Machine learning in building models of models' uncertainty

D.P. Solomatine,
D.L. Shrestha, N. Kayastha,

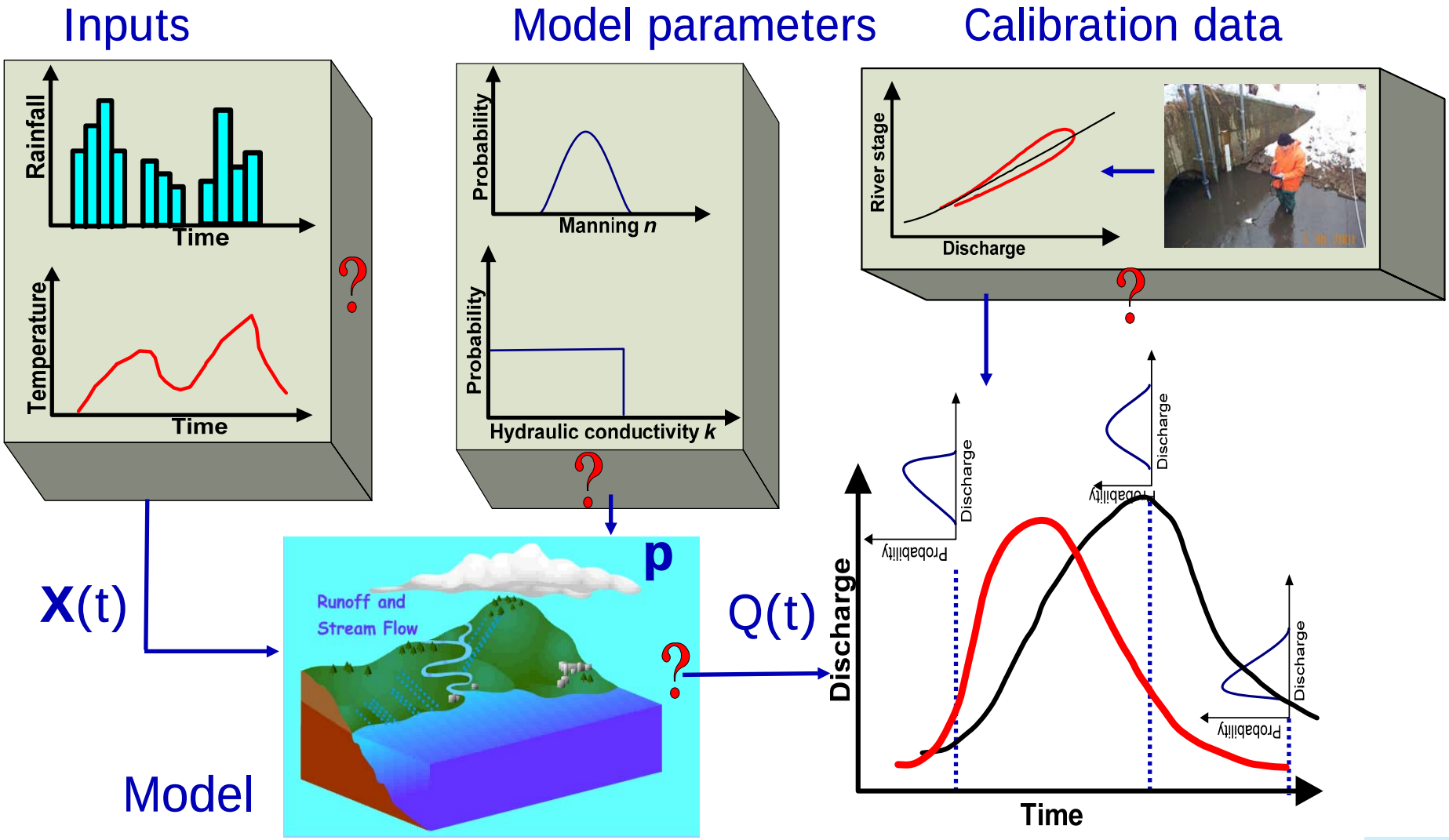
UNESCO-IHE
Institute for Water Education

Outline

- Introduction
- Methodology
- Case study
- Experimental results
- Conclusions
- Future works

Introduction

Sources of uncertainty in hydrological modelling



Uncertainty analysis methods considered

- 1. UNNEC = machine learning model of the *past errors of the optimal process model* is built

- 2. MLUE = machine learning model of the *process model's Monte Carlo simulation results* is built

1. UNEEC method

UNcertainty Estimation based on local Errors and Clustering

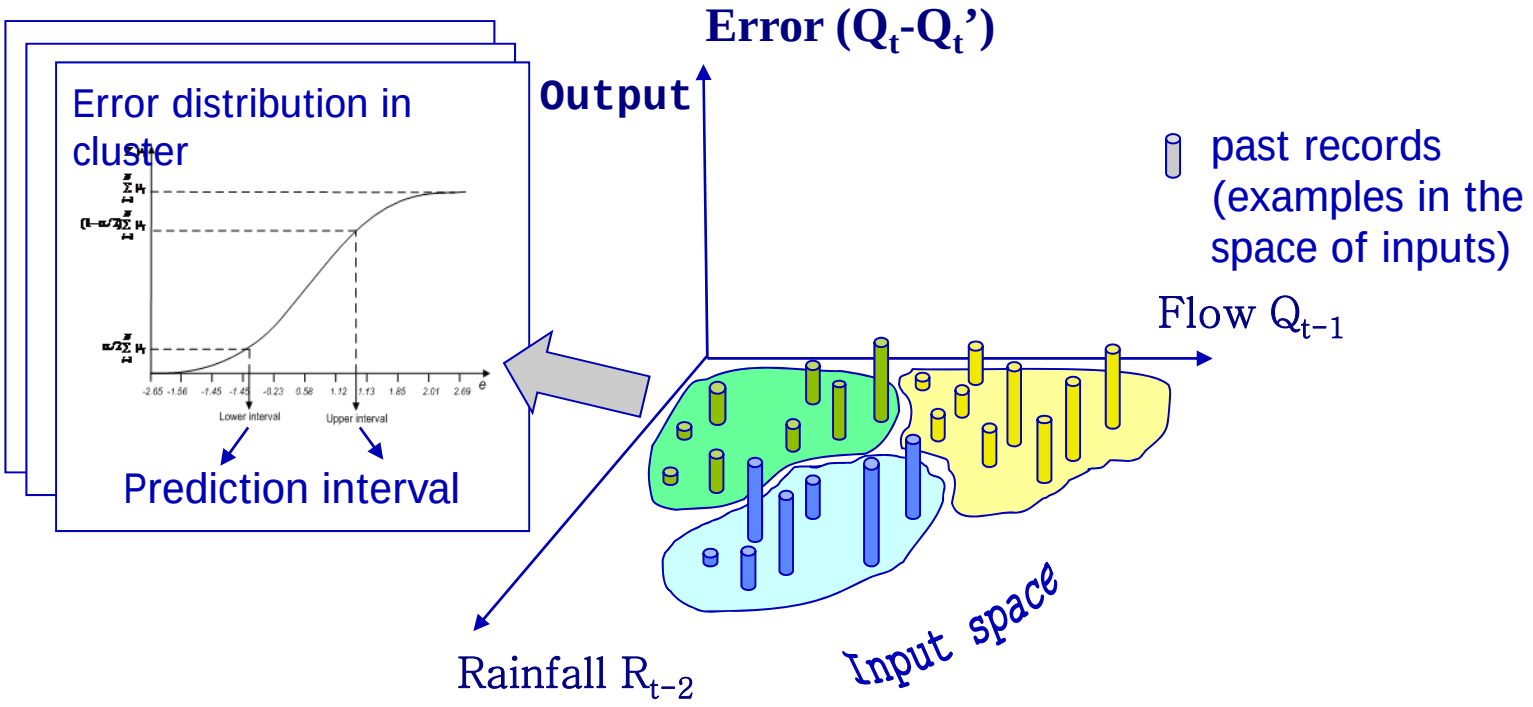
- machine learning model of the *past errors of the optimal process model* is built

D.P. Solomatine, D.L. Shrestha. A novel method to estimate model uncertainty using machine learning techniques. *Water Resources Res*. 45, W00B11, doi:10.1029/2008WR006839, 2009.

