Can a hydrological model be robust and efficient at the same time ?

A multicriteria crash test to assess the limit of model robustness across flow ranges

Paul Royer-Gaspard, Vazken Andréassian, Guillaume Thirel,
Charles Perrin, François Bourgin

INRAE, Antony, France

EGU2020

Take home messages

- **We propose a crash test to identify model performance trade-offs in multi-objective parameter selection of rainfall-runoff models**

- The crash test is applied to GR4J on 382 French catchments, with bias and robustness metrics calculated over three flow ranges

- Compromises between simulation ability over three flow ranges strongly limit model robustness

- Model robustness may be overestimated by studies focusing on average streamflows

- This diagnostic scheme can help developing polyvalent rainfall-runoff models

Structure of the display

Choose between quick or detailed presentation

- **Link to the easy-to-read content (2 minutes reading)**
  Key points

- Links to in-depth details of the study

Key points

Quick presentation of the study

KEY POINTS

Rationale of the approach

- Rainfall-runoff models lack of robustness in a changing climate context

- Specific model developments should focus on this issue, by:
  - Finding better calibration techniques (Fowler et al., 2016; 2018)
  - Preventing model complexity from disturbing robust parameter calibration (Andréassian et al., 2012)
  - Improving processes plausibility (Fowler et al., 2020)

- This study presents a crash test to identify model weak spots based on a multi-objective framework including robustness metrics over different ranges of streamflow

- The crash test is applied on GR4J in a large set of French catchments

KEY POINTS

Methodology

- The crash test is based on an extensive exploration of the parameter space to avoid questioning calibration issues

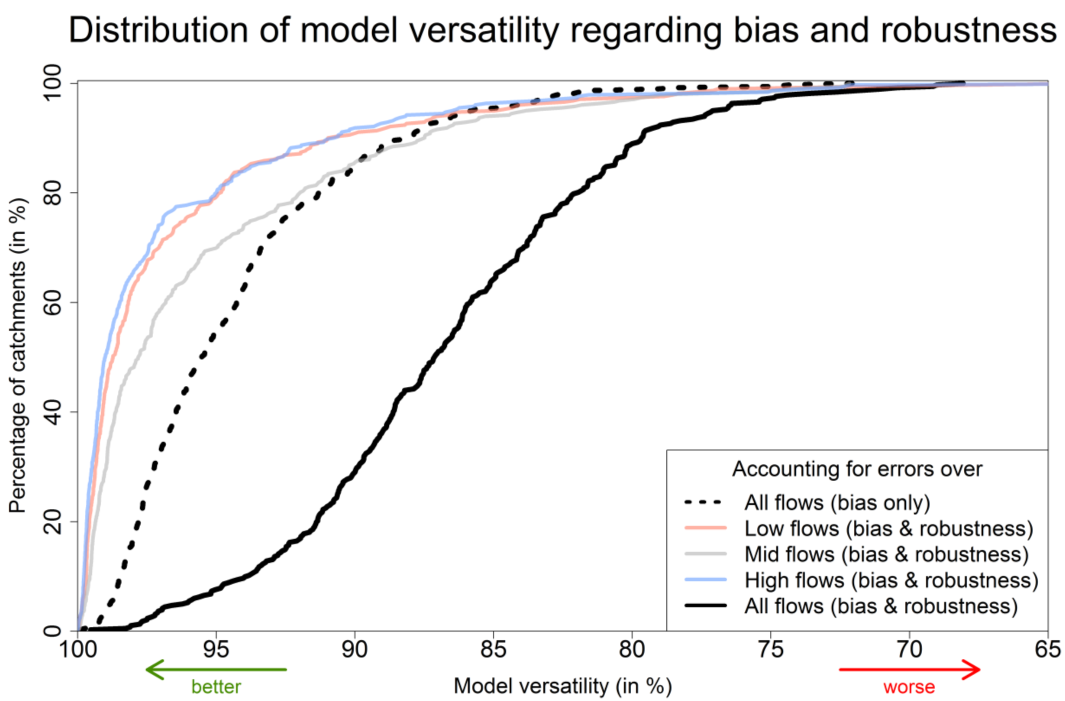
1. Most behavioural parameter sets → **2.** Model overall performance → **3.** Incompatible performance requirements → **4.** Improvable model processes

- Six metrics evaluate model's ability to provide unbiased and robust simulations over low flows, mid flows and high flows

- Model structural flaws are identified by analyzing:
  - Performance requirements that the model cannot match simultaneously
  - Model parameters to find incompatible patterns

KEY POINTS

Results



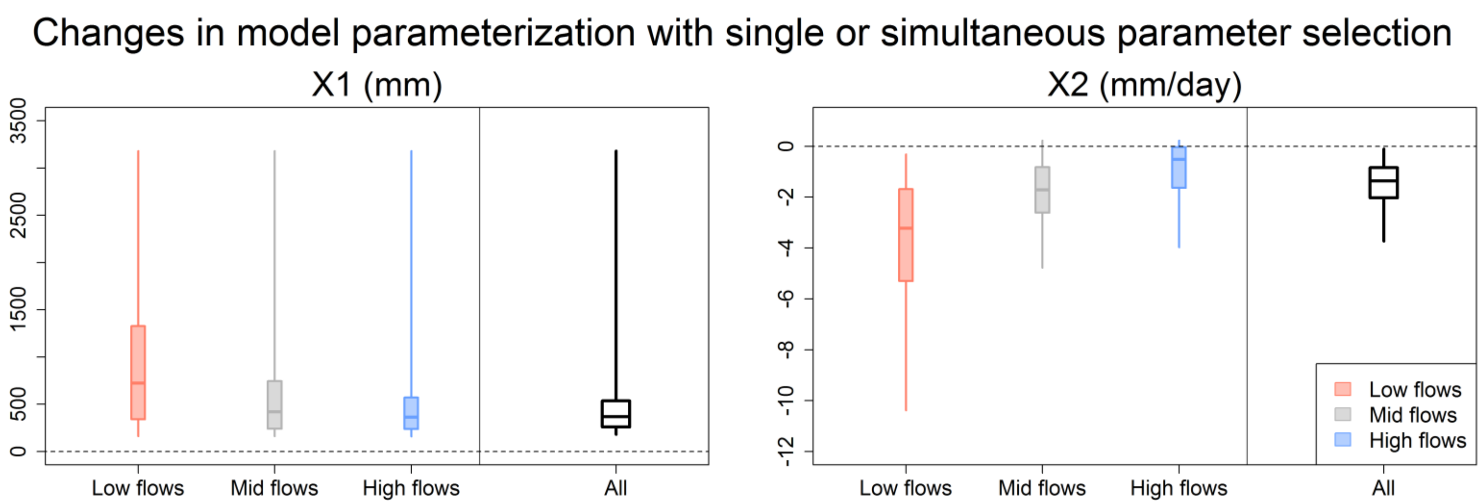
✓ Model's ability to provide robust and unbiased simulations of either low flows, mid flows or high flows is correct

