UNIL | Université de Lausanne

Ensemble Project

Universities of Basel, Bern, Lausanne & Neuchâtel (Switzerland)
Swiss Federal Institute of Technology, Zürich (Switzerland)
Stanford University (USA)

Functional error modeling for Bayesian inference in hydrogeology

Laureline Josset

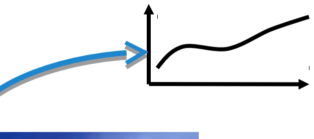
PhD supervisor Prof. Ivan Lunati

Institute of Earth Sciences
University of Lausanne

| le savoir vivant |

Challenges in groundwater problems

Motivation



Typical question:
What is the concentration of contaminant in the drinking water?

Problem:
Many uncertainties in the aquifer properties

Solution: Monte Carlo approaches

Uncertainty quantification, inversion, history matching, …

Challenges in groundwater problems
Monte Carlo approaches

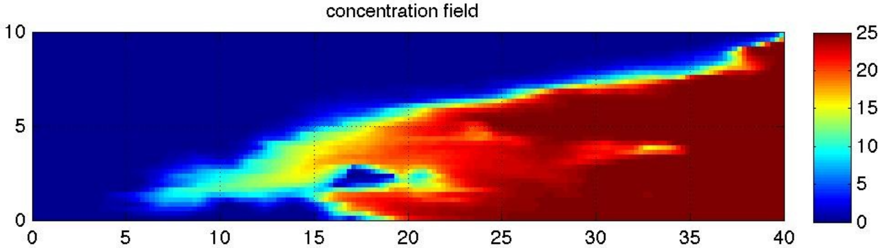
Description of the uncertainty on the permeability field

- Generate multiple geostatistical realizations
  - Based on prior knowledge
  - Methods: object-based, multipoint statistics, process-based, …

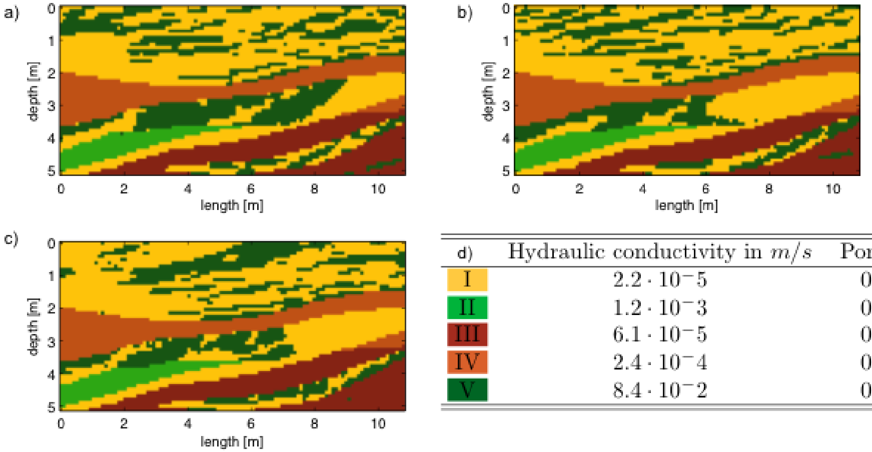
“Truth” inspired from the Herten test case (Bayer et al. 2011)

Issue

- Not the quantity of interest!
- Flow simulation for each of the realizations
  - Typical order: 10^3-10^5 simulations
  - > Untractable computational cost



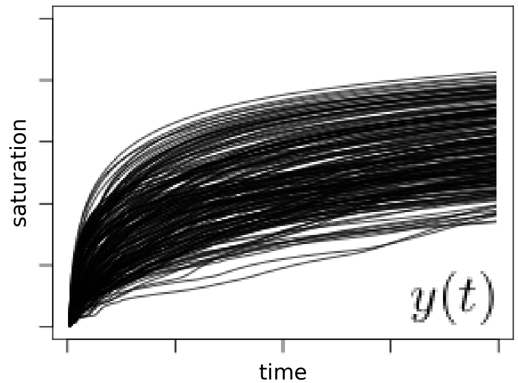
Simulation of saline intrusion



d)	Hydraulic conductivity in m/s	Porosity
I	$2.2 \cdot 10^{-5}$	0.21
II	$1.2 \cdot 10^{-3}$	0.32
III	$6.1 \cdot 10^{-5}$	0.13
IV	$2.4 \cdot 10^{-4}$	0.16
V	$8.4 \cdot 10^{-2}$	0.25

3 examples of geostatistical realizations generated using Direct Sampling (Mariethoz et al. 2010)

How to simulate flow?



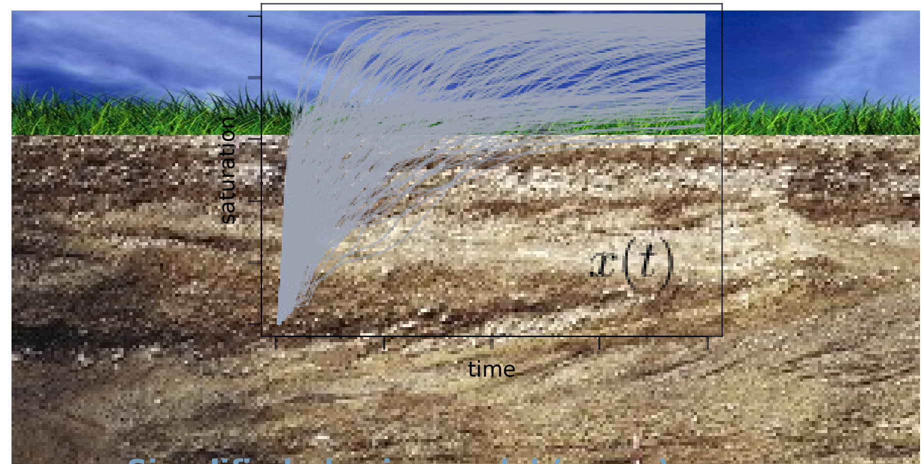
Exact model

- Full physics flow simulation
- Too costly
- Impossible to solve systematically for all geostatistical realizations
- Only for a few of them

Example: two-phase problem

$$\nabla \cdot \left[\left(\frac{k_n(S)}{\mu_n} + \frac{k_w(1-S)}{\mu_w} \right) k \nabla p \right] = 0$$