UNEEC: assumptions, constraints

- Assumptions
  - Model error is an indicator of the model uncertainty
  - Model error depends on the current condition of a natural system and can be predicted
  - Model errors are similar for similar conditions
- Constraints
  - Model structure and parameters are fixed
  - Need to re-train the error model with the changes in the catchment characteristics (e.g. land use change)
  - Data hungry, more data are needed for reliable results

Idea 1: local modelling of errors

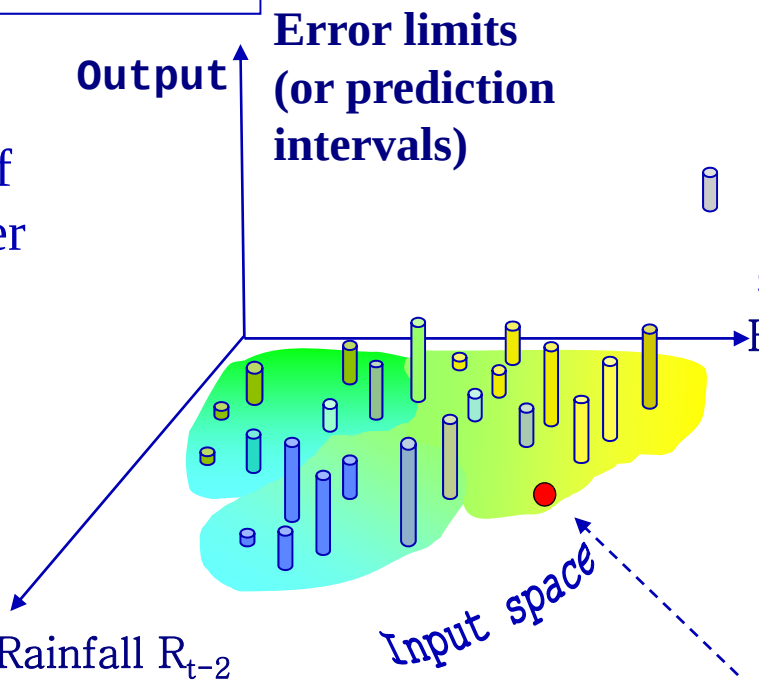


Idea 2: Use fuzzy clustering of examples to generate training data sets

$$PI^{L}_{example} = \sum_{clus=1}^{Nclus} \mu_{clus,\,example}\; PIC^{L}_{clus}$$

- $\mu_{clus,,example}$ is the membership grade of the *example* to cluster *clus*

Eager learning (ANN or M5 model tree)



past records (examples in the space of inputs)

Train regression (ANN) models:

$PI^{L} = f_{L}(X)$

$PI^{U} = f_{U}(X)$

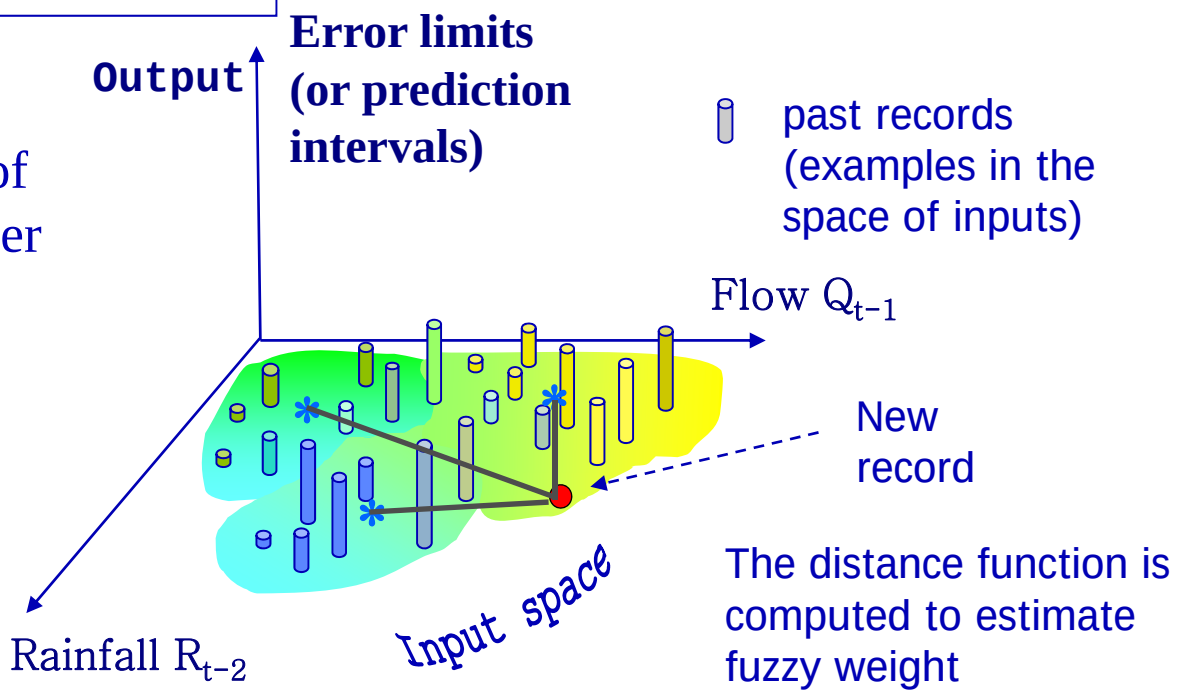
New record. The trained f^{L} and f^{U} models will estimate the prediction interval

Using instance-based learning

$$PI^{L}_{example} = \sum_{clus=1}^{Nclus} \mu_{clus,\,example}\; PIC^{L}_{clus}$$

- $\mu_{clus,,example}$ is the membership grade of the *example* to cluster *clus*

Instance based learning



The distance function is computed to estimate fuzzy weight

UNEEC details. Step 1: clustering

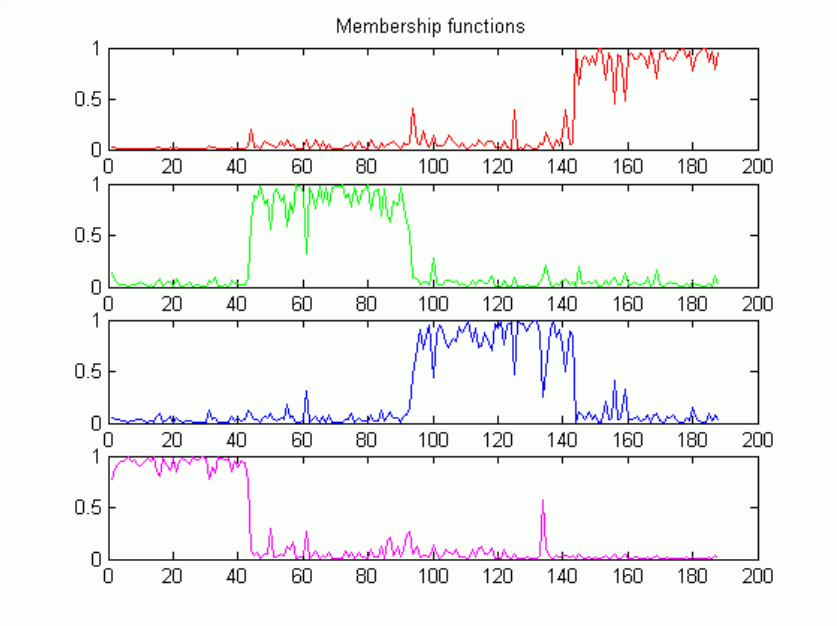
- Clustering (finding groups of data in the space characterising hydro-meteo condition): K-means clustering, fuzzy C-means clustering

Obj. function $$\min(U,V)\left\{J_m(U,V)=\sum_{j=1}^{c}\sum_{i=1}^{N}\mu_{i,j}^{m}D_{i,j}^{2}\right\}$$