❑ Simultaneous performance requirements over multiple ranges of flow yields to severe performance compromises

KEY POINTS

Results

- The two parameters controlling model water balance in GR4J (X_1, X_2) suffer contradictory constraints to match either low flows or high flows



- We suggest that the production store of the model may struggle to represent consecutive dry years while reacting fast enough to strong rainfall events

KEY POINTS

End

- Go to conclusion
- Go to table of contents
- Go to detailed display

Details

In-depth description of the methodology
and the results

Introduction

- Rainfall-runoff models lack robustness in changing climate conditions (e.g. Thirel et al., 2015)

- Models calibrated on wet (dry) periods and validated on dry (wet) periods tend to overestimate (underestimate) average streamflow (e.g. Coron et al., 2012)

- Authors generally raise the need to improve models structure to overcome « excessive » process simplification (e.g. Coron et al., 2014; Fowler et al., 2020)

- Imperfect calibration also contributes to the general lack of robustness, yielding to wrongfully discard models that are actually robust in some cases (Fowler et al., 2016)

Introduction

- Calibration issues are aggravated by compensation for data errors or for irrelevant model structure, by suboptimal algorithm, inadequate objective functions, model complexity…
  (Beven, 2006; Andréassian et al., 2012; Fowler et al., 2018)

- These issues can be alleviated by improving model structure toward :
  - Increased versatility over various flow ranges (to limit overcalibration)
  - Improved plausibility in process representation
  - Structural simplicity (to limit equifinality and suboptimality)

- Therefore, model improvements should focus on structural weak points strongly compromising its performance, to avoid excessive complexification

- Guidelines are thus required to advance diagnosis of model structures.

Scope of the study

- This study presents a crash test to assess performance compromises caused by structural weaknesses of conceptual rainfall-runoff models

- Compromises are evaluated with regard to a multi-objective framework including:
  - Bias metrics over three ranges of flow (low flows, mid flows, high flows)
  - Specifically designed robustness metrics over the three ranges of flow

- The parameter space is extensively explored to avoid questioning calibration issues such as choice of objective function or optimization algorithm

- The crash test is applied to GR4J in 382 French catchments

Principles of the approach

1. Seek parameters set exhibiting the highest versatility, i.e. reaching the most decent performances in a multi-objective framework
2. Evaluate the severity of compromises between model performance metrics
3. Identify the performance metrics for which decent scores cannot be matched simultaneously
4. Assess model processes involved in complementary performance requirements and thus limiting model versatility

1. Most behavioural parameter sets → **2.** Model overall performance → **3.** Incompatible performance requirements → **4.** Improvable model processes

Data and methods

Data and model

- GR4J (Perrin et al. 2003)
  - 4 parameters
  - Water balance is controlled by X_1 (soil moisture accounting) and X_2 (groundwater intercatchment exchange)
- 382 French catchments (appendix)
  - Almost unregulated
  - Variety of physical and hydroclimatic conditions

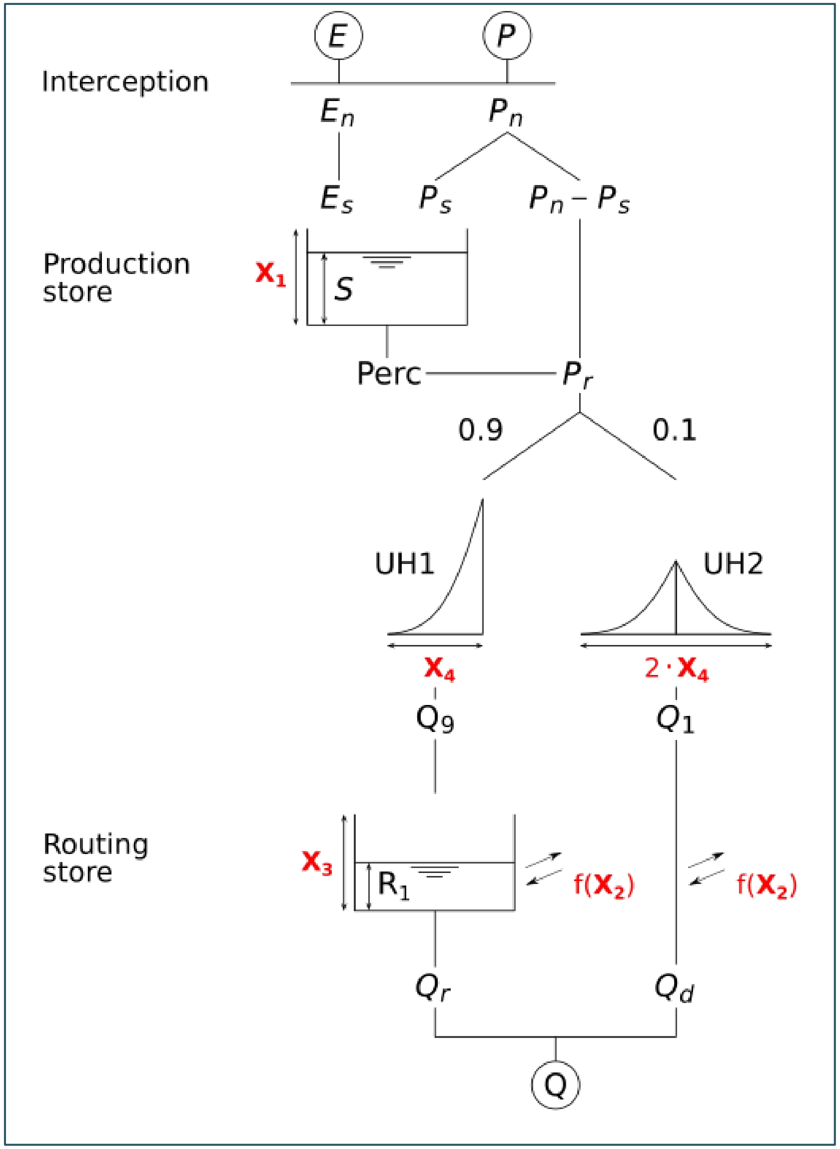


Fig1. GR4J structure

Data and methods

Performance metrics

3 bias metrics

- Bias over low flows ($Q \leq Q_{20\%}$)
- Bias over mid flows ($Q_{20\%} < Q < Q_{80\%}$)
- Bias over high flows ($Q_{20\%} \leq Q$)