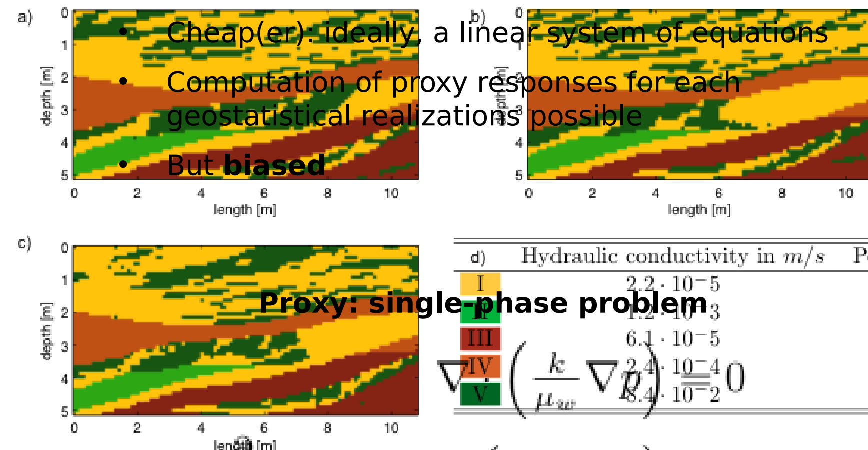
$$\frac{\partial}{\partial t}(\phi S) - \nabla \cdot \left(\frac{k_n(S)}{\mu_n} k \nabla p \right) = 0$$



“Truth” inspired from the Herten test case (Bayer et al. 2011)

Simplified physics model (proxy)

- Approximation of the physical processes
- Cheap(er): ideally, a linear system of equations
- Computation of proxy responses for each geostatistical realizations possible
- But **biased**



d)	Hydraulic conductivity in m/s	Porosity
I	$2.2 \cdot 10^{-5}$	0.21
II	$1.2 \cdot 10^{-3}$	0.32
III	$6.1 \cdot 10^{-5}$	0.13
IV	$2.4 \cdot 10^{-4}$	0.16
V	$8.4 \cdot 10^{-2}$	0.25

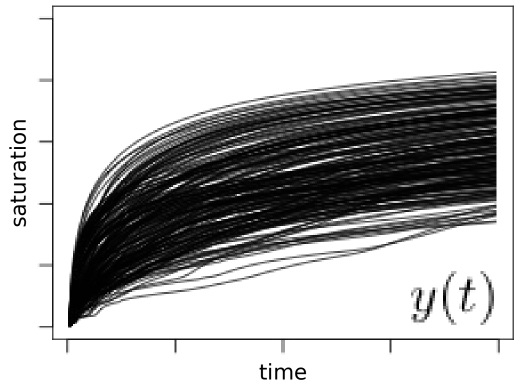
Proxy: single-phase problem

$$\nabla \cdot \left(\frac{k}{\mu_w} \nabla p \right) = 0$$

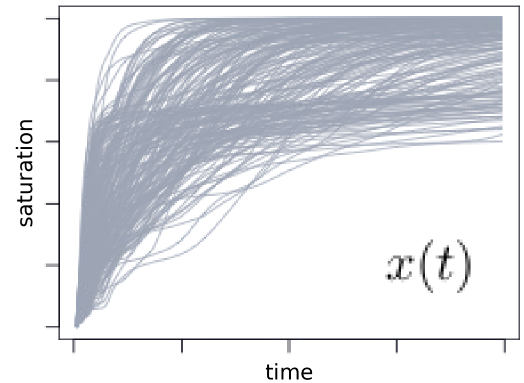
$$\frac{\partial}{\partial t}(\phi S) - \nabla \cdot \left(S \frac{k}{\mu_w} \nabla p \right) = 0$$

3 examples of geostatistical realizations generated using Direct Sampling (Mariethoz et al. 2010)

How to simulate flow?



Mapping between the two models



Exact model

- Full physics flow simulation
- Too costly
- Impossible to solve systematically for all geostatistical realizations
- Only for a few of them

Simplified physics model (proxy)

- Approximation of the physical processes
- Cheap(er): ideally, a linear system of equations
- Computation of proxy responses for each geostatistical realizations possible
- But **biased**

Error model

- To “recover” the missing physics
- Mapping between curves = regression model

How? Existing solutions:

→ On a learning set of realizations

→ Using Functional PCA

(Ramsay et al. 2006, 2009)

- Concurrent linear model

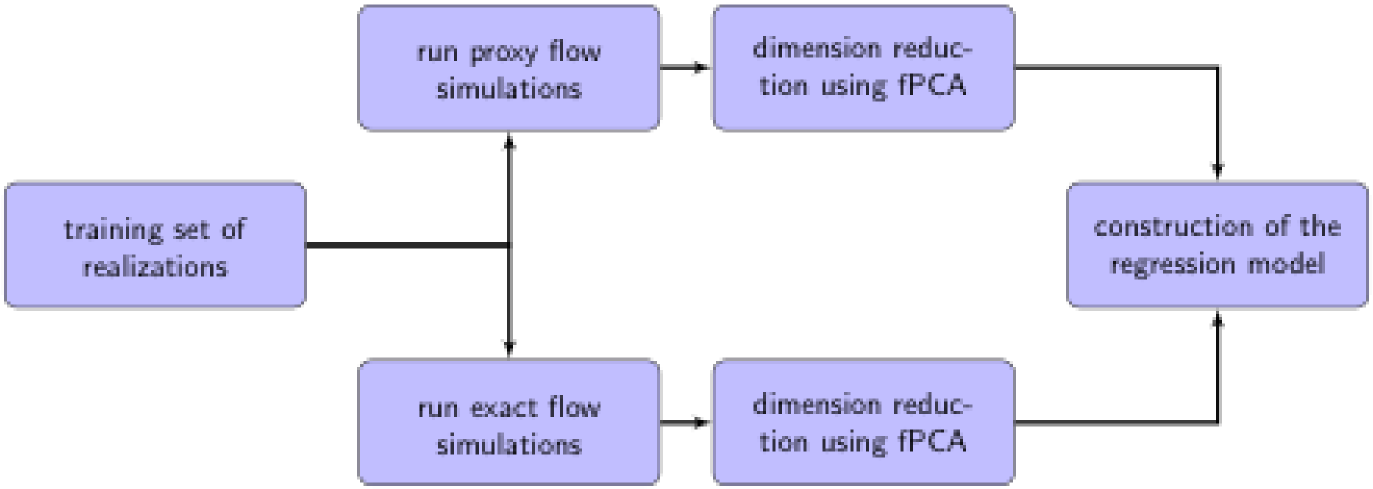
$$y_i(t) = \beta_0(t) + \beta_1(t)x_i(t) + \epsilon_i(t)$$