Distance $$D_{i,j}^{2}=\left\|x_i-v_j\right\|_A^2$$

Degree of Fuzzification $$m\geq 1$$

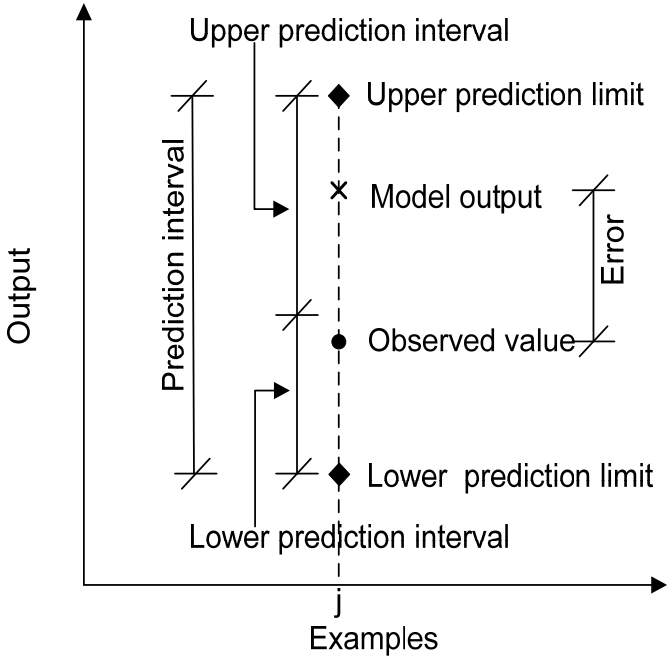
Constraint $$\sum_{j=1}^{c}\mu_{i,j}=1,\forall i$$

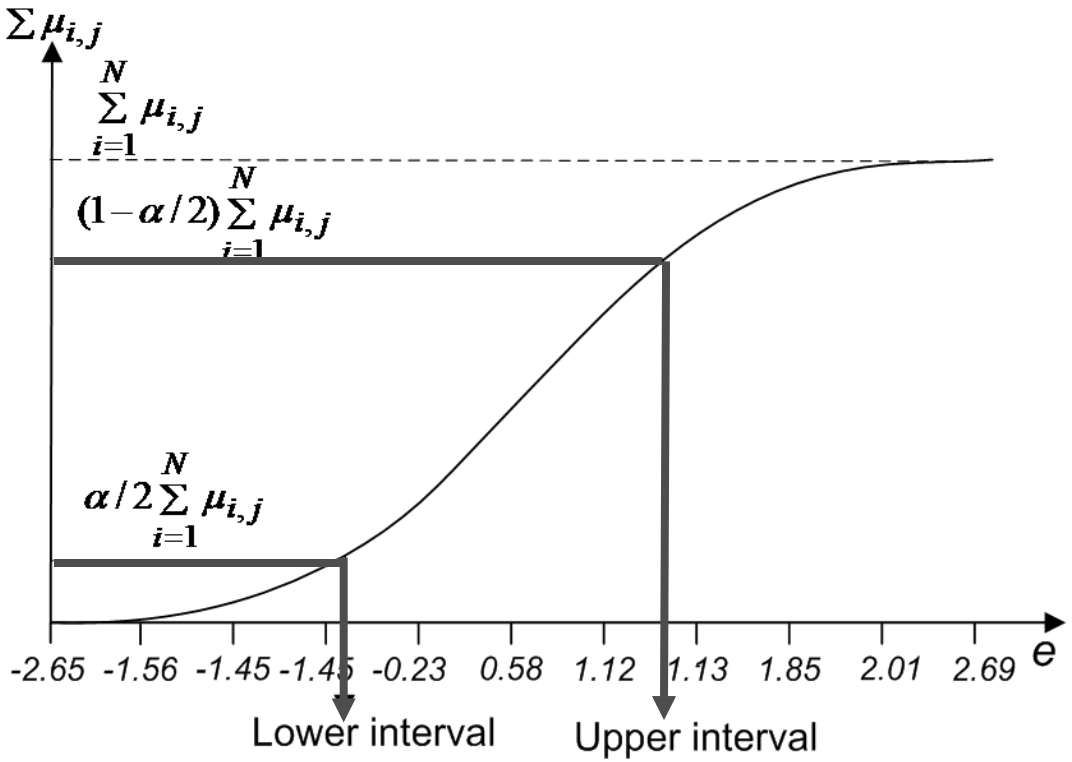


UNEEC details. Step 2: Determining Prediction Interval (PI) for each cluster

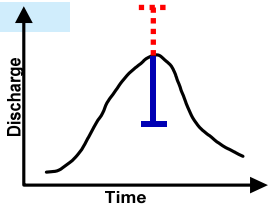
$$PIC_j^L = e_i \Rightarrow i : \sum_{k=1}^{i} \mu_{i,j} < \alpha/2 \sum_{i=1}^{N} \mu_{i,j}$$

$$PIC_j^U = e_i \Rightarrow i : \sum_{k=1}^{i} \mu_{i,j} < (1-\alpha/2) \sum_{i=1}^{N} \mu_{i,j}$$





UNEEC details.
Step 3, 4, 5: Building and using the model



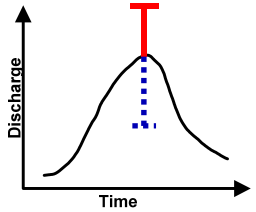
$$PI_i^L = \sum_{j=1}^{c} \mu_{i,j}\, PIC_j^L$$

$$PI^L = f_u^L(\mathrm{X_u})$$

$$PI^L = f_u^L(\mathrm{X_v})$$

$$PL_i^L = \underbrace{\hat{y}_i + PI_i^L}$$

Independent Computation



$$PI_i^U = \sum_{j=1}^{c} \mu_{i,j}\, PIC_j^U$$

$$PI^U = f_u^U(\mathrm{X})$$

$$PI^U = f_u^U(\mathrm{X_v})$$

$$PL_i^U = \underbrace{\hat{y}_i + PI_i^U}$$

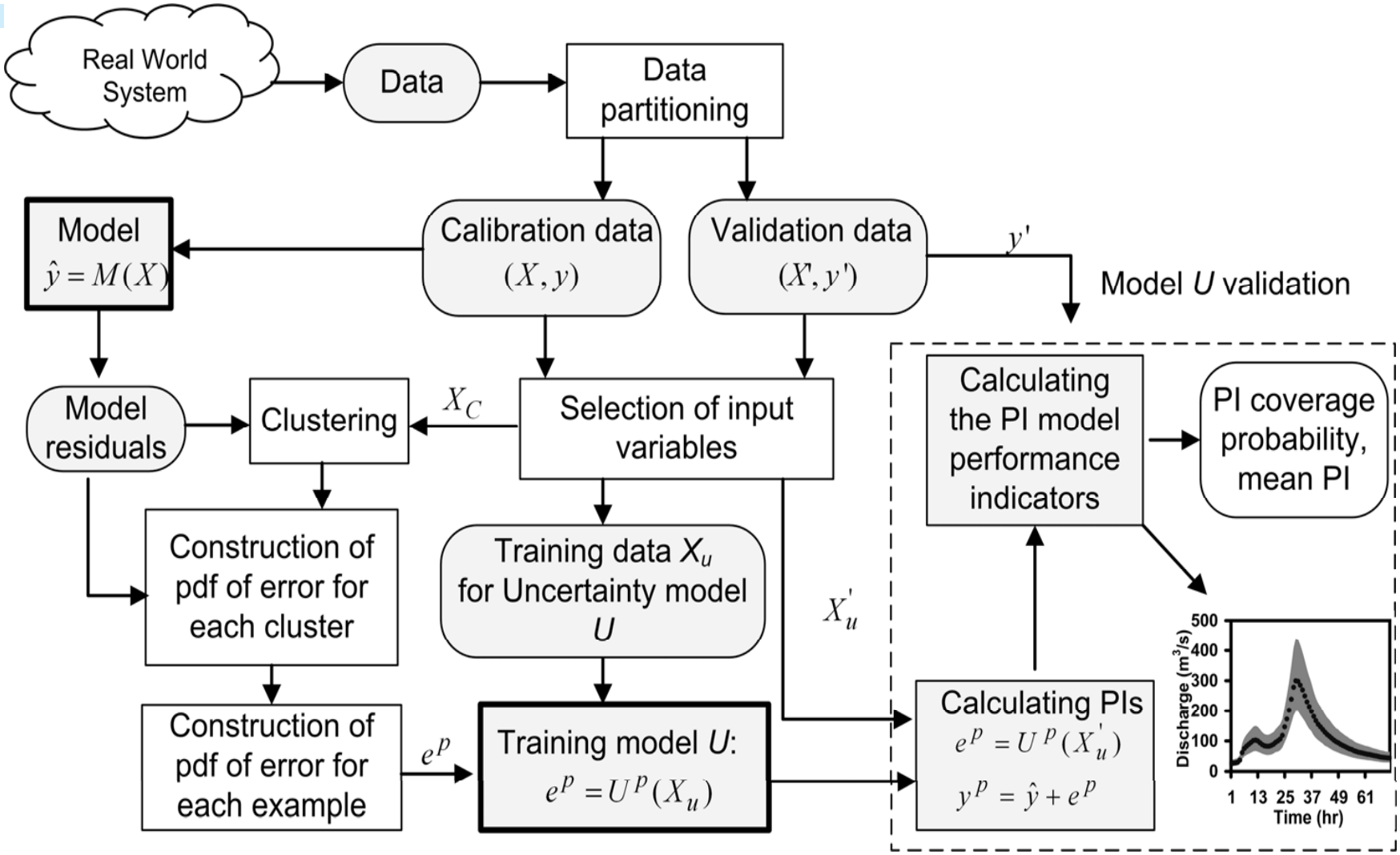
Step 3: Generation of Prediction intervals for each example

