Assessing the performance and robustness of two conceptual rainfall-runoff models on a worldwide sample of watersheds

Thibault Mathevet, Hoshin Gupta, Charles Perrin, Vazken Andréassian, Nicolas Le Moine

Assessing the performance and robustness of two conceptual rainfall-runoff models on a worldwide sample of watersheds

Thibault MATHEVET[1], Hoshin GUPTA[2], Charles PERRIN[3], Vazken ANDRÉASSIAN[3], Nicolas LE MOINE[4]
thibault.mathevet@gmail.com

[1] Visiting scientist at Hydrology & Water Resources Dep., Univ. of Arizona (2013/2014)
Modélisation économique des aménagements hydroélectriques, EDF
[2] Hydrology & Atmospheric sciences Dep., Univ. of Arizona
[3] Université Paris-Saclay, INRAE, UR HYCAR, Antony, France
[4] SorbonneUniversité, UMR Metis, Paris, France

Acknowledments

Vazken Andréassian (IRSTEA)

Hoshin Gupta

MOPEX 2004, Paris

Charles Perrin (IRSTEA)

Backgroud of this study

« How different are different models ? »

« I sometimes trust more *my* model than the observations »

« If *my* model can't make it, allmost no model can make it »

- **Question 1: How statistically comparable (based on a detailed evaluation procedure) are the simulation performances of two models?**
- **Question 2: Is the simulation performance of the models essentially identical when provided with the same observational information?**
- **Question 3:** Are differences in model performance dependent on watershed characteristics or on hydrometeorological processes?

Journal of Hydrology 585 (2020) 124698

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

Assessing the performance and robustness of two conceptual rainfall-runoff models on a worldwide sample of watersheds

Thibault Mathevet[a,*,1], Hoshin Gupta[b], Charles Perrin[c], Vazken Andréassian[c], Nicolas Le Moine[d]

[a] EDF-DTG (Électricité de France), 134 chemin de l'Etang, 38950 Saint Martin le Vinoux, France
[b] Department of Hydrology and Atmospheric Sciences, University of Arizona, Tucson, AZ, USA
[c] Université Paris-Saclay, INRAE, UR HYCAR, Antony, France
[d] SorbonneUniversité, UMR Metis, Paris, France

ARTICLE INFO

This manuscript was handled by A. Bardossy, Editor-in-Chief, with the assistance of Roger Moussa, Associate Editor

Keywords:
Hydrological modeling
Large-sample hydrology
Calibration
Evaluation
Diagnostics
Kling-Gupta efficiency

ABSTRACT

To assess the predictive performance, robustness and generality of watershed-scale hydrological models, we conducted a detailed multi-objective evaluation of two conceptual rainfall-runoff models (the GRX model, based on the GR4J model, and the MRX model, based on the MORDOR model), of differing complexity (with respectively, 5 and 11 free parameters in the rainfall-runoff module, and 4 and 11 free parameters in the snow module). These models were compared on a large sample of 2050 watersheds worldwide. Our results, based on the three components of the Kling-Gupta Efficiency metric (KGE), indicate that both models provide (on average) similar levels of performance in evaluation when calibrated with KGE, for water balance (mean bias lower than 2%), time-series variability (mean variability bias lower than 2%) and temporal correlation (mean correlation around 0.83). Further, both models clearly suffer from lack of robustness when simulating water balance, with a significant increase of the proportion of biased simulations over the evaluation periods (absolute bias lower than 2% in calibration and lower than 20% in evaluation for 80% of the watersheds). Simulation performance depend more on the hydro-meteorological conditions of a given period than on the complexity of the model structure. We also show that long-term aggregate statistics (computed on the overall period) can fail to reveal considerable sub-period variability in model performance, thereby providing inaccurate diagnostic assessment of the predictive model performance. Typically the median absolute bias is lower than 8% in evaluation, but the median maximum bias can be as high as 50% within a subperiod, for both models, when calibrated with KGE.

1. Introduction

Rainfall-Runoff (RR) models are widely used for a broad range of research and operational objectives, from hypothesis testing to improving process understanding to streamflow prediction for flood design. Whatever the application, hydrologists and modelers share a particular interest in: i) the efficiency, robustness and realism of model structures (and their consequent simulations); ii) the generality (transposability) of model structures across locations (i.e. ability to be efficient in a variety of hydroclimatic contexts); and iii) methods for parameter identification (Gupta et al., 2014). To achieve these objectives, a variety of strategies for model development and specification have been pursued, ranging from detailed site-specific investigations to more general studies. The term robustness is often used to describe some expected model properties in a broad sense. Here, robustness is understood as the capability of a model to hold a certain level of performance in changing conditions, i. e. independently from the input/output information used for calibration. Robustness is usually assessed by comparing the difference of evaluation metrics under changing conditions (typically from calibration to evaluation periods, but also from dry to wet conditions, etc.).

The investigations discussed in this paper are rooted in the past experience of the authors with RR model intercomparison studies (Perrin et al., 2001, Perrin et al., 2003, Perrin et al., 2008; Le Moine et al., 2007; Pushpalatha et al., 2011, 2012; Coron et al., 2012, 2014), as well as investigations into diagnostic model identification procedures (Gupta et al 2008, 2009, Gupta et al., 2012; Yilmaz et al., 2008; Martinez and Gupta, 2010, 2011; de Vos et al., 2010; Pokhrel et al., 2012).

* Corresponding author.
E-mail address: thibault.mathevet@edf.fr (T. Mathevet).
[1] Visiting research scholar at Hydrology and Atmospheric Sciences, University of Arizona, in 2014.

https://doi.org/10.1016/j.jhydrol.2020.124698
Received 24 October 2019; Received in revised form 22 January 2020; Accepted 14 February 2020
Available online 19 February 2020

Why large sample hydrology ?

- **Improving understanding:**
  more rigorous testing and comparison of competing model hypotheses and structures on common grounds;

- **Improving the robustness of generalizations**:
  allowing statistical analyses of model performances and avoid giving too much weight to outliers;

- **Facilitating classification, regionalization and model transfer**:
  gathering a wide diversity of hydrometeorological contexts, enabling testing classification and regionalisation strategies;

- **Supporting the estimation of uncertainties:**
  establishing the predictive capabilities and performance of hydrological models on a variety of hydrometeorological contexts.

Hydrol. Earth Syst. Sci., 18, 463–477, 2014
www.hydrol-earth-syst-sci.net/18/463/2014/
doi:10.5194/hess-18-463-2014

Hydrology and Earth System Sciences
Open Access

Large-sample hydrology: a need to balance depth with breadth

H. V. Gupta[1], C. Perrin[2], G. Blöschl[3], A. Montanari[4], R. Kumar[5], M. Clark[6], and V. Andréassian[2]
[1]Department of Hydrology and Water Resources, The University of Arizona, Tucson, AZ, USA
[2]Irstea, Hydrosystems and bioprocesses Research Unit (HBAN), Antony, France
[3]Institute of Hydraulic Engineering and Water Resources Management, Vienna University of Technology, Vienna, Austria
[4]Department DICAM, University of Bologna, Bologna, Italy
[5]UFZ – Helmholtz Centre for Environmental Research, Leipzig, Germany
[6]Hydrometeorological Applications Program, Research Applications Laboratory, Boulder, CO, USA

Correspondence to: H. V. Gupta (hoshin.gupta@hwr.arizona.edu)

Received: 26 June 2013 – Published in Hydrol. Earth Syst. Sci. Discuss.: 12 July 2013
Revised: 18 December 2013 – Accepted: 26 December 2013 – Published: 6 February 2014

Abstract. A holy grail of hydrology is to understand catchment processes well enough that models can provide detailed simulations across a variety of hydrologic settings at multiple spatiotemporal scales, and under changing environmental conditions. Clearly, this cannot be achieved only through intensive place-based investigation at a small number of heavily instrumented catchments, or by empirical methods that do not fully exploit our understanding of hydrology. In this opinion paper, we discuss the need to actively promote and pursue the use of a "large catchment sample" approach to modeling the rainfall–runoff process, thereby balancing depth with breadth. We examine the history of such investigations, discuss the benefits (improved process understanding resulting in robustness of prediction at ungauged locations and under change), examine some practical challenges to implementation and, finally, provide perspectives on issues that need to be taken into account as we move forward. Ultimately, our objective is to provoke further discussion and participation, and to promote a potentially important theme for the upcoming Scientific Decade of the International Association of Hydrological Sciences entitled Panta Rhei.

1 Introduction

"Because almost any model with sufficient free parameters can yield good results when applied to a short sample from a single catchment, effective testing requires that models be tried on many catchments of widely differing characteristics, and that each trial cover a period of many years." (Linsley, 1982).

1.1 Motivations for developing large-sample hydrology

A holy grail of hydrological science is to achieve a degree of process understanding that enables construction of models that are capable of providing detailed and physically realistic simulations across a variety of different hydrologic environments, and at multiple spatial and temporal scales (Nash and Sutcliffe, 1970; Klemeš, 1986a; Michel et al., 2006). With its focus on reducing predictive uncertainty, the Prediction in Ungauged Basins (PUB) initiative of the International Association of Hydrological Sciences (IAHS) has helped move the culture of hydrologic science closer to this objective, and away from a reliance upon universal models applied via recalibration at each study location (Hrachowitz et al., 2013).

This move has deep roots (see Linsley, 1982), but has become considerably stronger since the 1999 IAHS meeting in Birmingham where the idea was extensively discussed. It has helped drive the search for improved understanding of the hydrological cycle, and for modeling approaches that

a. achieve the three R's (reliability, robustness and realism);

b. have greater generality and transposability; and for which

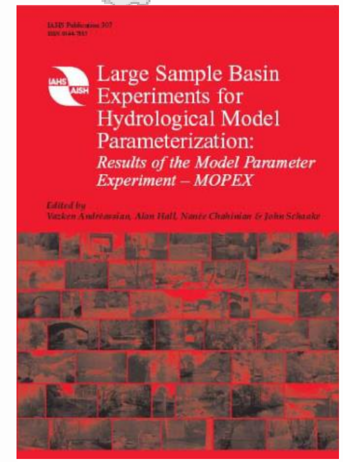
c. the parameters can be more easily specified from data.

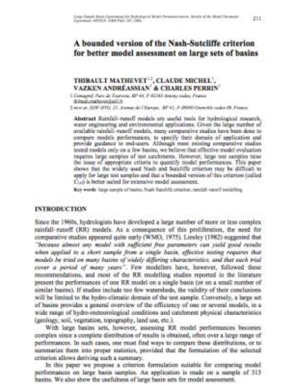
Published by Copernicus Publications on behalf of the European Geosciences Union.

Why large sample hydrology ?

- **Improving understanding:**
  1 – *What are the respective performances of different RR model structures ?*
  2 – *Are the performances of RR structure dependant of watershed caracteristics, climatological or hydrological processes ?*

- **Improving the robustness of generalizations:**
  - How to properly compare two (*n*) RR model structures ?
  - How can I state than two (*n*) RR structures are different ?





211

A bounded version of the Nash-Sutcliffe criterion for better model assessment on large sets of basins

THIBAULT MATHEVET[1,2], CLAUDE MICHEL[1], VAZKEN ANDRÉASSIAN[1] & CHARLES PERRIN[1]

Abstract Rainfall-runoff models are useful tools for hydrological research, water engineering and environmental applications. Given the large number of available rainfall-runoff models, many comparative studies have been done to compare models performances, to specify their domain of application and provide guidance to end-users. Although most existing comparative studies tested models only on a few basins, we believe that effective model evaluation requires large samples of test catchments. However, large test samples raise the issue of appropriate criteria to quantify model performances. This paper shows that the widely used Nash and Sutcliffe criterion may be difficult to apply for large test samples and that a bounded version of this criterion (called C_{2M}) is better suited for extensive model assessment.