- Fully functional linear model

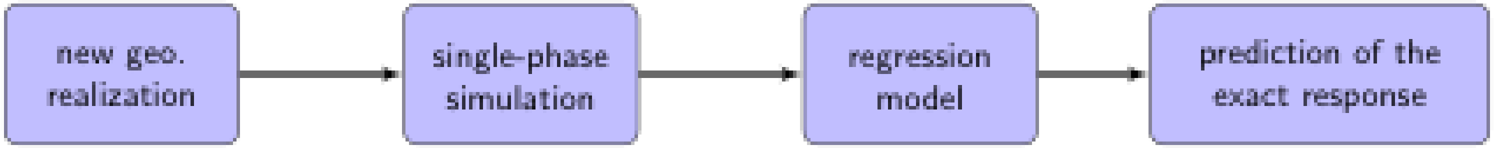
$$y_i(t) = \beta_0(t) + \int \beta_1(s,t)x_i(s)ds + \epsilon_i(t)$$

Workflow

Training phase of the error model



Prediction of the of the error model



Leak of oily pollutant

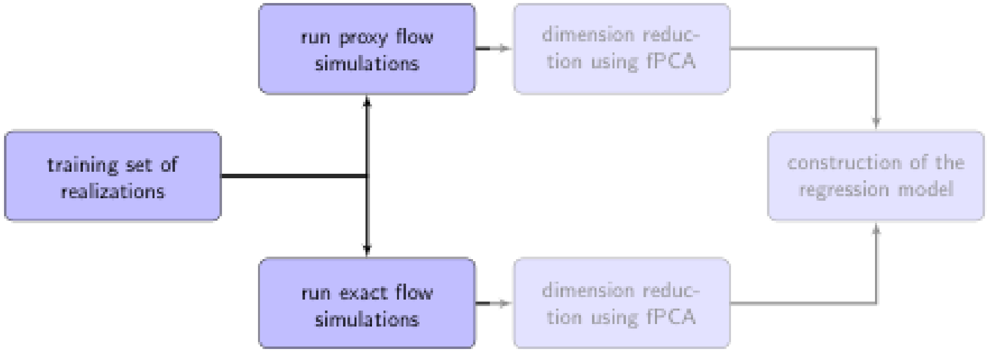
Drinking well

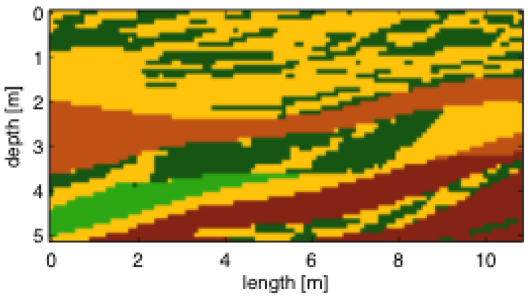
Description of the uncertainty:
1000 geostatistical realizations