Step 4: Building the uncertainty Model

Step 5: Using the uncertainty Model

Model Outputs with uncertainty bounds

UNEEC methodology



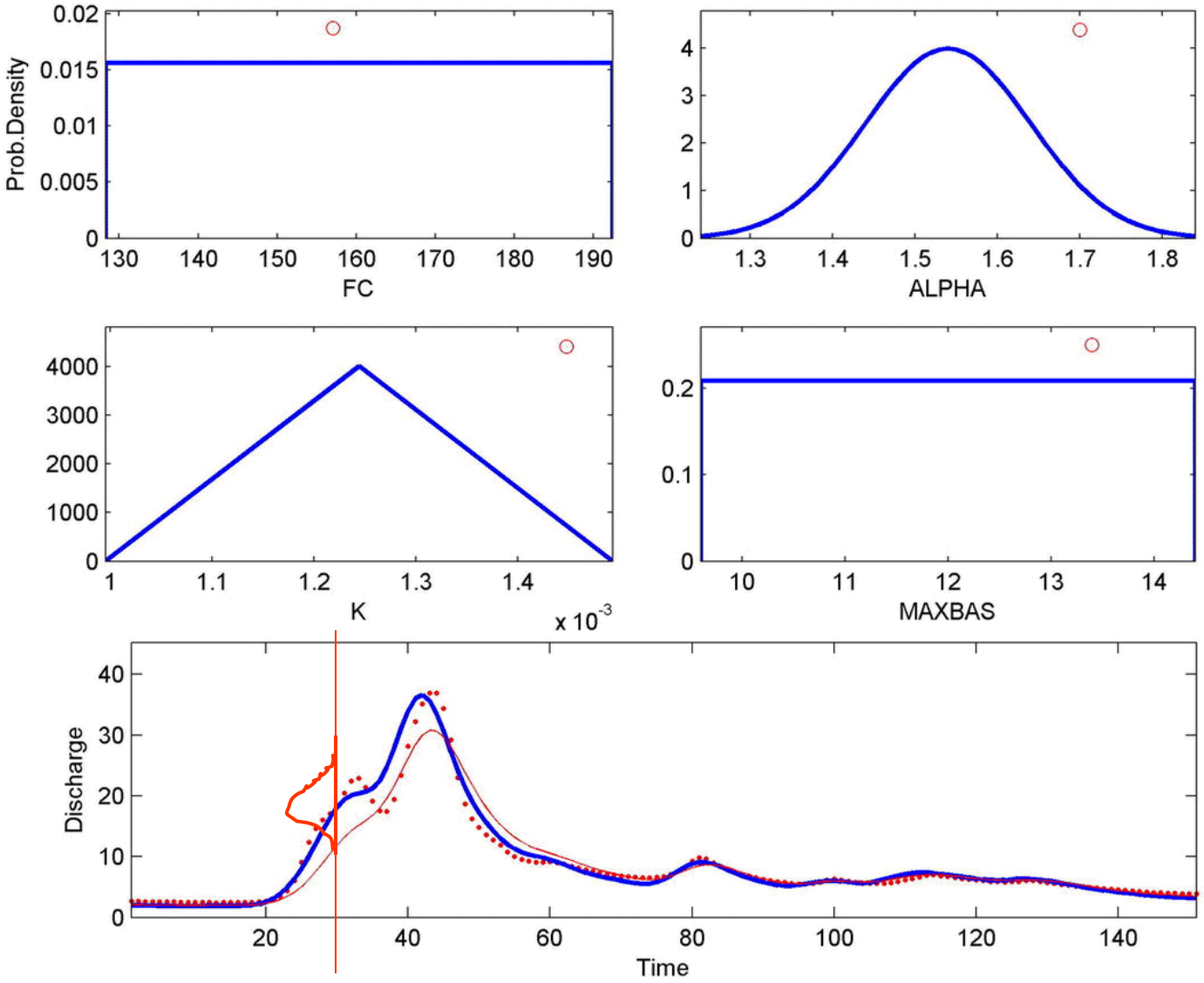
2. MLUE method

Machine Learning in Uncertainty Estimation

- machine learning model of the *process model's Monte Carlo simulation results* is built

D. L. Shrestha, N. Kayastha, and D. P. Solomatine. A novel approach to parameter uncertainty analysis of hydrological models using neural networks. *Hydrol. Earth Syst. Sci.*, 13, 1235–1248, 2009.

Monte Carlo simulation of parametric uncertainty



Monte Carlo simulation of parametric uncertainty

- Consider the model M calculating y (e.g., discharge)
  - $y(t) = M(\mathbf{X}(t), \mathbf{p})$
  - where $\mathbf{X}(t)$ = vector of inputs (precipitation, temperature etc) known for $t = 1,..., T$
  - p = vector of parameters (soil properties, roughness, etc)
- Monte Carlo approach:
  - sample N parameter vectors $\mathbf{p}_i$
  - run the model for each of them $y_i(t) = M(\mathbf{X}(t), \mathbf{p}_i)$ and generate N outputs (leave some of them if GLUE used)
  - assess distribution of $Q_i(t)$ for each time moment t (or its parameters - mean, variance, prediction intervals, quantiles)
- The problem:
  - How to assess the parametric uncertainty of the model M for $t = T+1$ when new input data $\mathbf{X}(t+1)$ is fed?

Issues with MC

- Issues with re-running MC for new inputs:
  - 1) convergence of the Monte Carlo simulation is very slow (O(N^-0.5)) so larger number of runs needed to establish a reliable estimate of uncertainties
  - 2) number of simulation increases exponentially with the dimension of the parameter vector ((O(n^d)) to cover the entire parameter domain
- Idea:
  - encapsulate the results of MC simulation in a machine learning model

MLUE Methodology

Methodology (1)

- Consider the sources of the uncertainty analysis to be conducted within the framework of Monte Carlo simulation
- Execute the MC simulations to generate the data

$$y_i(t) = M(\mathbf{X}(t), \mathbf{p}_i)$$

- Estimate the uncertainty measures of the MC realizations, e.g., mean, variance, prediction intervals, quantiles
  - In this study, we use two quantiles (say, 5% and 95%), forming the prediction interval PI

Methodology (2)

- Analyze the dependency of the uncertainty measures (quantiles) on the *input and state variables* of the hydrological model
  - we used Correlation and Average mutual information analysis
- Select the input variables for machine learning model based on the dependency analysis
- Train the *machine learning model U* to predict the uncertainty measures of MC realizations $PI = U(\mathbf{X})$
- Validate machine learning model *U* by estimating the uncertainty measures with the “new” input data

