Automatic measured data validation applied on hydraulic and water quality parameters in sewer systems

Nemanja Branisavljević[1]
Dušan Prodanović[1]
Zoran Kapelan[2]

[1]University of Belgrade
[2]University of Exeter

Research challenges

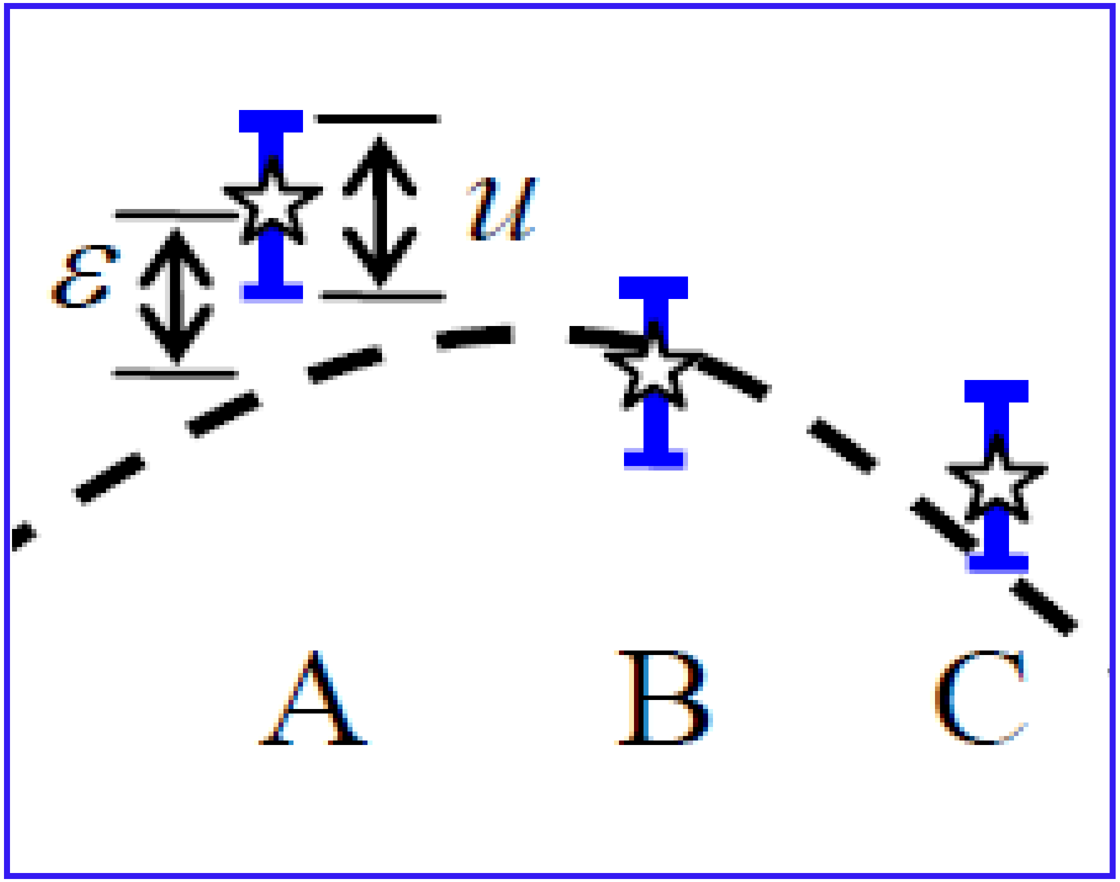
Challenge 1:
Large amount of data can not be validated traditionally

Challenge 2:
Data validation procedure have to be capable to use any kind of information available

Challenge 3:
Data validation procedure have to be independent on type or number of relations between data

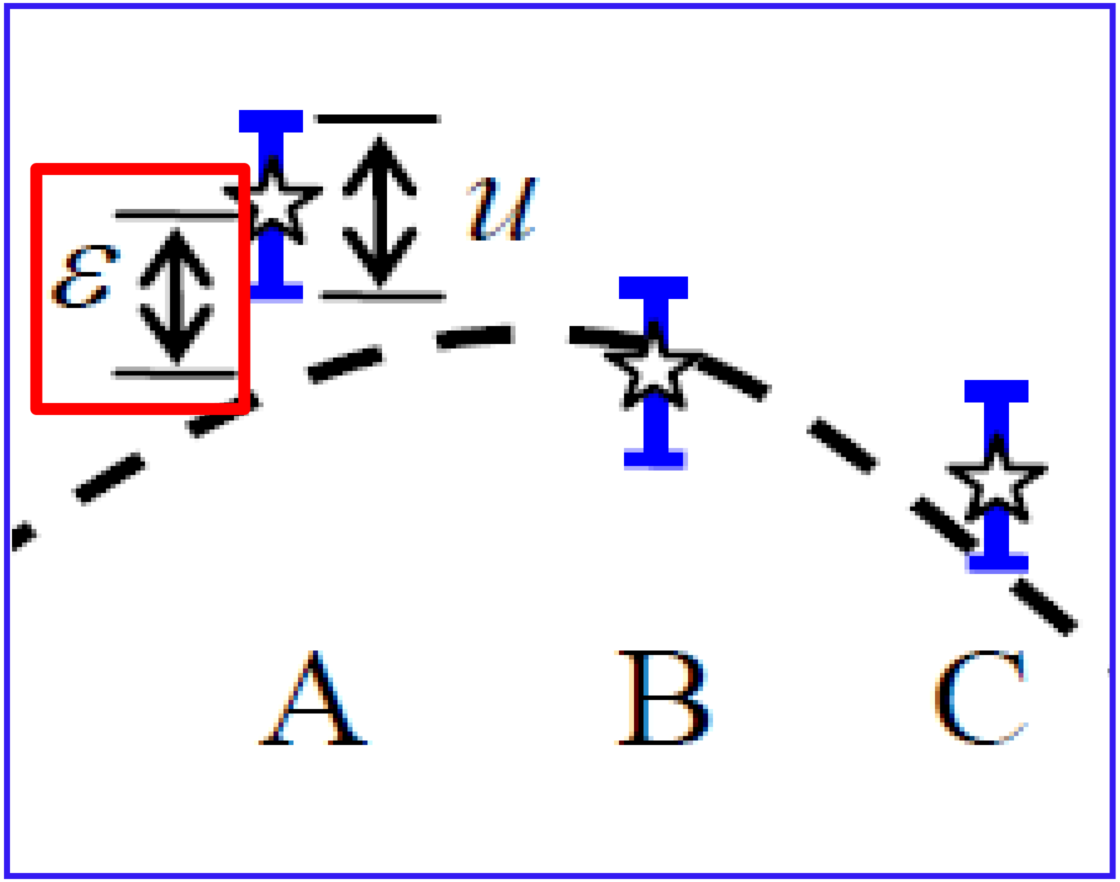
Measured data values

- Exact value (------) – not available
- u - uncertainty
- ε - error



Measured data values

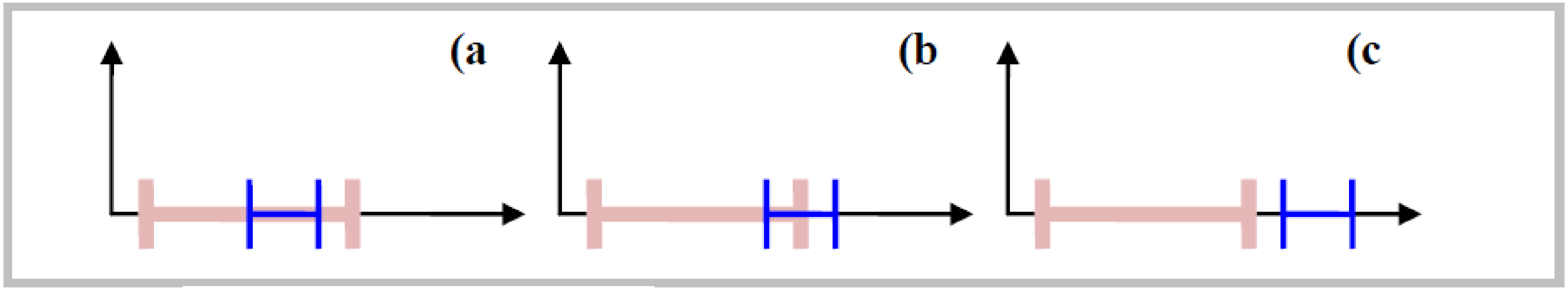
- Exact value (------) – not available
- u - uncertainty
- ε - error