ILLUSTRATION 1

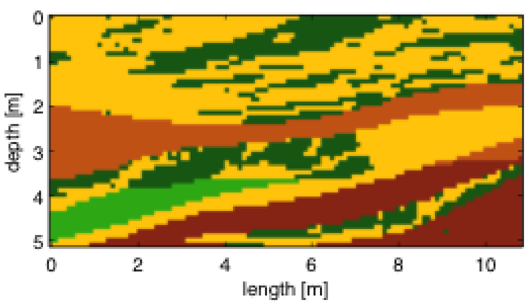
UNCERTAINTY QUANTIFICATION

Workflow





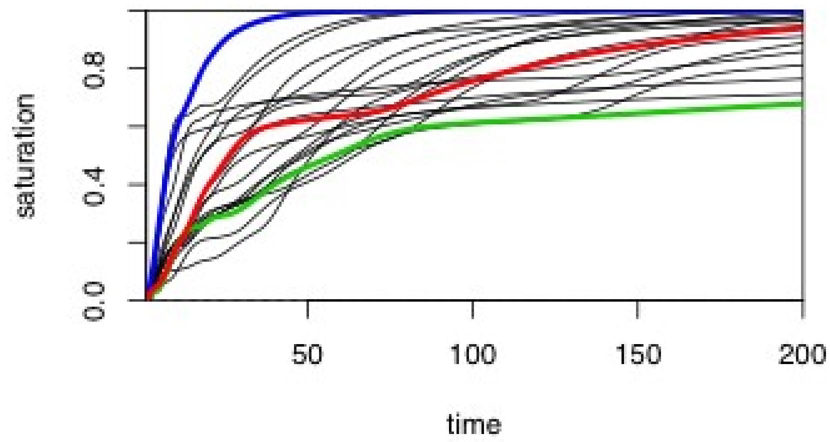
Blue curves



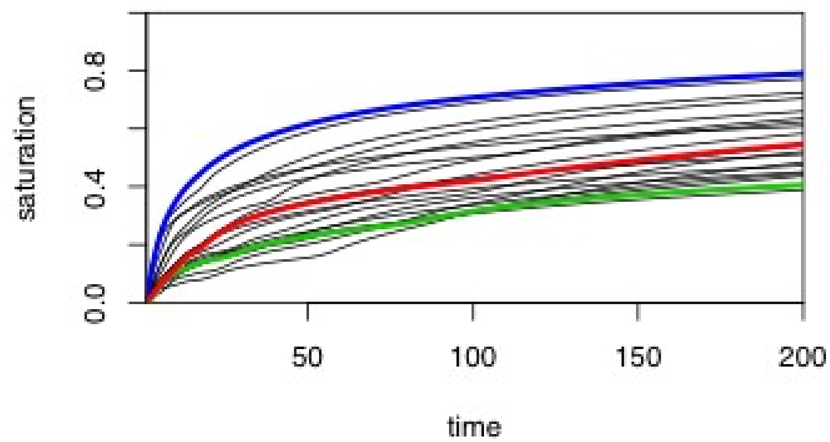
Green curves

Training set of 20 realizations

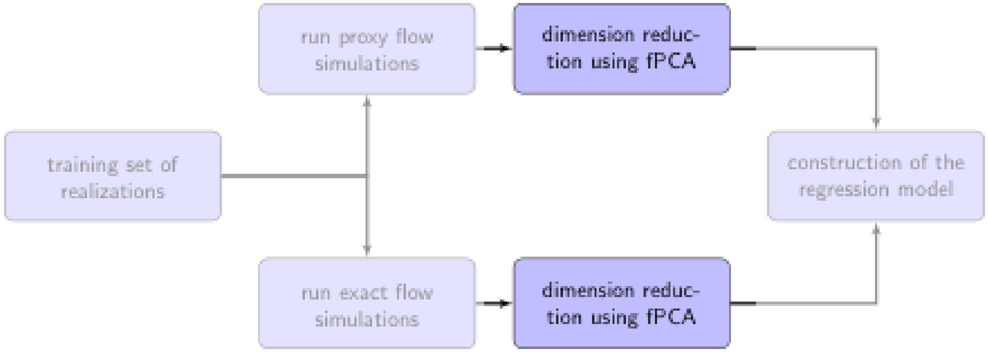
a) proxy breakthrough curves

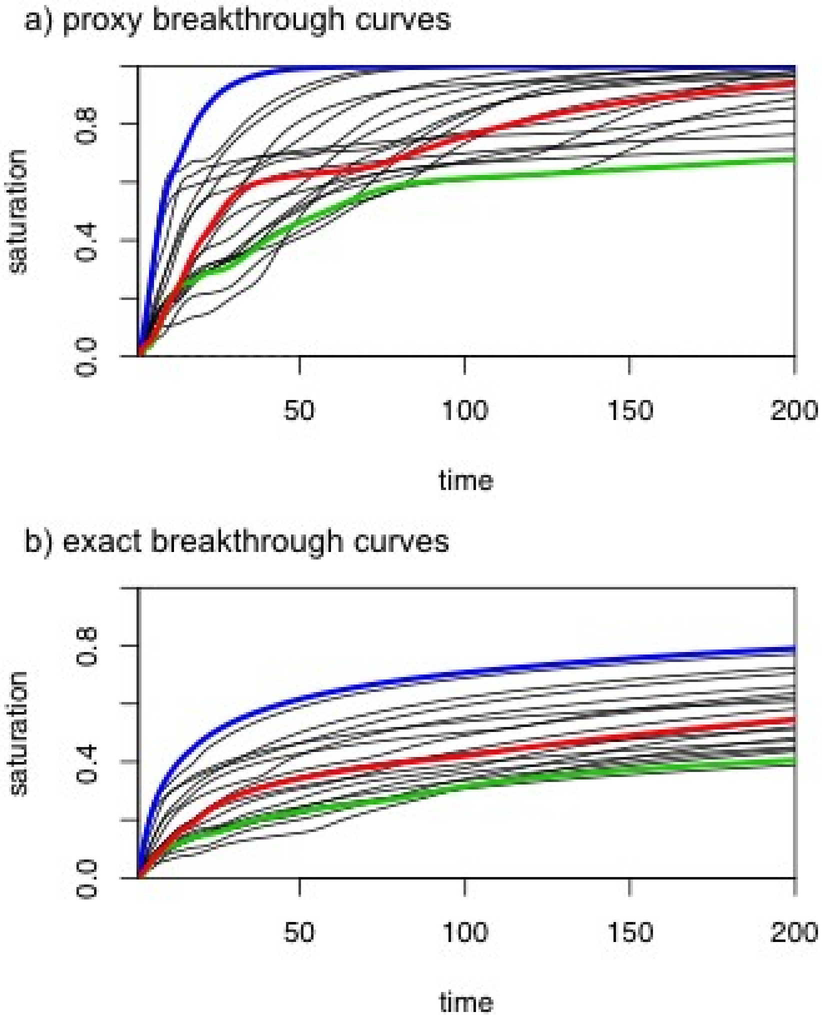


b) exact breakthrough curves



Workflow





FPCA $$x_i(t) \approx \bar{x}(t) + \sum_{j=1}^{N} s_{ij}\zeta_j(t)$$

Principal components (or harmonics) $\zeta_j(t)$ that maximises

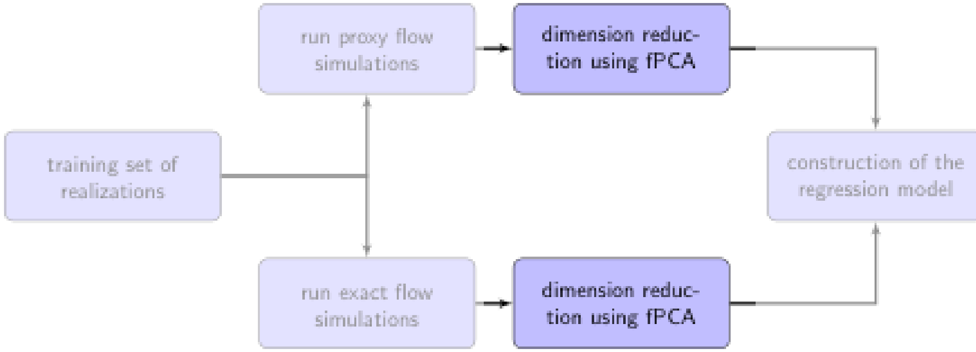
$$d_i = \mathrm{var}\Big(\int \zeta_i(t)[x_j(t) - \bar{x}(t)]dt \Big)$$

Principal components scores

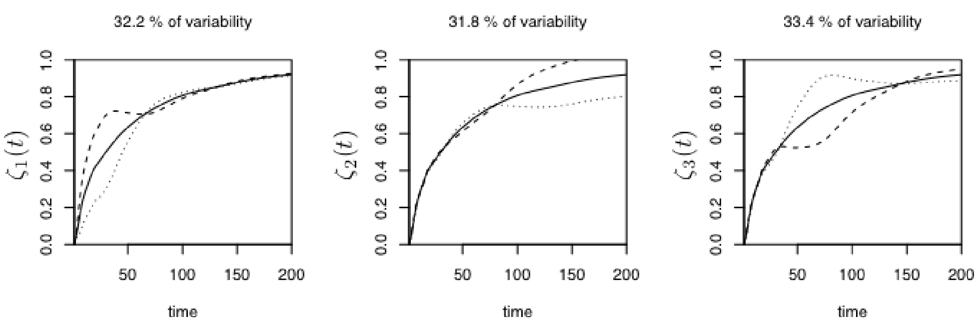
$$s_{ij} = \int [x_i(t) - \bar{x}(t)]\zeta_j(t)dt$$