Validation

- Measuring predictive capability of uncertainty model *U* (measures the accuracy of uncertainty models in approximating the quantiles of the model outputs generated by MC simulations)
  - Coefficient of correlation (r) and root mean squared error (RMSE)
- Measuring the statistics of the uncertainty estimation (i.e. goodness of the model U as uncertainty estimator)
  - Prediction interval coverage probability (PICP) and mean prediction interval (MPI) (Shrestha & Solomatine 2006, 2008)

$$PICP = \frac{1}{n}\sum_{t=1}^{n} C$$

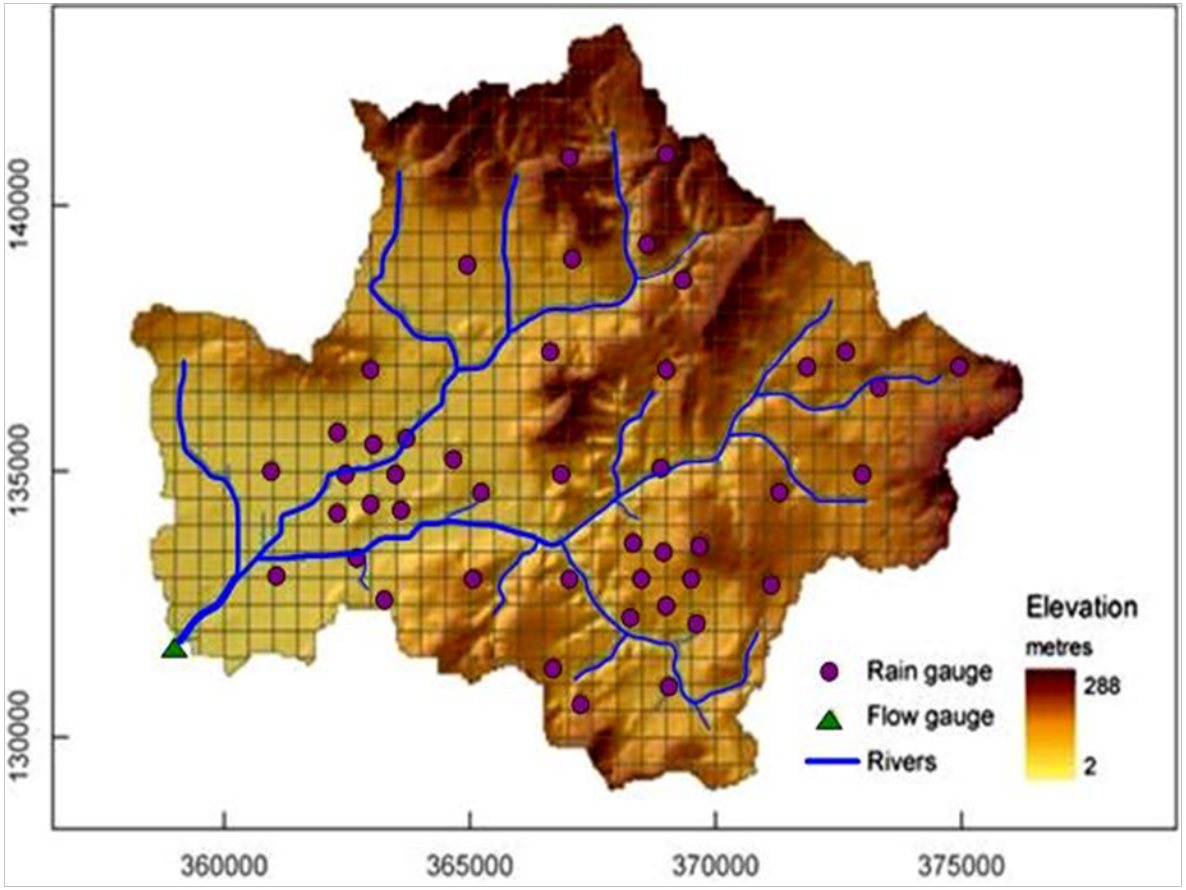
$$\text{with } C = \begin{cases} 1, & PL_t^L \le y_t \le PL_t^U \\ 0, & \text{otherwise} \end{cases}$$

$$MPI = \frac{1}{n}\sum_{t=1}^{n}(PL_t^U - PL_t^L)$$

- Visualizing such as scatter and time plot of the prediction intervals obtained from the MC simulation and their predicted values

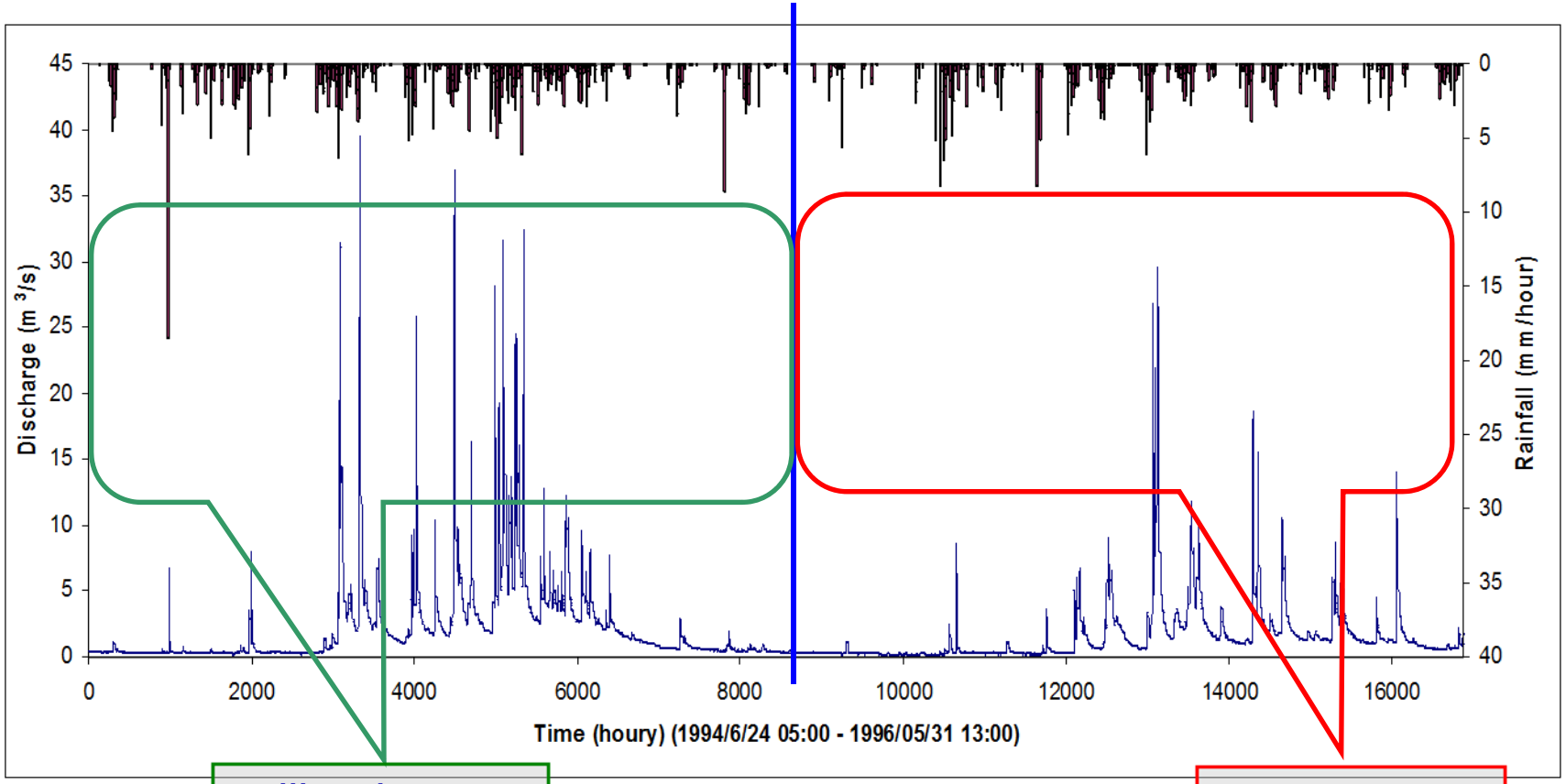
Application

Study area: Brue catchment, UK



- Catchment area: 135 km^2
- Location: south west of England
- Average annual rainfall: 867 mm
- Mean river flow: 1.92 m3/s
- Calibration data: 24/06/94-24/06/95
- Validation data: 24/06/95-31/05/96

