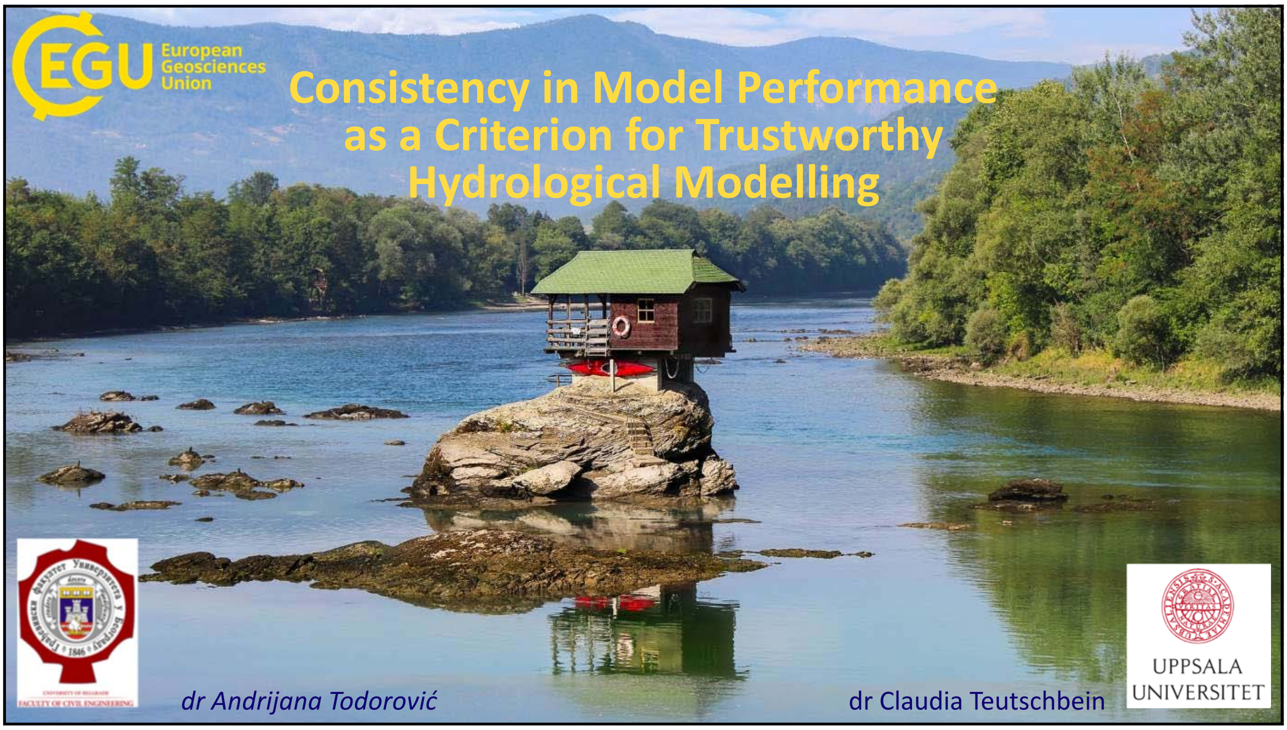
EGU22-11463

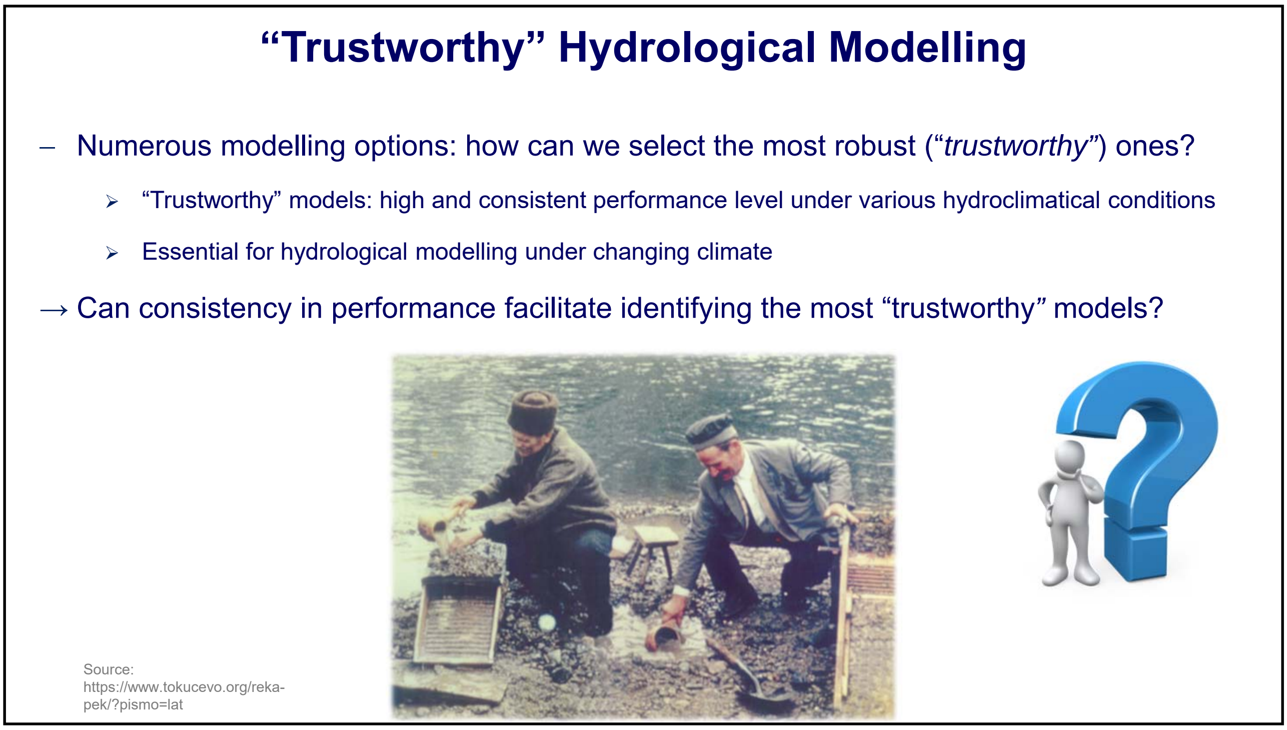


Consistency in Model Performance as a Criterion for Trustworthy Hydrological Modelling

dr Andrijana Todorović

dr Claudia Teutschbein

1



“Trustworthy” Hydrological Modelling

– Numerous modelling options: how can we select the most robust (“*trustworthy*”) ones?

- ➢ “Trustworthy” models: high and consistent performance level under various hydroclimatical conditions
- ➢ Essential for hydrological modelling under changing climate

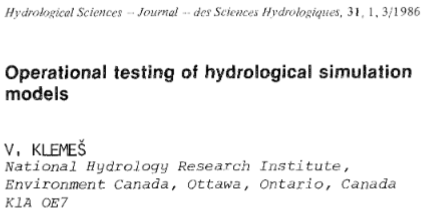
→ Can consistency in performance facilitate identifying the most “trustworthy” models?

