IAHS AISH ICSW STAHY ICWRS

HW06 workshop

Expert judgement vs. statistical goodness-of-fit for hydrological model evaluation: Results of experiment

Charles Perrin, Vazken Andréassian
and Louise Crochemore

With special thanks for the *expert* contribution of
Laurent Coron, Julien Lerat, Roger Moussa, Frédéric Hendrickx,
Marie Bourqui, Simon Gascoin, Laura Farrant,
Simon Seibert, Uwe Ehret, John Ewen

XXV IUGG General Assembly,
3-4 July 2011, Melbourne

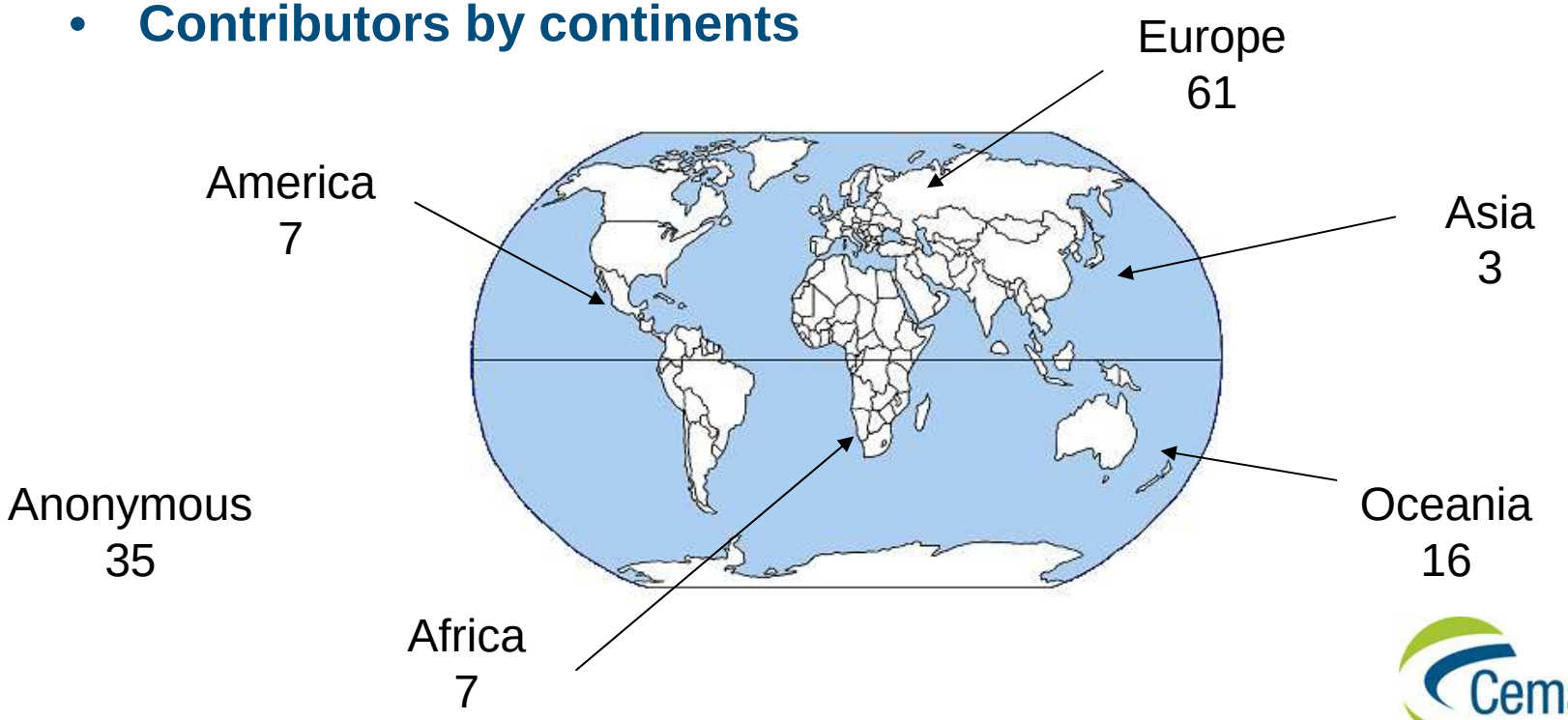
Cemagref
Sciences, eaux & territoires

Part 1.
Expert judgement

Answers to the survey

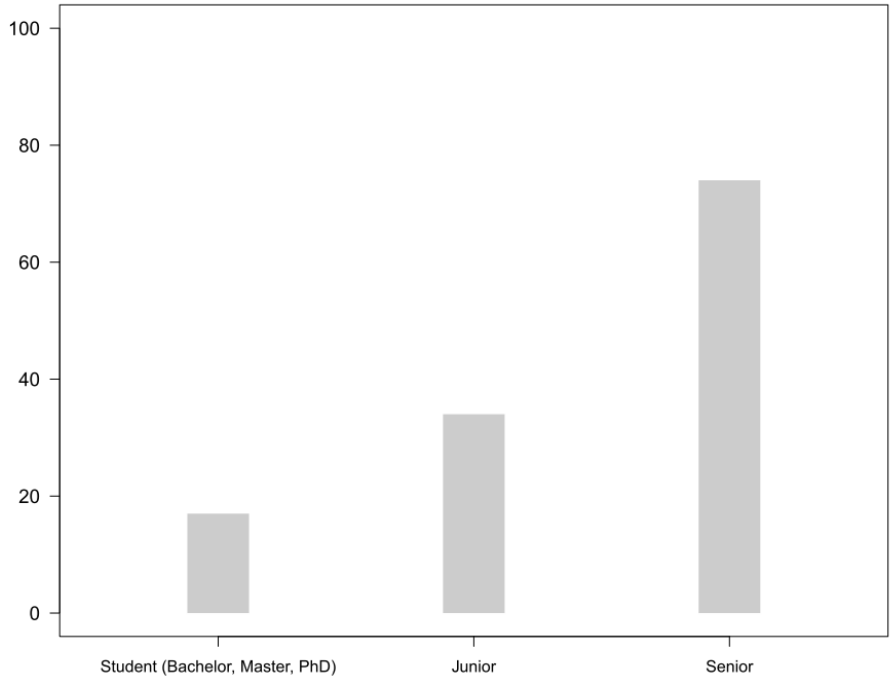
- **Answers analyzed**
  - **Web site:** **90**
  - **Yesterday's session:** **39**
  - **Total:** **129**

- **Contributors by continents**

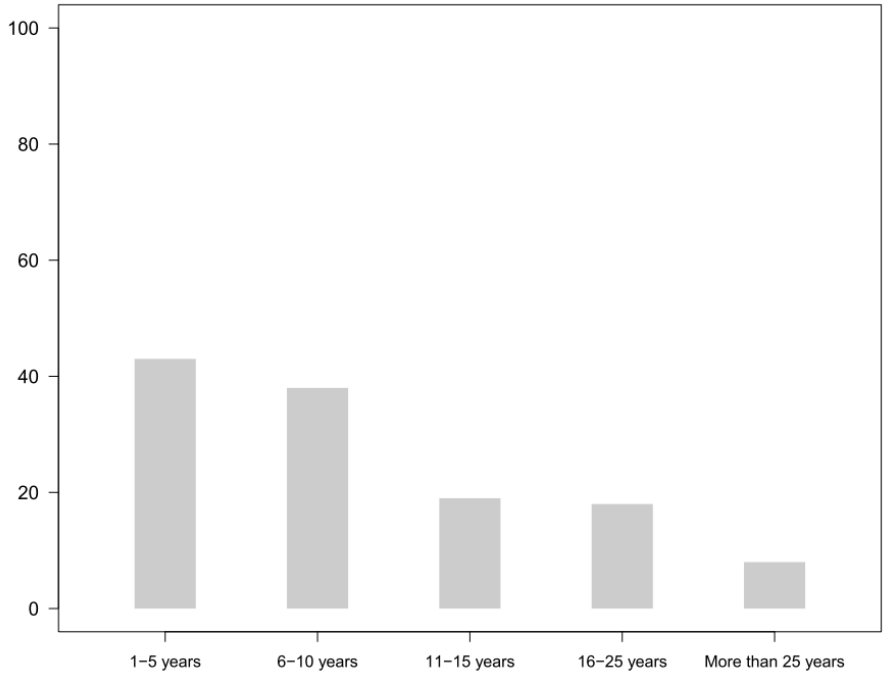


Who are the experts?