Key words large sample of basins; Nash-Sutcliffe criterion; rainfall-runoff modelling

INTRODUCTION

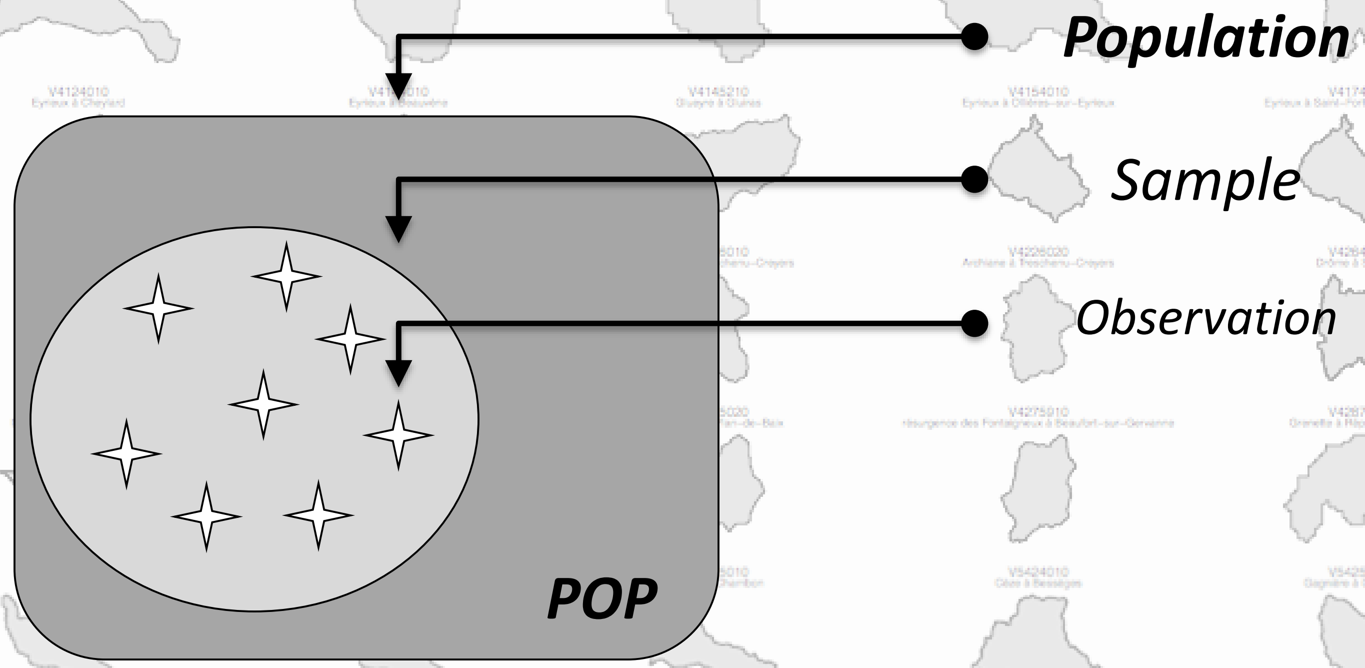
Since the 1960s, hydrologists have developed a large number of more or less complex rainfall–runoff (RR) models. As a consequence of this proliferation, the need for comparative studies appeared quite early (WMO, 1975). Linsley (1982) suggested that *"because almost any model with sufficient free parameters can yield good results when applied to a short sample from a single basin, effective testing requires that models be tried on many basins of widely differing characteristics, and that each trial cover a period of many years"*. Few modellers have, however, followed these recommendations, and most of the RR modelling studies reported in the literature present the performances of one RR model on a single basin (or on a small number of similar basins). If studies include too few watersheds, the validity of their conclusions will be limited to the hydro-climatic domain of the test sample. Conversely, a large set of basins provides a general overview of the efficiency of one or several models, in a wide range of hydro-meteorological conditions and catchment physical characteristics (geology, soil, vegetation, topography, land use, etc.).

With large basins sets, however, assessing RR model performances becomes complex since a complete distribution of results is obtained, often over a large range of performances. In such cases, one must find ways to compare these distributions, or to summarize them into proper statistics, provided that the formulation of the selected criterion allows deriving such a summary.

In this paper we propose a criterion formulation suitable for comparing model performances on large basin samples. An application is made on a sample of 313 basins. We also show the usefulness of large basin sets for model assessment.

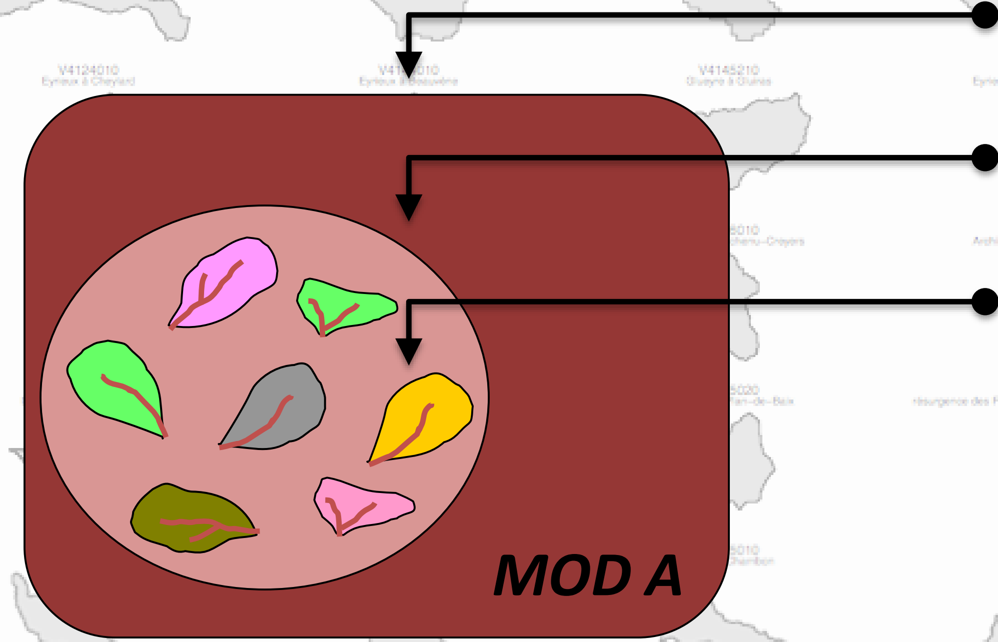
Why large sample hydrology ?

- Allowing **statistical** comparison of RR model structures
  - **Infer the properties of a population from a sample of observations**



Why large sample hydrology ?

- Allowing **statistical** comparison of RR model structures
  - **Infer the properties of a population from a sample of observations**



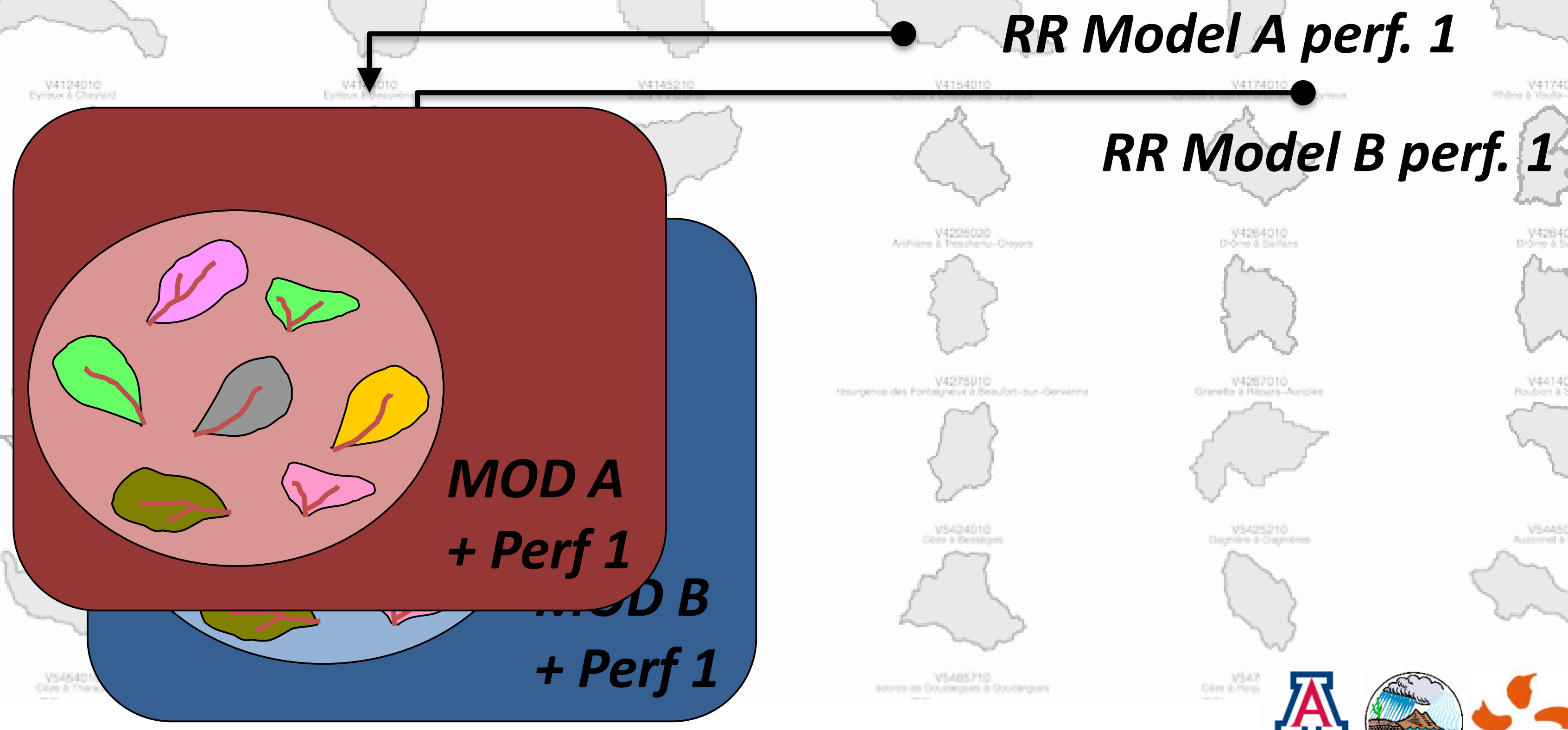
RR Model A performance

Distribution of perf. on a sample of n waterhseds

Performance on Watershed i

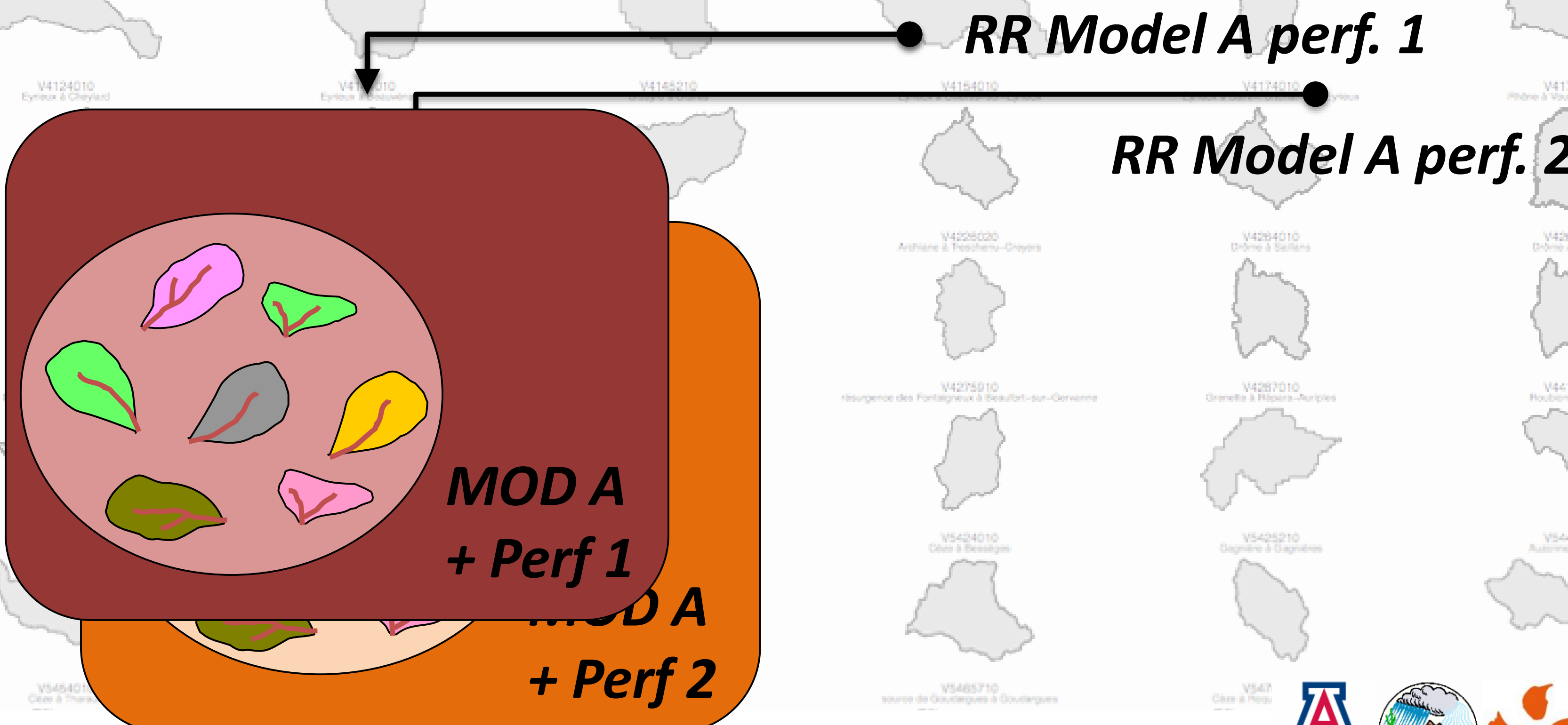
Why large sample hydrology ?