Proportion of data explained by the i^{th} harmonics $$\frac{d_i}{\sum d_j}$$

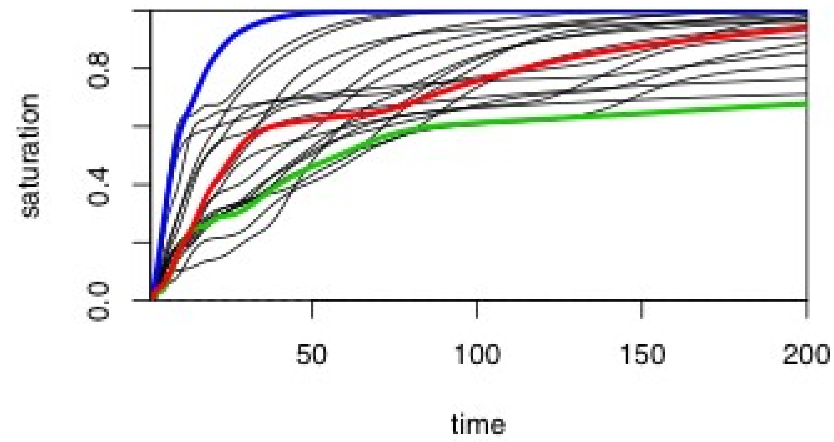
Workflow



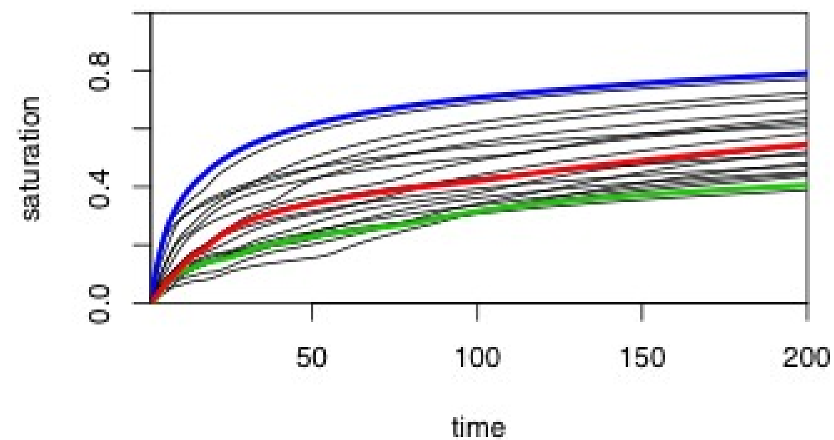
a) proxy harmonics



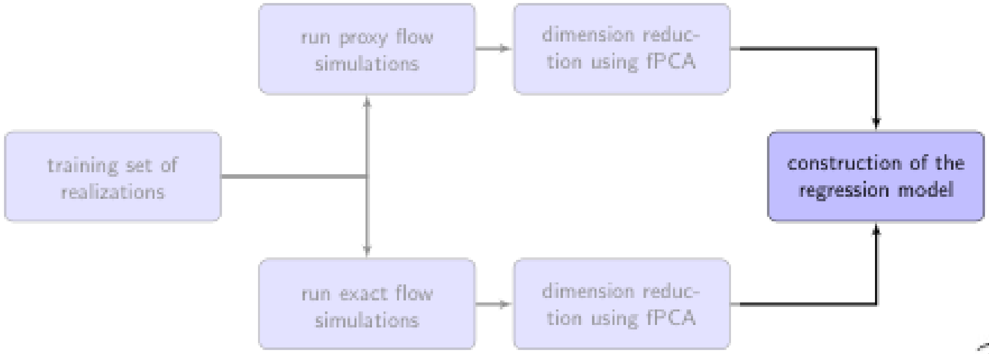
a) proxy breakthrough curves



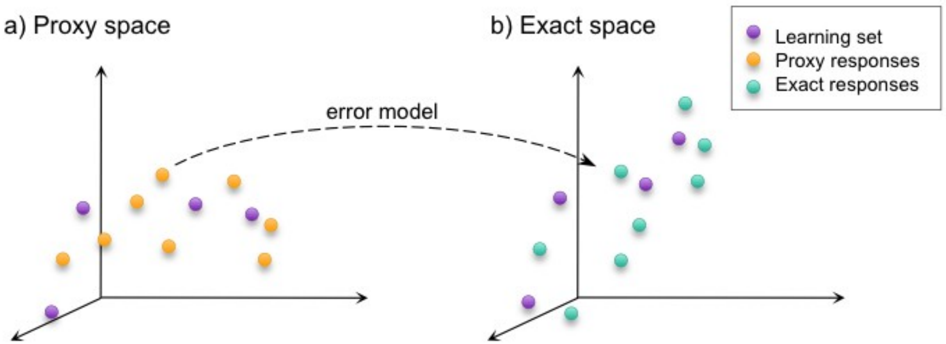
b) exact breakthrough curves

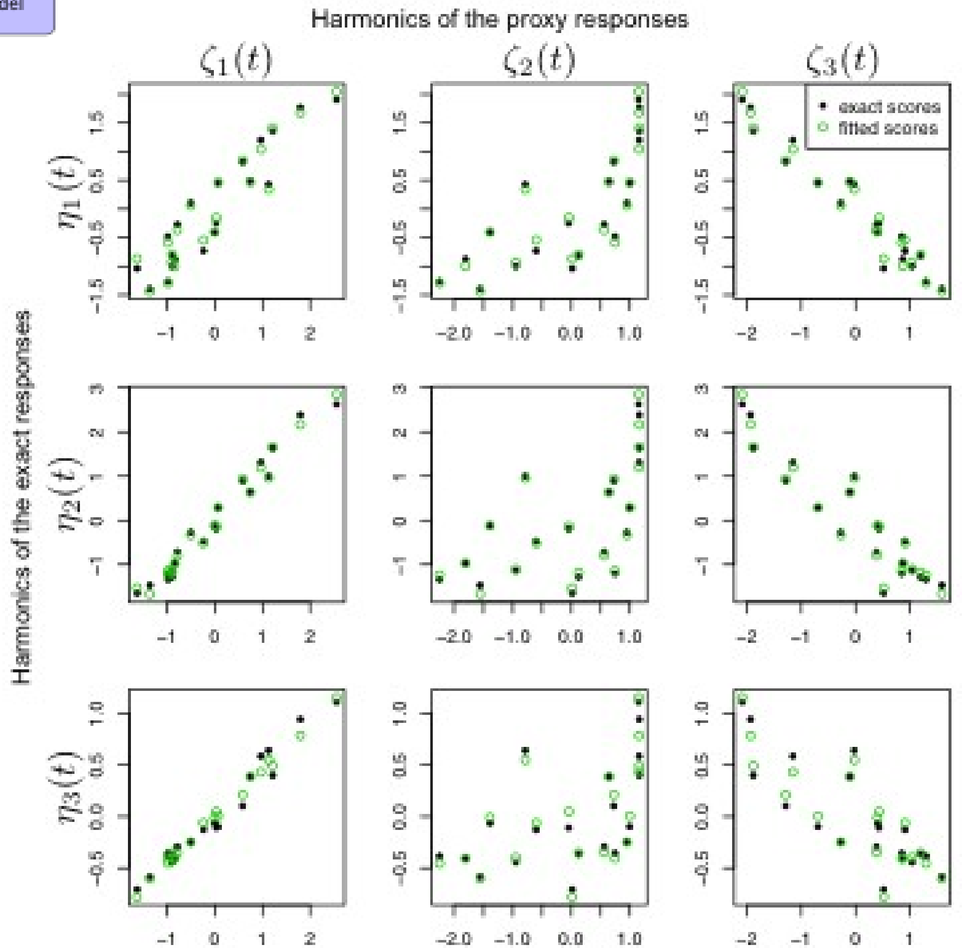


Workflow



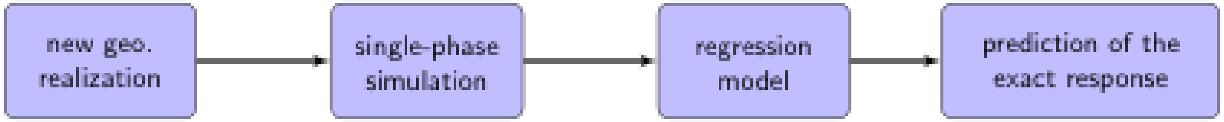
$$x_i(t) \approx \bar{x}(t) + \sum_{j=1}^{3} s_{ij}\zeta_j(t)$$