Status



Experience in modelling

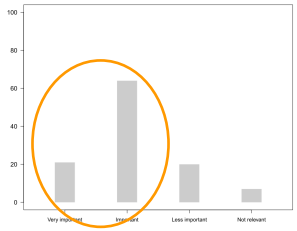


Cemagref
Sciences, eaux & territoires

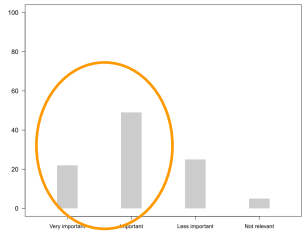
Criteria chosen by experts for model evaluation

Low flows

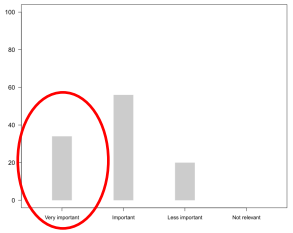
Very important | Important | Less important | Not relevant



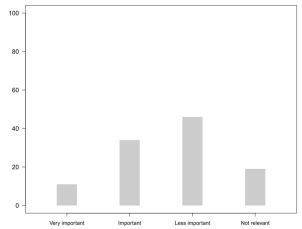
Mean volume



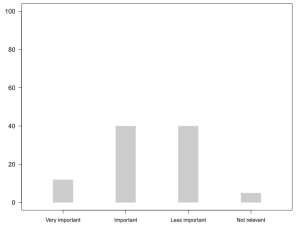
Timing



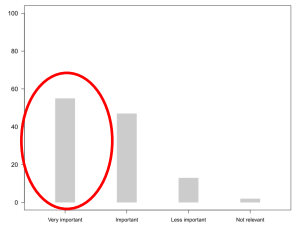
Magnitude of extremes



Duration



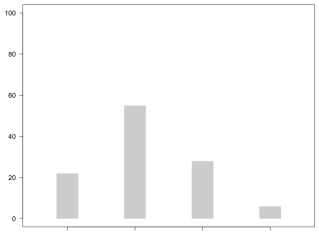
Rising limb



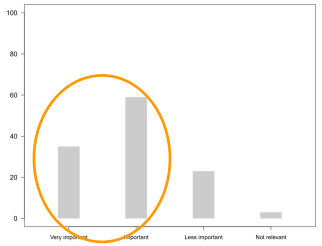
Recession

High flows

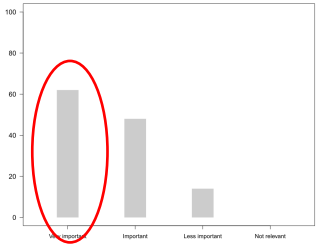
Very important | Important | Less important | Not relevant



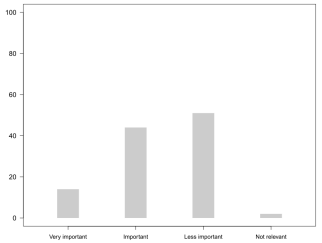
Mean volume



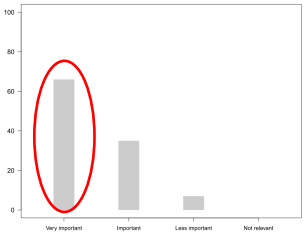
Timing



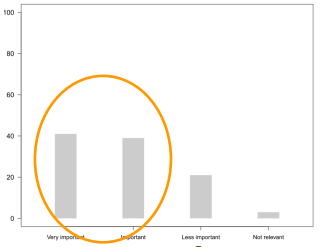
Magnitude of extremes



Duration



Rising limb

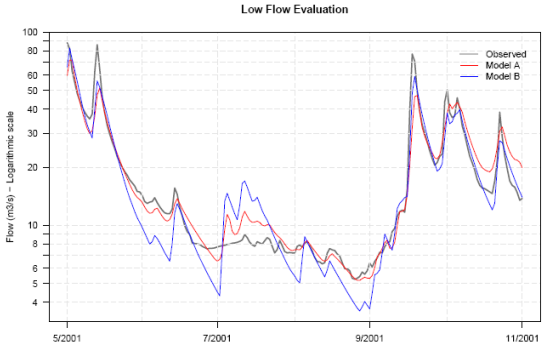


Recession

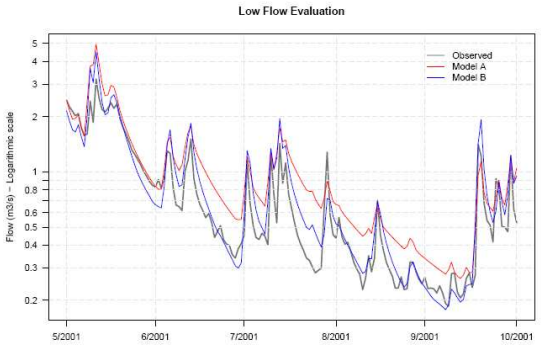
Agreement between experts

Model comparison

Low flows

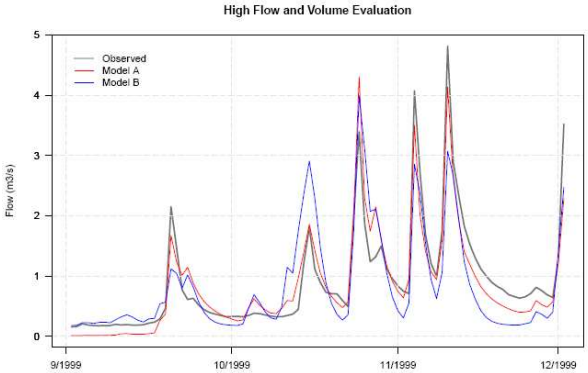


A better than B for 99% of judges

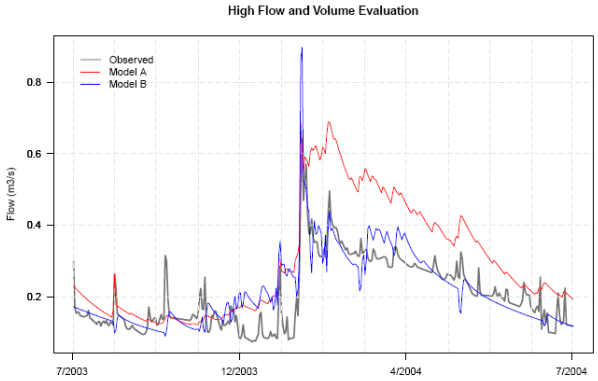


B better than A for 95% of judges

High flows



A better than B for 97% of judges

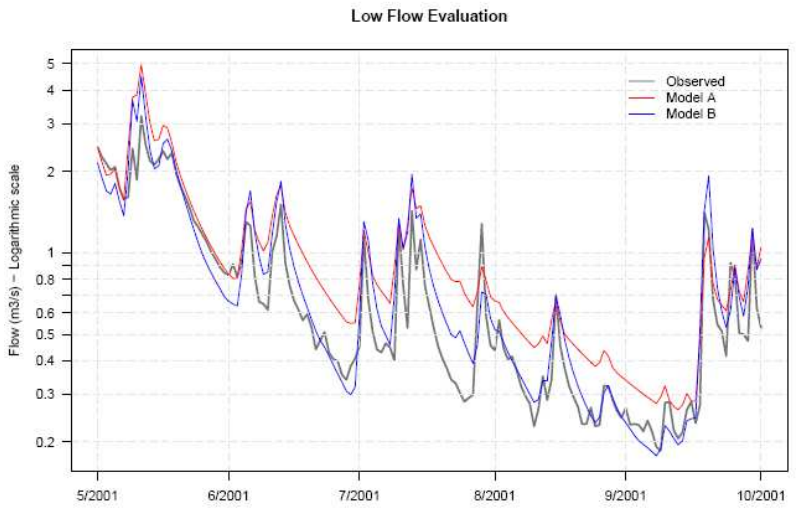


B better than A for 97% of judges

Agreement between experts

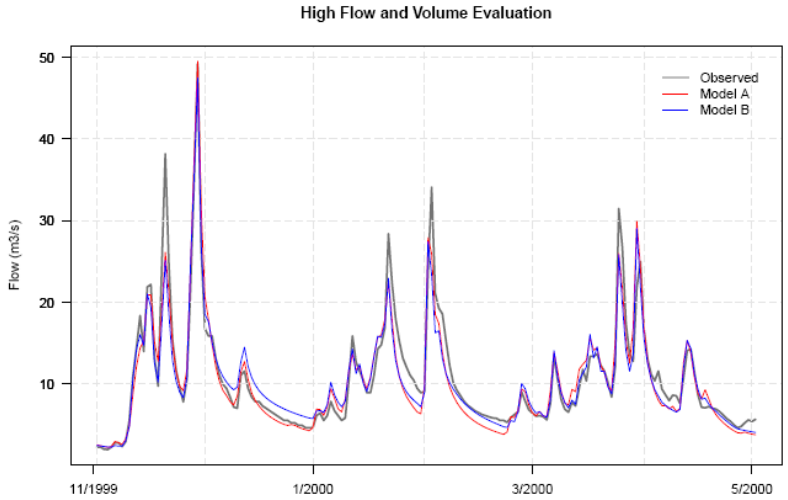
Model rating

Low flows



Very good + Good	64%
Slightly Good	20%
Average	11%
Slightly Poor	2%
Very Poor + Poor	2%