- Allowing **statistical** comparison of RR model structures
  - **Infer the properties of a population from a sample of observations**



Why large sample hydrology ?

- Allowing **<u>statistical</u>** comparison of RR model structures
  - **<u>Infer the properties of a population from a sample of observations</u>**



Insights from previous studies (1/2)

- **Perrin et al. 2001 :**
  - 20 RR model structures, +400 watersheds, daily time-step, NSE ;
  - Complex models suffers from a lack of robustness and 4-6 free parameters seems sufficient to give the « best » results ;

- **Mathevet et al., 2006 :**
  - 4 RR model structures, +300 watersheds, hourly time-step, NSE + modification ;
  - NSE do not allow robust statistical comparisons ;
  - Framework to state if two RR structures performances are significantly different or not ;

- **Coron et al. 2011 :**
  - 3 RR model structures, +200 watersheds, daily time-step, 2 performance metrics ;
  - RR model are extremely dependent to climatic conditions during calibration and have a strong lack of robustness when evaluated on contrasted climatic periods ;

THE UNIVERSITY OF ARIZONA.

Insights from previous studies (2/2)

- **Fenizia et al. 2011, Kavetski et al. 2011 :**
  - SUPERFLEX : flexible modeling framework, with a collection of conceptual structures and constitutive functions ;
  - Hypothese of a better representation of underlying « true » hydrological processes ;

- **van Esse et al. 2013 (including Perrin & Fenizia) :**
  - 30 RR model structures, +200 watersheds, hourly time-step, 4 performance metrics ;
  - **Allmost no difference between a flexible modeling (SUPERFLEX) and a fixed modeling (GR4H) framework ;**

- **Gupta et al. 2009, Gupta & Kling 2011:**
  - Nash-Sutcliffe Efficiency is not an accurate objective function for RR model calibration ;
  - Bias on the water balance and the variability of streamflows ;
  - Introduction of the Kling-Gupta Efficiency (KGE) ;

- **To be updated**

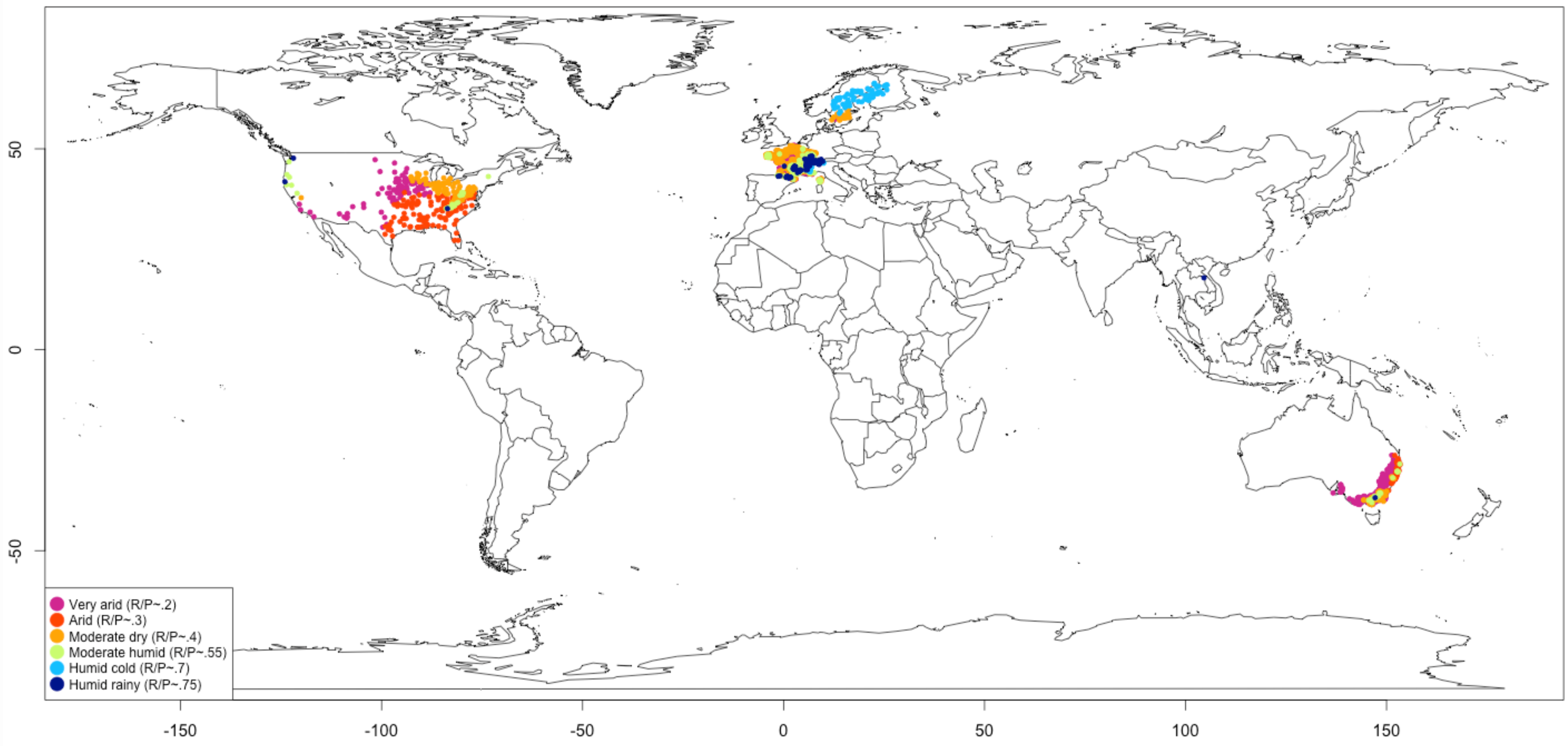
Experimental design (1/4)

- **A (very) large sample of watersheds :**
  - Collect samples allready used in litterature *(Chiew et al., 2000; Duan et al., 2006; Le Moine et al., 2008 ; Vaze et al. 2010; Coron et al., 2011 ; Valery et al., 2009 & 2010; Nicolle et al. 2014 ; Top-Down modeling working group)*;
  - French national projects (PEMHYCE : Nicolle et al. 2014; R2D2 : Kuentz, 2013) ;
  - « My » sample at EDF ;

- **2050 watersheds worldwide** (+ ~200 not used):
  - France, USA, Australia (80%);
  - Switzerland, Sweden, UK, Laos, Italy (20%) ;

- **Since this study :**
  - Many open-source & unified hydrometeorological samples ;
  - **Camels** initiatives largy supported by N. Addor & colleagues (USA, UK, NZ, Chile, Brasil, Australia, etc.) ;

Experimental design (1/4)



Experimental design (2/4)

- **2 Rainfall-Runoff model structures :**
  - Used in many different comparative studies since 2004 ;
  - Statistically the most efficient among 20 different RR on hundreds of watersheds;

GRX (IRSTEA/Cemagref, Paris)

•Empirical development on 100 to 1000 of watersheds worldwide
• 2 buckets
• 5 free parameters
• Undergroud exchanges function
• PET based on Tair and extra-terrestrial radiation
• Snow : 2 buckets & 4 free param.

MRX (EDF / Grenoble)

•Conceptual develoment on <10 watersheds in the Alps
• 4 buckets
• 11 free parameter
• No Undergroud exchanges function
• « optimised » PET

•Snow : 2 buckets & 11 free param.

THE UNIVERSITY OF ARIZONA.

Experimental design (3/4)

- Evaluation metrics :
  - **NSE : Nash-Sutcliffe efficiency** *(Nash & Sutcliffe., 1970)*

$$NSE_Q = 1 - \frac{MSE}{\sigma_Q^2} = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}\left(Q_i - \hat{Q}_i\right)^2}{\frac{1}{n}\sum_{i=1}^{n}\left(Q_i - \overline{Q}\right)^2}$$

  - **KGE : Kling-Gupta efficiency** (Gupta et al., 2009)

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

$$\beta = \frac{\overline{\hat{Q}}}{\overline{Q}} \qquad \alpha = \frac{\hat{\sigma}_Q}{\sigma_Q}$$

Linear correlation

And also (Kling et al., 2012): $\gamma = \frac{\hat{\sigma}_Q}{\sigma_Q} \Big/ \frac{\overline{\hat{Q}}}{\overline{Q}} = \frac{\alpha}{\beta}$

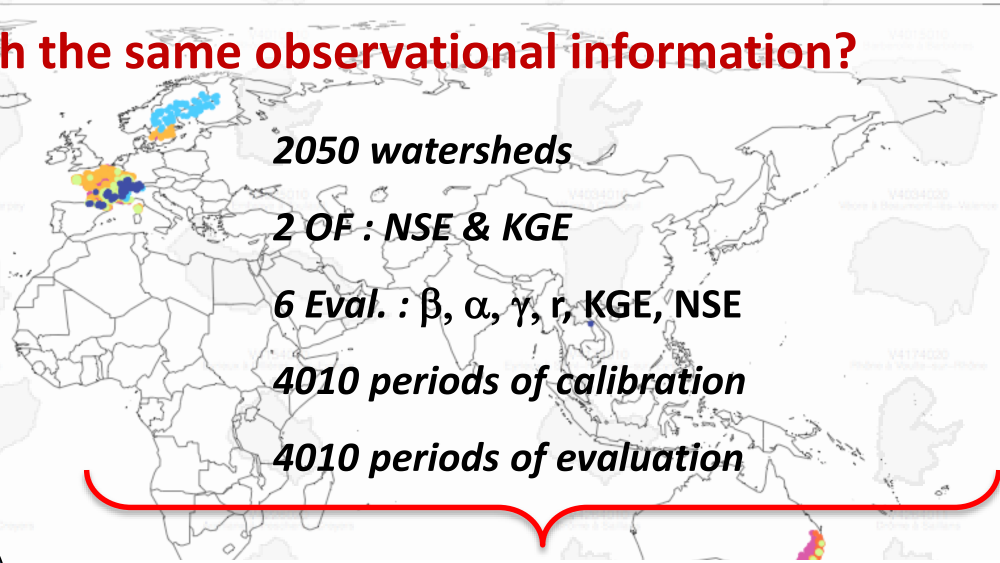
Experimental design (4/4) :

- **Classical Split-sample test** (Klemes, 1986) : 2 periods of calibration and 2 periods of validation

Calibration (P1)	**Evaluation (P2)**
Calibration (P2)	**Evaluation(P1)**

Question 1: How statistically comparable (based on a detailed evaluation procedure) are the simulation performances of two models? &

Question 2: Is the simulation performance of the models essentially identical when provided with the same observational information?