Data validation – detecting errors if they exist

Comparing measured and predicted values

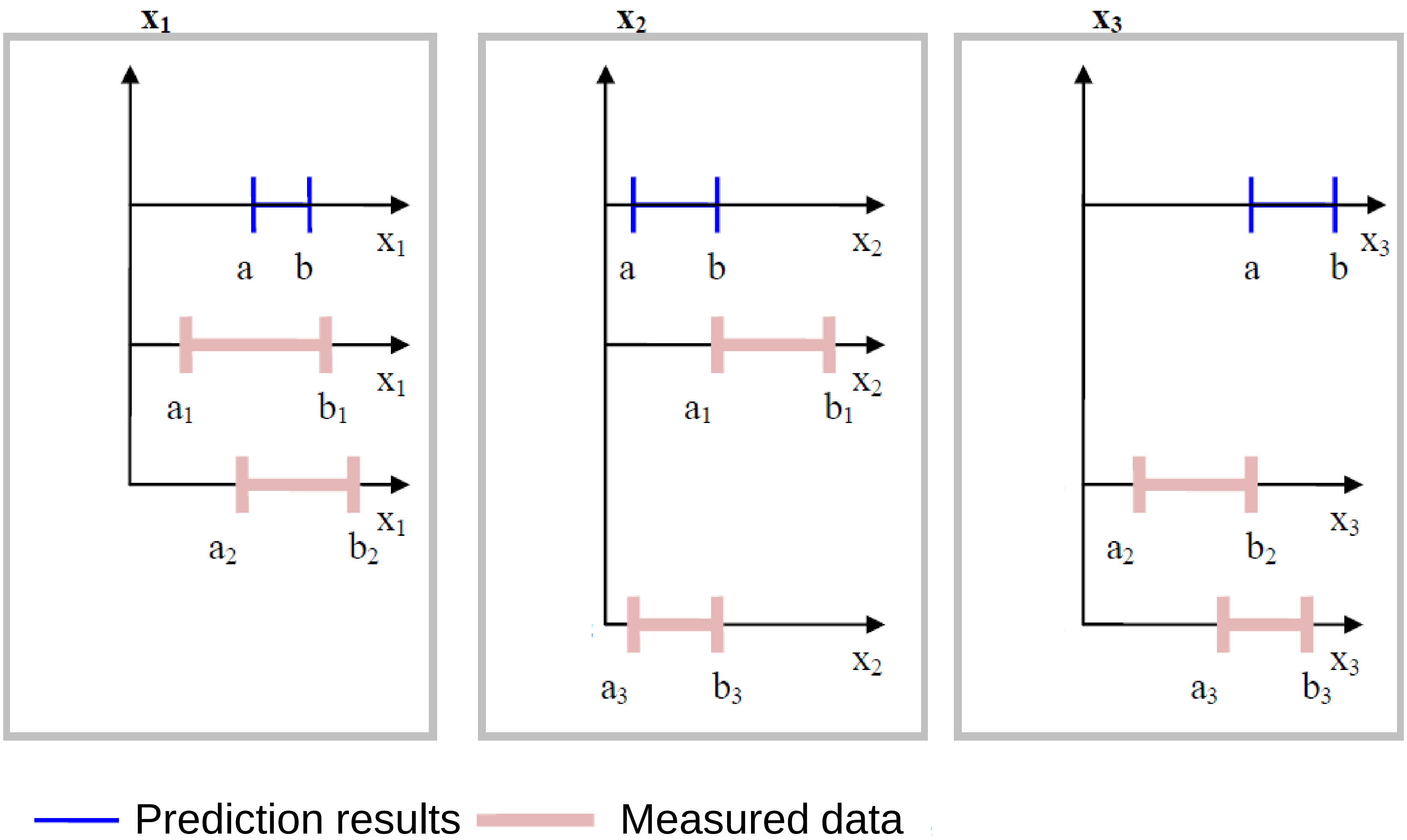
Measured and calculated values are presented with intervals



Prediction results Measured data

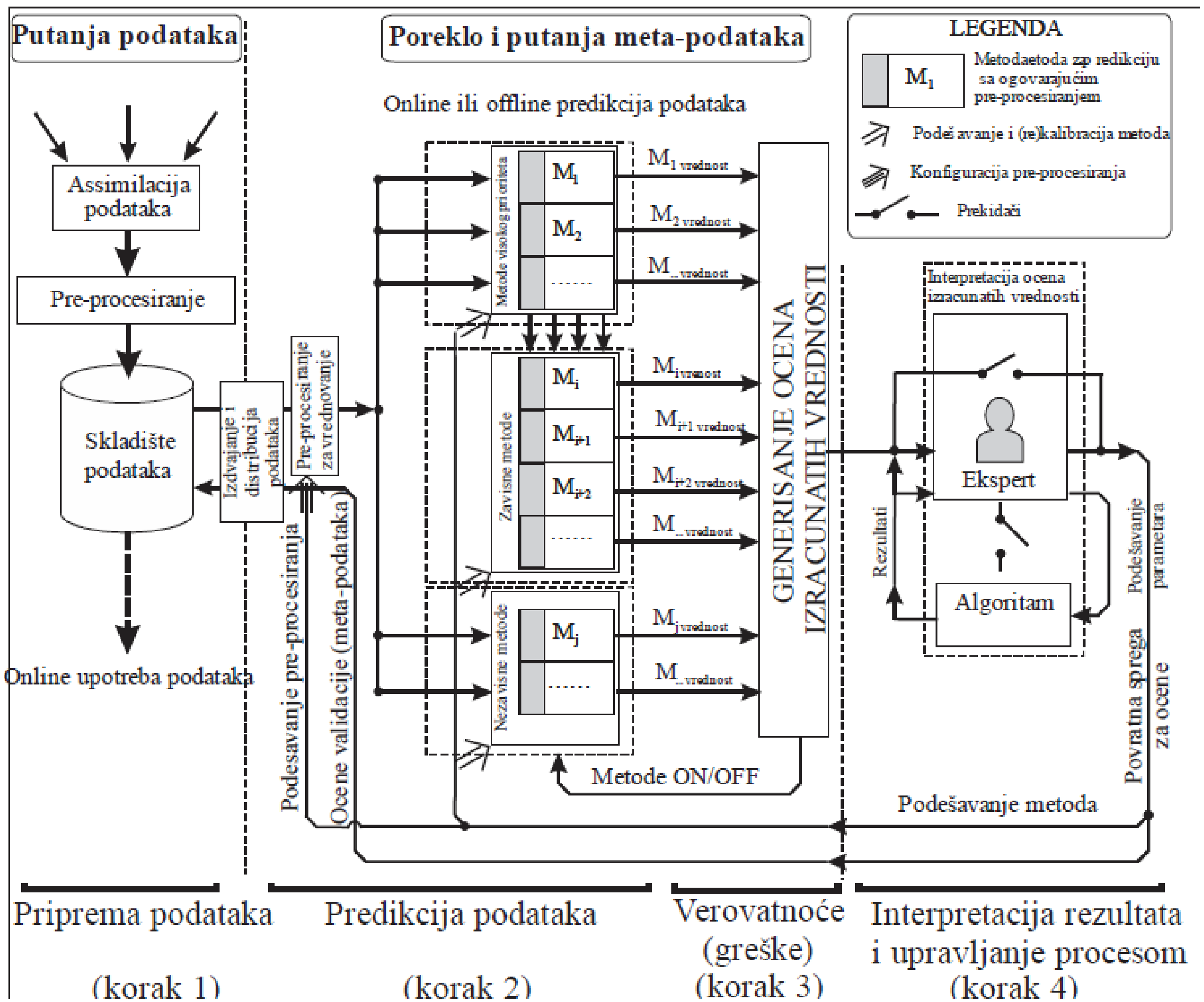
Enables detection of errors in data – data validation