High flows

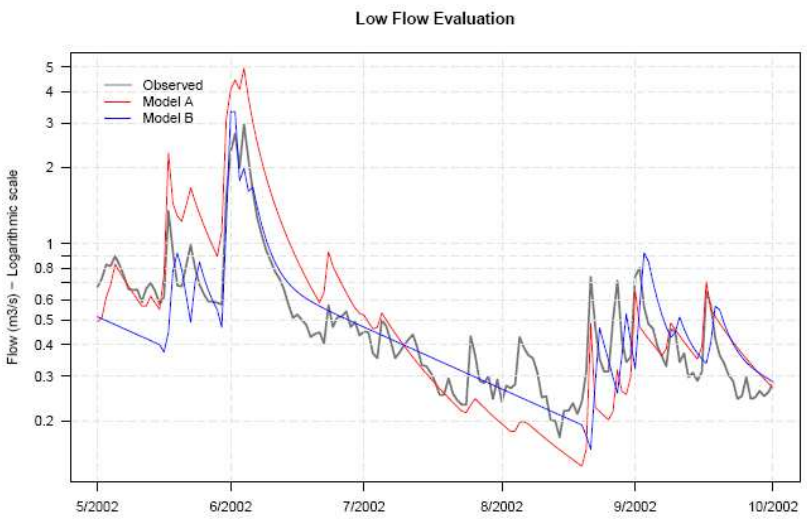


Very good + Good	77%
Slightly Good	16%
Average	5%
Slightly Poor	0%
Very Poor + Poor	1%

Disagreement between experts

Model comparison

Low flows

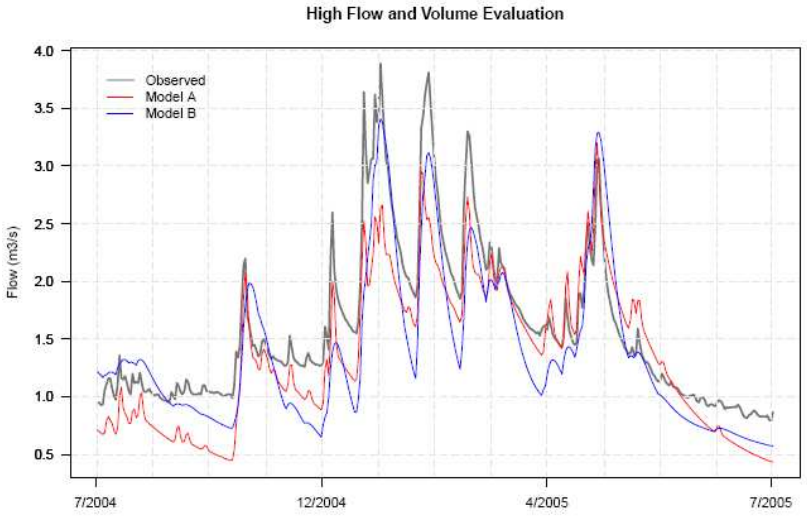


A better than B for 30% of judges

B better than A for 40% of judges

A and B equivalent for 40% of judges

High flows



A better than B for 27% of judges

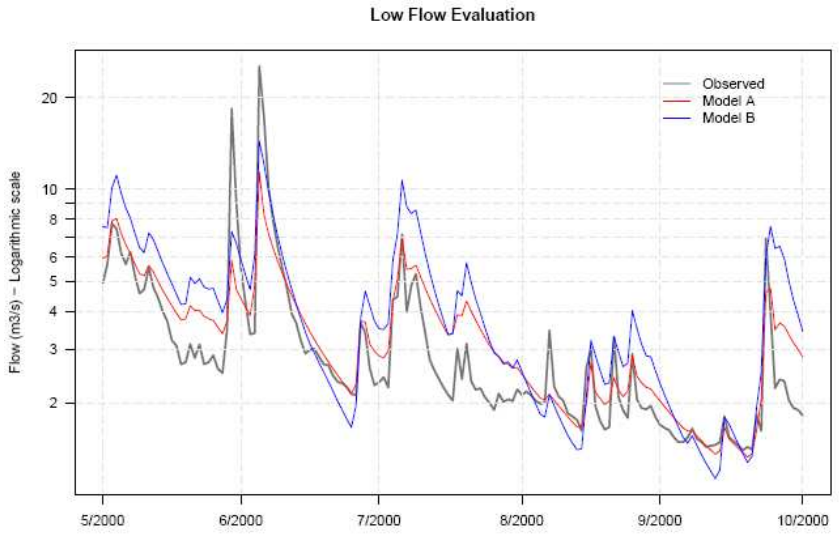
B better than A for 38% of judges

A and B equivalent for 35% of judges

Disagreement between experts

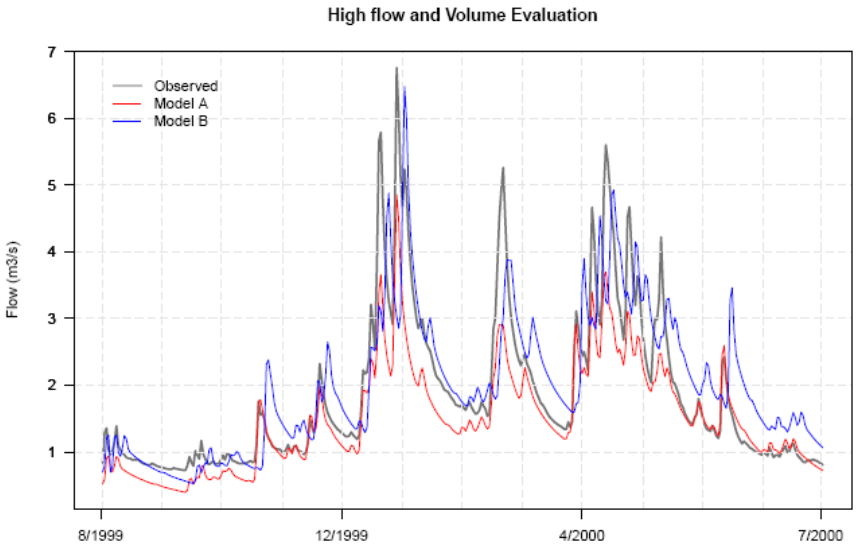
Model rating

Low flows



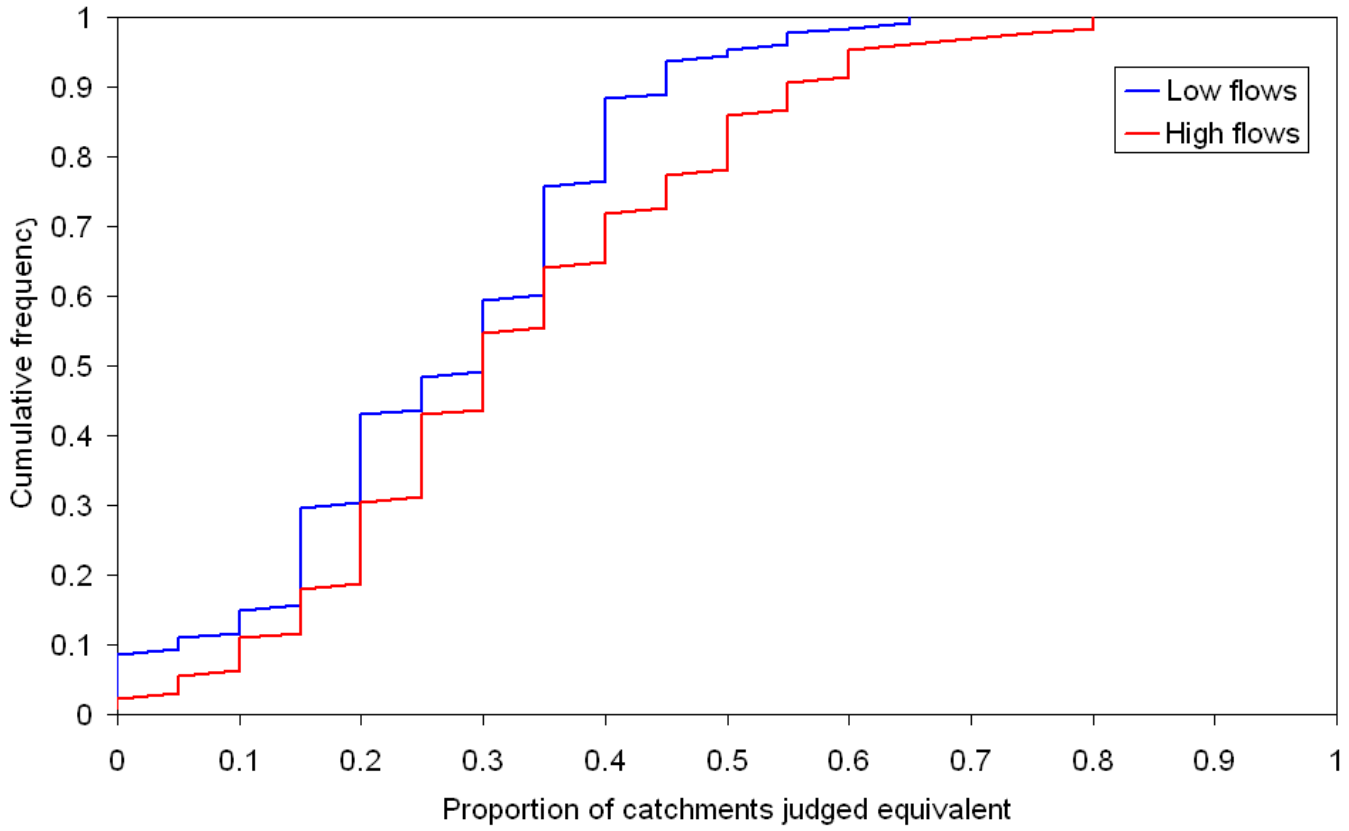
Very good + Good	10%
Slightly Good	29%
Average	22%
Slightly Poor	23%
Very Poor + Poor	15%