Source: https://www.tokucevo.org/reka-pek/?pismo=lat

2

Consistency in Model Performance?

- Consistency in performance is evaluated by applying SST, DSST, or an extension thereof
  - Model performance over the full calibration period is considered



Hydrological Sciences – Journal – des Sciences Hydrologiques, 31, 1, 3/1986

Operational testing of hydrological simulation models

V. KLEMEŠ
National Hydrology Research Institute,
Environment Canada, Ottawa, Ontario, Canada
K1A OE7

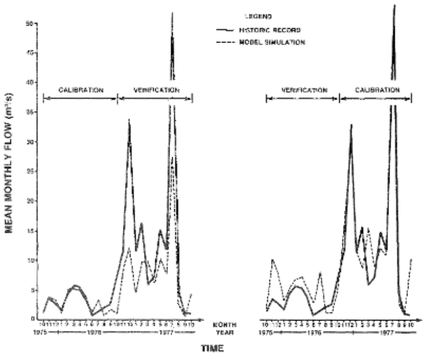
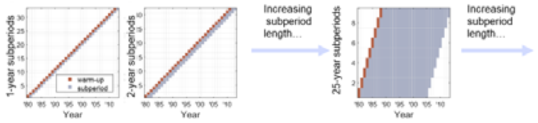
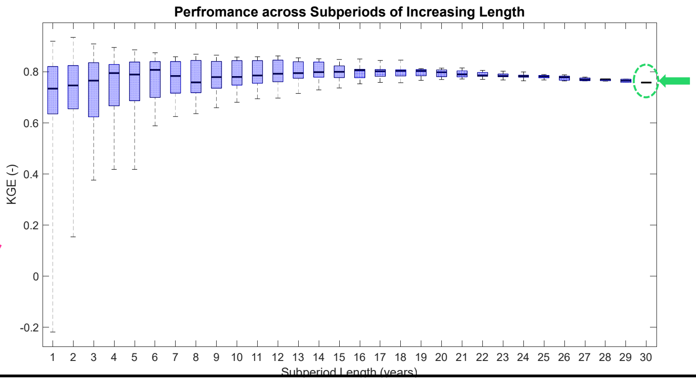