Comparing measured and predicted values -complex system-

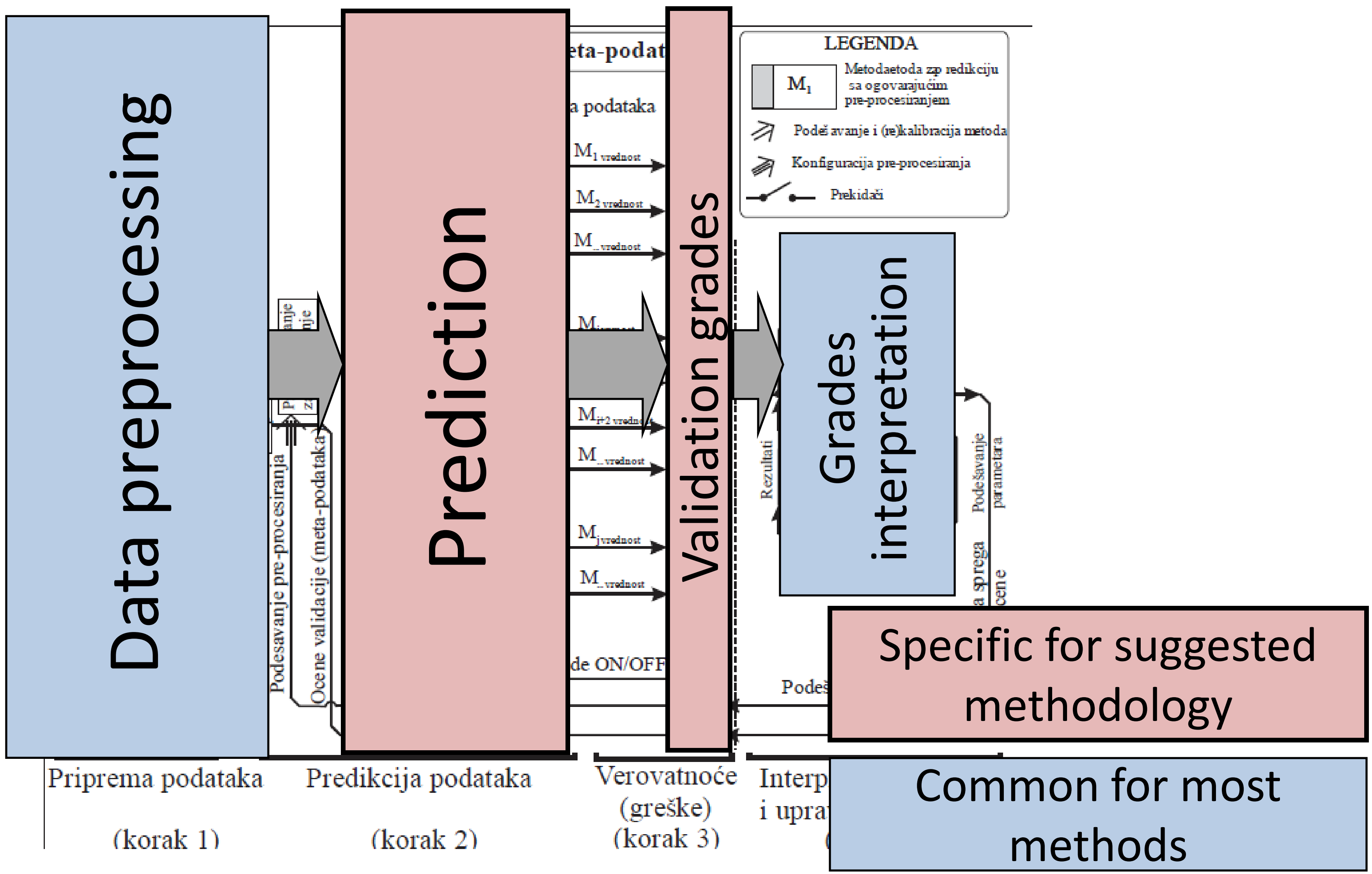


Suggested data validation methodology

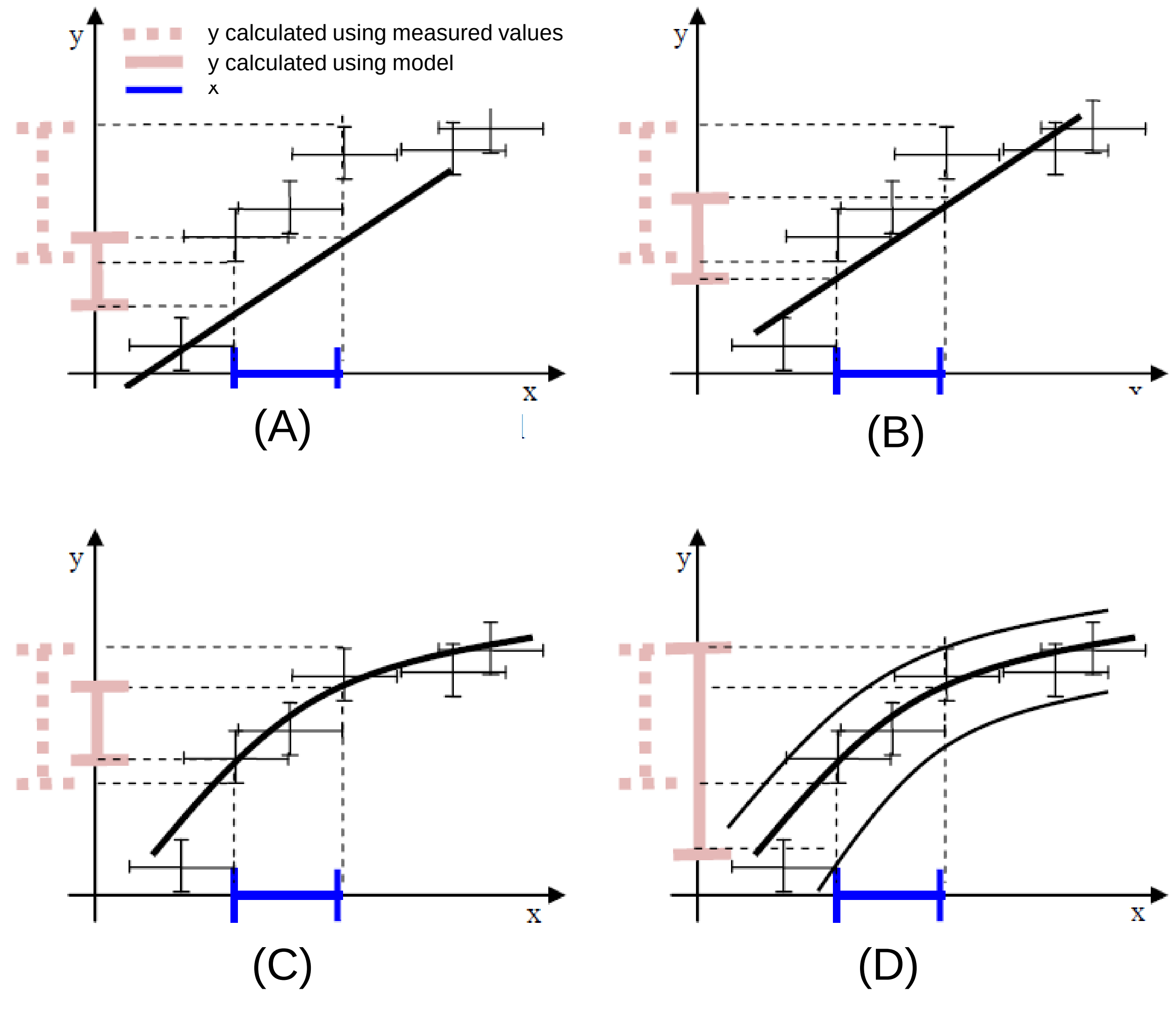
sorry for this mess on the screen – will make it more readable...