→ bias values are accounted in absolute terms compared to 1, as follows:

$$Bias = \left|\frac{Q_{sim}}{Q_{obs}} - 1\right| * 100\%$$

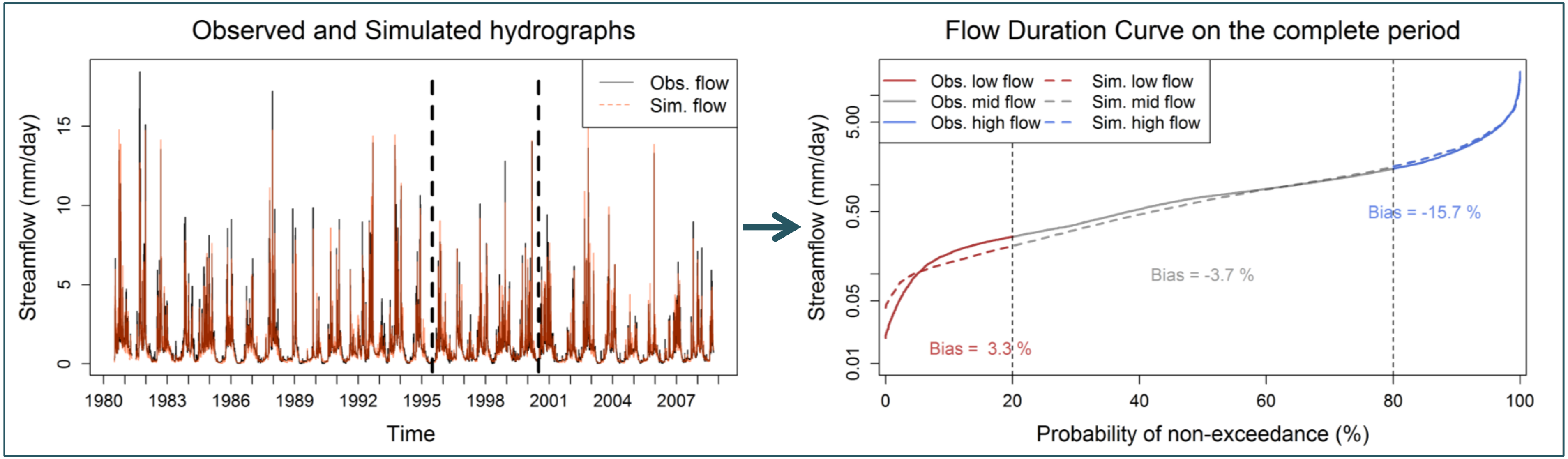


Fig2. Computation of bias metrics

Data and methods

Performance metrics

3 bias metrics

- Bias over low flows ($Q \leq Q_{20\%}$)
- Bias over mid flows ($Q_{20\%} < Q < Q_{80\%}$)
- Bias over high flows ($Q_{20\%} \leq Q$)

→ bias values are accounted in absolute terms compared to 1, as follows:

$$Bias = \left|\frac{Q_{sim}}{Q_{obs}} - 1\right| * 100\%$$

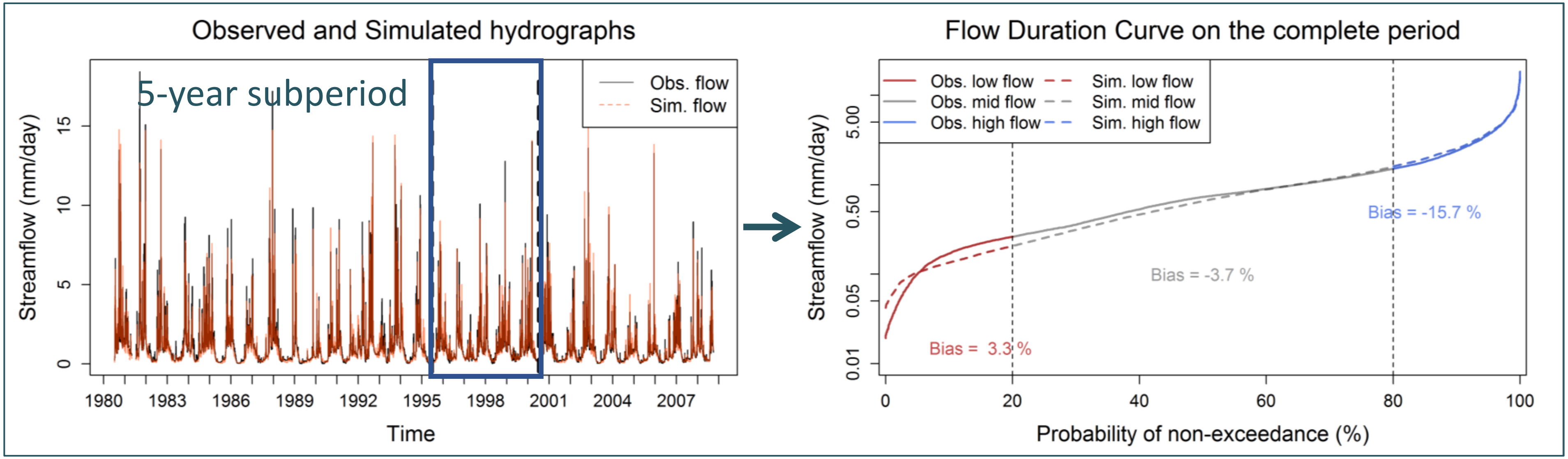


Fig2. Computation of bias metrics

Data and methods

Performance metrics

3 robustness metrics

- Robustness over low flows ($Q \leq Q_{20\%}$)
- Rob. over mid flows ($Q_{20\%} < Q < Q_{80\%}$)
- Robustness over high flows ($Q_{80\%} \leq Q$)

→ Computed as the average of the variations of model bias over sliding 5-year subperiod (inspired from Coron et al., 2014)
→ In other words, computed as the absolute area of the colored shaded zones in the figure below