High flows



Very good + Good	22%
Slightly Good	25%
Average	26%
Slightly Poor	13%
Very Poor + Poor	12%

Experts who hesitate



Cemagref
Sciences, eaux & territoires

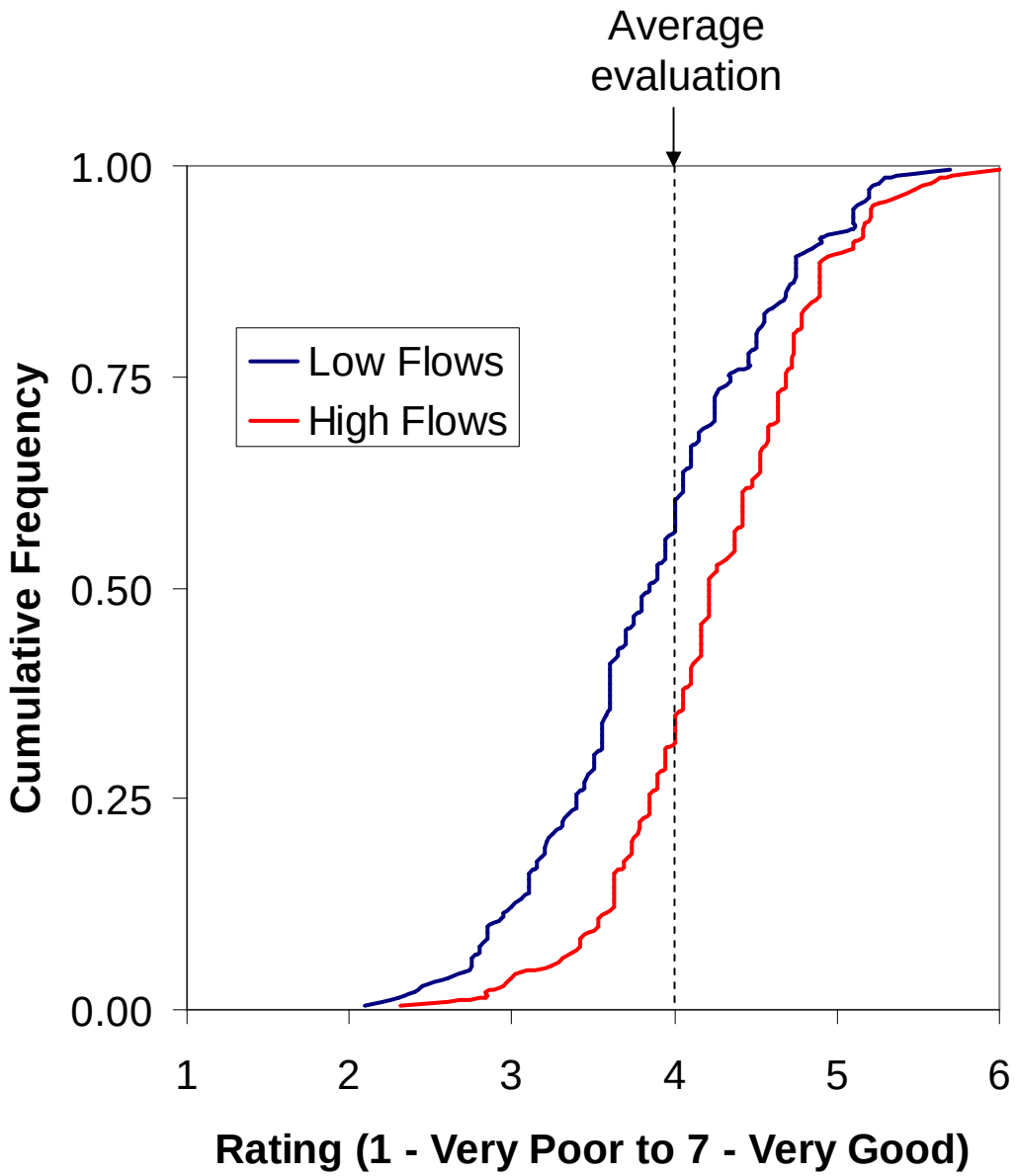
Variability in rating

- **We transformed the qualitative scale into a numeric scale, in order to compute an average note for each expert**
  - ***1: Very Poor***
  - ***2: Poor***
  - ***3: Slightly Poor***
  - ***4: Average***
  - ***5: Slightly Good***
  - ***6: Good***
  - ***7: Very Good***

- **The idea is to get a generic idea of the distribution of optimism & pessimism among experts**

Cemagref
Sciences, eaux & territoires

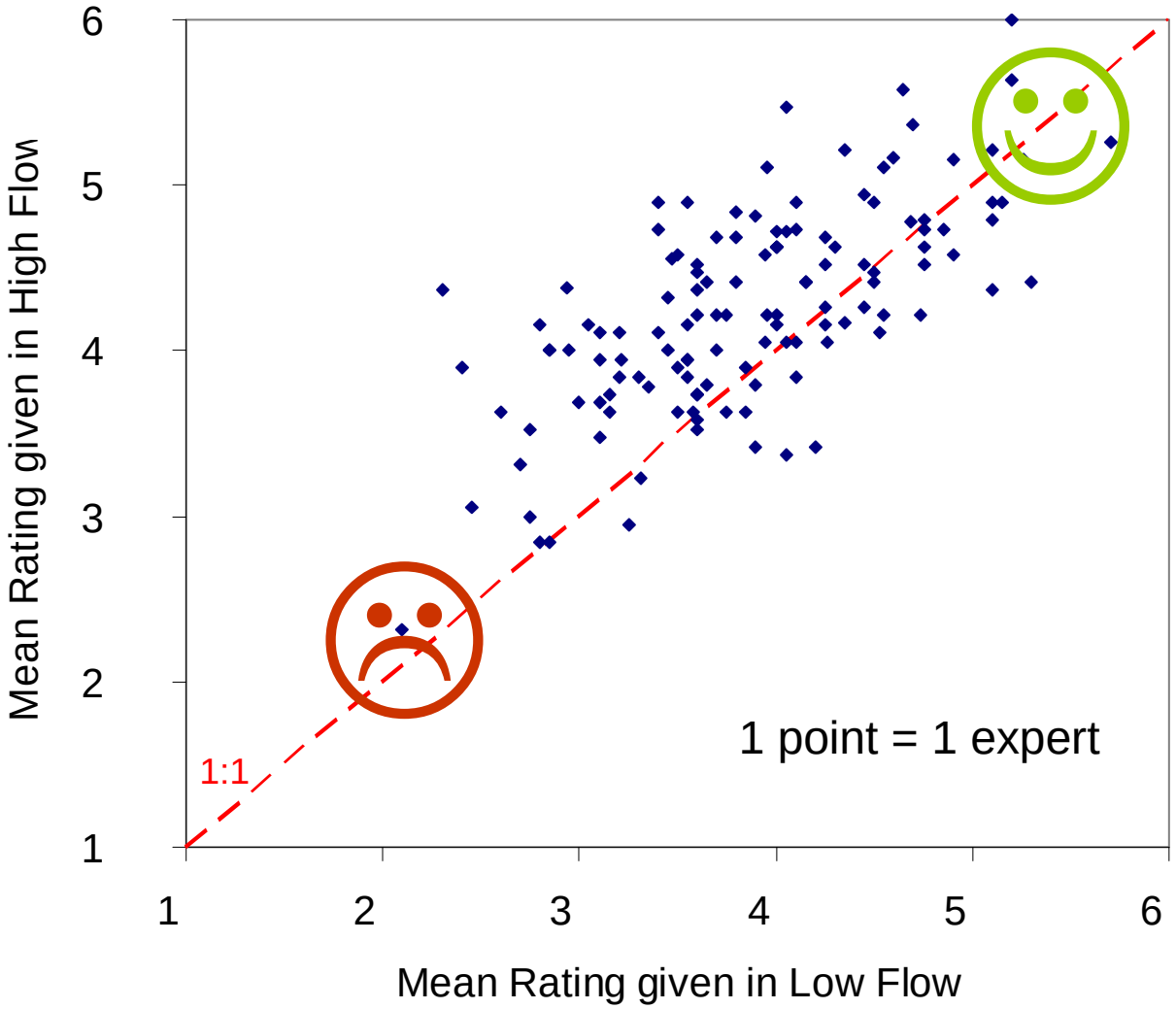
Optimism and pessimism: general view



- Good spread between optimistic and pessimistic modellers
- Judges more optimistic on high flows