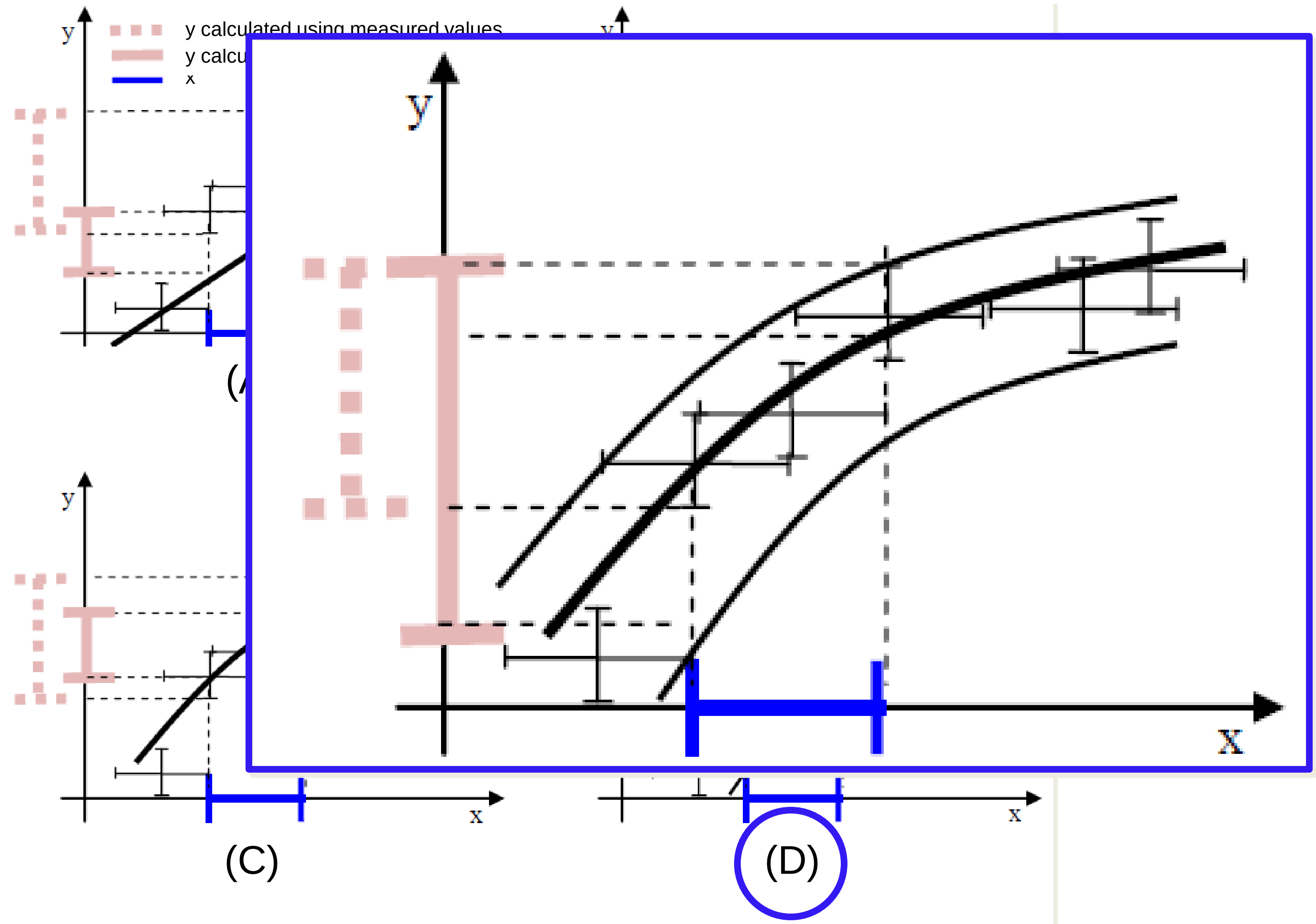
Suggested data validation methodology



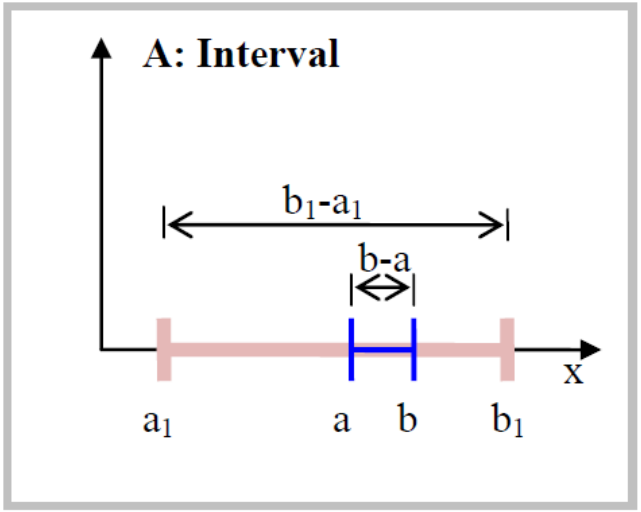
Data prediction
-relation errors and uncertainty-