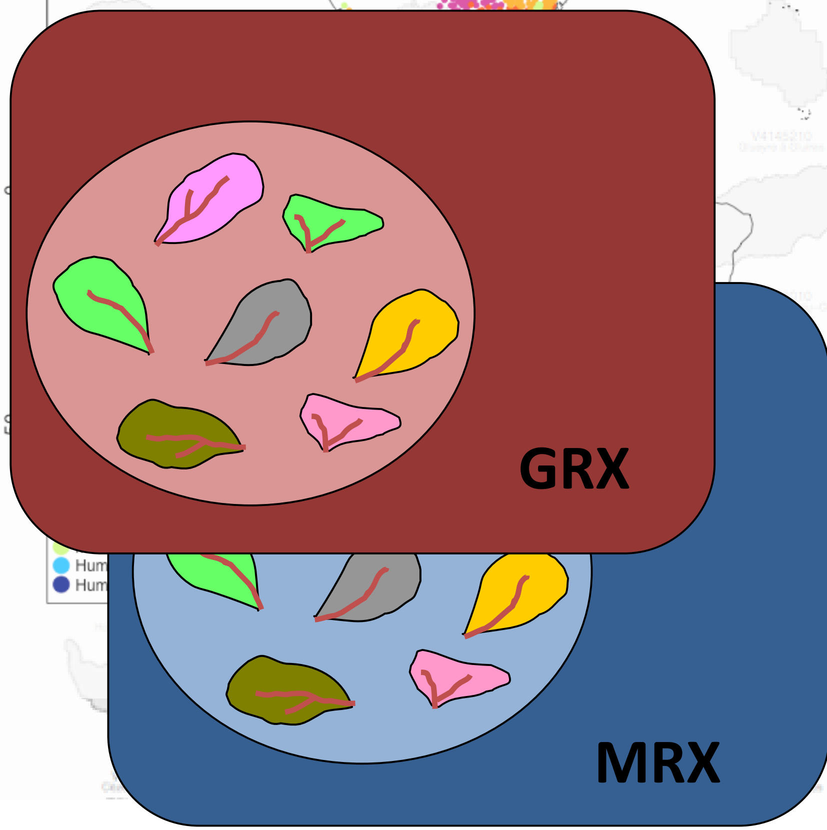
2050 watersheds

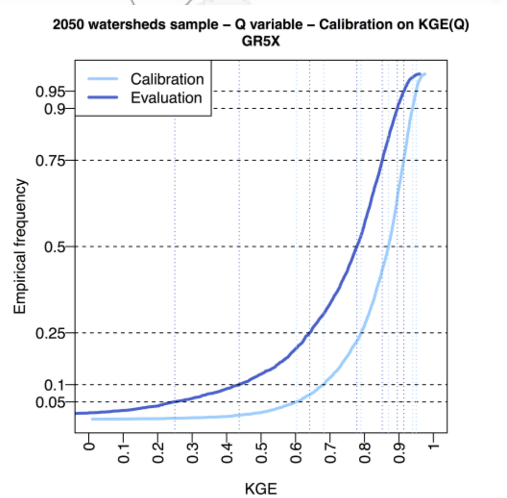
2 OF : NSE & KGE

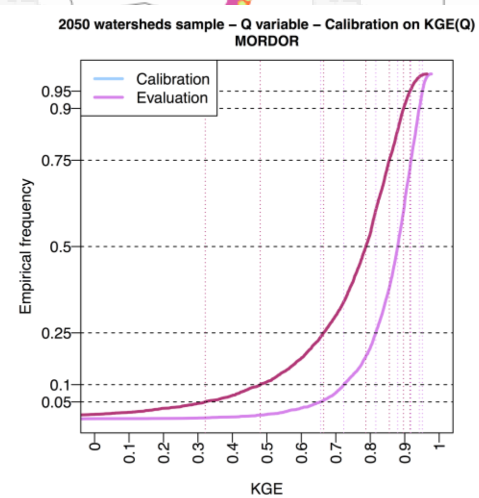
6 Eval. : β, α, γ, r, KGE, NSE

4010 periods of calibration

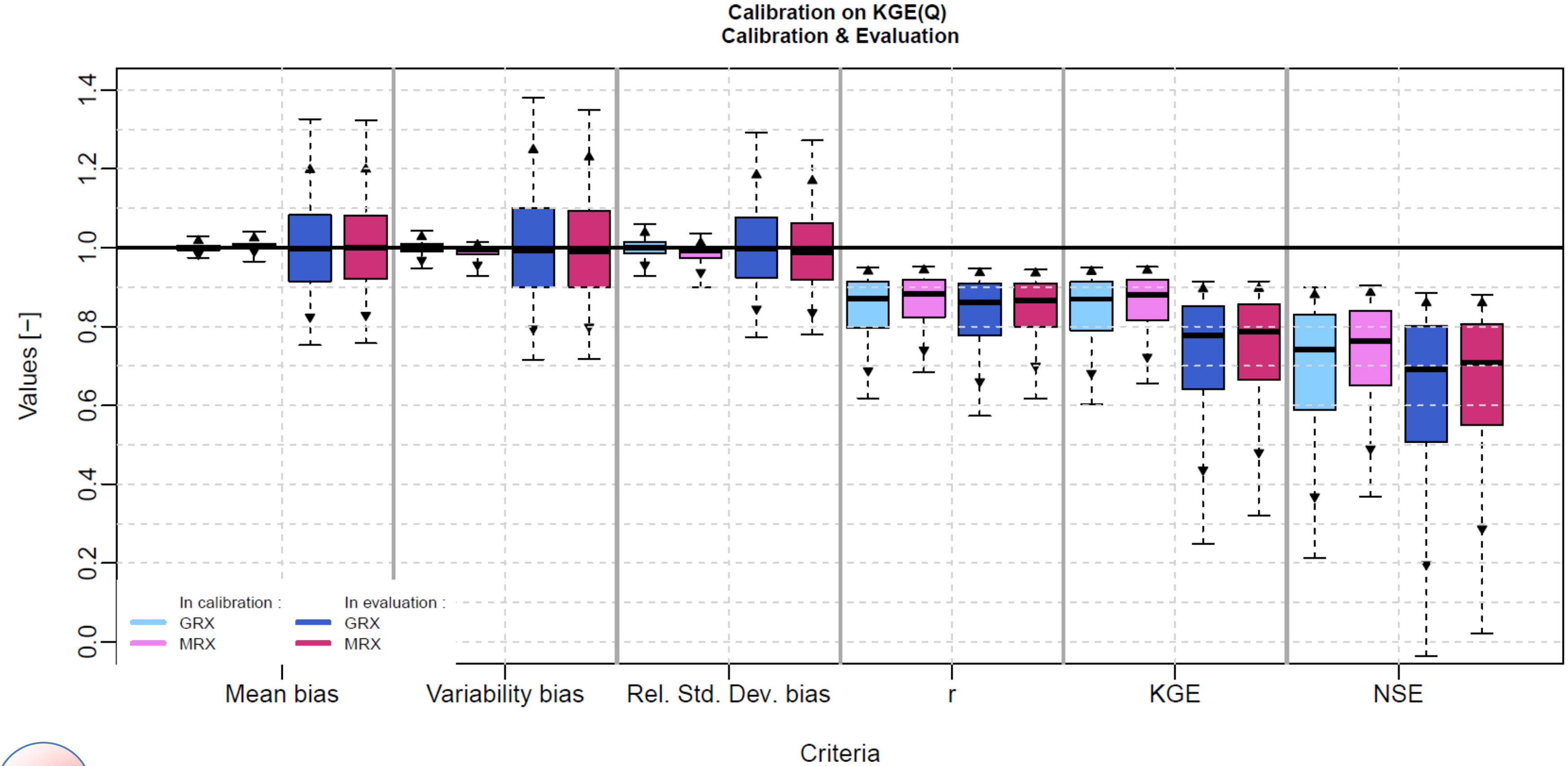
4010 periods of evaluation







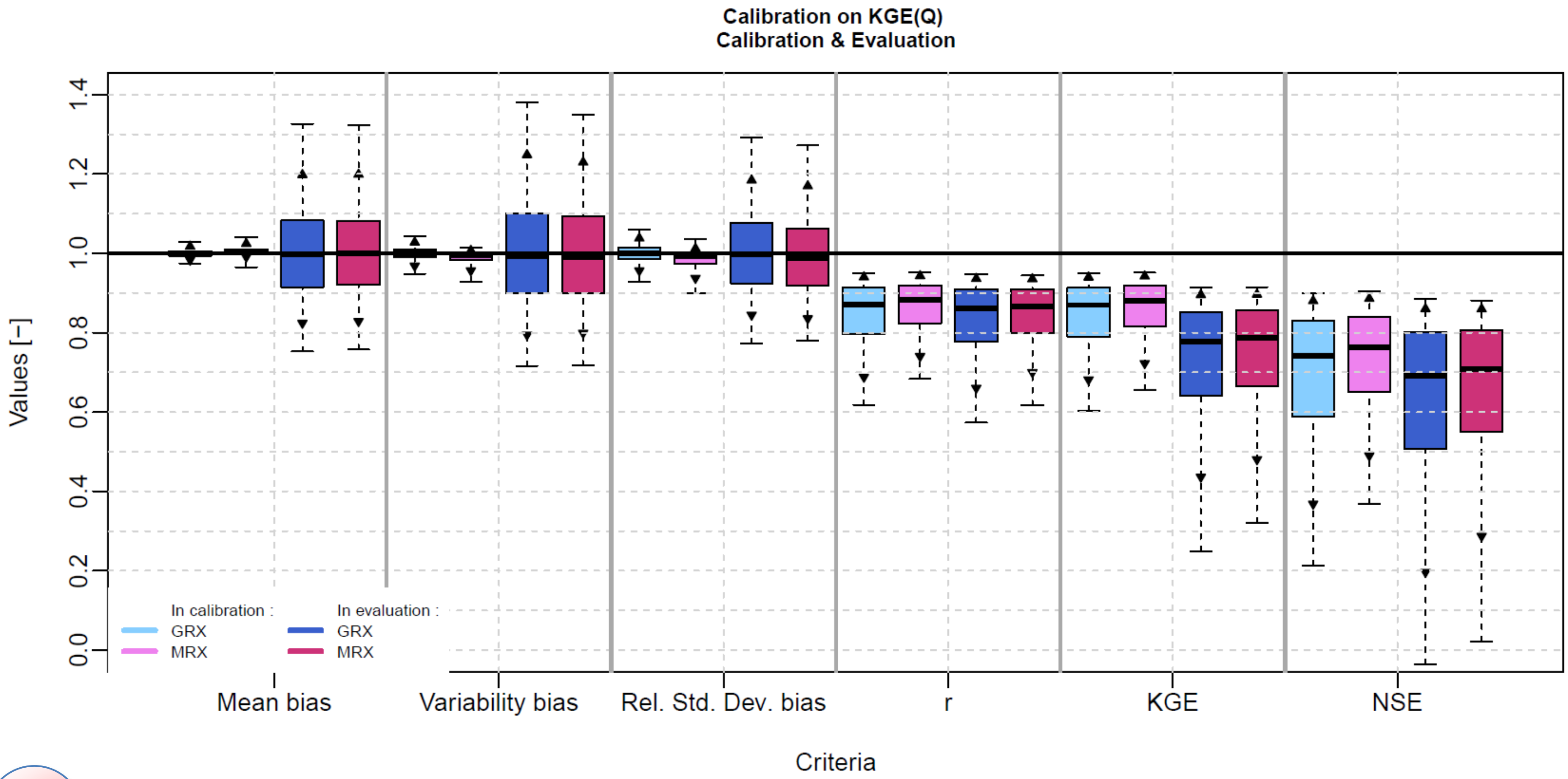
Results : Boxplots



0

How to read these boxplots ?