Fig. 1 Example of differential split-sample test for a simulation model for monthly flows using satellite-based information on cloud cover as input; for the Gers River at Layrac, France (Strübing, 1984).

Consistency in Model Performance?

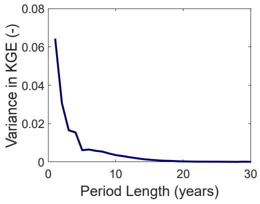
- Large variations in the model performance across different parts of the record period
  - Subperiods of increasing lengths, shifted by one year





Commonly, only performance over the full calibration period is considered

Considerable variability in performance across the subperiods



Catchments and Data

- Analyses are conducted in 3 unimpaired catchments from different climatic regions
  - The Kolubara catchment in Serbia, and the Getebro and Ytterholmen catchments in Sweden

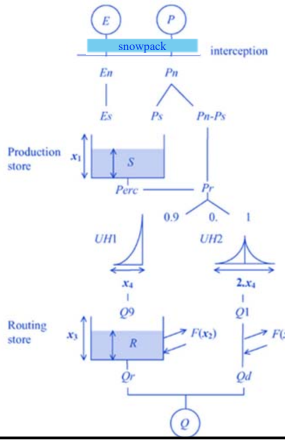
- Daily data over 60-year long record periods: precipitation, temperature and flows
  - PET is calculated for daily temperature by applying the Hamon method
  - Increase in temperatures in all catchments over the record period

Catchment	Köppen-Geiger Clim. Zone	Latitude (ºN)	A (km^2)	Elevation (m a.s.l.)	P (mm/yr)	T (ºC)	Q (mm/yr)	Runoff coeff. (-)	AI = PET/P (-)	Record period
Kolubara	Cfa	44.28	995	444.9	772.2	11.2	285.4	0.370	1.02	1954-2013
Getebro	Dfb	56.99	1333	183.0	669.7	6.4	224.6	0.335	1.20	1961-2020
Ytterholmen	Dfc	66.16	1012	254.8	676.8	0.4	371.1	0.548	1.23	1961-2020

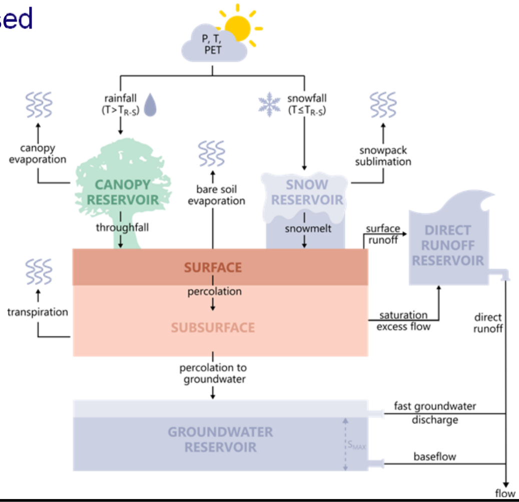
5

Hydrological Models

- Hydrological simulations with the GR4J (6) and 3DNet-Catch (23) hydrological models
  - Both models include a snow routine
- Spatially-lumped model setups are used



Perrin et al. (2003)
Bai et al. (2021)



Todorović et al. (2019)

6

Taking into Account Consistency in Model Performance (1)

- 20,000 parameter sets are created from the uniform prior distributions by applying LHS
- The performance of the parameter sets
  - Multi-temporal performance: each set is ranked according to *KGE* in each subperiod within the full calibration period (30 water years)
    - 1- through 30-year long subperiods are considered
  - Performance in the evaluation period (the 2nd half of the full record period)