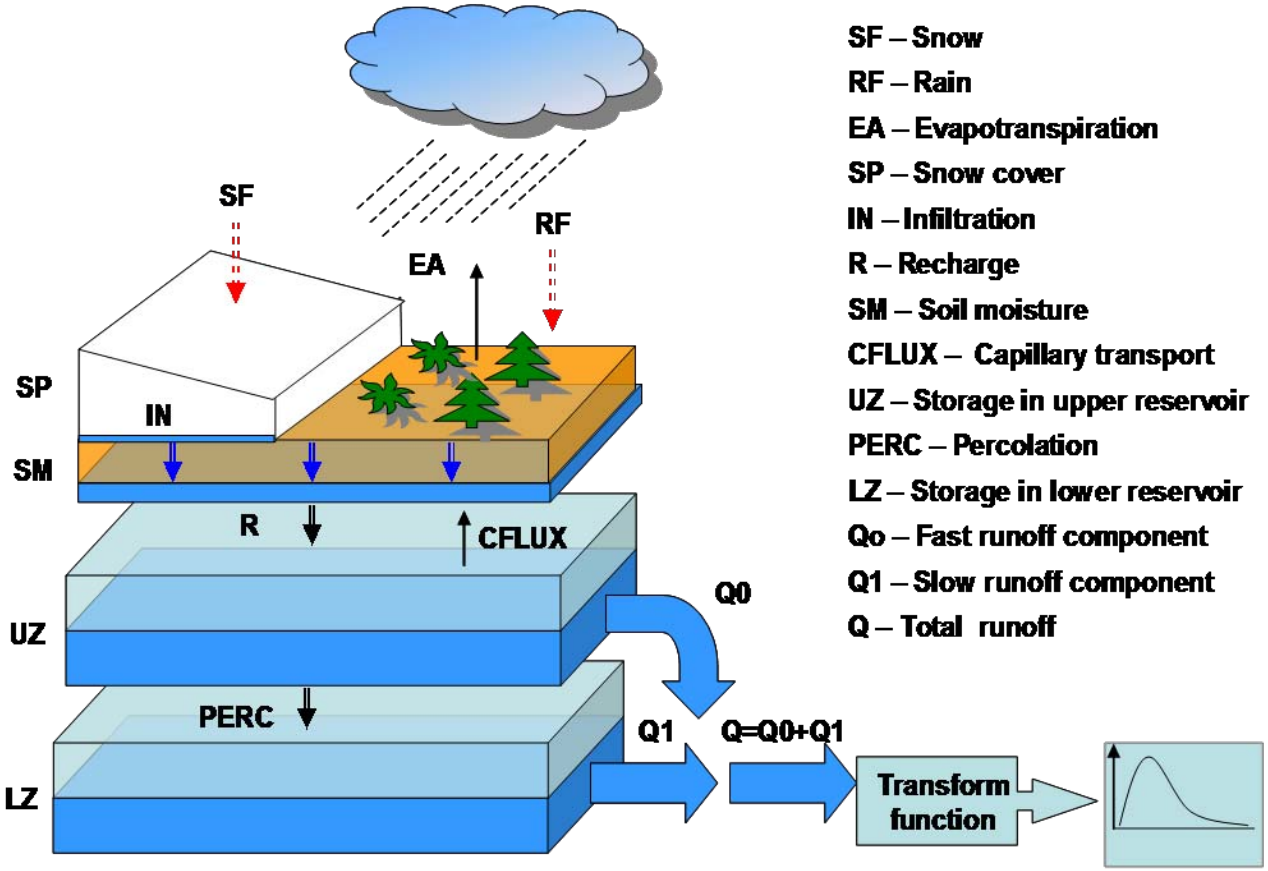
Study area: Brue catchment, UK



Calibration data (8760 points):

Validation data (8217 points):

Conceptual Hydrological model HBV



Data Analysis

- Analysis of dependency btw various combinations of the input variables and the output
  - Correlation
  - Average mutual information (AMI) between *REt* and PIs,

    (optimal lag time is around 7-9 hours).
  - Additional analysis of the correlation and AMI between the PIs and observed discharge *Qt* are carried out. (i.e. with the lag of 0, 1, 2) have very high correlation with the PIs.

Experimental setup

- MC simulation (MLUE)
  - 9 Parameters of HBV model are sampled uniformly from the feasible ranges
  - Nash-Sutcliffe coefficient of efficiency (CE) is used as error measure
  - Convergence – stabilized after 10,000 (75,000 runs made)
  - Only 25,000 “good” models considered (rejection threshold is set to 0) to compute prediction quantiles

Experimental setup

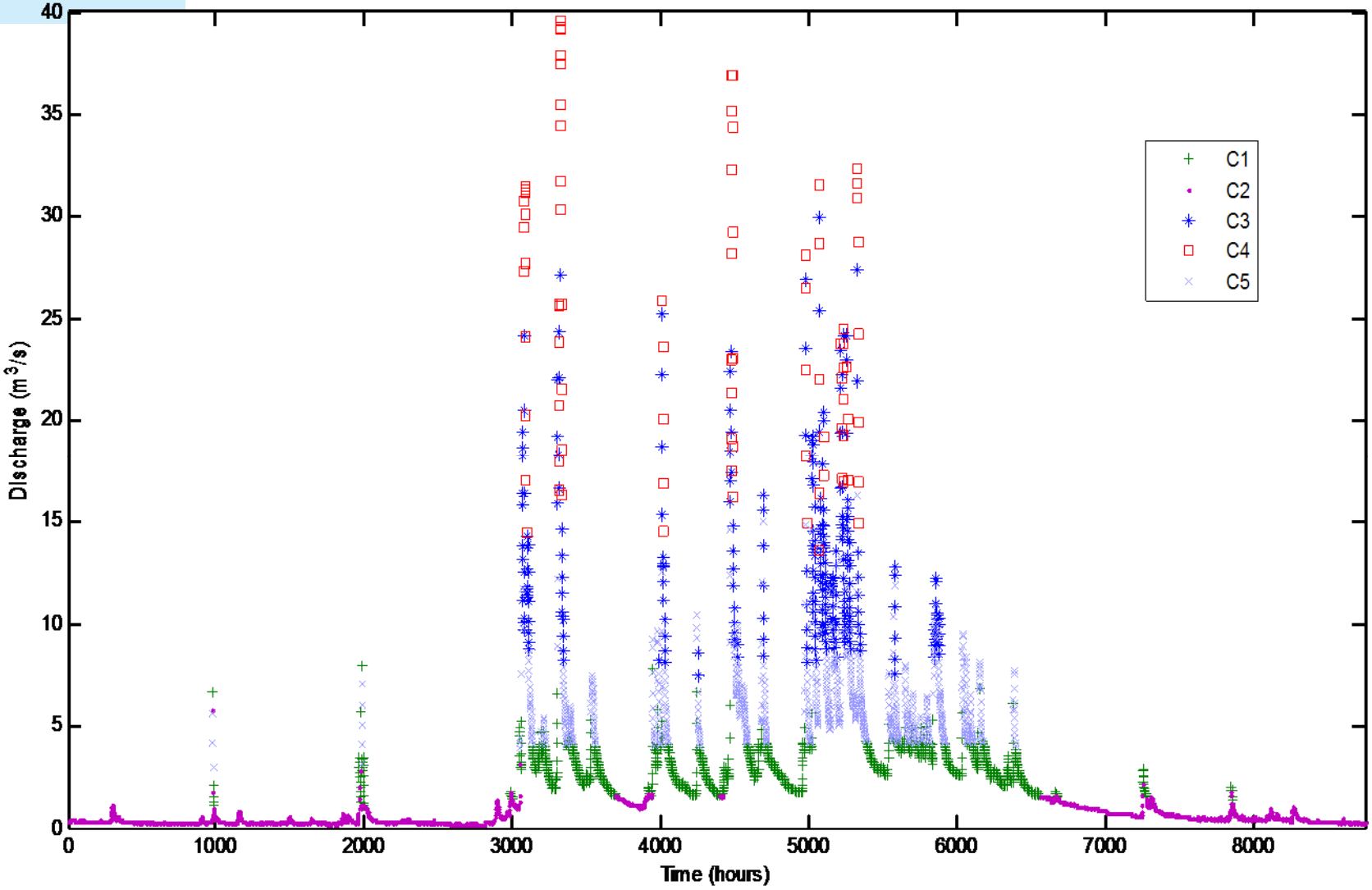
- Machine learning model U (MLUE)
  - $PI = U(RE_{t-5a}, Q_{t-1}, \Delta Q_{t-1})$
    - PI - lower or upper prediction intervals,
    - RE_{t-5a} - average of RE_{t-5}, RE_{t-6}, RE_{t-7}, RE_{t-8}, and RE_{t-9}
    - $\Delta Q_{t-1} - Q_{t-1} - Q_{t-2}$.
  - Input variables were selected based on the analysis of their relatedness to output error (average mutual information)