Cemagref
Sciences, eaux & territoires

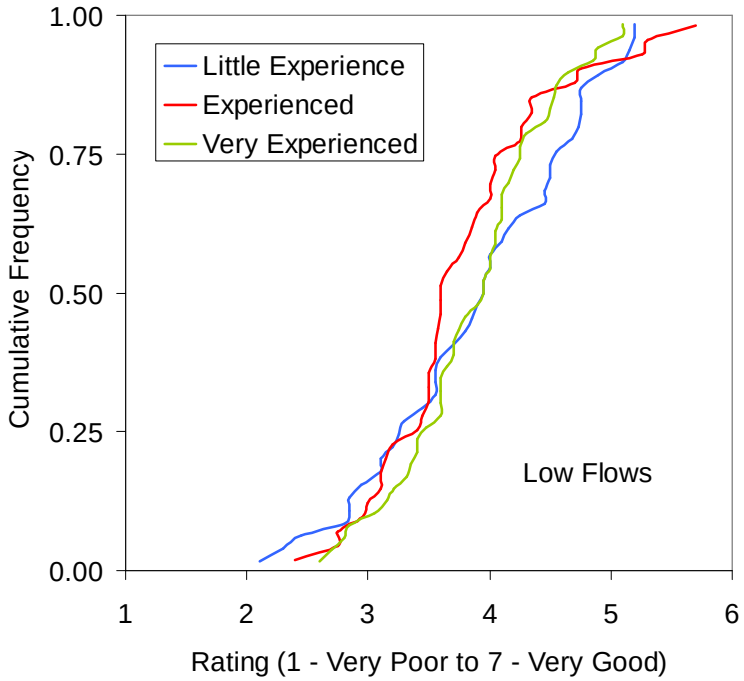
LF ratings vs HF ratings

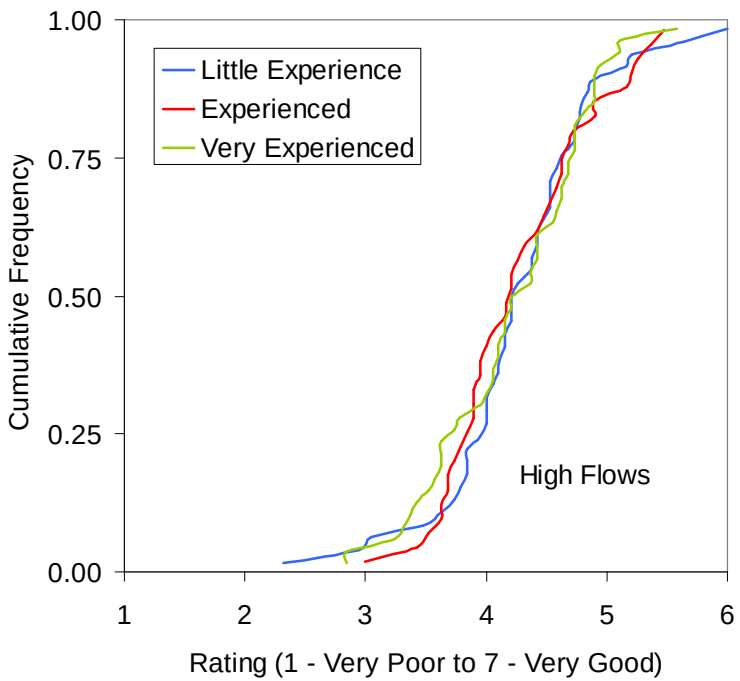


- Quite good correlation between optimism (pessimism) in high and low flows
- One expert seems desperate on the quality of simulation...
- while others feel very enthusiastic

Cemagref
Sciences, eaux & territoires

Optimism and pessimism among expert: impact of experience





- Experience does not seem to have a strong influence on rating
- Experience modellers remain enthusiastic with time

Cemagref
Sciences, eaux & territoires

Learning from outliers

- **Some « outliers » experts are present in the room**

- **Do they wish to stand up and explain their point of view ?**

Cemagref
Sciences, eaux & territoires

What can we learn from multiple survey takers ?

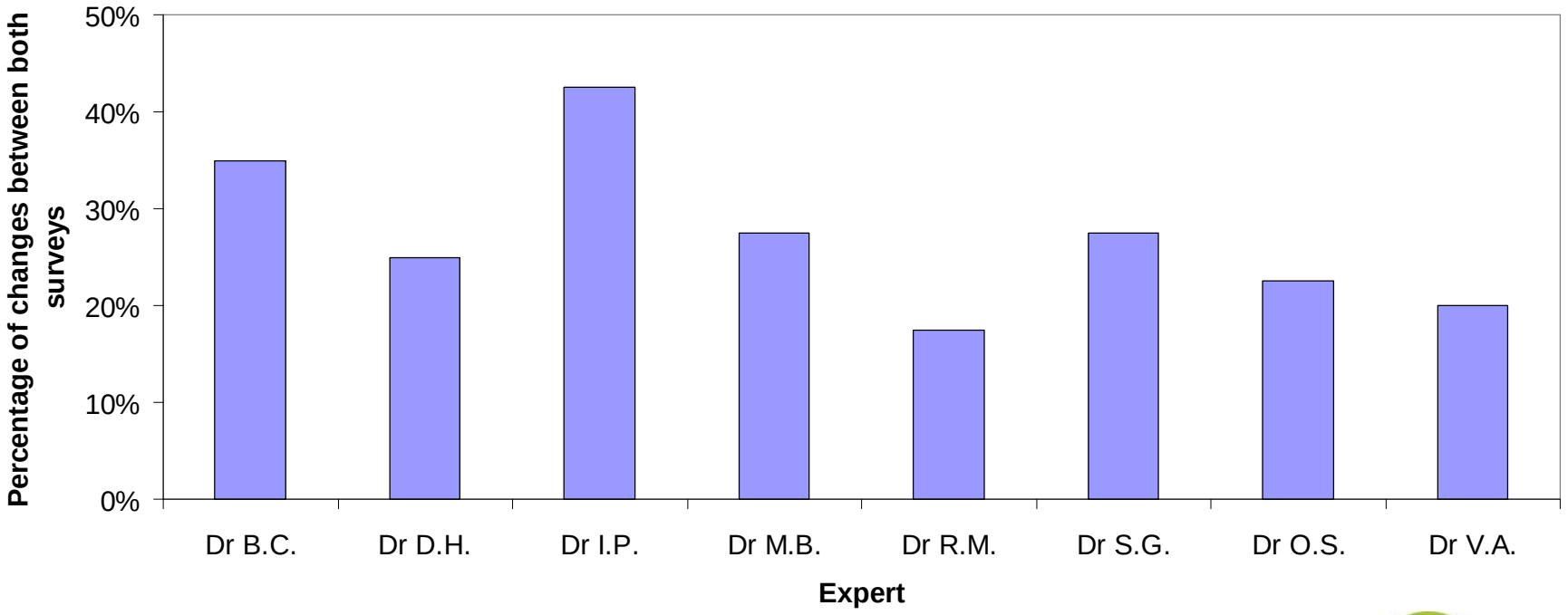
- **We analyzed the survey results of 8 participants who took the survey twice (Internet and yesterday's session)**