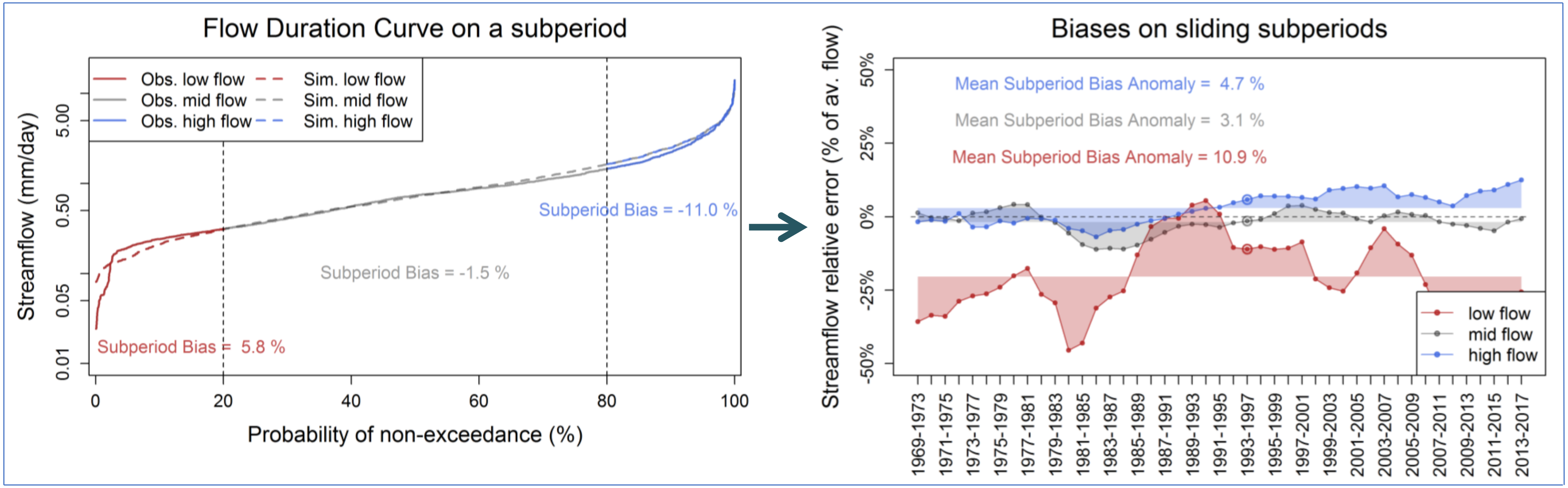


Fig3. Computation of robustness metrics

Data and methods

Performance metrics

- In summary, model performance is evaluated against 6 metrics targeting various ranges of flow and water balance over different timescales

Metric	Targeted timescale	Targeted range of flows	Optimum
Bias over low flows ($\boldsymbol{B_{lf}}$)	$>$ 20 years	$Q < Q_{20\%}$	0
Bias over mid flows ($\boldsymbol{B_{mf}}$)	$>$ 20 years	$Q_{20\%} < Q < Q_{80\%}$	0
Bias over high flows ($\boldsymbol{B_{hf}}$)	$>$ 20 years	$Q > Q_{80\%}$	0
Robustness over low flows ($\boldsymbol{R_{lf}}$)	5 years	$\tilde{Q} < \tilde{Q}_{20\%}$	0
Robustness over mid flows ($\boldsymbol{R_{mf}}$)	5 years	$\tilde{Q}_{20\%} < \tilde{Q} < \tilde{Q}_{80\%}$	0
Robustness over high flows ($\boldsymbol{R_{hf}}$)	5 years	$\tilde{Q} > \tilde{Q}_{80\%}$	0

- In the following, we apply the crash test on GR4J with the 6 metrics
- Various multi-objective framework have been tested by selecting subsets within the set of 6 metrics

1. Seek most behavioural parameter sets

Methodology

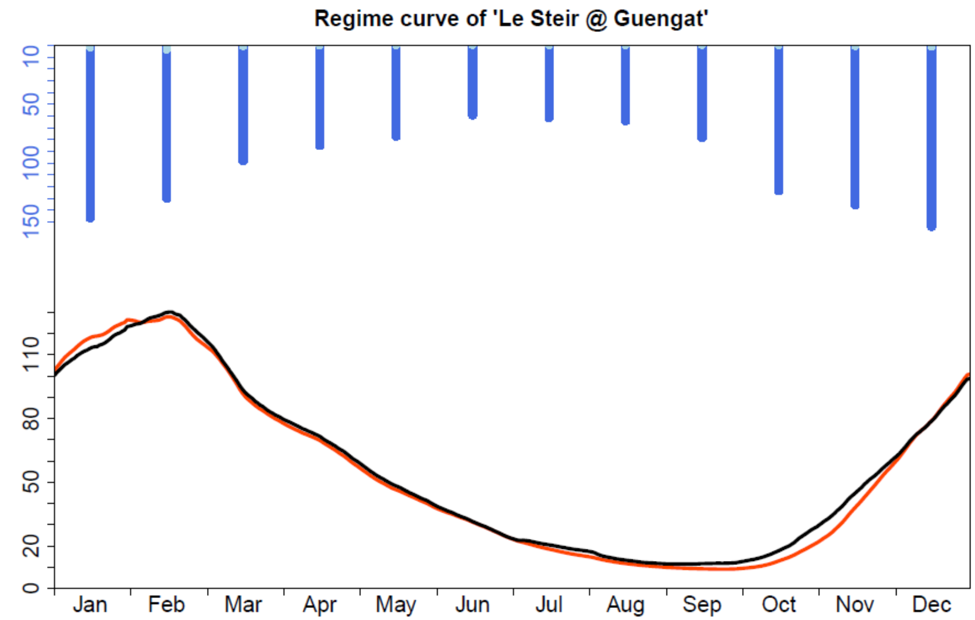
- On each catchment, we test N *parameter sets* and compute their performance on K *metrics*
(In this study, $N = 10.000$ and $K = 6$)
- For each metric k, each parameter set i is associated to a quantile of performance $q_{i,k}$ regarding its ranking among all the parameter sets
- Ideally, there exists a $best$ parameter set reaching $\forall k, q_{best,k} = 100\%$ meaning that it is better than all other parameter sets for each metric
- In most cases though, this parameter set does not exist
- Thus, we define the *most behavioural parameter set* as the one maximizing $\underset{k}{\mathrm{mean}}(q_{i,k})$
- The *versatility* score of the model is computed as $\hat{q} = \max_i \left[\underset{k}{\mathrm{mean}}(q_{i,k})\right]$
- In an ideal case, $\hat{q} \rightarrow 100\%$