Data prediction
-relation errors and uncertainty-



Validation grades
-data probability-



Prediction results

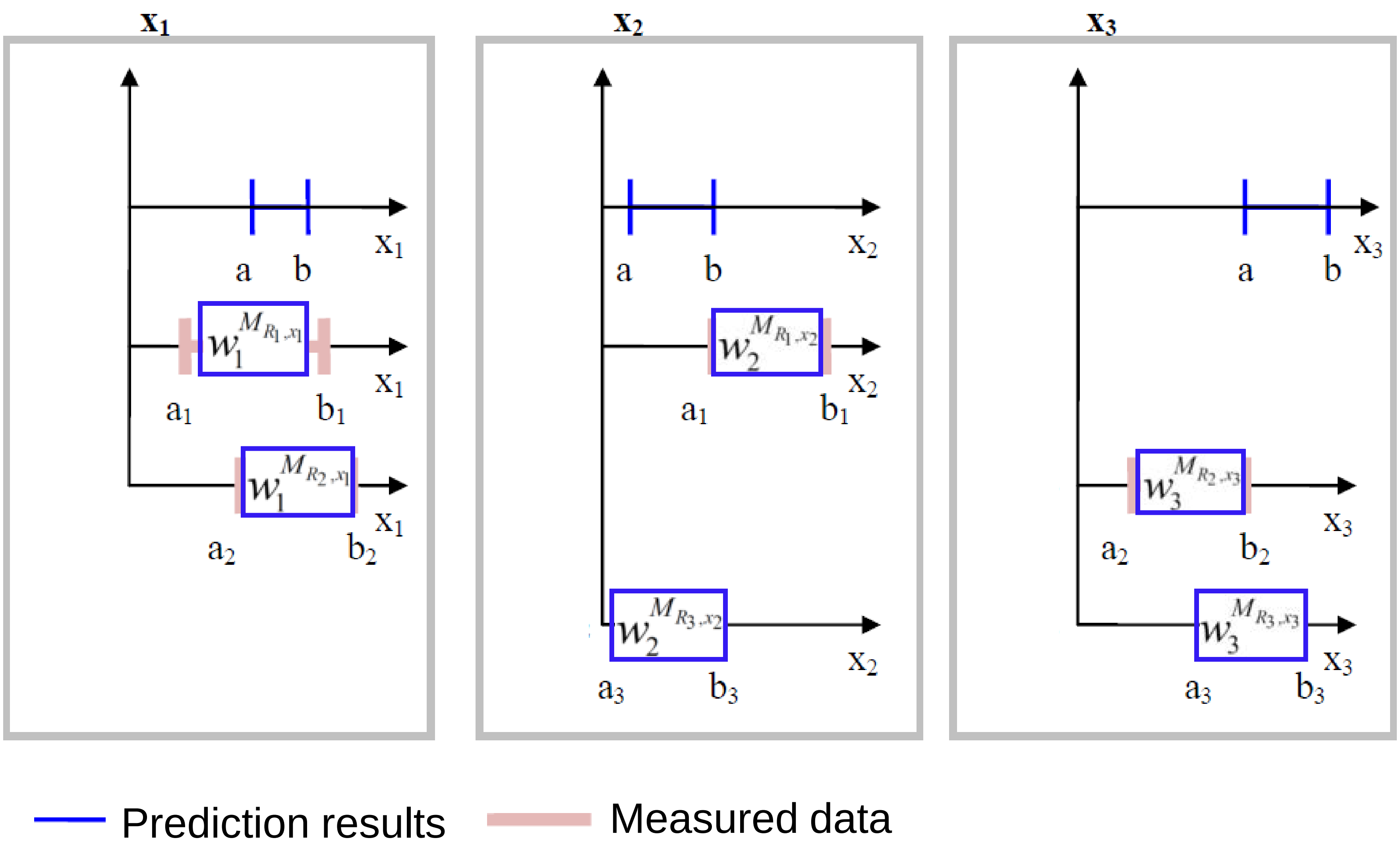
Measured data

$$p\left(\left[a,b\right]\mid\left[a_1,b_1\right]\right)=\frac{p\left(\left[a,b\right]\cap\left[a_1,b_1\right]\right)}{p\left(\left[a_1,b_1\right]\right)};$$

$$p\left(x_i \mid x_i^{M_j}, X_{x_i}^{M_j}\right)= p\left(\left[a,b\right]\mid\left[a_1,b_1\right]\right)=\begin{cases}1-\dfrac{\max\left(0,a-a_1\right)+\max\left(0,b_1-b\right)}{\left(b_1-a_1\right)}, & a_1\le b \\ & b_1\ge a \\ 0 & \end{cases}$$

$$\max\left(p\left(x_i \mid x_i^{M_j}, X_{x_i}^{M_j}\right)\right)=\max\left(P\left(\left[a,b\right]\mid\left[a_1,b_1\right]\right)\right)=\frac{b-a}{b_1-a_1}$$

Validation grades
-weighted average of data probabilities-



Validation grades
-likelihood of calculated values-

Weights– likelihood of predicted value

$$\boxed{w_i^{M_{R_j,x_i}} = p\left(x_i^{M_{R_j,x_i}} \mid X_{x_i}^{M_{R_j,x_i}}\right)}$$

Maximization of a sum of likelihoods - EM algorithm

$$\boxed{J = \max \sum_i \sum_j p\left(x_i^{M_{R_j,x_i}} \mid X_{x_i}^{M_{R_j,x_i}}\right)}$$

$\mathbf{E}$ step:

$$p\left(x_i^{M_{R_j,x_i}} \mid X_{x_i}^{M_{R_j,x_i}}\right) = \sum_{X_{x_i}^{M_{R_j,x_i}}} \left(\frac{\partial M_{R_j,x_i}\left(X_{x_i}^{M_{R_j,x_i}}, \theta\right)}{\partial X_{x_i}^{M_{R_j,x_i}}} \right)^{-1} P_{x_i}$$

$\mathbf{M}$ step:

$$P_{x_i} = \sum_j p\left(x_i \mid x_i^{M_{R_j,x_i}}, X_{x_i}^{M_{R_j,x_i}}\right) \times \frac{p\left(x_i^{M_{R_j,x_i}} \mid X_{x_i}^{M_{R_j,x_i}}\right)}{\sum_j p\left(x_i^{M_{R_j,x_i}} \mid X_{x_i}^{M_{R_j,x_i}}\right)}$$

Validation grades
-grades-

1. Not normalized $$\boxed{x_i^{grade} = \sum_j w_i^{M_{R_j,x_i}} \times p\left(x_i \mid x_i^{M_{R,x_i}}, X_{x_i}^{M_{R,x_i}}\right)}$$ (*between 0 and a<1*)

2. Normalized $$\boxed{x_i^{grade} = p_{x_i}^{norm} = \left(\sum_{x_i} w_i^{M_j} \frac{p\left(x_i \mid x_i^{M_{R_j,x_i}}, X_{x_i}^{M_{R_j,x_i}}\right)}{\max\left(p\left(x_i \mid x_i^{M_{R_j,x_i}}, X_{x_i}^{M_{R_j,x_i}}\right)\right)}\right)}$$ (*between 0 and 1*)

Evaluating validation systems

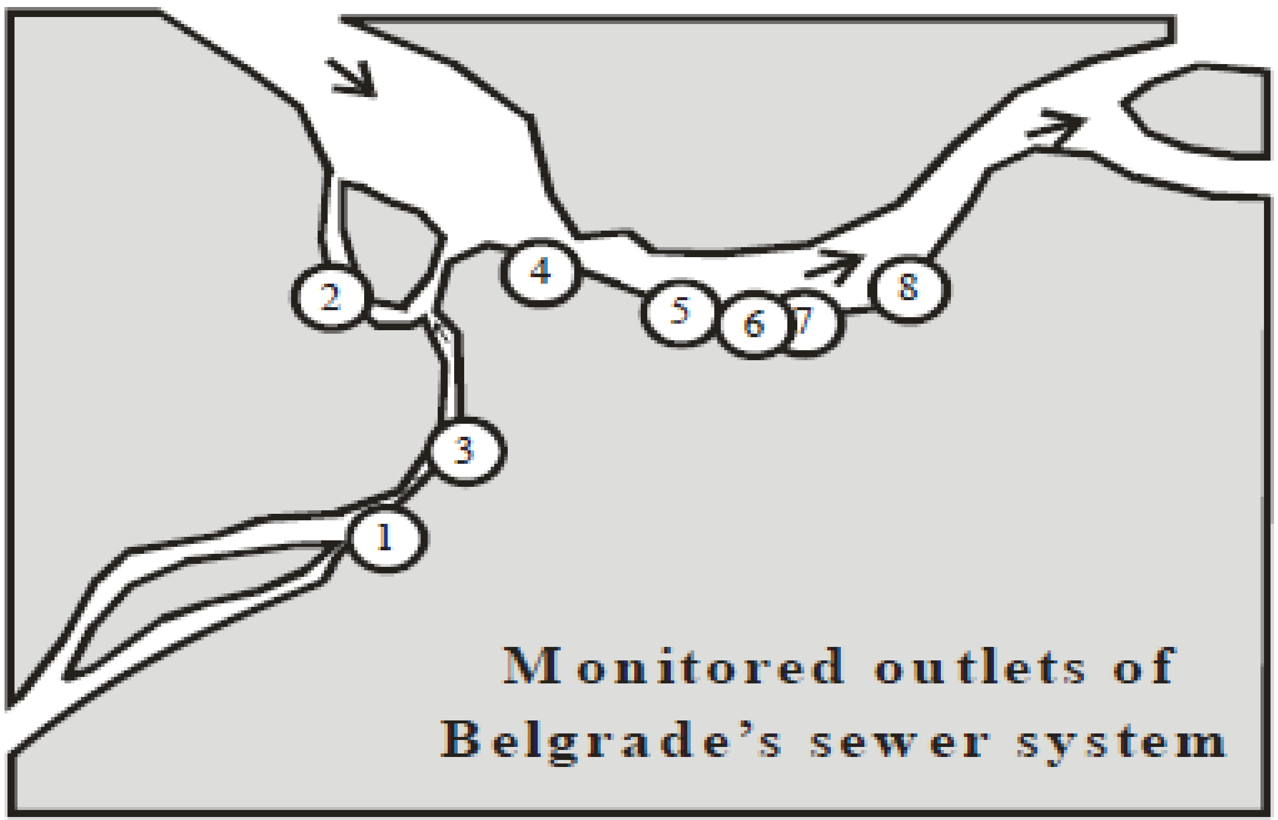
$$p = \frac{N_{registered}}{N_{anomalies} + N_{missed} + N_{registered\ nonanomalies}}$$