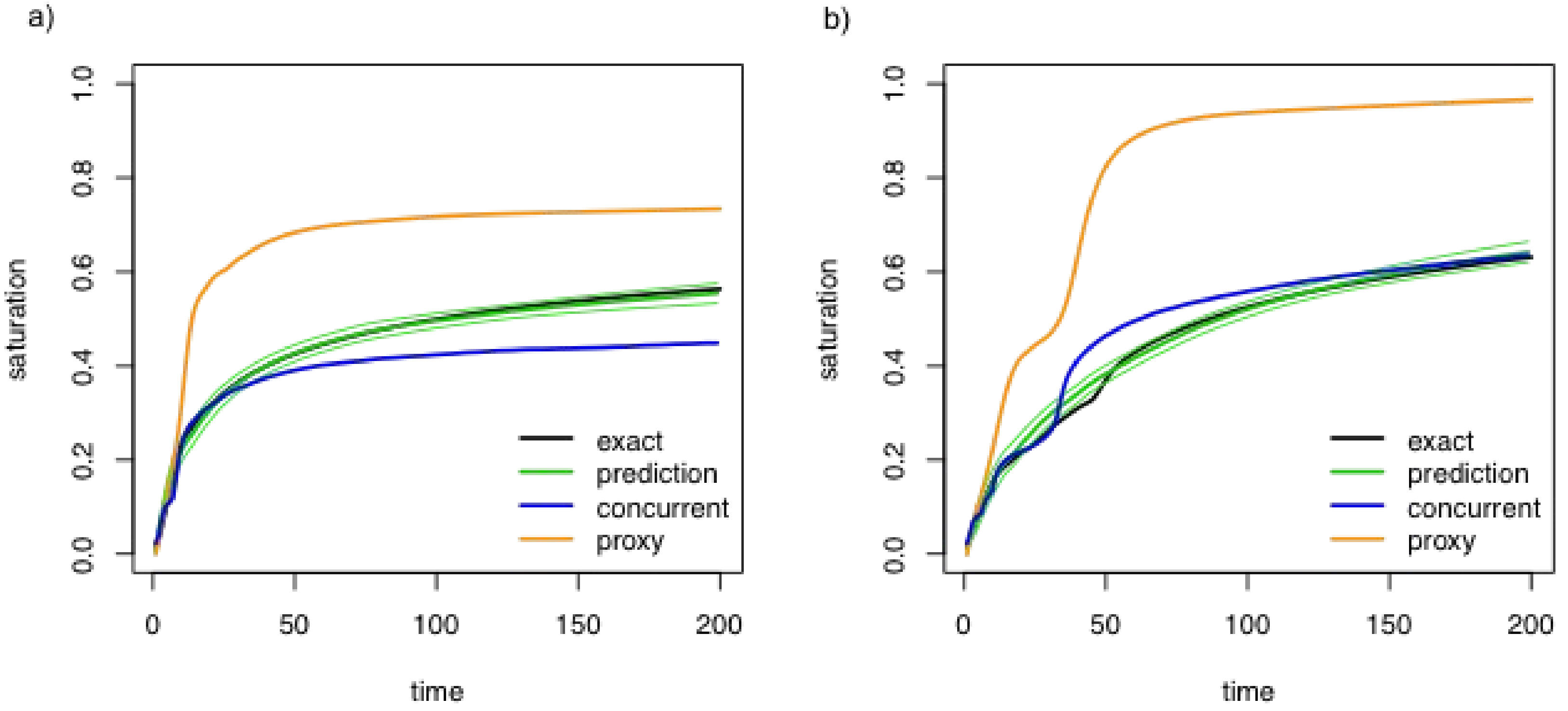




laureline.josset@unil.ch - MASCOT-NUM 2015

Two examples of predictions

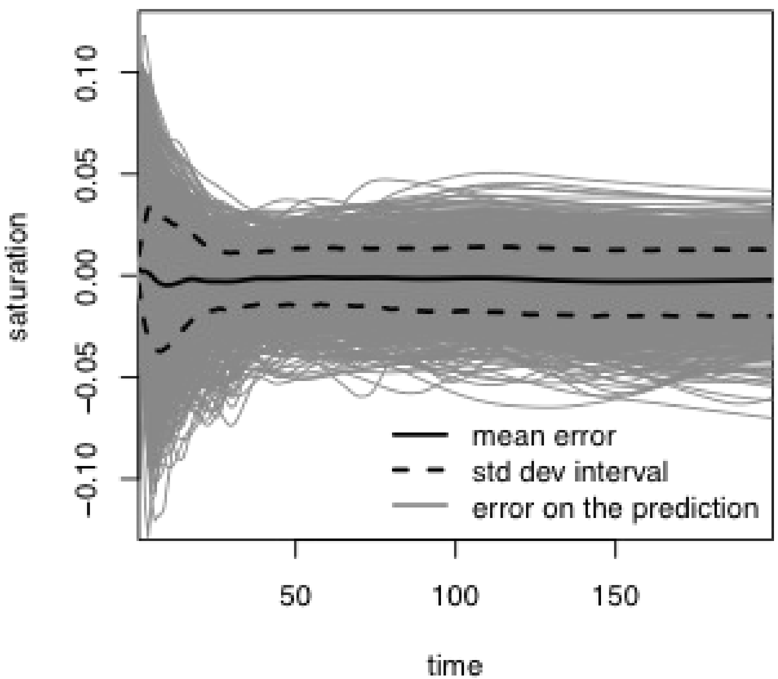




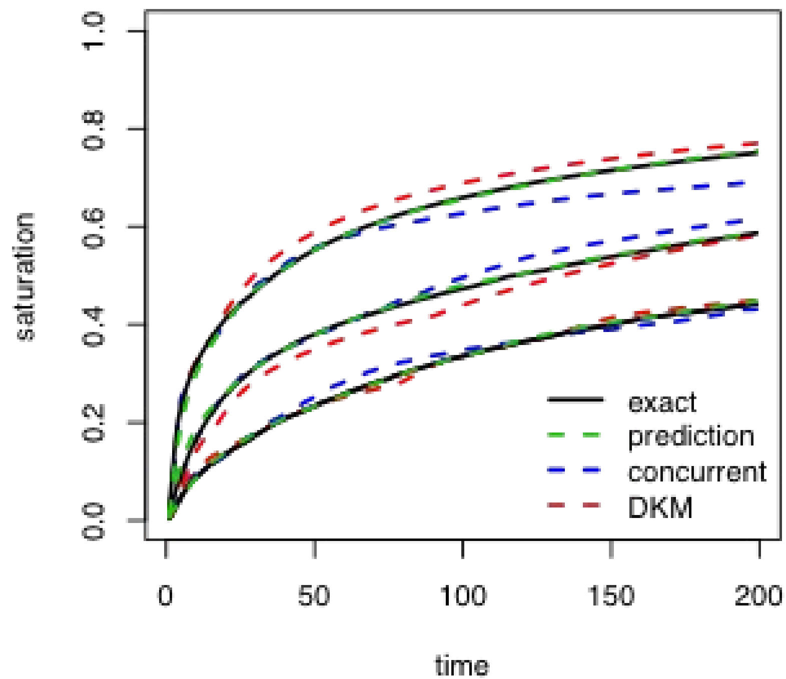
Prediction of the ensemble
1000 realizations

new geo. realization → single-phase simulation → regression model → prediction of the exact response