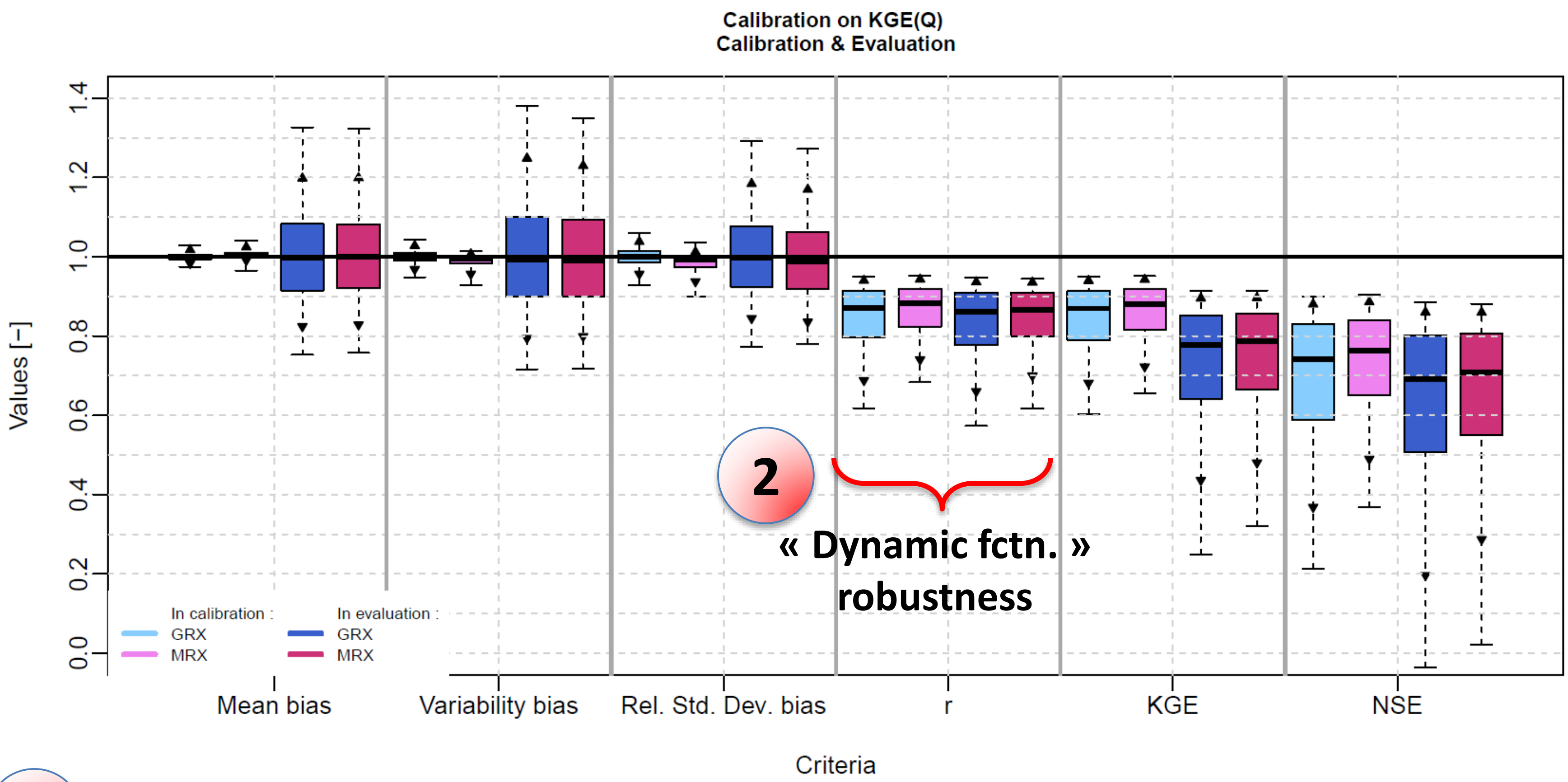
Results : Boxplots



1

WB & Var. lack of robustness

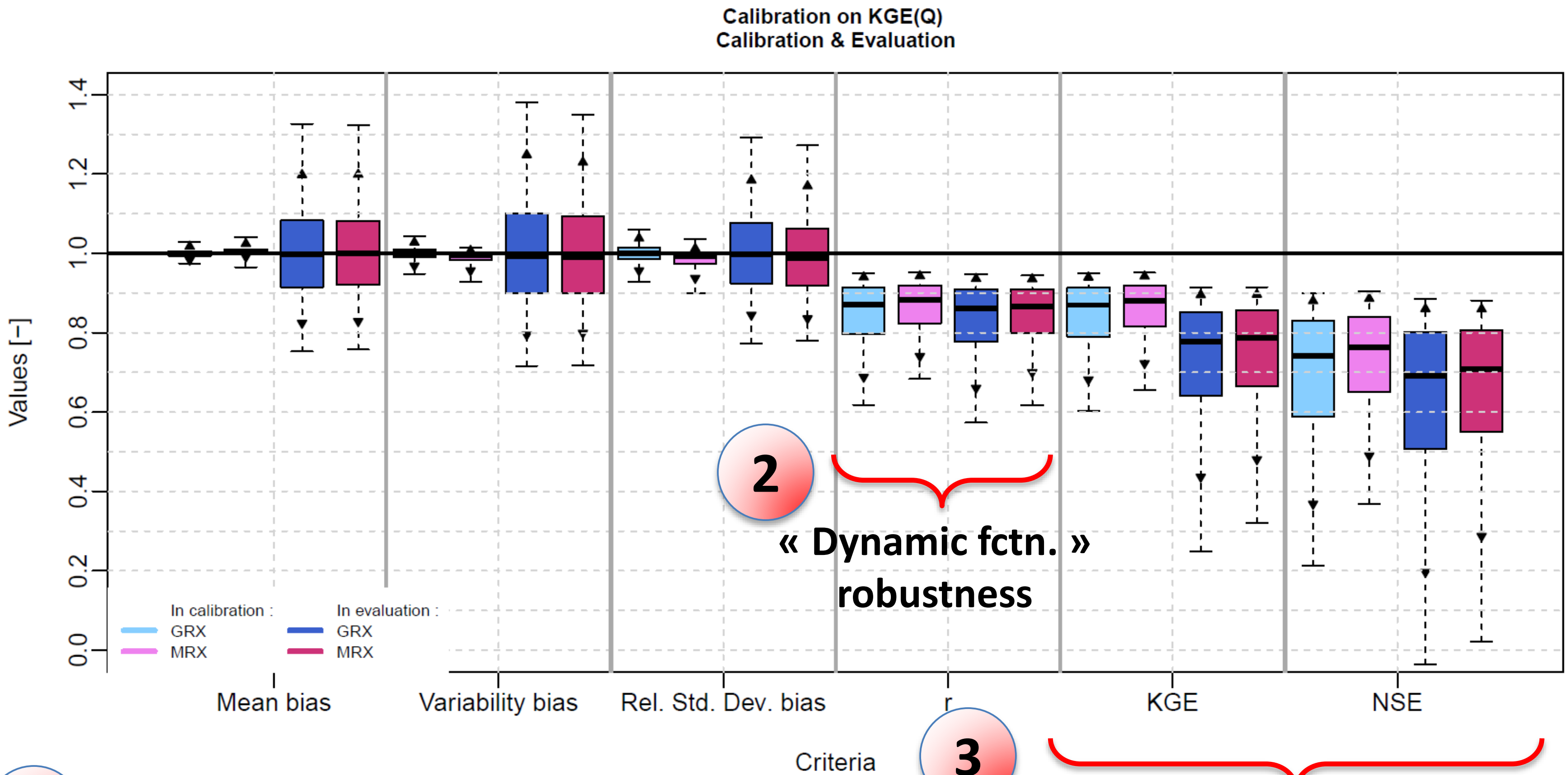
Results : Boxplots



1

WB & Var. lack of robustness

Results : Boxplots



1

WB & Var. lack of robustness

2

« Dynamic fctn. » robustness

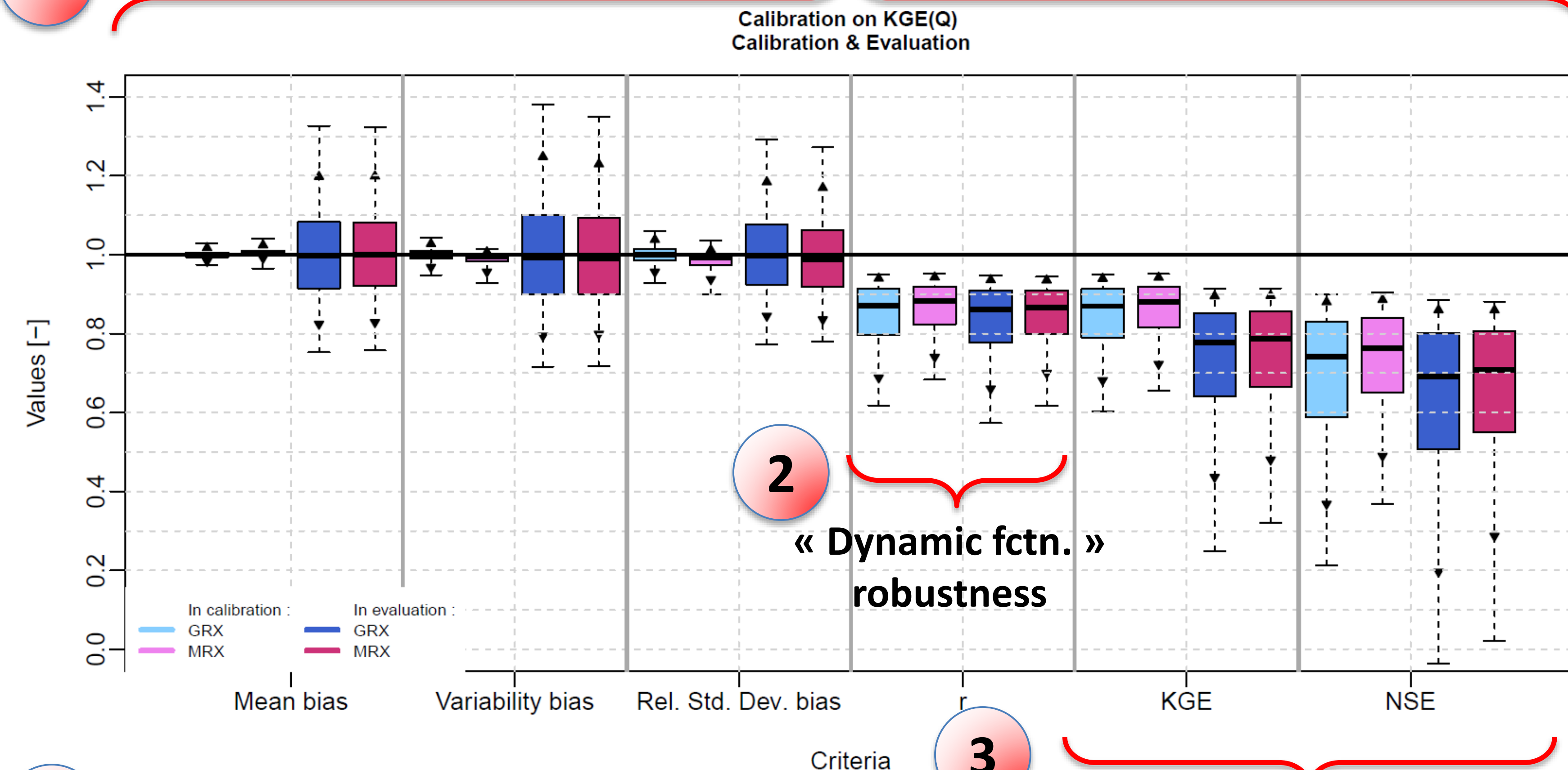
3

Impact of WB & Var. lack of robustness

Results : Boxplots

4

Very similar results