Catchment	Calibration	Evaluation
Kolubara	1955-1985	1985-2013
Getebro	1962-1992	1992-2020
Ytterholmen	1962-1992	1992-2020

$$KGE = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2}$$

$$r = \frac{\sum (Q_{\mathrm{obs},i} - \bar{Q}_{\mathrm{obs}})(Q_{\mathrm{sim},i} - \bar{Q}_{\mathrm{sim}})}{\sqrt{\sum (Q_{\mathrm{obs},i} - \bar{Q}_{\mathrm{obs}})^2 \sum (Q_{\mathrm{sim},i} - \bar{Q}_{\mathrm{sim}})^2}}$$

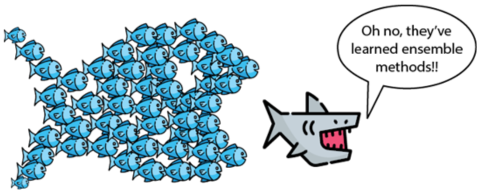
$$\alpha = \frac{\hat{S}_{Q_{\mathrm{sim}}}}{\hat{S}_{Q_{\mathrm{obs}}}} \qquad \beta = \frac{\bar{Q}_{\mathrm{sim}}}{\bar{Q}_{\mathrm{obs}}}$$

Taking into Account Consistency in Model Performance (2)

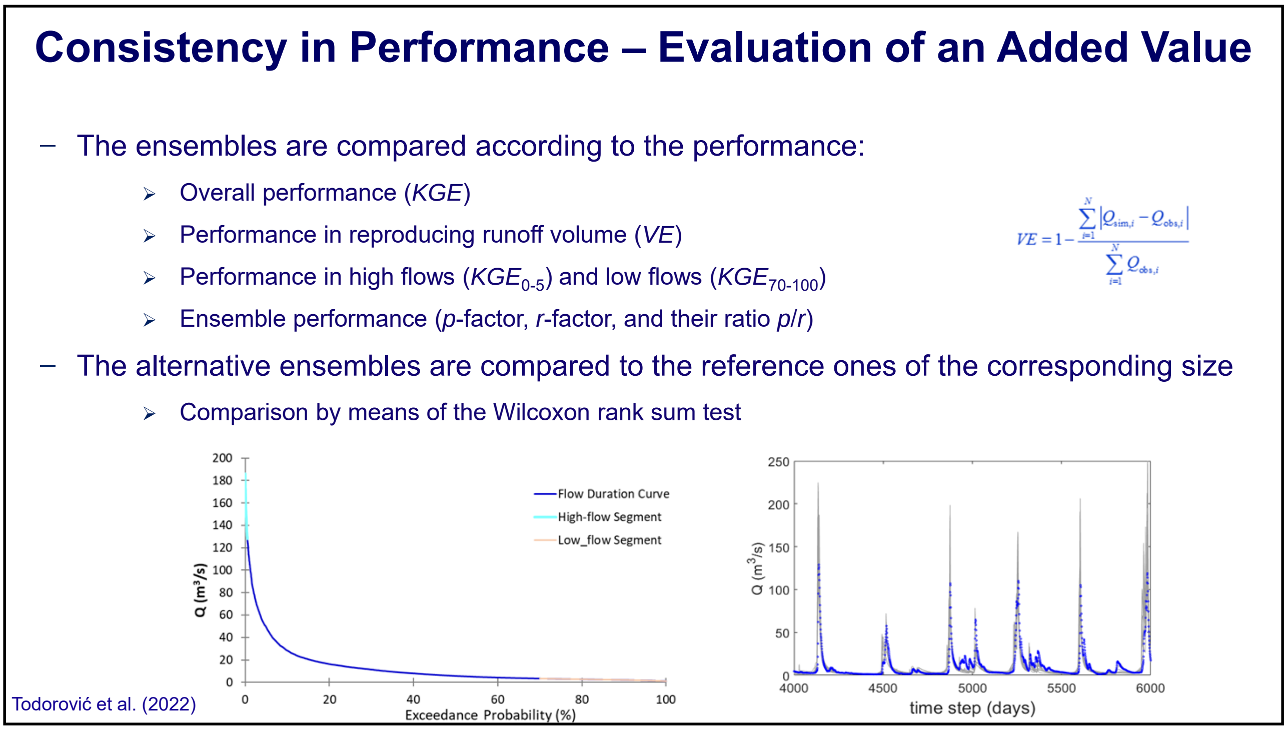
- The ensembles are created from the parameter sets:
  1) with the largest *KGE* values in the full calibration period (REFERENCE)
  2) with the highest mean rank in performance across sub-periods (RANK - MEAN)
  3) with the highest minimum rank in performance across sub-periods (RANK - MIN)