- **The idea is to get a generic idea of the consistency of conclusions**

Cemagref
Sciences, eaux & territoires

Simulations judged « equivalent » at first may then be judged differently…

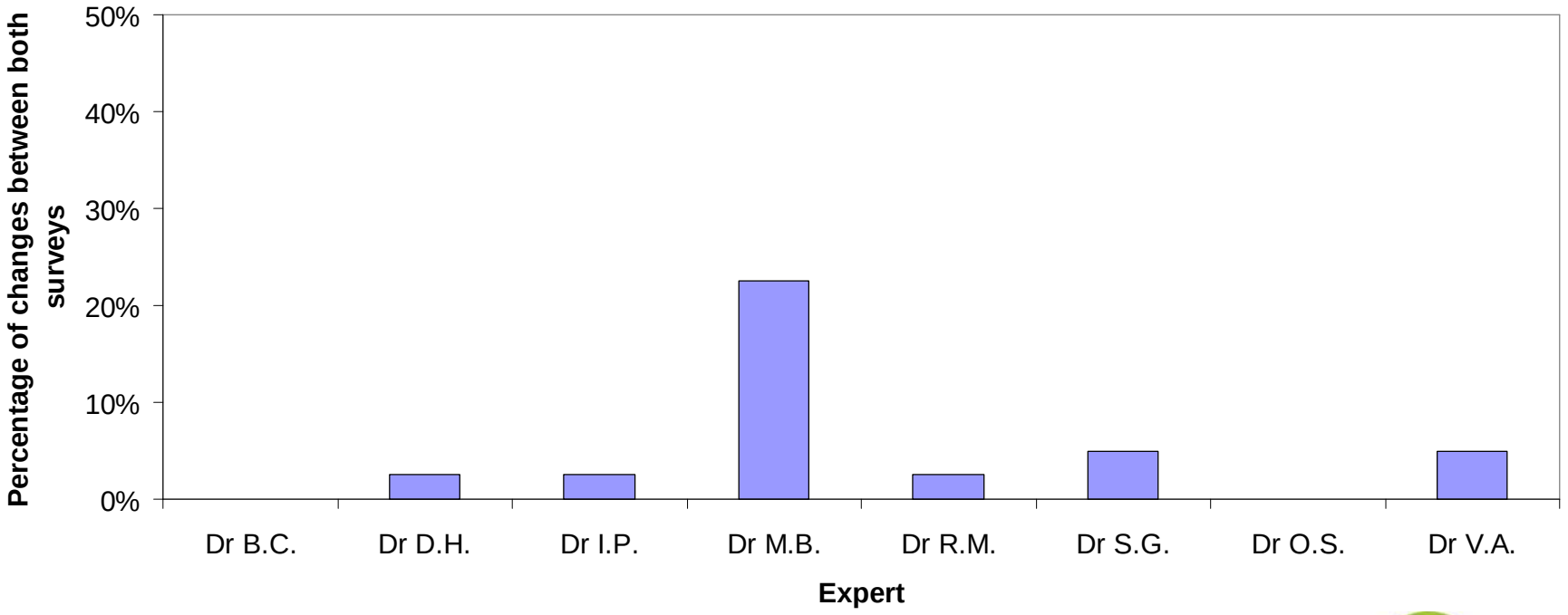
Evaluation 1 : Model A = Model B
Evaluation 2 : Model A <> Model B



Cemagref
Sciences, eaux & territoires

… but otherwise they are very consistent in their judgement

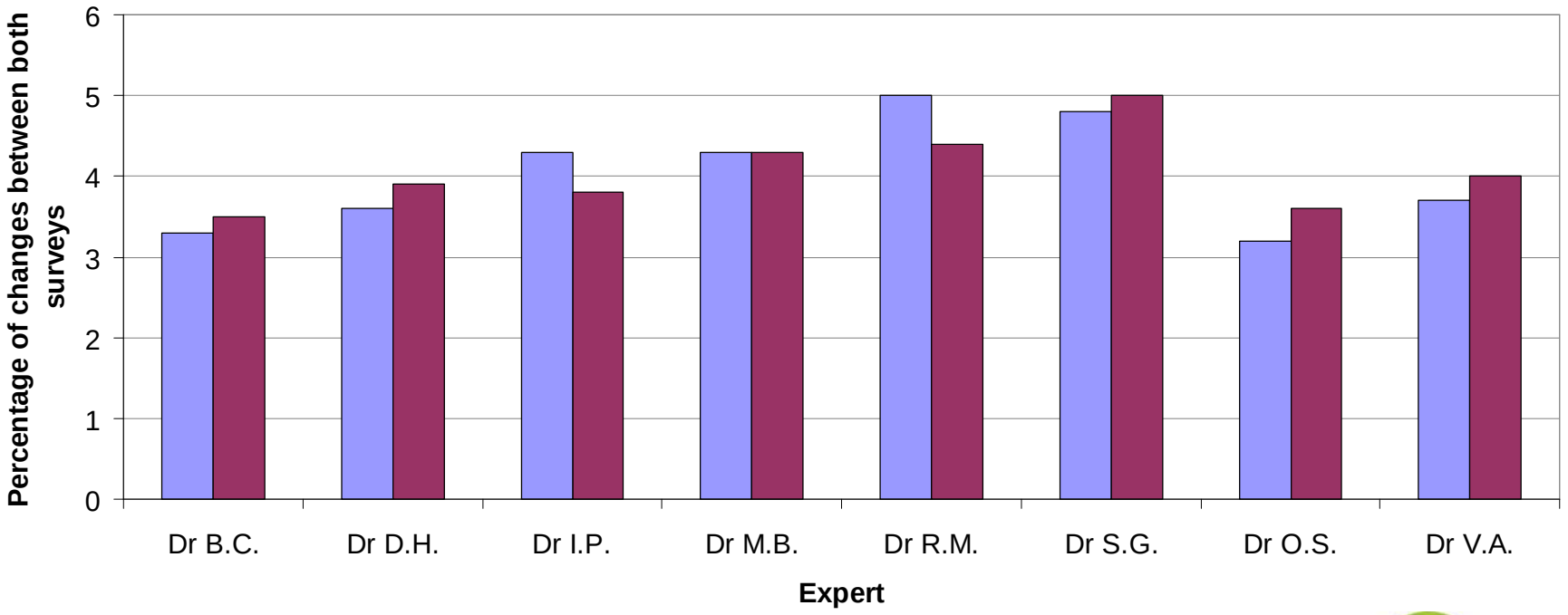
Evaluation 1 : Model A > Model B
Evaluation 2 : Model A < Model B



Cemagref
Sciences, eaux & territoires

Optimism and pessimism are rather stable overall…

Average note given:
1-Very poor to 7-Very good



Cemagref
Sciences, eaux & territoires

… but some strong differences may exist

Name	Max. difference of judgement
Dr BC	2
Dr DH	2
Dr IP	3 : P – SG
Dr MB	2
Dr RM	2
Dr SG	2
Dr OS	4 : P – G
Dr VA	3 : SP – G

Part 2.
Numerical criteria

Set of numerical criteria

- **Analysis of:**
  - **Existing reviews** (Smith, et al., 2004; Clarke, 2008; Dawson et al., 2007, 2010)
  - **Guidelines** (ASCE, 1993)
  - **Studies on criteria** (Willmott, 1981; Krause et al., 2005; Jachner et al., 2007; Legates and McCabe, 1999; Gupta et al., 2009; Ehret and Zehe, in press; Ewen, in press)
  - **Model comparisons** (WMO, 1975,1986, 1992; Chiew et al., 1993; Smith et al., 2004)
- **Selection of 60 criteria (some of them based exactly on the same error)**

Cemagref
Sciences, eaux & territoires

Set of numerical criteria

Types of criteria:

- **Continuous / Event-based**

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - S_i)^2} \qquad PDIFF = \max_{i\in[1,n]}(O_i) - \max_{i\in[1,n]}(S_i)$$

- **Absolute / Relative**

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|O_i - S_i| \qquad EI = 1 - \frac{\sum_{i=1}^{n}(O_i - S_i)^2}{\sum_{i=1}^{n}(O_i - \overline{O})^2}$$

- **Different types of errors**

$$MCE = \frac{1}{n}\sum_{i=1}^{n}(O_i - S_i) \qquad MAE = \frac{1}{n}\sum_{i=1}^{n}|O_i - S_i| \qquad MSE = \frac{1}{n}\sum_{i=1}^{n}(O_i - S_i)^2$$

Set of numerical criteria

Types of criteria:

- **Different types of transformations on flows**

$$EIR = 1 - \frac{\sum_{i=1}^{n}\left(\sqrt{O_i} - \sqrt{S_i}\right)^2}{\sum_{i=1}^{n}\left(\sqrt{O_i} - \frac{1}{n}\sum_{j=1}^{n}\sqrt{O_j}\right)^2}$$

$$EILN = 1 - \frac{\sum_{i=1}^{n}\left(\ln(O_i + \varepsilon) - \ln(S_i + \varepsilon)\right)^2}{\sum_{i=1}^{n}\left(\ln(O_i + \varepsilon) - \frac{1}{n}\sum_{j=1}^{n}\ln(O_j + \varepsilon)\right)^2}$$

$$EI_{rel} = 1 - \frac{\sum_{i=1}^{n}\left(\frac{O_i - S_i}{O_i}\right)^2}{\sum_{i=1}^{n}\left(\frac{O_i - \overline{O}}{\overline{O}}\right)^2}$$