$$\mathrm{AMI} = \sum_{i,j} P_{XY}(x_i, y_j) \log_2 \left[\frac{P_{XY}(x_i, y_j)}{P_X(x_i) P_Y(y_j)} \right]$$

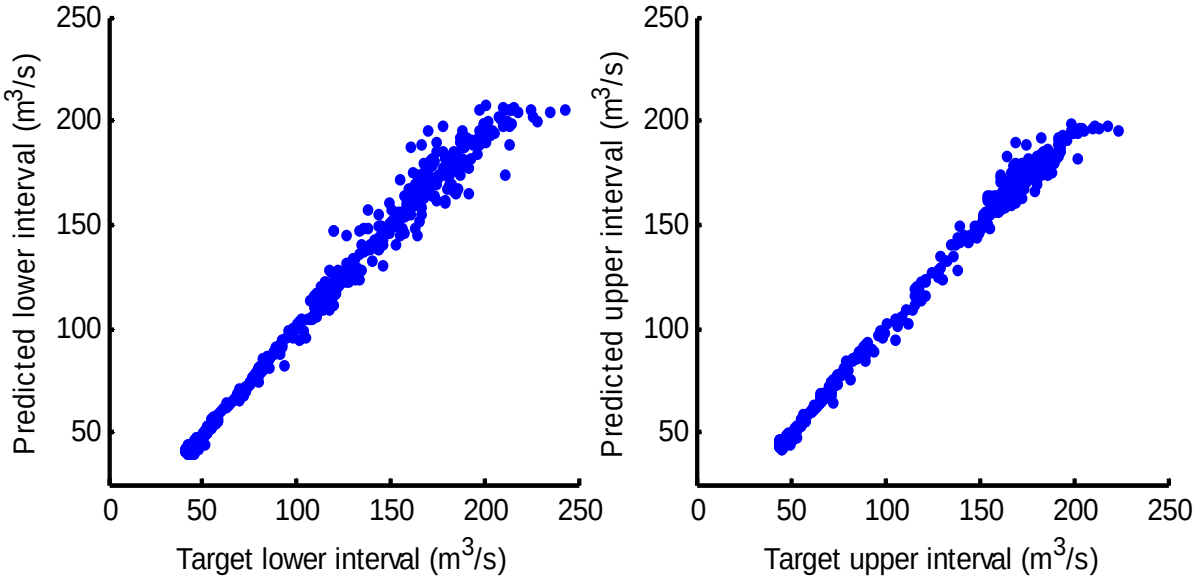
- Methods:
  - M5 model trees,
  - locally weighted regression
  - MLP neural networks

Results

UNEEC: Clustering result example

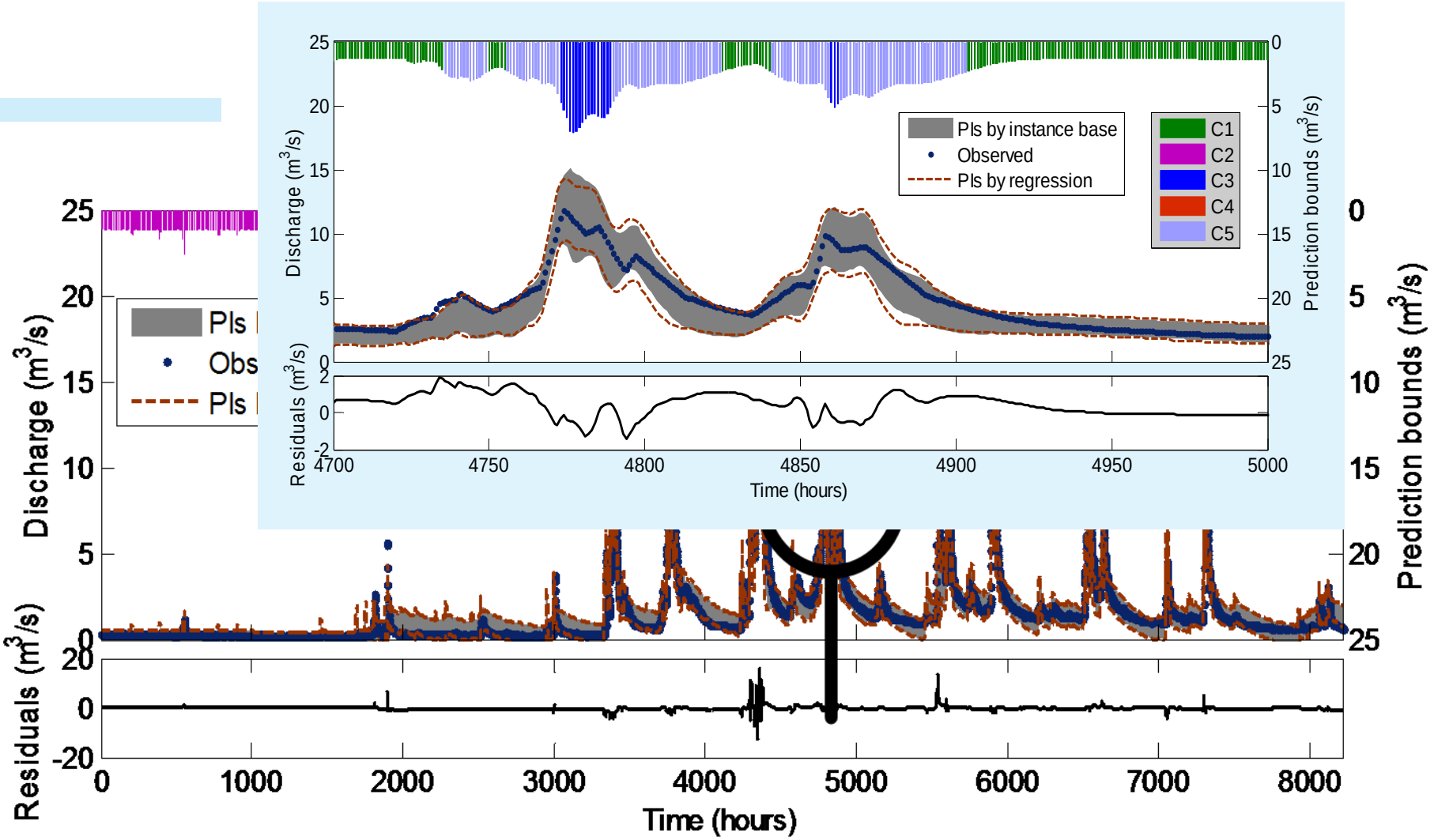


UNEEC: Performance (MLP ANN)