2

« Dynamic fctn. » robustness

3

Impact of WB & Var. lack of robustness

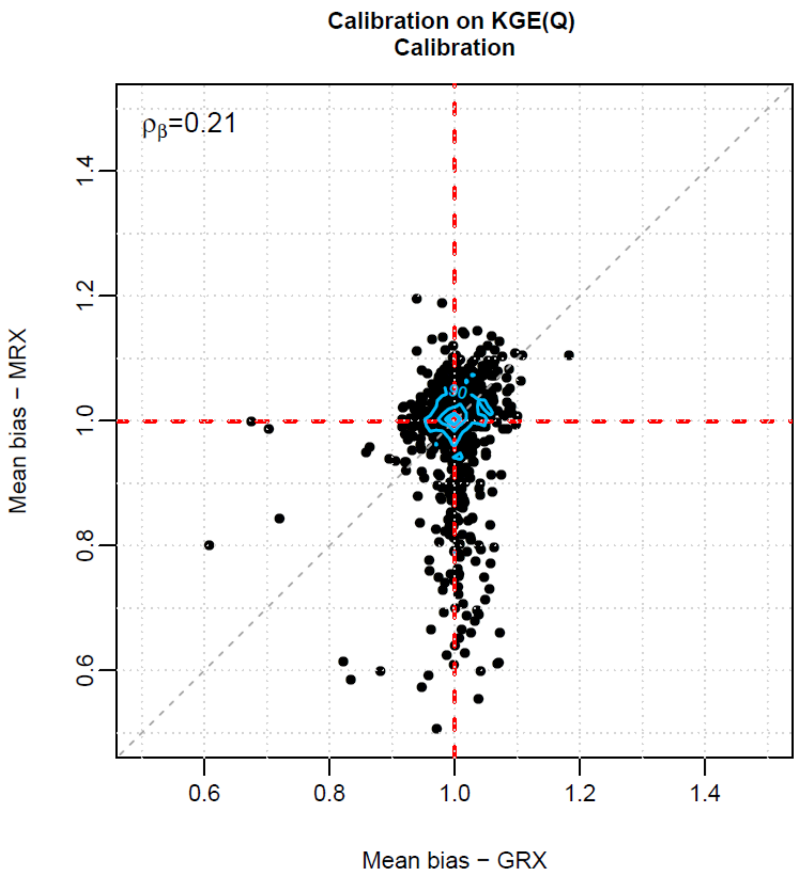
1

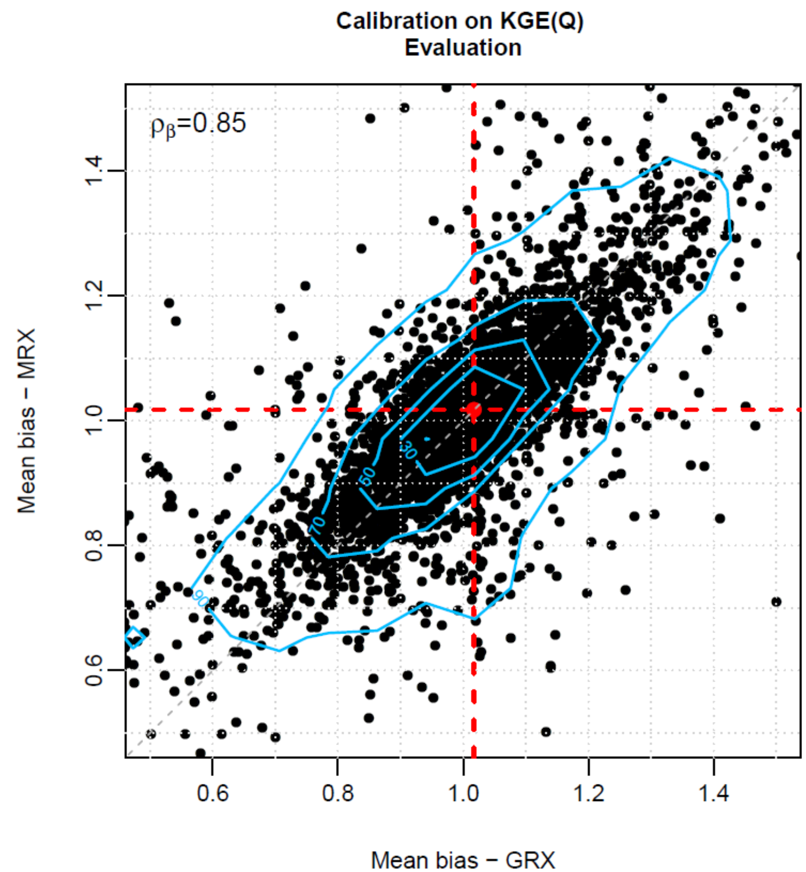
WB & Var. lack of robustness

THE UNIVERSITY OF ARIZONA.

eDF

Results : Scatterplots





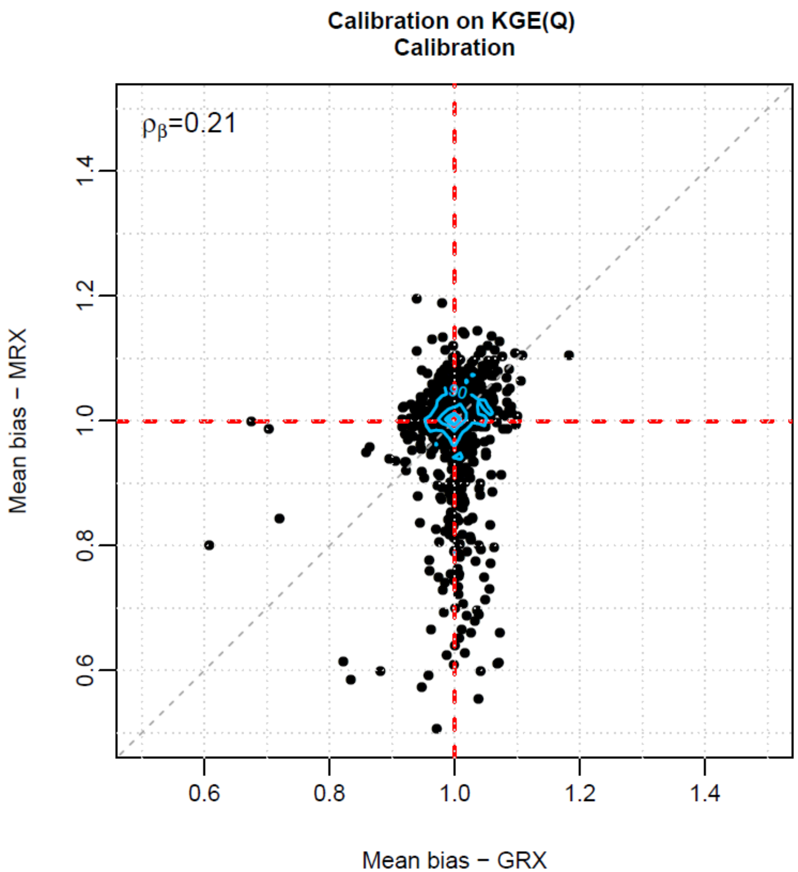
0

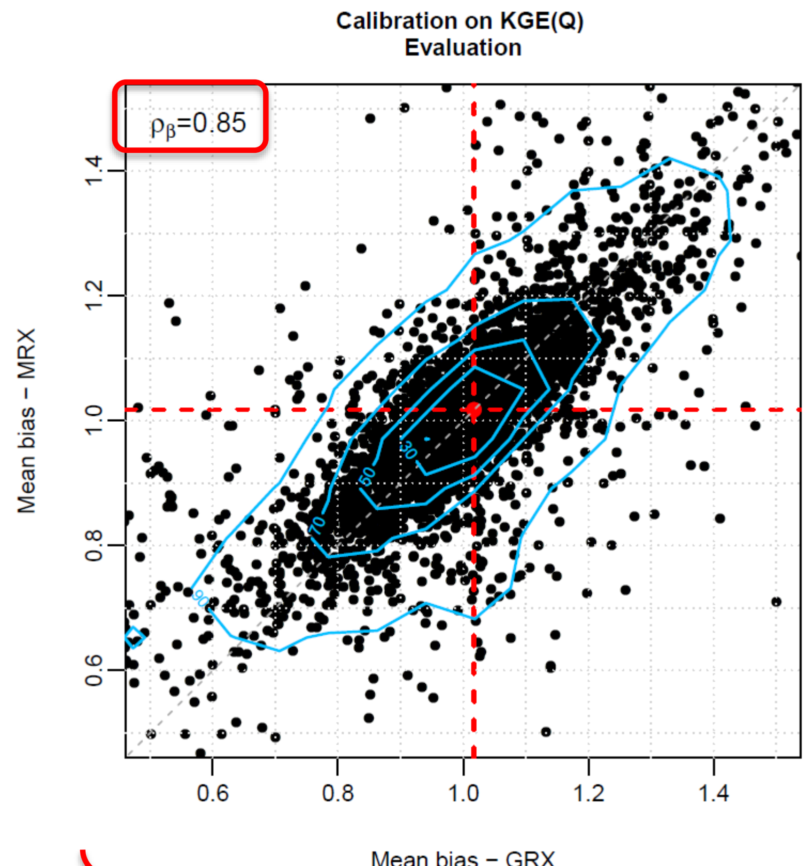
How to read these scatterplots ?

THE UNIVERSITY OF ARIZONA.