Comparing to Kappa index of coincidence:

- Higher penalty to missed anomalies, and
- Lower penalty to correct values interpreted as anomalies

Case study

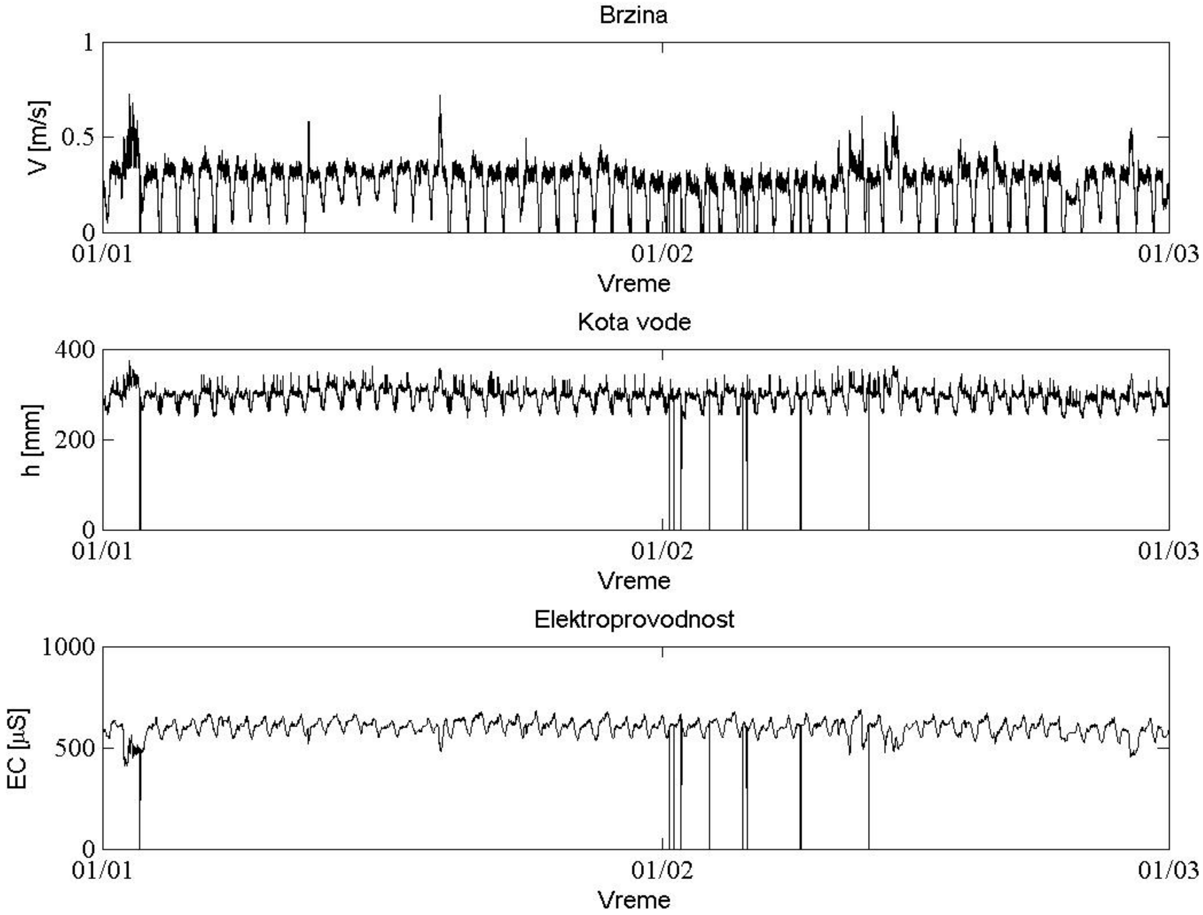
Measurements of hydraulic and water quality parameters in Belgrade sewer CSO



Measurements in Belgrade sewer CSO

Measured variables: Velocity (**V**), water depth (**h**), electro-conductivity(**EC**)

Period: 1/1/2007-1/3/2007



Relations between measured variables

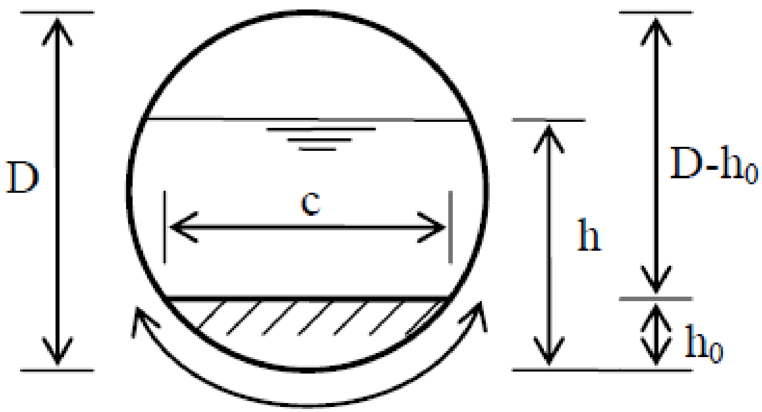
	V	h	EC	V^{t-1}	h^{t-1}	EC^{t-1}
R_1	$V^t = f_{V^t}^{R_1}\left(h^t\right)$	$h^t = f_{h^t}^{R_1}\left(V^t\right)$		$V^{t-1} = f_{V^{t-1}}^{R_1}\left(h^{t-1}\right)$	$h^{t-1} = f_{h^{t-1}}^{R_1}\left(V^{t-1}\right)$	
R_2		$h^t = f_{h^t}^{R_2}\left(EC^t\right)$	$EC^t = f_{EC^t}^{R_2}\left(h^t\right)$		$h^{t-1} = f_{h^{t-1}}^{R_2}\left(EC^{t-1}\right)$	$EC^{t-1} = f_{EC^{t-1}}^{R_2}\left(h^{t-1}\right)$
R_3	$V^t = f_{V^t}^{R_3}\left(EC^t\right)$		$EC^t = f_{EC^t}^{R_3}\left(V^t\right)$	$V^{t-1} = f_{V^{t-1}}^{R_3}\left(EC^{t-1}\right)$		$EC^{t-1} = f_{EC^{t-1}}^{R_3}\left(V^{t-1}\right)$
R_4	$V^t = f_{V^t}^{R_4}\left(V^{t-1}\right)$			$V^{t-1} = f_{V^{t-1}}^{R_4}\left(V\right)$		
R_5		$h^t = f_{h^t}^{R_5}\left(h^{t-1}\right)$			$h^{t-1} = f_{h^{t-1}}^{R_5}\left(h\right)$	
R_6			$EC^t = f_{EC^t}^{R_6}\left(EC^{t-1}\right)$			$EC^{t-1} = f_{EC^{t-1}}^{R_6}\left(EC\right)$