  "alternative" ensembles

- Three different ensemble sizes are considered: 1% (200), 5% (1000) and 10% (2000)
- This procedure is applied in each catchment and with both models



https://livebook.manning.com/book/grokking-machine-learning/chapter-12/

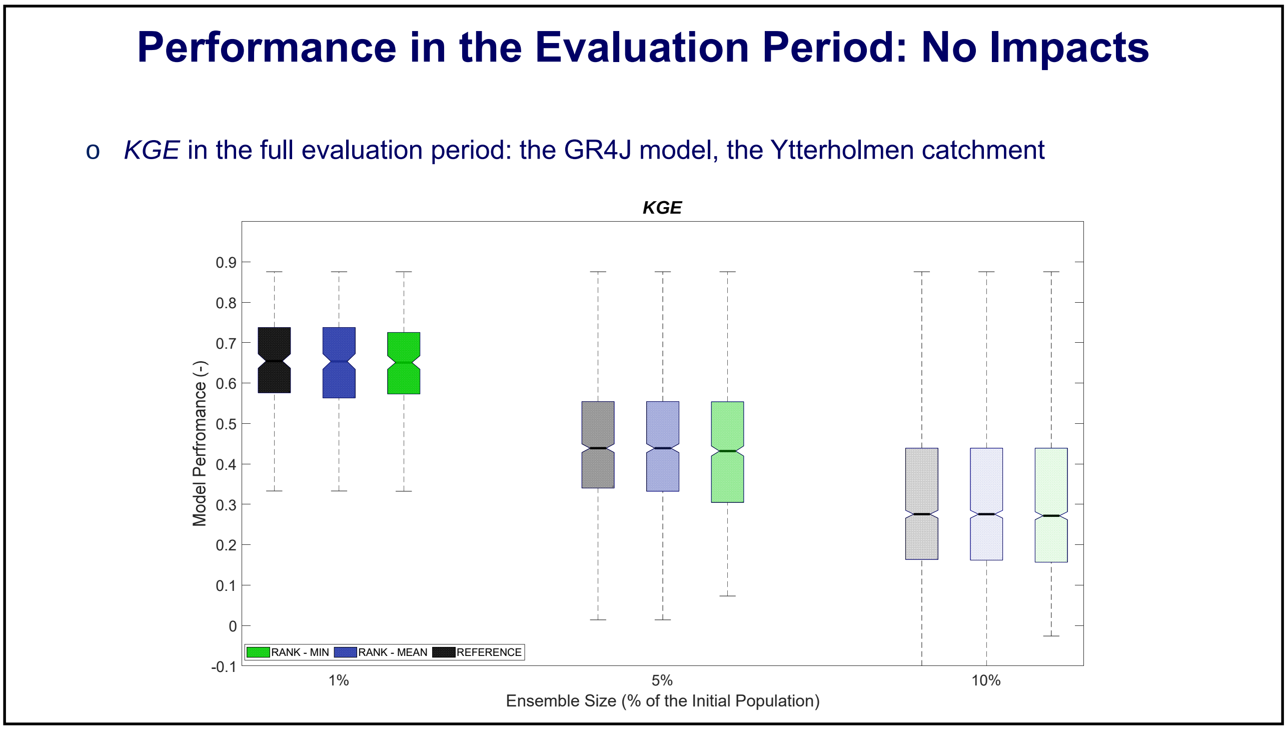


Consistency in Performance – Evaluation of an Added Value

- The ensembles are compared according to the performance:
  - Overall performance (*KGE*)
  - Performance in reproducing runoff volume (*VE*)
  - Performance in high flows ($KGE_{0\text{-}5}$) and low flows ($KGE_{70\text{-}100}$)
  - Ensemble performance (*p*-factor, *r*-factor, and their ratio *p*/*r*)
- The alternative ensembles are compared to the reference ones of the corresponding size
  - Comparison by means of the Wilcoxon rank sum test

$$VE = 1 - \frac{\sum_{i=1}^{N} \left| Q_{\mathrm{sim},i} - Q_{\mathrm{obs},i} \right|}{\sum_{i=1}^{N} Q_{\mathrm{obs},i}}$$