1. Seek most behavioural parameter sets

Example on the Steir River @ Guengat

- The model almost perfectly simulates the river regime
- Model biases curve on sliding subperiods is flat
- ➢ No performance compromises

(Model versatility is around 98%)



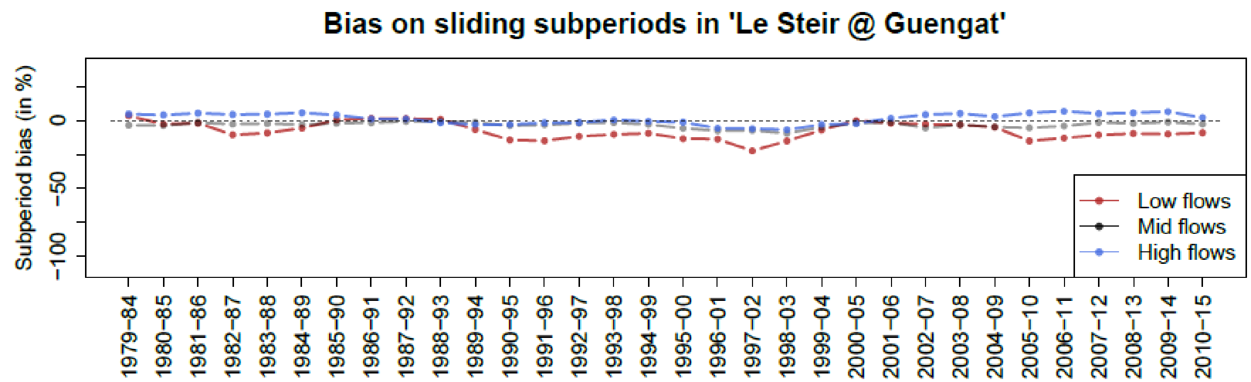


Fig4. Example 1

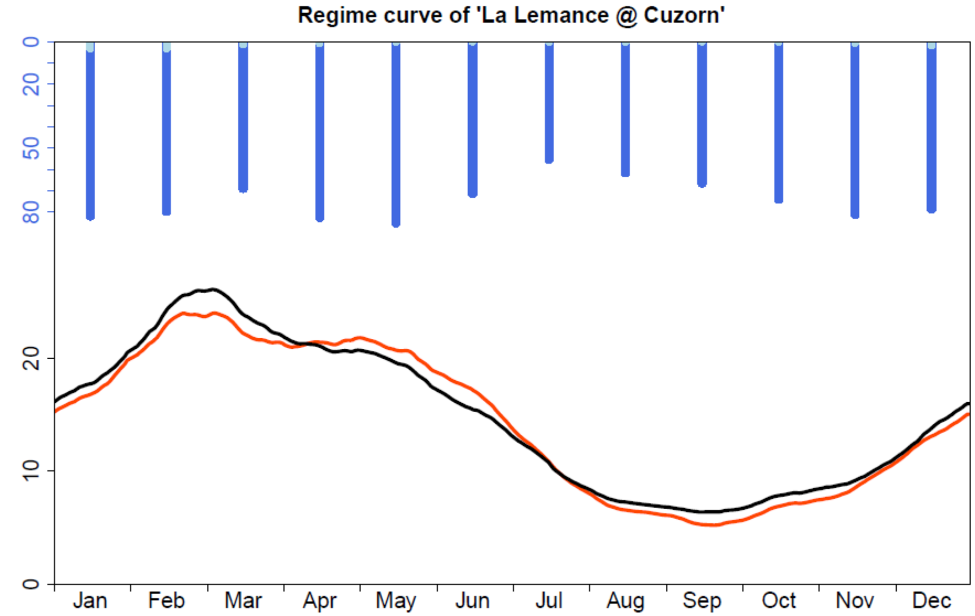
1. Seek most behavioural parameter sets

Example on the Lemance River @ Cuzorn

- The model struggles to simulate the river regime
- Model biases curve on sliding subperiods is not flat (for low flows especially)

➢ Strong performance compromises

(Model versatility is around 80%)



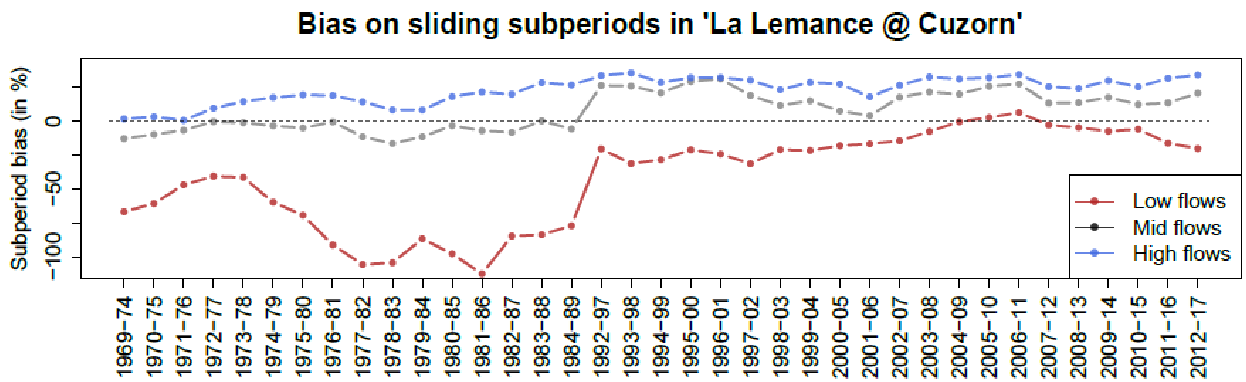


Fig5. Example 2

2. Model overall performance

Model versatility

- GR4J versatility scores are computed with different combinations of metrics (Fig6) (all 3 biases; 2 low, mid or high flows metrics; all 6 metrics)

✓ Model's ability to match bias and robustness of the same range of flow is quite good

✓ Model's ability to provide unbiased simulations over the three ranges of flow is also correct

❑ Model versatility is though severely affected if the six metrics are accounted for