eDF

Results : Scatterplots

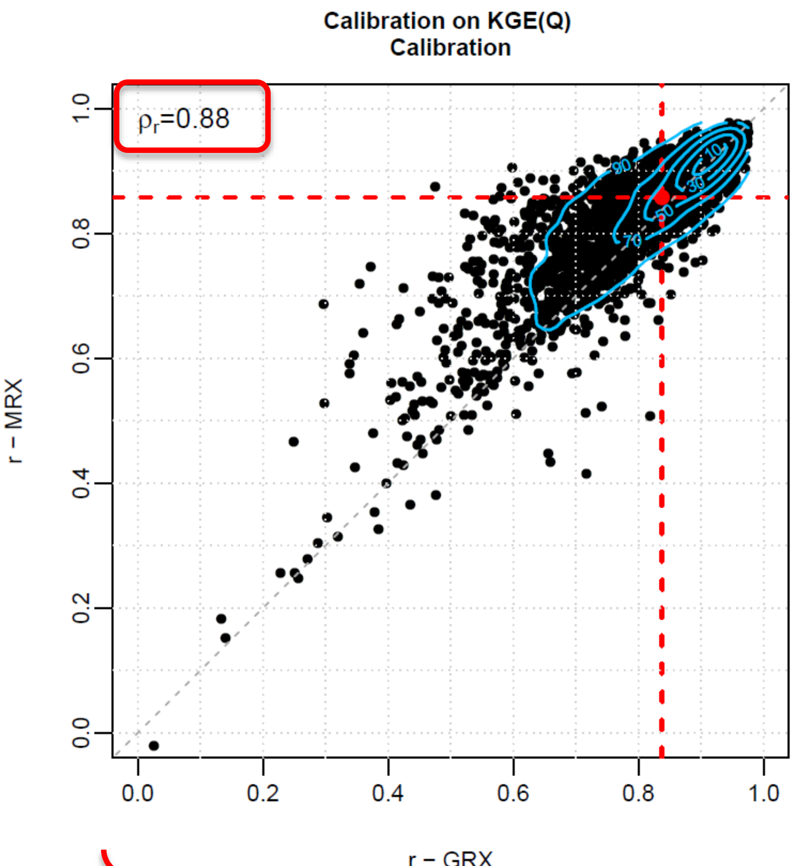


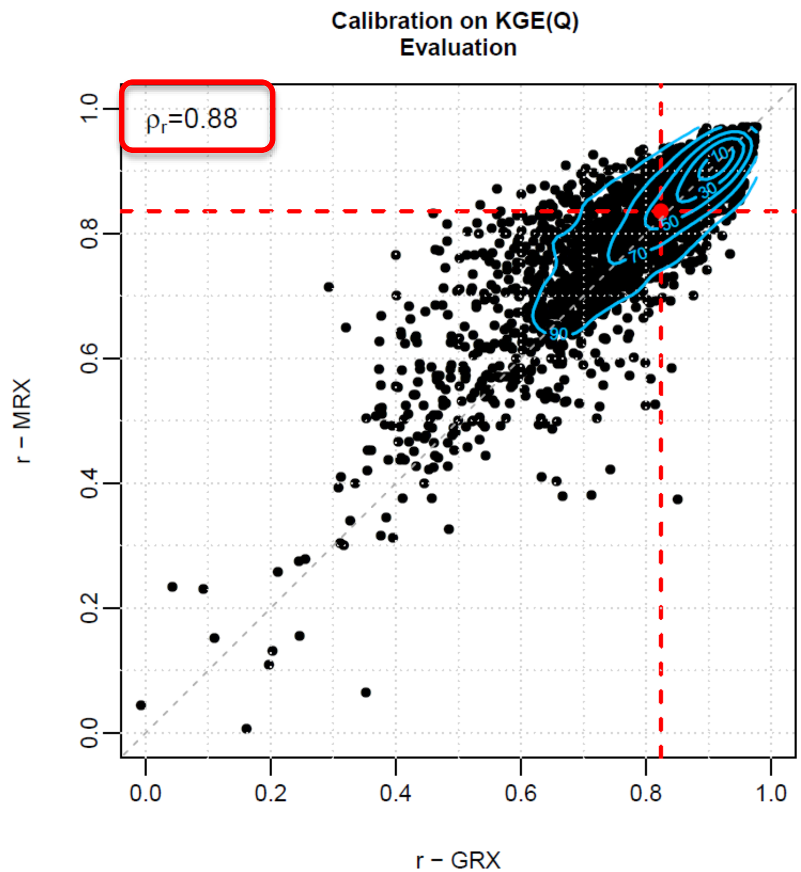


1

Strongly correlated behaviour for β, α, γ

Results : Scatterplots



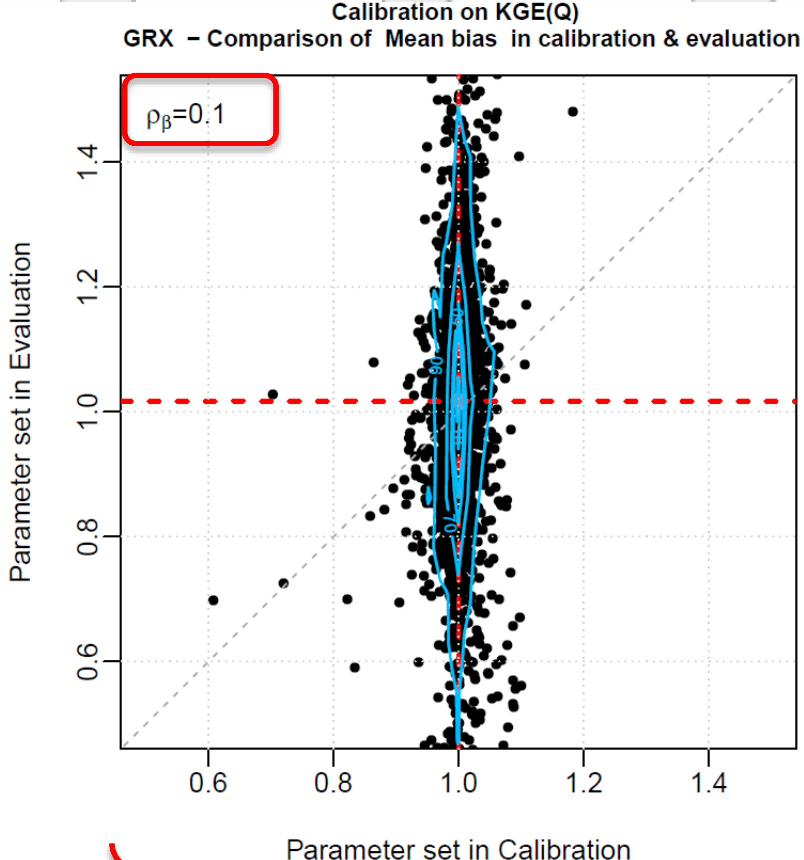


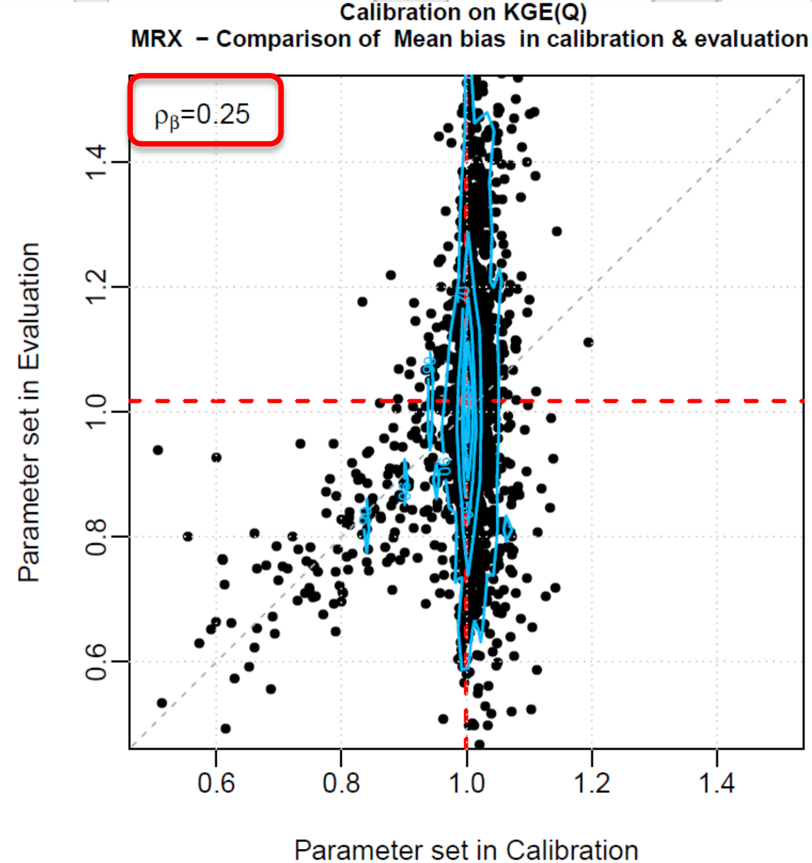
2

Strongly correlated behaviour for r, KGE & NSE

THE UNIVERSITY OF ARIZONA

Results : Scatterplots



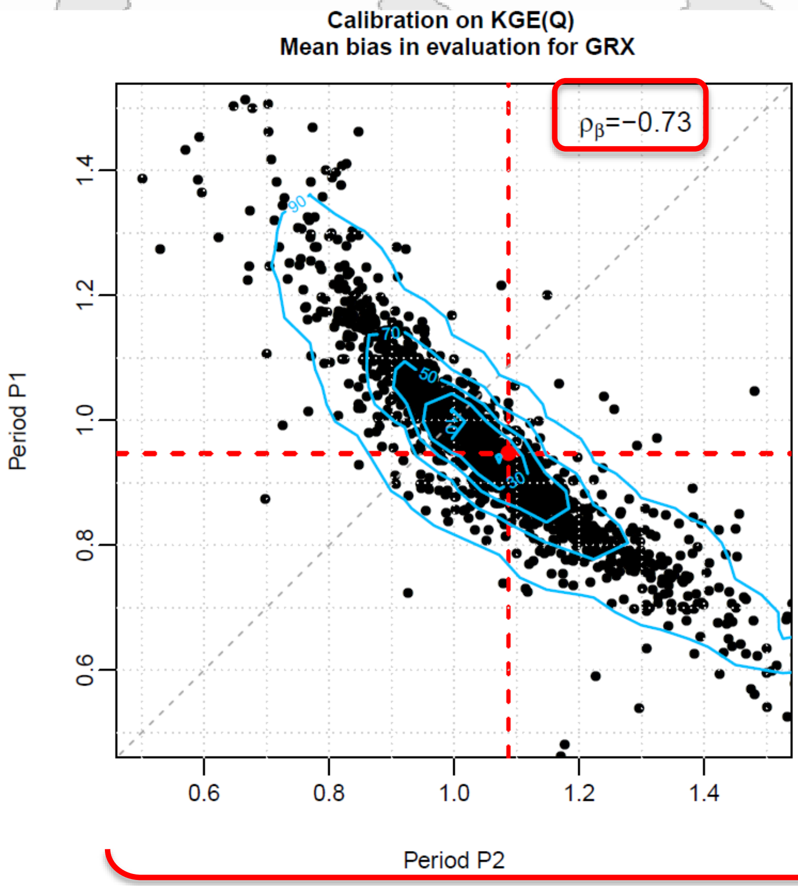


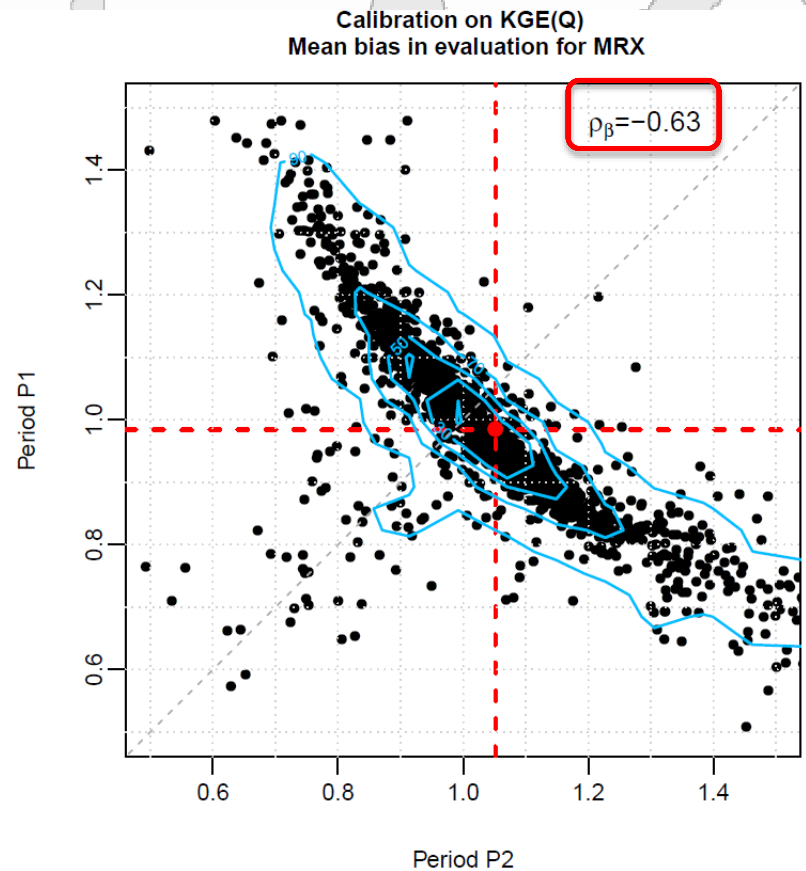
3

Very poor robustness for β, α, γ

THE UNIVERSITY OF ARIZONA.
eDF

Results : Scatterplots

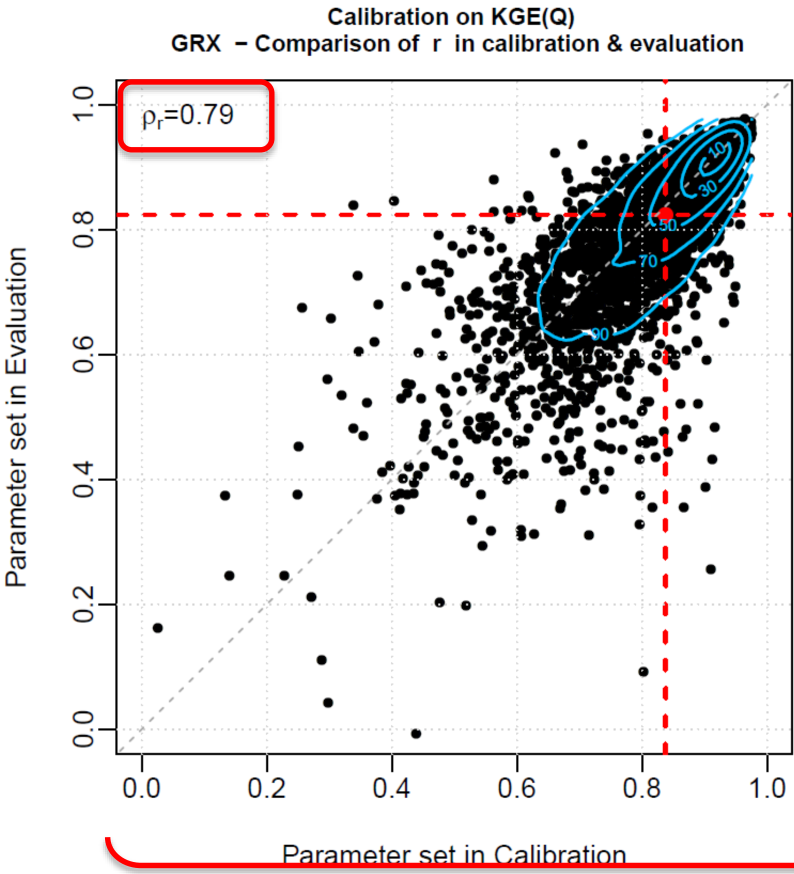


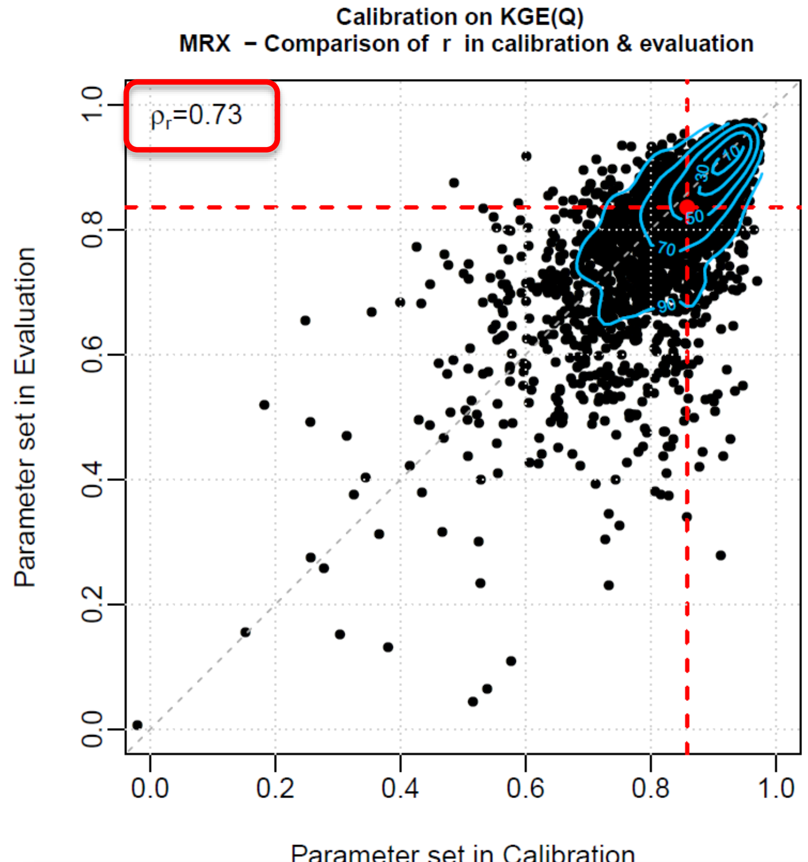


3+

Anti-corralated behaviour
for β on different evaluation periods

Results : Scatterplots

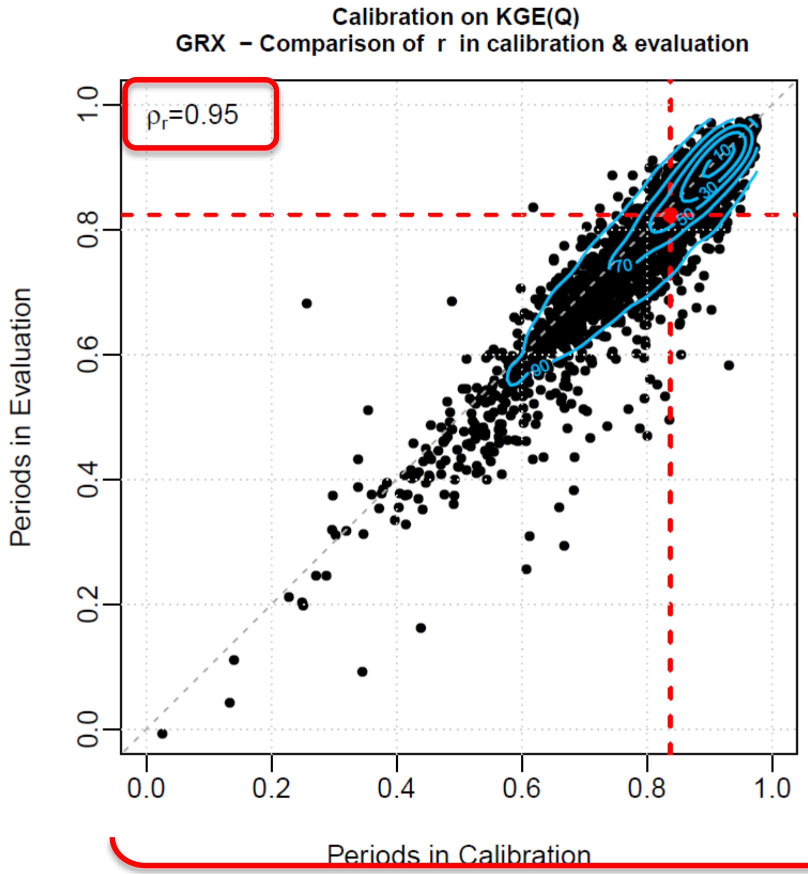


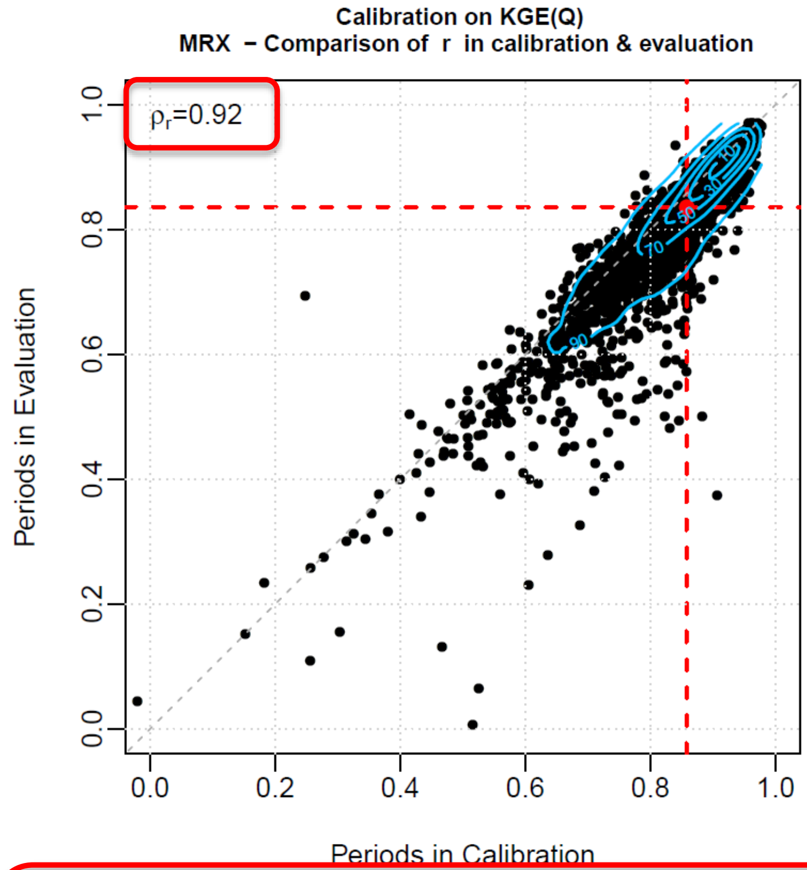


4

Very robust and consistent performances of parameter set (whatever the period)

Results : Scatterplots





5 **Extremely robust and consistent performances on periods (whatever the parameter set)**

THE UNIVERSITY OF ARIZONA. eDF

Results : Synthesis

1 **Both models suffer from a strong lack of robustness in the simulation of water balance and streamflow variability.** The water balance bias varies on the range ±10% for 50% of the watersheds, on the range ±20% for 80% of the watersheds. **However, both models are particularly robust concerning the representation of the dynamic functioning of the watershed.**

