necessary to ensure that the data are representative.
- The second most important hydrologic variable for modelling is the discharge resulting from rainfall, which can be considered as an integrator of all catchment-scale processes.
- Rating curve – determined under normal flow conditions and often extrapolated for high flows
- There are other hydrologic processes, such as evaporation and evapo-transpiration, infiltration, interception, and depression storage, that contribute to the basin-scale hydrologic system, and their inclusion requires some approximations and assumptions while their exclusion results in over-simplification.
- Another factor that contributes to the uncertainty is the noise that is inherently present in all types of measured data.

Challenges in the choice of hydrologic models – Data issues….

- In addition to hydrologic data, geometric, topographic, geologic, and land use data are needed for distributed type of models.
- On a local scale, such data can be found in limited situations. The resolutions vary and depend upon the region and the catchment.
- On a global or regional scale, remotely-sensed topographic data are available, particularly from satellite observations. Their resolutions also vary, but the publicly available data sets do rarely have resolutions finer than 1 km x 1 km horizontally, and a few 10's of meters vertically.
- The results of any distributed model that uses such coarse data will have inherent errors of the same order or higher, than those of the input topographic data.

Challenges in the choice of hydrologic models : Modelling issues

- Conceptual, or Physics based, or Data driven ?
- Stochastic or deterministic?
- Linear or non-linear? – Linear assumption makes subsequent analysis and application simple, but in many instances, it is far from reality. Non-linear assumption makes the problem more realistic but at a cost.
- Stationary or non-stationary? – In stochastic modelling, assumption of stationarity makes analysis and application simpler, but with human influence in the hydrological system, the stationarity assumption no longer holds in many situations. For, example, the influence of climate change.

Challenges in the choice of hydrologic models : Modelling issues…

- Data driven models are relatively simple to formulate and easy to implement. Since data contain all the information about the system, their use is quite logical.
- Physics-based models consider the catchment processes from a physics point of view, but their formulation, calibration and implementation are quite resource and expertise demanding. So far, no fully physics-based model has been successfully applied to a catchment without making drastic assumptions and simplifications.
- There are also conceptual semi-distributed models that attempt to lump system characteristics on a small scale.

Challenges in the choice of hydrologic models : Scale issues

- Governing equations
  - St.-Venant equations for overland flow
  - Richard's equation for soil water flow
  - Diffusion type equation based on Darcy's law and continuity for groundwater flow
  - Horton, Philip, Green & Ampt etc. for infiltration.
  - Mass transfer, aerodynamic, or combination type equations for evaporation
- Are the governing equations valid in the scale of typical distributed models?

Challenges in the choice of hydrologic models – Parameters and their calibration issues

- Physically identifiable and measurable parameters vs. optimized parameters
- Spatially and temporally homogeneous or non-homogeneous?
- Global and local optima - Popular global search methods include population-evolution-based search strategies, such as the Shuffled Complex Evolution (SCE) algorithm (Duan et al., 1993) and Genetic Algorithm (GA) (Wang, 1991).
- Single objective function vs. multi-objective function.
- Based on the original SCE algorithm, recent studies have led to the development of the Shuffled Complex Evolution Metropolis (SCEM) and the Multi-Objective Shuffled Complex Evolution Metropolis (MOSCEM) algorithms (Vrugt et al., 2003a, 2003b).
- Direct comparison of these methods would be helpful in selecting the most suitable calibration algorithm from the extensively used shuffled complex evolution family of algorithms.

Challenges in the choice of hydrologic models : Parameters and their calibration issues

- Parsimony
- Optimized vs. trial and error adjusted
- Objective function – single vs. multi
- Local optimum vs. global optimum
- No distributed model that accounts for catchment heterogeneities and spatially varying inputs has yet been calibrated using field measured data
- Most “distributed models” which start with laws of physics end up as data driven ones calibrated using some optimization technique thereby defeating the very purpose of adopting such an approach

Challenges in the choice of hydrologic models: Parameters and their calibration issues….

- Equi-finality (Ludwig von Bertalnffy, 1968) for multi-parameter optimization
  - Same final result may be arrived at from different initial conditions and in different ways
  - Two models are said to be equi-final if they lead to equally acceptable results. It is a key concept to assess how uncertain hydrological predictions are.
  - A state *A* (a set of target parameters) is said to be Pareto optimal if there is no other state *B* dominating the state *A* with respect to a set of objective functions.
  - A state *A* dominates a state *B*, if *A* is better than *B* in at least one objective function and not worse with respect to all other objective functions.
  - There is no unique set of parameter values, but rather a feasible parameter space from which a Pareto set of optimal solutions is sought

Concluding remarks

- One should take note of the saying that "all models are wrong, but some are useful", and exercise careful judgment in choosing a model or a modelling approach for a specific purpose.

Thank you
for your attention!