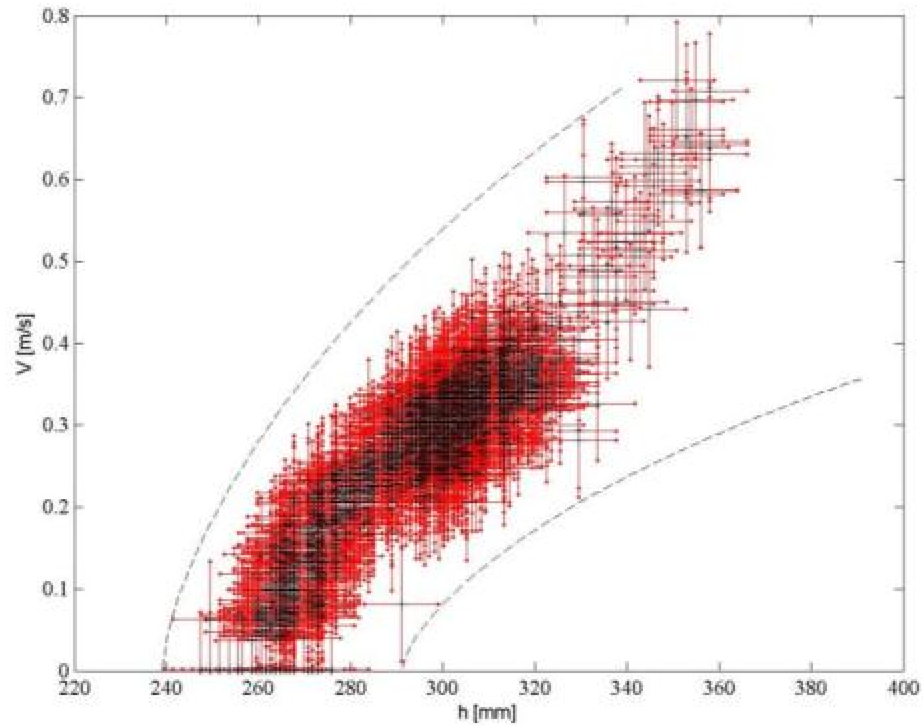
R1 – Chezy-Manning equation

R2, **R3** – Water quality model of EC reduction

R4, **R5**, **R6** – AR(1) models

Chezy-Manning equation





$$V(h_0, C, h) = C \times R(h_0, h)^{2/3}$$

$$C = \sqrt{I_d}/n$$

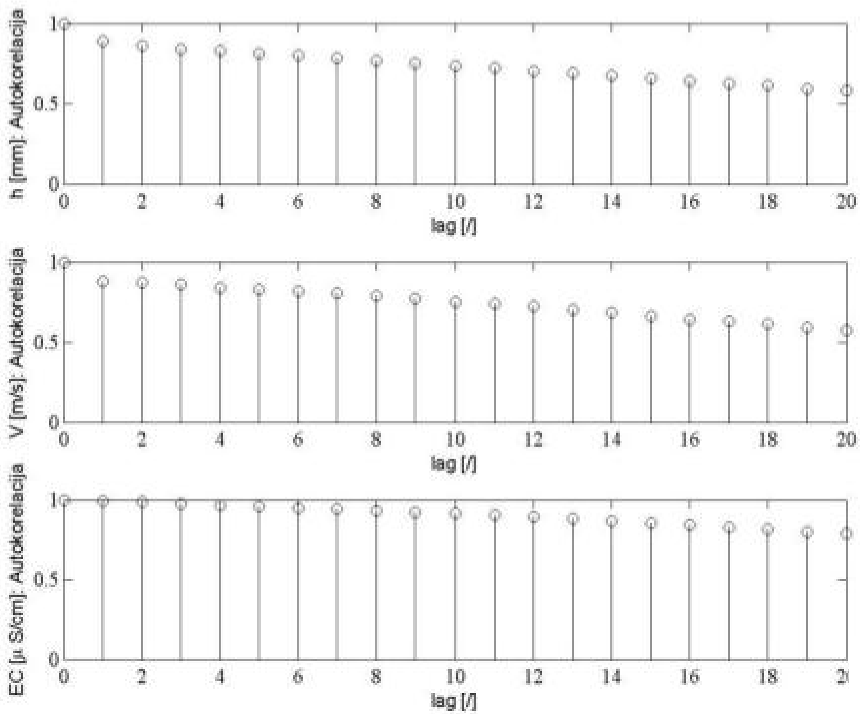
$$M_{R_1,V} : V(h) = [3.8, 1.9] \times R([239, 291], h)^{2/3}$$