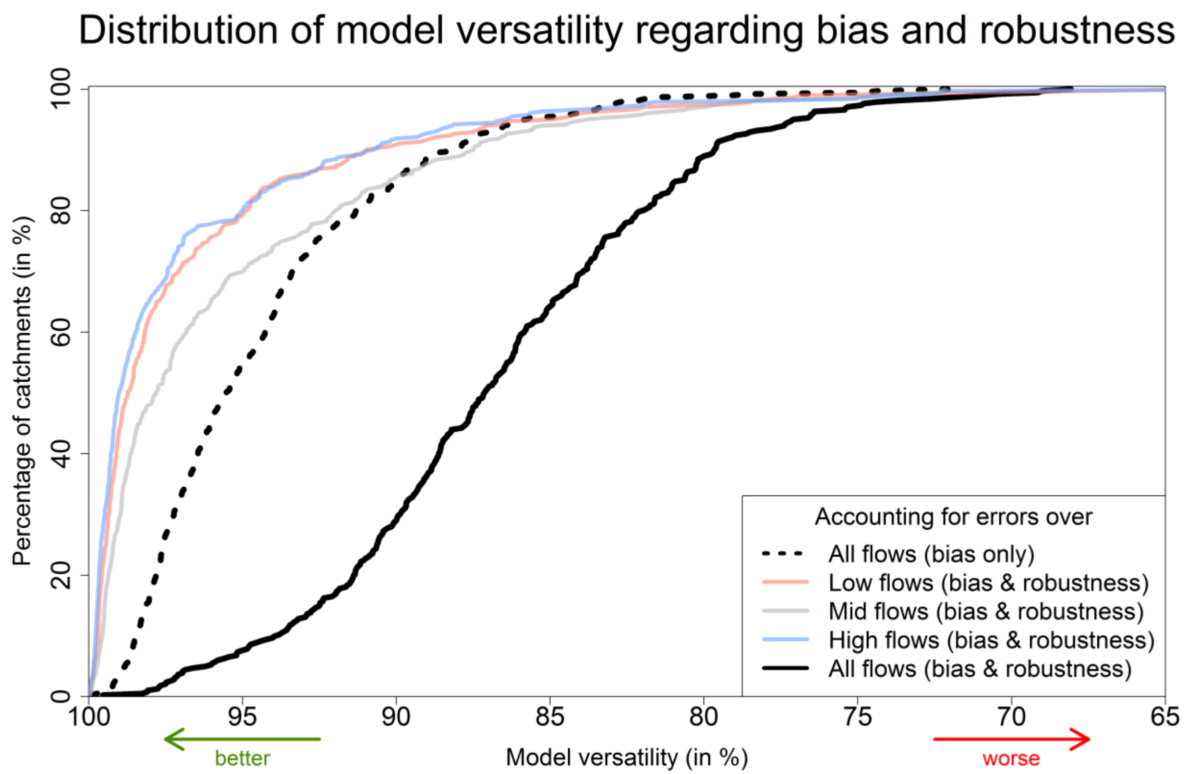


Fig6. Model versatility

2. Model overall performance

Model actual performance

- How does bias and robustness of the model evolve with different combinations of metrics used to select the behavioural parameter sets ?

✓ If parameter selection is based on a single metric, GR4J can perform well for this same metric (e.g. no bias)

❑ Bias is very sensitive to perf. compromises if more metrics are used in param. Selection

❑ Model robustness is altered if more metrics are used in parameter selection, but is rather insensitive the choice of these metrics

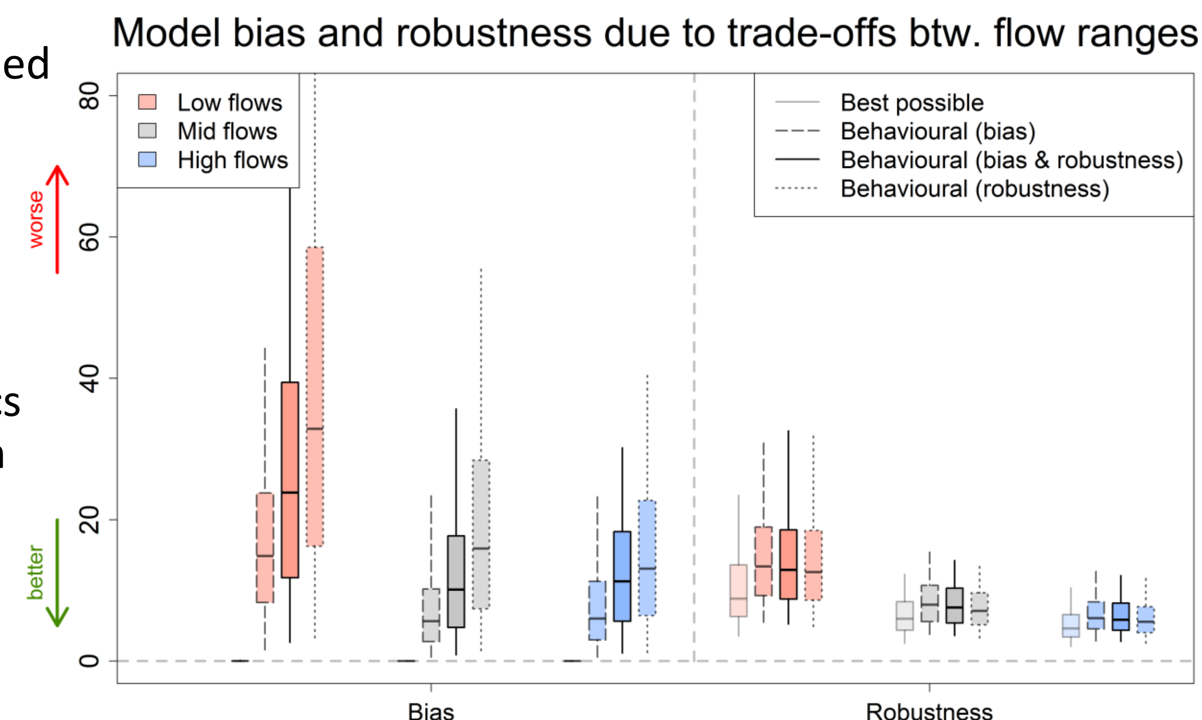


Fig7. Model performances

2. Model overall performance

Summary

- Our results so far show that:
  - Compromises in model bias over the three ranges of flow yields moderate biases
  - Compromises between bias and robustness exacerbates the difficulty for the model to simultaneously match multiple performance requirements
  - Model robustness over a specific range of flow is only slightly improved by a single-metric parameter selection

- Which pairs of metrics trigger the most severe trade-offs between model performance ?

➢ In other words, do some pairs of metrics strictly prevent the model to reach higher versatility scores in many catchments ?