c) *Error curves: predicted – exact*



d) *Point-wise quantiles P10-50-90*



Good prediction of the point-wise quantiles

Prediction for each of the curves ✉ useful beyond UQ

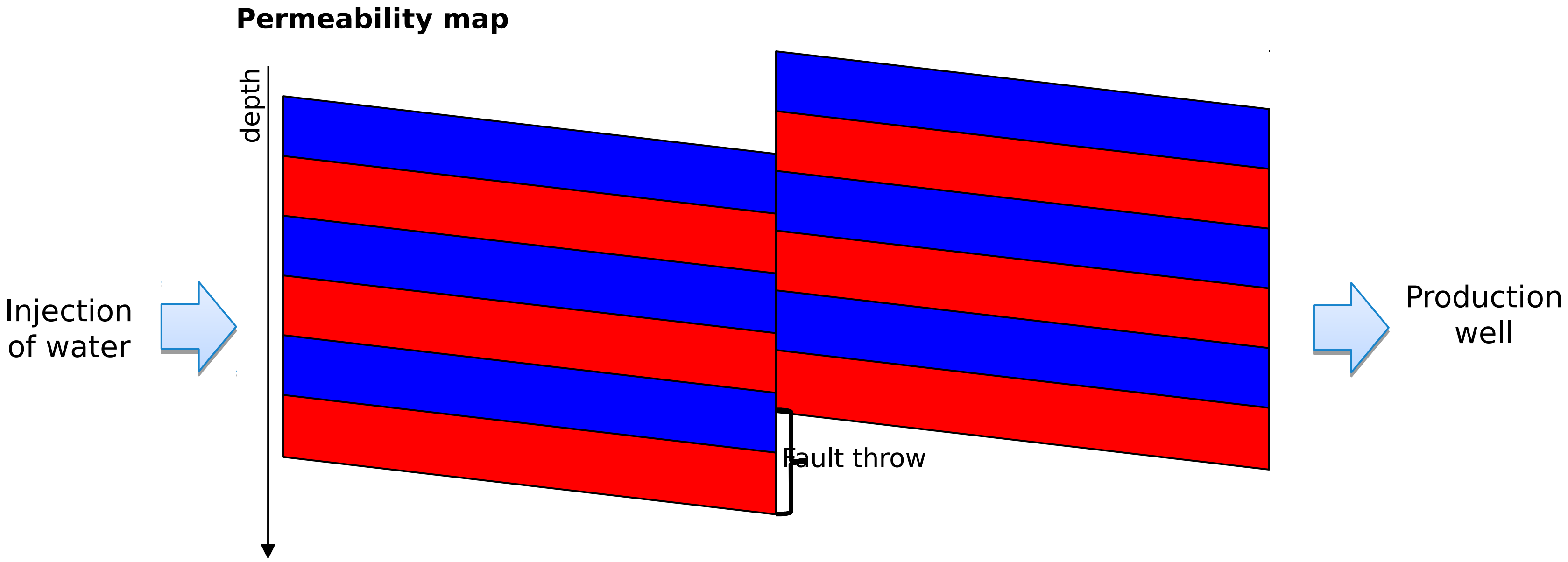
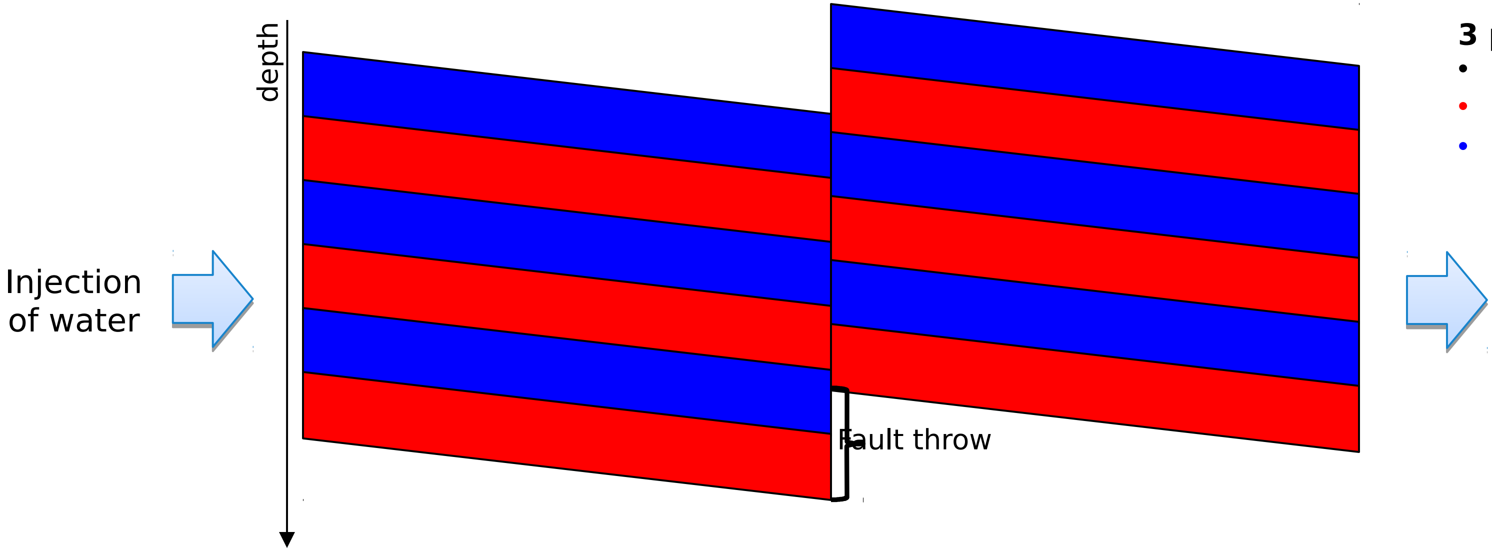