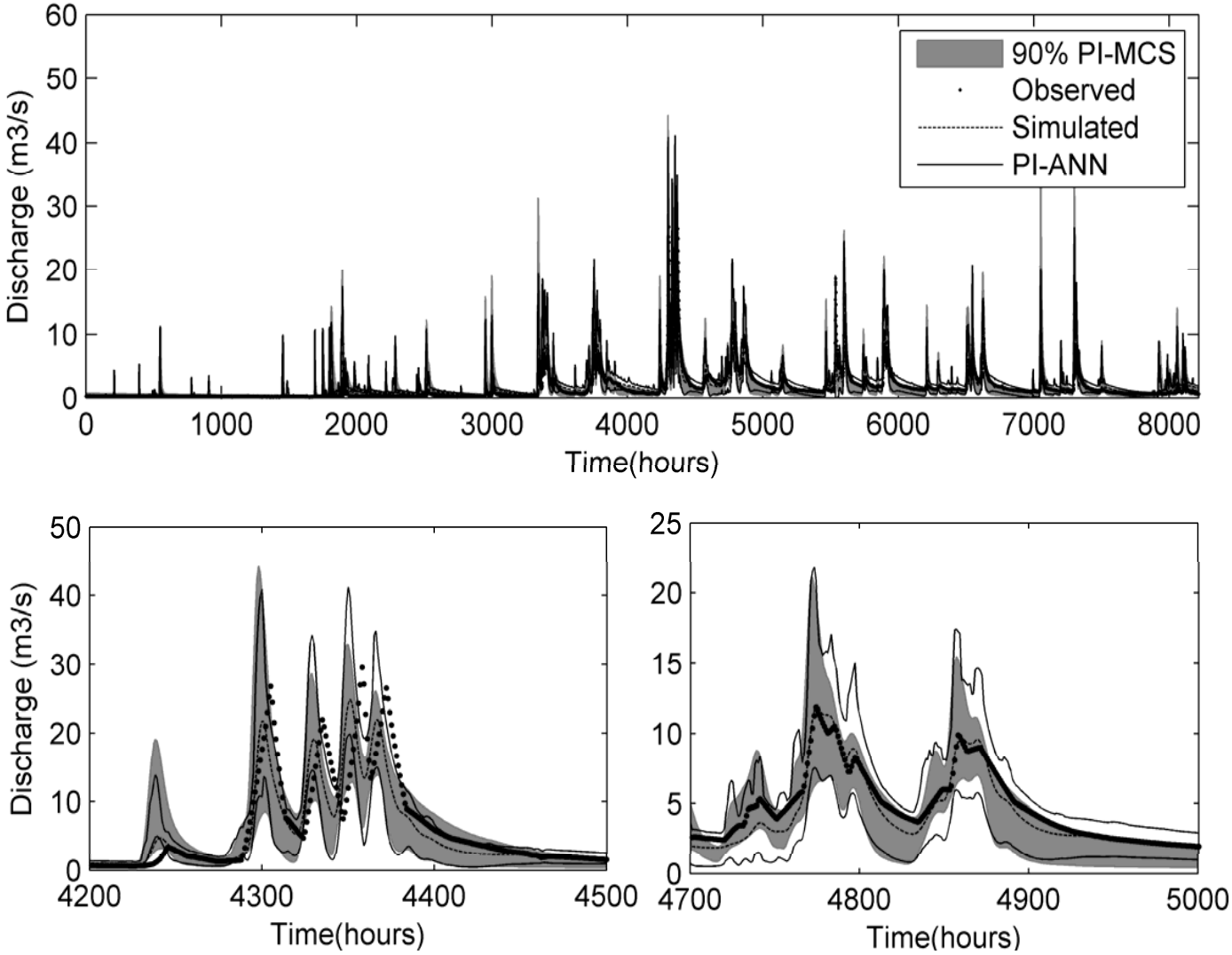


Prediction interval	Data set	Mean	Std. dev.	RMSE	Corr. coef.
lower	training	110.91	53.6	5.9582	0.9937
	CV	112.18	52.64	6.0852	0.9934
	training+CV	111.35	53.32	5.9582	0.9937
upper	training	115.16	55.11	3.9002	0.9975
	CV	116.69	54.18	3.9332	0.9974
	training+CV	115.66	54.79	3.9002	0.9975

UNEEC: Estimation of uncertainty bounds



MLUE: Estimation of prediction intervals



MLUE: Performances

- *Predictive capability*

	Corr C		RMSE	
	PI^L	PI^U	PI^L	PI^U
MT	**0.841**	**0.792**	**0.614**	**1.641**
LWR	**0.822**	**0.798**	**0.643**	**1.604**
ANN	**0.847**	**0.806**	**0.584**	**1.568**

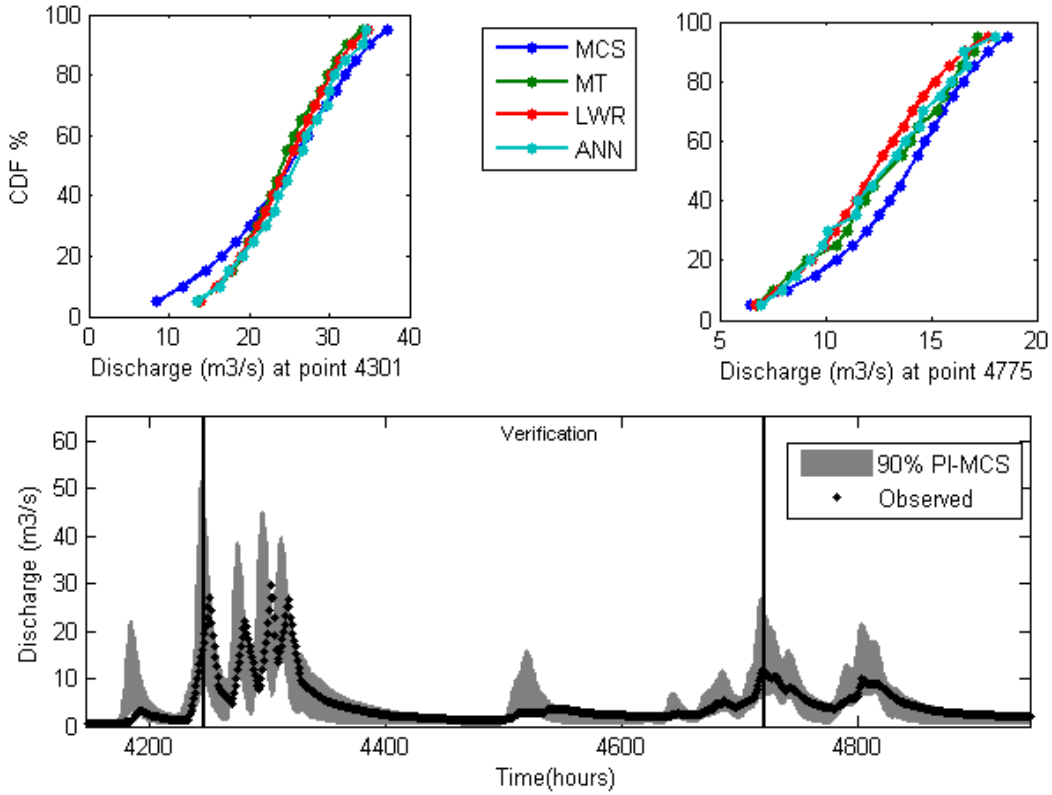
- *Goodness of uncertainty measures*

	MCS	MT	LWR	ANN
PICP %	77.24	66.97	**75.16**	65.54
MPI m^3/s	2.09	**2.03**	1.93	1.96

MCS = Monte Carlo
MT = M5 Model tree
LWR = local weighted regression
ANN =MLP neural network

Extensions

- Estimation of several quantiles 5%, 10%:10%:90%, 95%
  - i.e. estimating cdf of MC realizations by machine learning models



Use of Machine learning methods: conclusions

- Machine learning methods are able to replicate:
  - Past performance of a process model
  - Results of Monte-Carlo simulations
- The methods are computationally efficient and can be used in real time application of various kinds
- The results demonstrate that the interpretable uncertainty estimates are generated
- *Future work:*
  - Other ML methods are to be tested
  - The methods can be applied in the context of other sources of uncertainty - input, structure, or combined

Advertisements...

Welcome to our SHORT COURSES on
Flood Risk Management (June, 3 weeks)
New data sources for flood modelling (September, 1 week)
and others

www.unesco-ihe.org/education

UNESCO-IHE
Institute for Water Education