3. Incompatible performance requirements

Methodology

- From the most behavioural parameter set, it is not possible to improve model performance somewhere without reducing model versatility score (the most behavioural parameter set is Pareto optimal)

- From the most behavioural parameter set, it is even possible that improving model performance for one metric systematically degrades another

- For example, in Fig8:
  - Metric 1 can be improved along with Metrics 2 or 3
  - Metrics 2 and 3 cannot
  - ➢ They are strictly *incompatible*

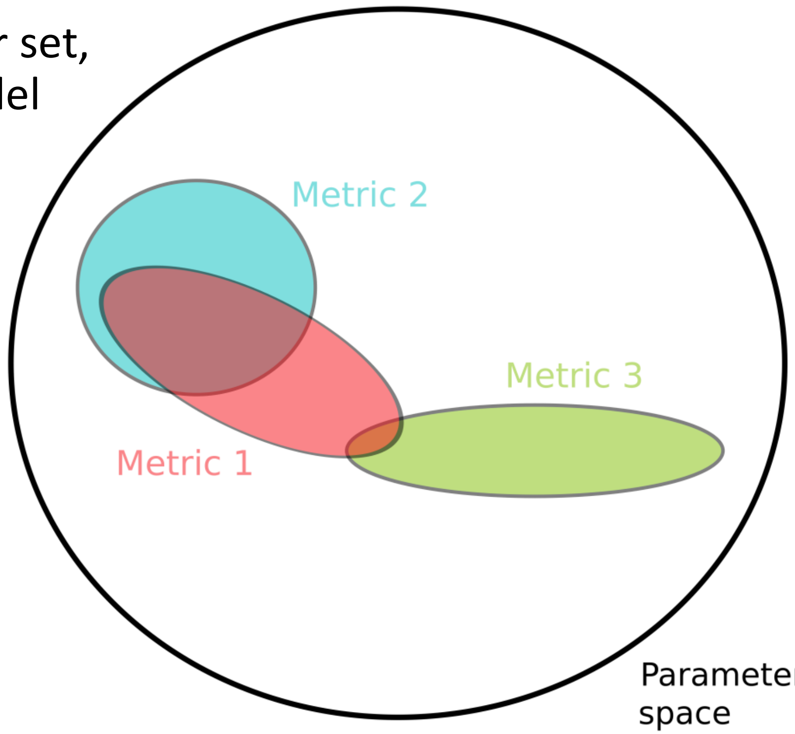


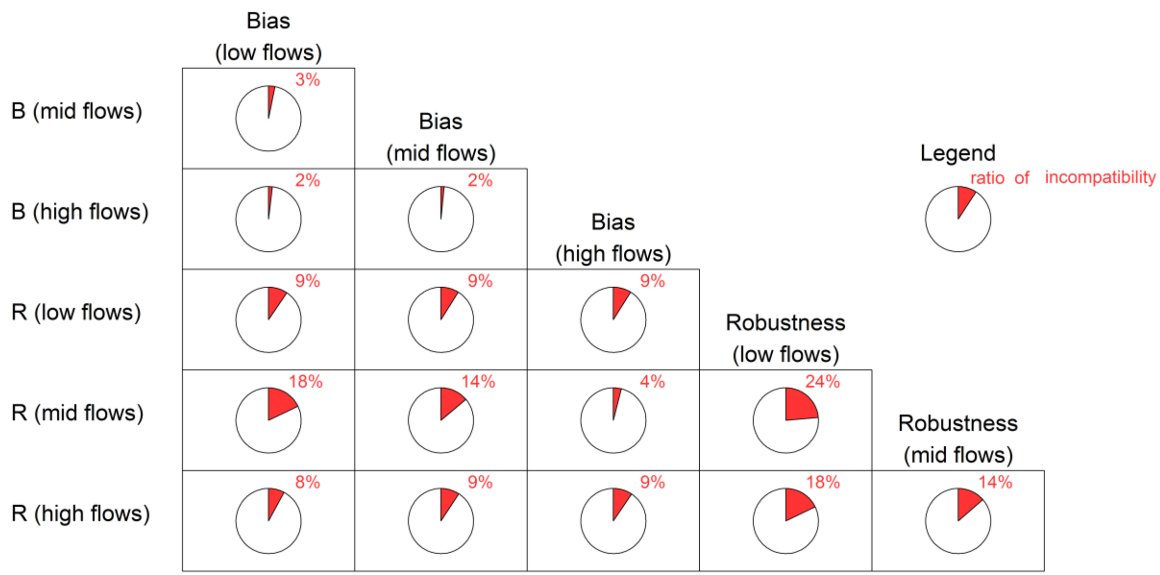
Fig8. Conceptual view of metrics incompatibility

3. Incompatible performance requirements

Results

- Which pairs of metrics trigger the most severe trade-offs between model performance ? Fig9 shows the percentage of catchments where a pair of metrics exhibit strict incompatibility

Metrics incompatibility across the set of catchments



✓ Biases show almost no strict incompatibility issues

❑ Robustness over low and mid flows are often incompatible with other metrics

❑ Robustness metrics are particularly often incompatible with another

➢ Three-metric-wise analyses further show that robustness issues strictly limit model versatility in 59% of the catchment set

3. Incompatible performance requirements

Summary

- The model struggles to simultaneously simulate interannual variations of the three ranges of flow
- It is though able to simulate the average low, mid and high flows on a long term perspective

Therefore

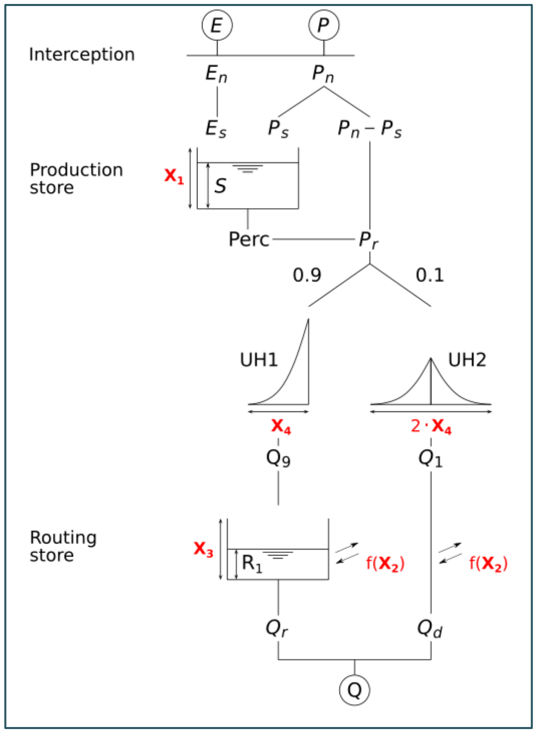
- ➢ Because most focused on average flows, it is possible that assessment studies of model robustness may have overestimated models robustness
- ➢ In GR4J's case, distinguishing model biases over different ranges of flow does not yield the model to crash

- In the following, we analyse parameter distribution when parameter selection is done with the 2 metrics associated to each range of flow (e.g. bias over low flows and robustness over low flows)

4. Improvable model processes

Incompatible constraints on the parameters

- How parameters are distributed if parameter selection is focused on different ranges of flow ?