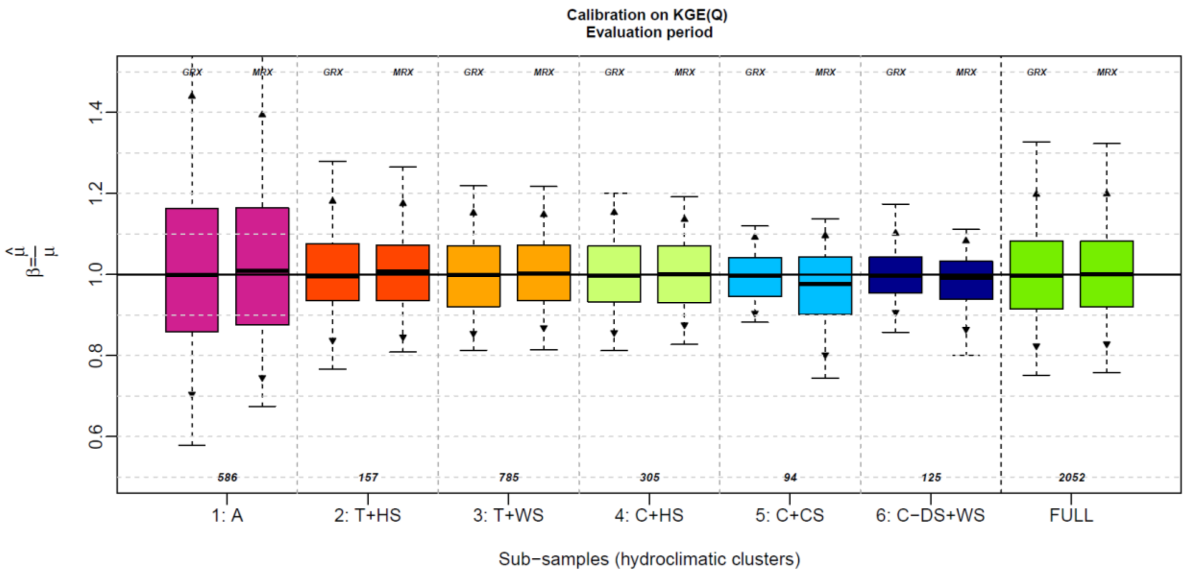
2 **The performance of both models is highly correlated (r ranging from 0.75 to 0.92), despite the strong difference of structure and complexity. This means that model performance correlation (between simulations provided by the two models) is at the same level as the correlation between each of the model simulations and the observations, suggesting that there is no significant difference in overall abilities of the two models across the range of watersheds used for testing.**

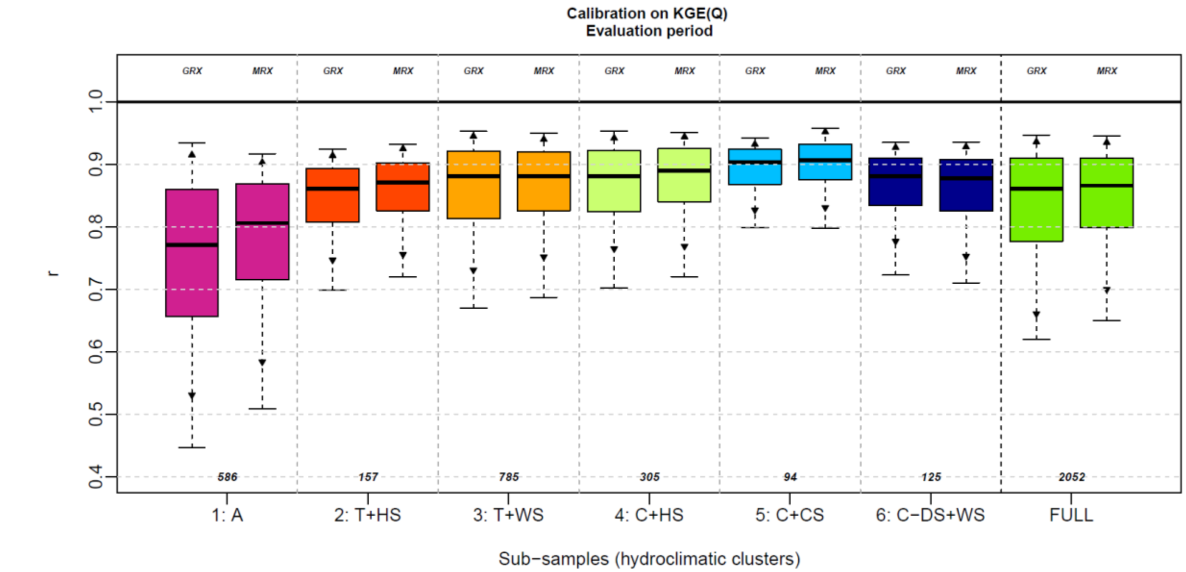
3 **Hence, it seems that differences in hydroclimatic conditions between calibration to evaluation periods play a more important role on the differences in performance from calibration to evaluation than differences in model structures do.**


V3335010
Sanne à Saint-Romain-de-Surieu
V3404310
Rival à Brézins
V3405010
Combe-Robert à Saint-Geoirs
V3424310
Rival à Beaufort
V3425810
Suzon à Saint-Barthélemy
V3434010
Collières à Saint-Rambert-d'Albon
V3504010
V3515010
V3515610
V3517010
V3524010
V3614010


Statistical analysis of RR performance
BABEL
MERCI

Question 3: Are differences in model performance dependent on watershed characteristics or on hydrometeorological processes?





Question 3: Are differences in model performance dependent on watershed characteristics or on hydrometeorological processes?