Todorović et al. (2022)

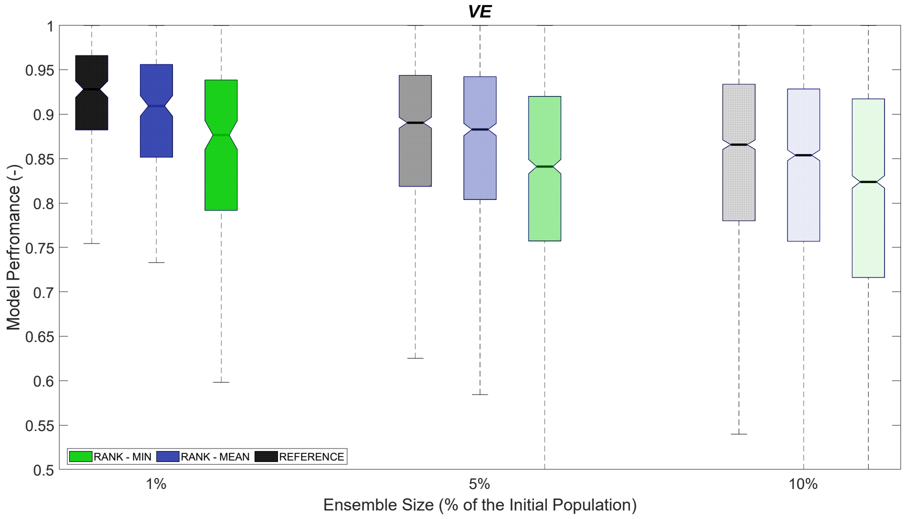


Performance in the Evaluation Period: No Impacts

- *KGE* in the full evaluation period: the GR4J model, the Ytterholmen catchment

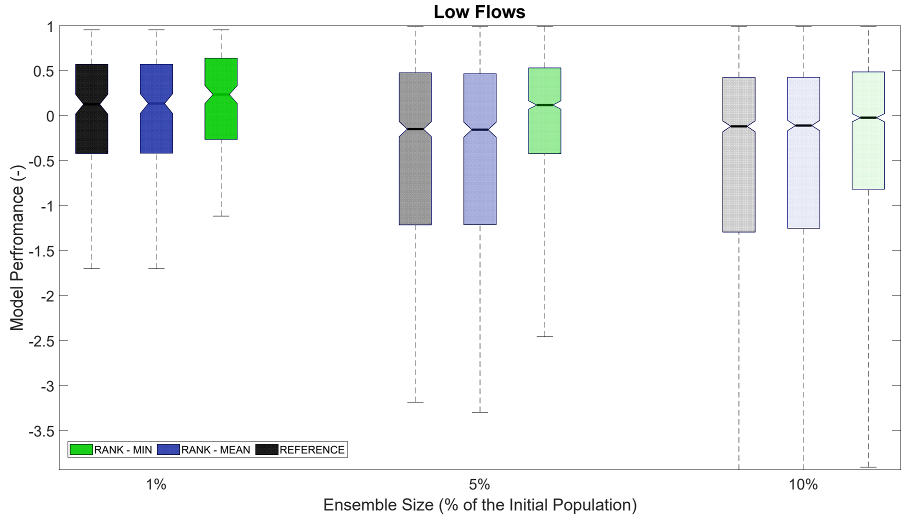
Performance in the Evaluation Period: Deterioration

- *VE* in the full evaluation period: the 3DNet-Catch model, the Getebro catchment



Performance in the Evaluation Period: Improvement

- Performance in low flows in the evaluation period: the GR4J model, the Ytterholmen catchment



Overall Performance in the Evaluation Period

- The alternative ensembles outperform the reference ones in some instances
  - Statistically significant differences in favour of the alternative ensembles according to the Wilcoxon-rank sum test (green triangles)
- High variability across performance indicators, models and catchments
  - Neither way of creating the alternative ensembles is shown superior to the other
  - Slightly higher frequency of improvement is obtained with ensembles with lower thresholds (10%)