- **Systematic errors**

$$ESC = \frac{100}{n}\underset{S_i \neq O_i}{sign}\left(S_i - O_i\right)$$

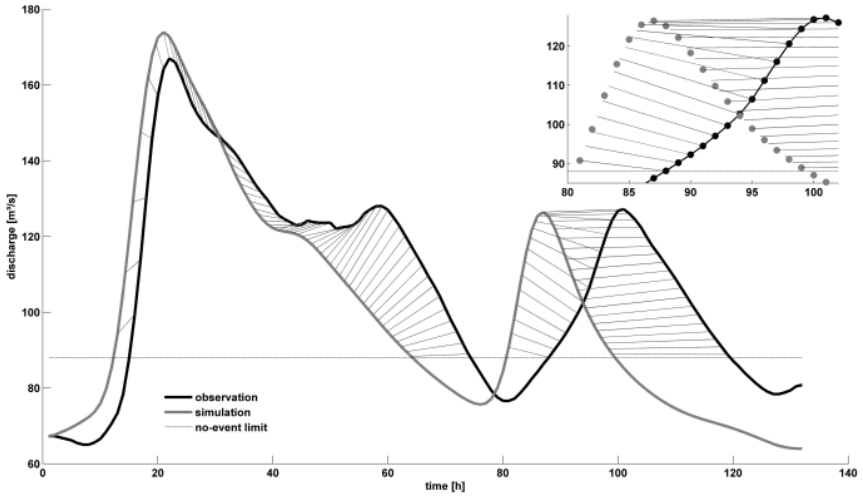
Cemagref
Sciences, eaux & territoires

Set of numerical criteria

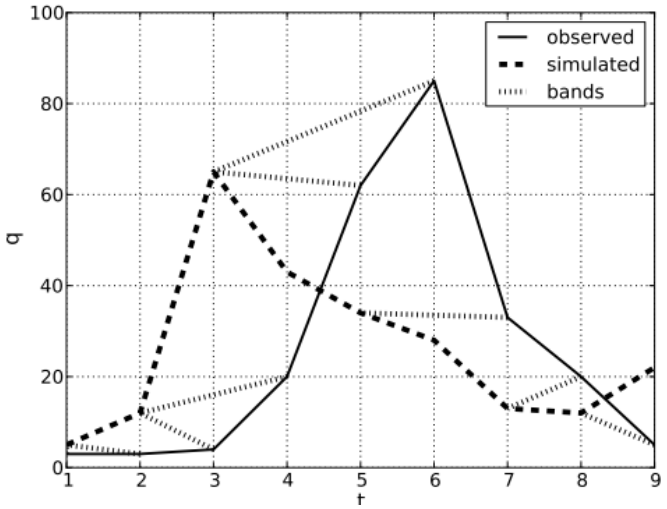
Types of criteria:

– **Criteria that mimic how the eye compares two curves**

Series distance (Ehrert and Zehe, in press)



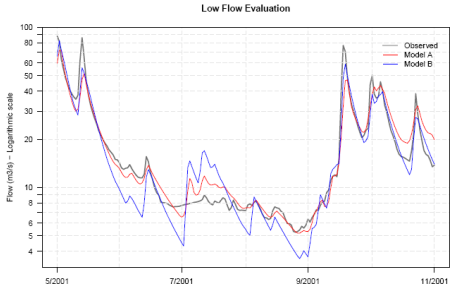
Pattern matching (morph-NSE) (Ewen, in press)

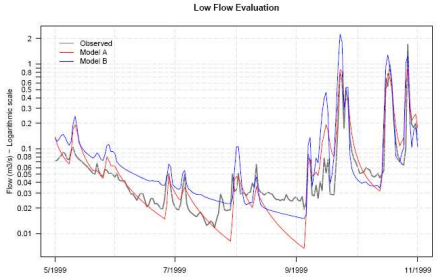


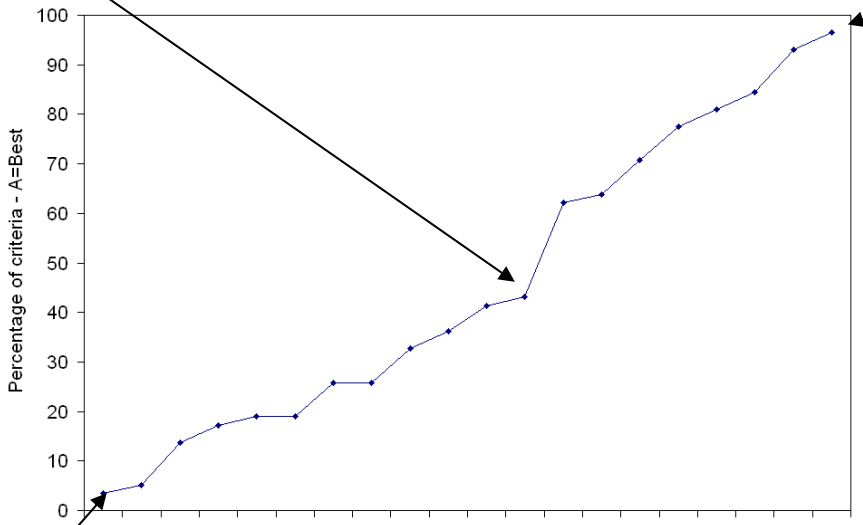
Cemagref
Sciences, eaux & territoires

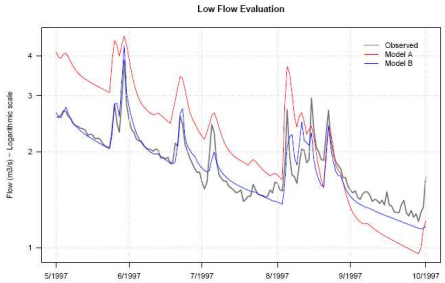
(Dis)agreement between criteria

- **Low flows**





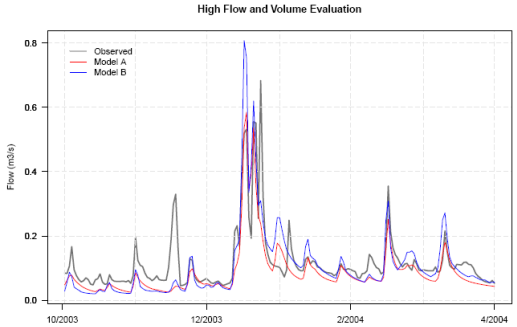


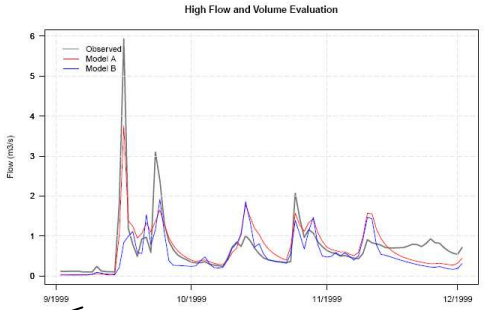


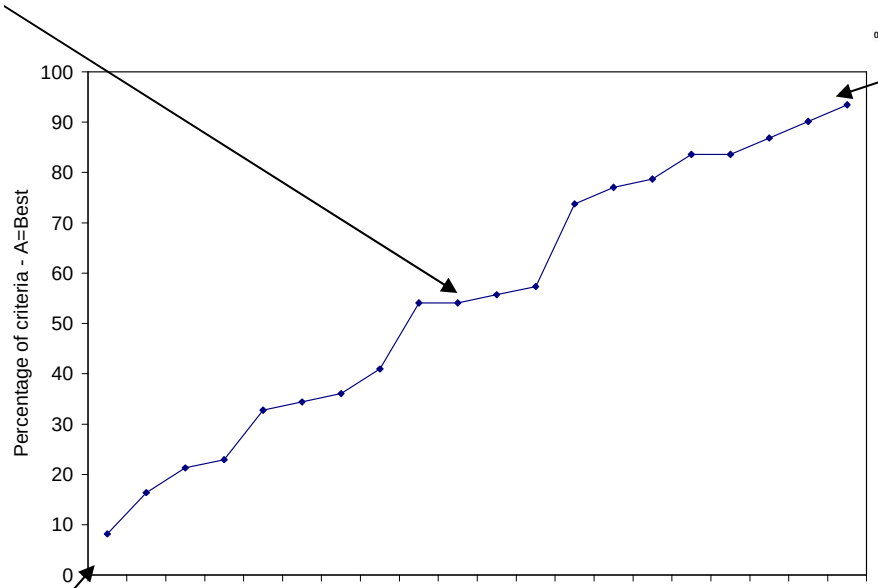
Cemagref
Sciences, eaux & territoires

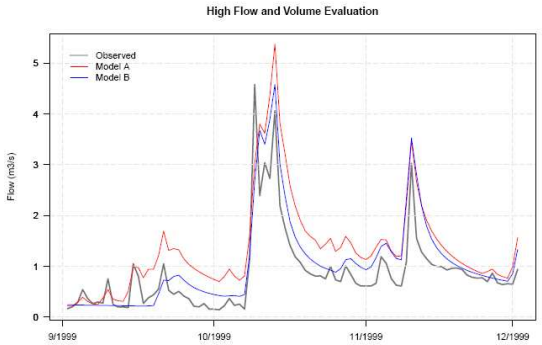
(Dis)agreement between criteria

- **High flows**









Cemagref
Sciences, eaux & territoires

Part 3.
Expert judgement vs. numerical criteria

Cemagref
Sciences, eaux & territoires

Agreement between expert judgement and numerical criteria

Agreement:

Percentage of cases where model comparison by experts and criteria give the same results in terms of model comparison

Low flows

Best matching criteria

RMLFV - Ratio of mean low flow volumes	56%
MdAPE - Median absolute percentage error	54%
MARE - Mean absolute relative error	53%
EILN - Nash-Sutcliffe efficiency index on log-flow	53%
EILF - Nash-Sutcliffe efficiency index on low flows	52%

Worst matching criteria

FTEI - Flood threshold exceedance index	31.88%
Rmod - Modified correlation coefficient	30.90%
KGE - Kling-Gupta Efficiency	30.49%
RMP - Ratio of mean flood peaks	30.14%
THREAT - Threat score	27.36%

Agreement between expert judgement and numerical criteria

Agreement:

Percentage of cases where model comparison by experts and criteria give the same results

High flows

Best matching criteria

MORPH-NSE	53%
ODMA	53%
MORPH-MAE	52%
MSES - Overall systematic error	51%
RMV - Ratio of mean flood volumes	50%

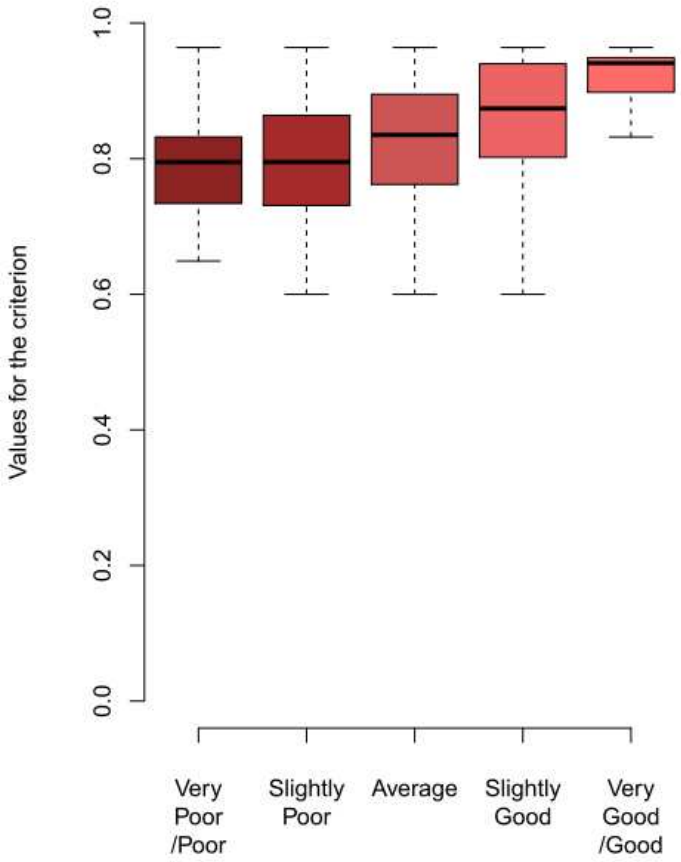
Worst matching criteria

RBFI - Ratio of base flow index	36%
MSDE - Mean squared derivative error	35%
MRE - Mean relative error*100	32%
RLFD - Ratio of low flow deficit	30%
ESC - Error sign count	27%

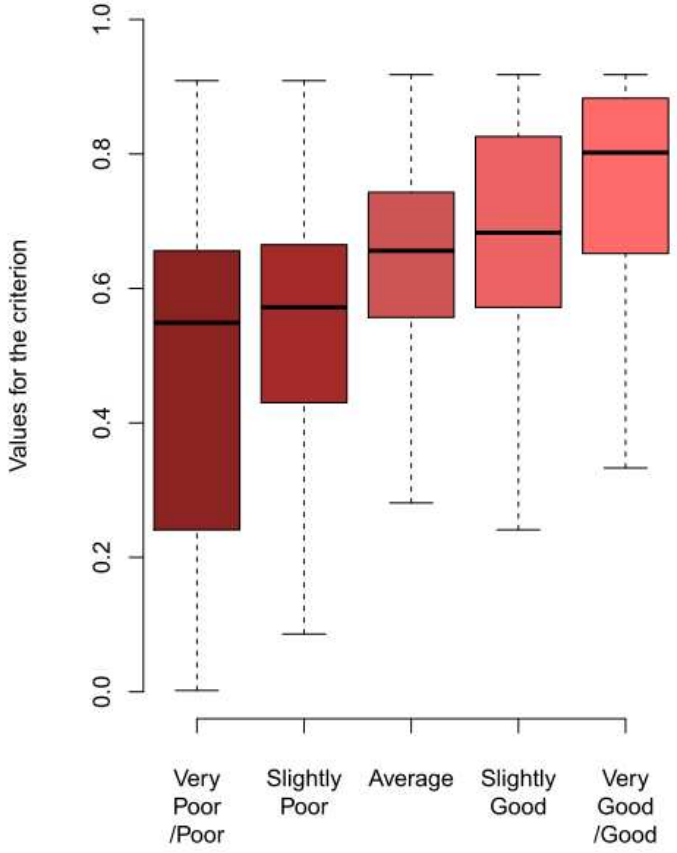
Cemagref
Sciences, eaux & territoires

Correspondence between expert rating and criteria values
Example on high flows

Morph-NSE

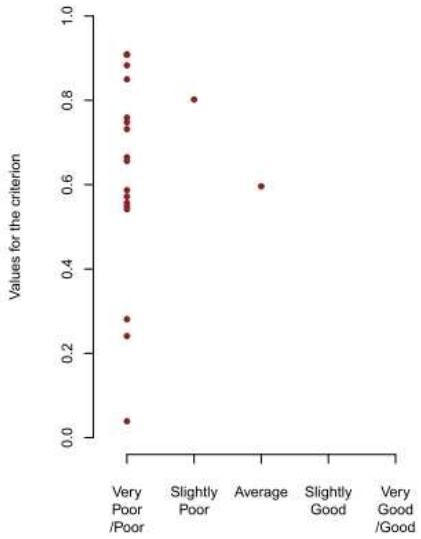


EI

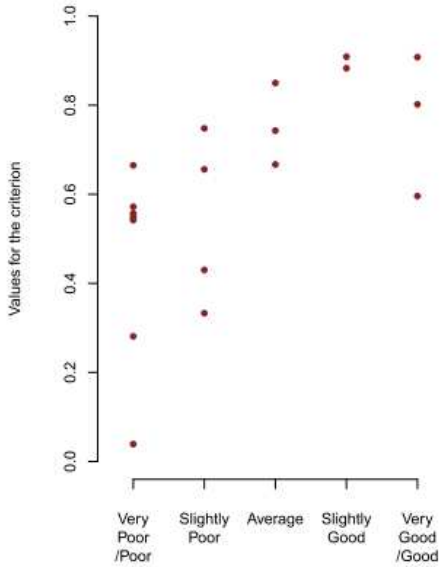


Cemagref
Sciences, eaux & territoires

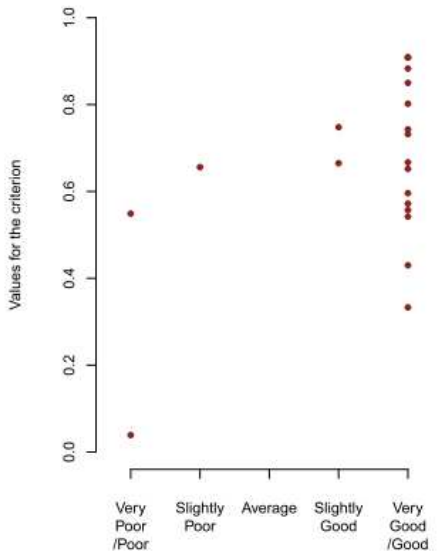
Correspondence between expert rating and criteria values: Behind the general trend, a large heterogeneity



Expert 1



Expert 2



Expert 3

Cemagref
Sciences, eaux & territoires

Conclusion & Perspectives

Cemagref
Sciences, eaux & territoires

Conclusions

- **Useful insights on which criteria better match the way the eye evaluates hydrographs**
  - **High flows: relevance of recently proposed criteria (series distance – pattern matching)**
  - **Low flows: relevance of flow volume and criteria based on transformations emphasizing low flows**
- **Large spread of results between experts, as well as various levels of agreement between criteria**
- **Difficult to find criteria that correspond to all judges**

Follow-up and perspectives

- **An individual detailed diagnosis sent to interested testers**
- **More in-depth analysis available in September**
- **Survey to renew for more specific objectives and possibly to extend to probabilistic prediction**

Cemagref
Sciences, eaux & territoires