Changes in model parameterization with single or simultaneous parameter selection

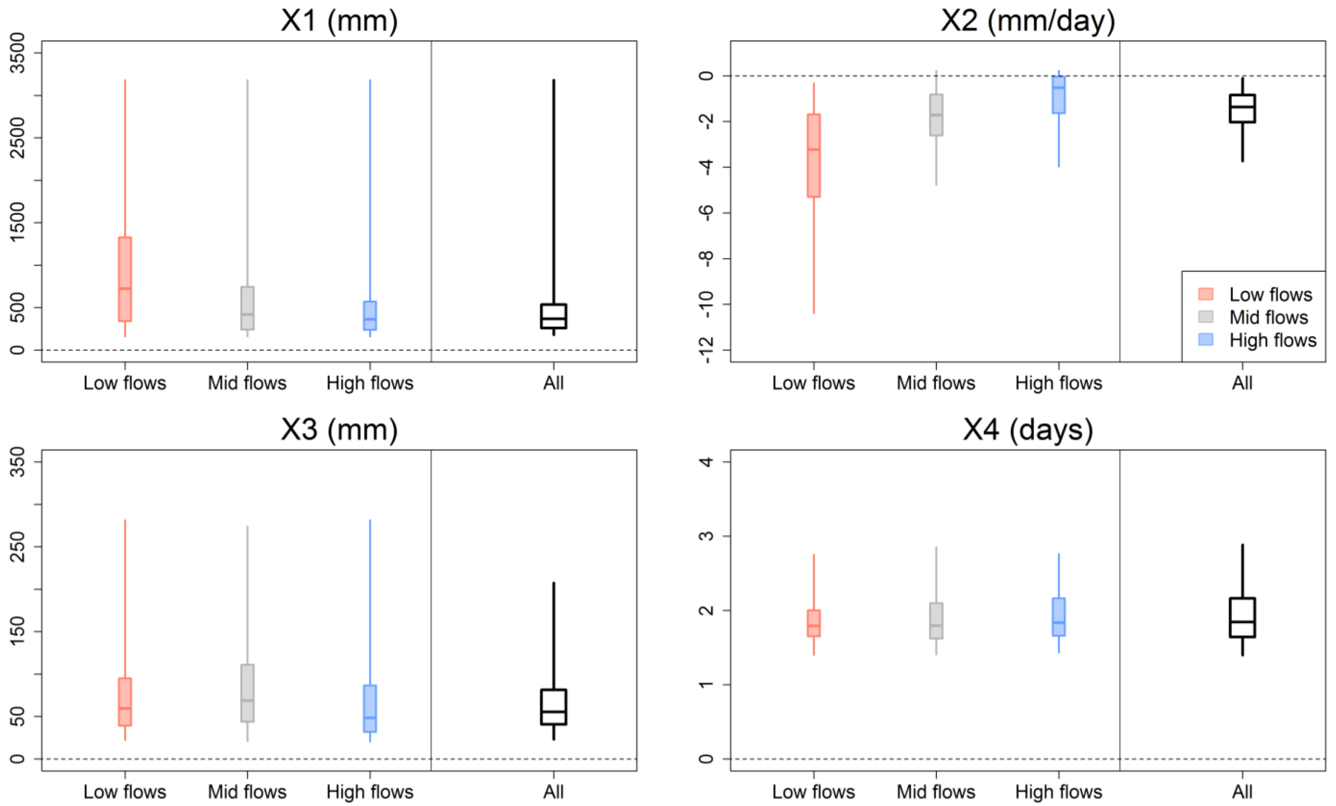


Fig10. Parameter distribution

4. Improvable model processes

Incompatible constraints on the parameters

- Can we interpret the obtained distribution ?

  - Better simulations over low flows demands significantly higher X_1 and very negative X_2 compared to mid and high flow metrics
  - Lower X_1 denote higher sensitivity of catchments wetness to incoming short term events, higher X_1 denote higher inertia to past conditions
  - Very negative X_2 denote higher water leakage through the routing store and thus that the model must get rid of water in excess

  - We suggest that
  - Higher X_1 help the model to represent successive dry years since it increases its inertia, but the model must compensate for higher water inputs in the routing store by very low X_2 values
  - Lower X_1 help the model to quickly react to rare and strong rainfall events and thus produce higher flood peaks, with X_2 closer to zero maintaining sufficient water level in the routing store before the floods

Conclusion

Summary

- We propose a crash test based on a **multiobjective framework to identify model structural flaws**
- The crash test is applied on **GR4J to test its ability to robustly simulate low flows, mid flows and high flows** on a large catchments set

- The model demonstrates a **correct ability to produce unbiased simulations** of multiple flows at the same time
- However, the model **struggles to provide robust simulations** over multiple ranges of flow
- The model **trades off its robustness over a range of flow for another**, thus compromising its overall skill
- Model parameters seem to suffer contradictory constraints during calibration, **indicating which parameterization should be improved in priority**

Conclusion

Perspectives

- A deeper analysis of GR4J’s states and fluxes should provide further insights of how model parameterization responds to incompatible performance requirements

- Model developments focused on the model interannual dynamics should be tested (Fowler et al., 2020)

- Explore the possibility to design a robustness-oriented calibration method for rainfall-runoff models based on the multi-objective framework tested in this study

Conclusion

Recommendations if you want to use the crash test at home

1. Find performance metrics as complementary as possible to stress model compromises

2. Select a catchments set as large and diverse as possible to robustly identify model weak spots

3. The rainfall-runoff model should not be too complex to avoid choosing between low computation time and dense exploration of the parameter space

4. If the model has too many parameters, explore the parameter space among a list of optimal parameter sets obtained in other catchments (Perrin et al., 2008) rather than by a Monte Carlo process