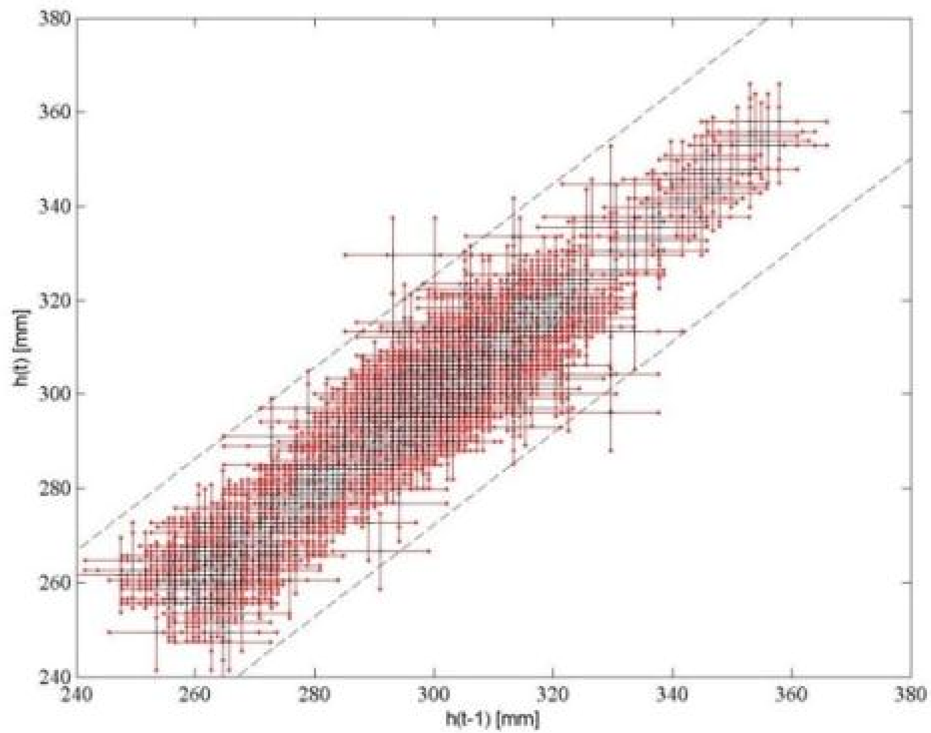
AR(1) models

$$x^t = ax^{t-1} + b$$

Autocorrelation coefficient

$$R_{ff}(\tau) = \int_{-\infty}^{\infty} f(t+\tau)\overline{f}(t)\,\mathrm{d}t$$



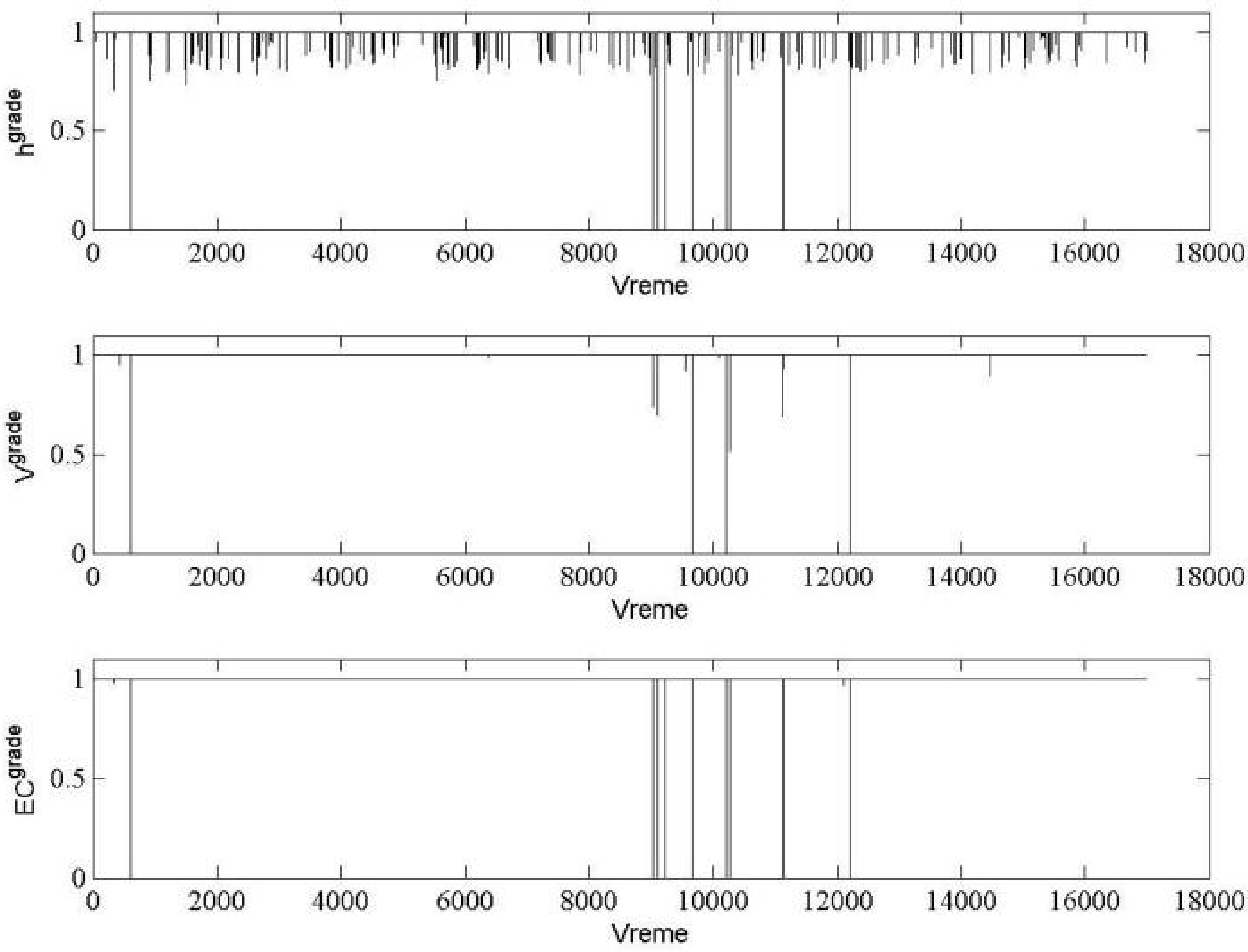


Model calibration results

x	h	V	EC
a	0.97427	0.97258	0.99712
$b=\left[\underline{b},\bar{b}\right]$	[-33, 52]	[-0.17, 0.19]	[-59, 49]

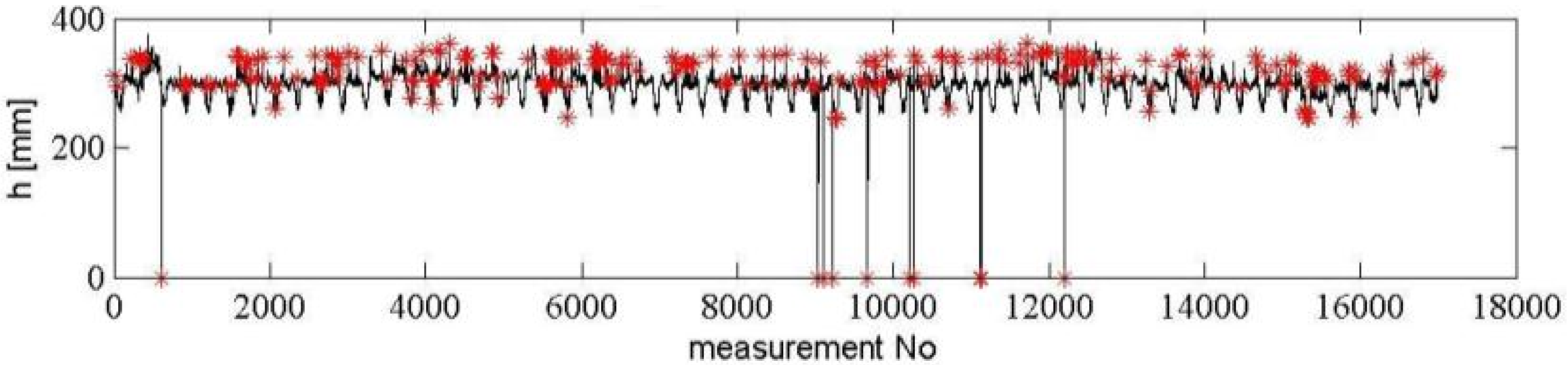
Results

-Validation grades-



Results
-comparison with manual data validation-

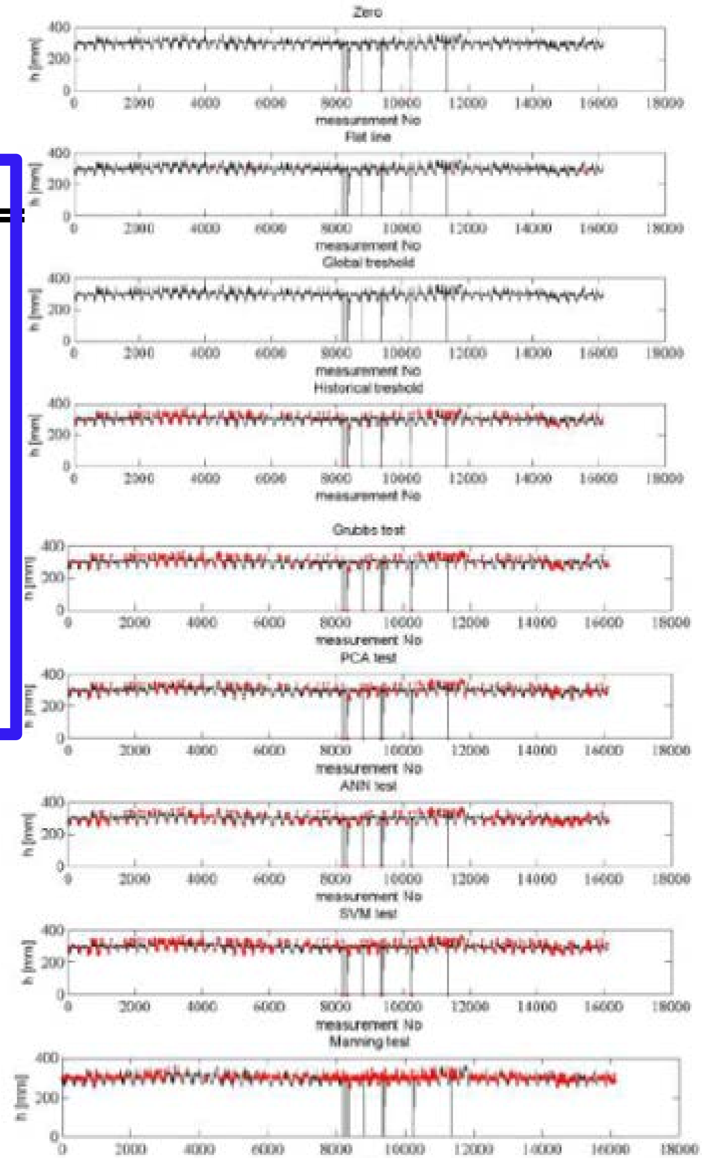
Threshold value= **0.99**



(No anomalies)	(No detected)	(No missed)	(No false)	p
244	219	25	73	0.64

Nine validation methods

	No anomalies	No detected	No missed	No false	p
M_1	244	24	220	0	0.052
M_2	244	0	244	28	0.000
M_3	244	0	244	0	0.000
M_4	244	202	42	220	0.399
M_5	244	187	57	334	0.294
M_6	244	224	20	359	0.360
M_7	244	109	135	896	0.085
M_8	244	237	11	483	0.321
M_9	244	71	173	2725	0.023



Averaging validation results:

$$p = \frac{N_{registered}}{N_{anomalies} + N_{missed} + N_{registered\ nonanomalies}} = \frac{195}{244+49+58} = \boxed{0.56}$$

Conclusions

- New theoretical framework for data validation
- No restrictions about number and type of additional information that may be used in the validation process
- No restrictions of type and number of relations that may be used
- The validation system may be evaluated and improved by improving the relations between the data values

Automatic measured data validation applied on hydraulic and water quality parameters in sewer systems

Nemanja Branisavljević[1]
Dušan Prodanović[1]
Zoran Kapelan[2]

[1]University of Belgrade
[2]University of Exeter