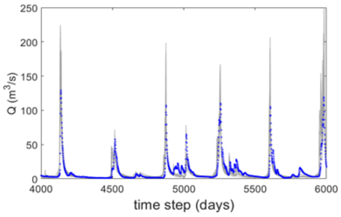
		3DNet-Catch								GR4J							
		KGE		*VE*		HF		LF		*KGE*		*VE*		HF		LF	
		MEAN	MIN	MEAN	MIN	MEAN	MIN	MEAN	MIN	MEAN	MIN	MEAN	MIN	MEAN	MIN	MEAN	MIN
Kolubara	1%	▲	●	▲	▲	●	●	●	▲	●	●	●	●	●	●	▲	▲
	5%	▲	●	▲	▲	●	●	●	●	●	●	●	●	●	●	▲	▲
	10%	▲	▲	●	●	▲	▲	●	●	▲	●	▲	●	▲	●	●	●
Gettebro	1%	●	●	●	●	●	●	▲	●	●	●	●	●	●	▲	●	●
	5%	●	●	●	●	●	●	▲	●	●	●	●	●	●	●	●	▲
	10%	●	●	▲	●	▲	▲	▲	▲	●	●	●	●	●	●	●	●
Ytterholmen	1%	●	●	●	●	●	▲	●	●	●	●	●	●	●	●	●	▲
	5%	●	●	●	●	●	▲	●	●	●	●	●	●	●	●	●	▲
	10%	●	●	▲	▲	●	●	▲	▲	●	●	●	●	●	●	●	●

13

Ensemble Performance in the Evaluation Period

- Generally similar performance of the three ensembles
  - In many cases, alternative ensembles have slightly higher values of the *p*/*r* ratios than the corresponding reference ensemble

			3DNet-Catch			GR4J		
			p- factor	*r*- factor	*p/r*	*p*- factor	*r*- factor	*p/r*
Gettebro	1%	Refrerence	60%	1.04	57%	**65%**	1.33	48%
		Median	60%	**1.02**	**59%**	64%	1.31	49%
		Minimum	**69%**	1.41	49%	64%	**1.25**	**51%**
	5%	Refrerence	75%	1.34	56%	77%	1.73	45%
		Median	74%	**1.32**	**56%**	77%	1.73	45%
		Minimum	**86%**	1.59	54%	**80%**	**1.54**	**52%**
	10%	Refrerence	79%	1.51	52%	85%	1.99	43%
		Median	78%	**1.48**	**53%**	**85%**	1.97	43%
		Minimum	**88%**	1.72	51%	85%	**1.77**	**48%**



14

Concluding Remarks and Outlook

- Multi-temporal performance can facilitate identification of “trustworthy” models in some cases
- Identification of the “trustworthy” models remains a challenge in hydrology
- Further research is needed
  - What exactly causes variability in the model performance across time scales?
  - How can we use multi-temporal performance to improve model structures or calibration strategies?
  - How does this variability behave in catchments with strong trends?…

https://img.freepik.com/free-photo/man-jumping-impossible-possible-cliff-sunset-background-business-concept-idea_1323-266.jpg?w=360

15

Thank you for your attention!

REFERENCES

Bai, P., Liu, X., & Xie, J. (2021). Simulating runoff under changing climatic conditions: A comparison of the long short-term memory network with two conceptual hydrologic models. Journal of Hydrology, 592, 125779. https://doi.org/10.1016/j.jhydrol.2020.125779

Klemeš, V. (1986). Operational testing of hydrological simulation models. *Hydrological Sciences*, *31*(1), 13–24. https://doi.org/10.1080/02626668609491024

Perrin, C., Michel, C., & Andre, V. (2003). Improvement of a parsimonious model for streamflow simulation. *Journal of Hydrology*, *279*(1–4), 275–289. https://doi.org/10.1016/S0022-1694(03)00225-7

Todorović, A., Stanić, M., Vasilić, & Plavšić, J. (2019). The 3DNet-Catch hydrologic model: Development and evaluation. *Journal of Hydrology*, *568*. https://doi.org/10.1016/j.jhydrol.2018.10.040

Todorović, Andrijana, Grabs, T., & Teutschbein, C. (2022). Advancing traditional strategies for testing hydrological model fitness in a changing climate. *Hydrological Sciences Journal*, 1–22. https://doi.org/10.1080/02626667.2022.2104646

European Geosciences Union

atodorovic@grf.bg.ac.rs

16