ILLUSTRATION 2

HISTORY MATCHING

IC Fault test case

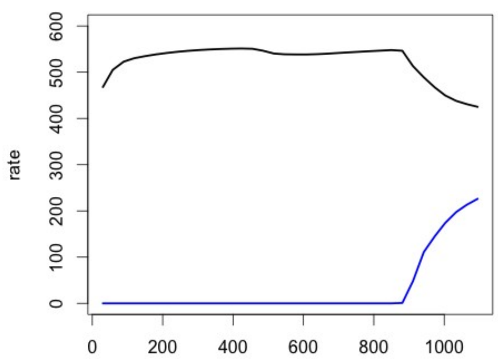
Permeability map



Imperial College Fault problem
Z Tavassoli, JN Carter, PR King (2004)

3 parameters:
- Fault throw = ?
- K_{high} = ?
- K_{low} = ?

Observed data:
- Oil production rate
- Water production rate



Goal:
Sample the parameters given the observed data

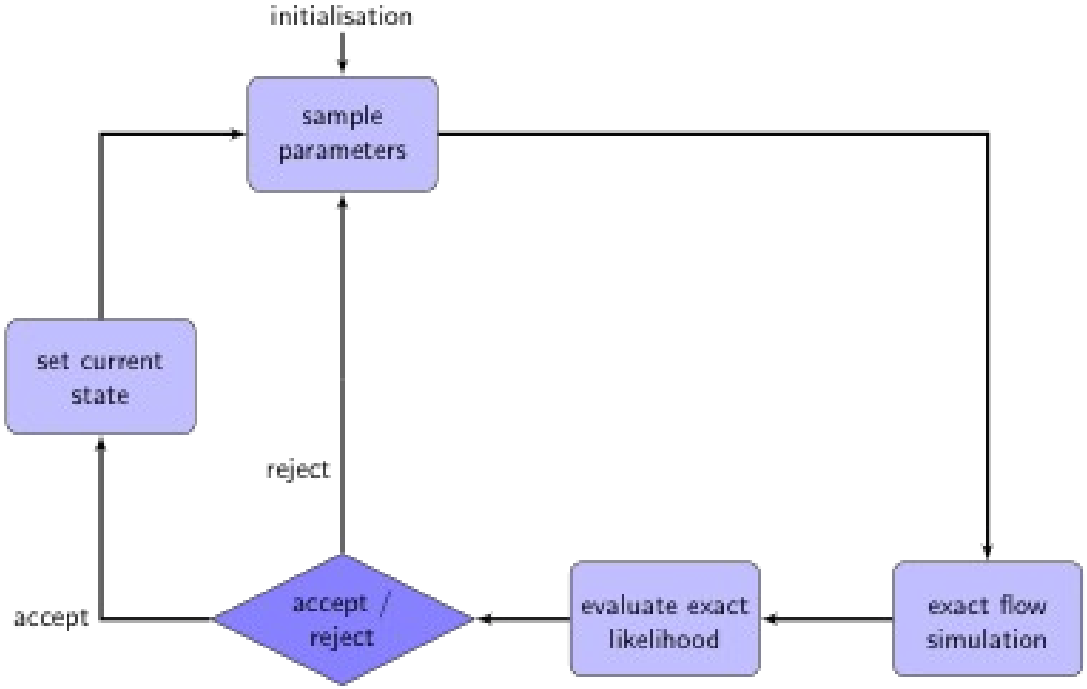
$$p(\theta|y) \propto \mathcal{L}(\theta; y)p(\theta)$$

Choice of simplified physics model: single-phase simulation

→ Provides information on the connectivity of the realizations

→ Cheap: pressure problem is solved only once

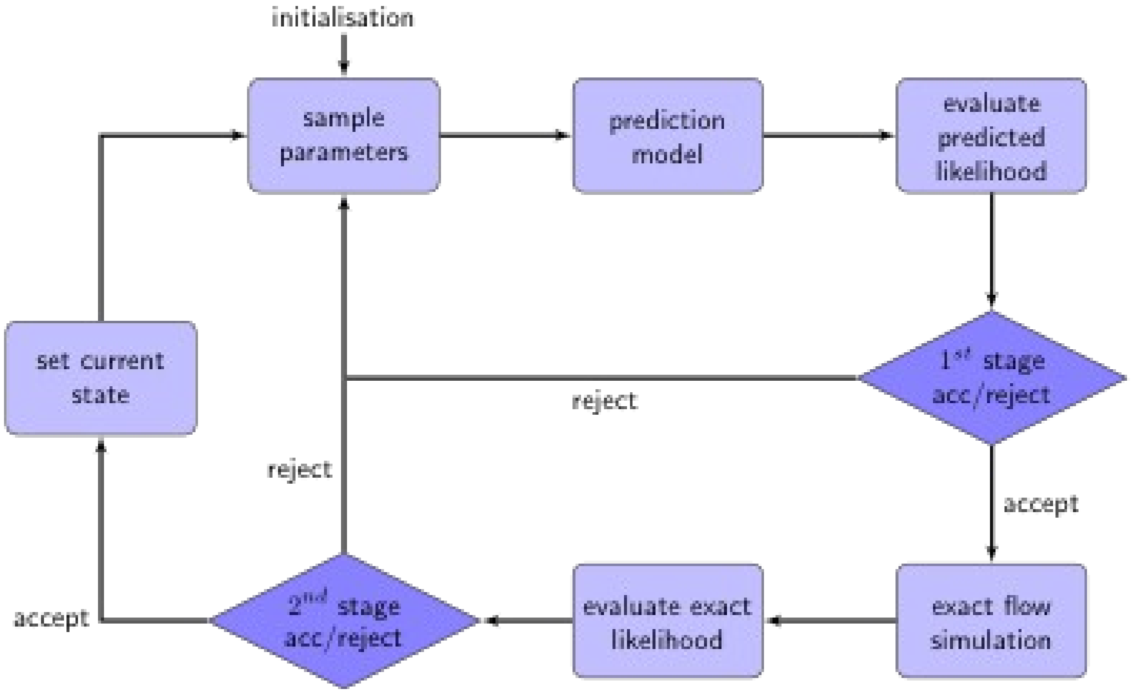
2-stage MCMC



Metropolis-Hastings

- To sample the posterior probability density function
- Typical application 10^5 iterations
- finite length chains should be able to explore all areas of the prior space
- Increase the step length of the chains?
  - Drastic reduction of the acceptance rate
  - High number of wasted simulations

2-stage MCMC



Metropolis-Hastings

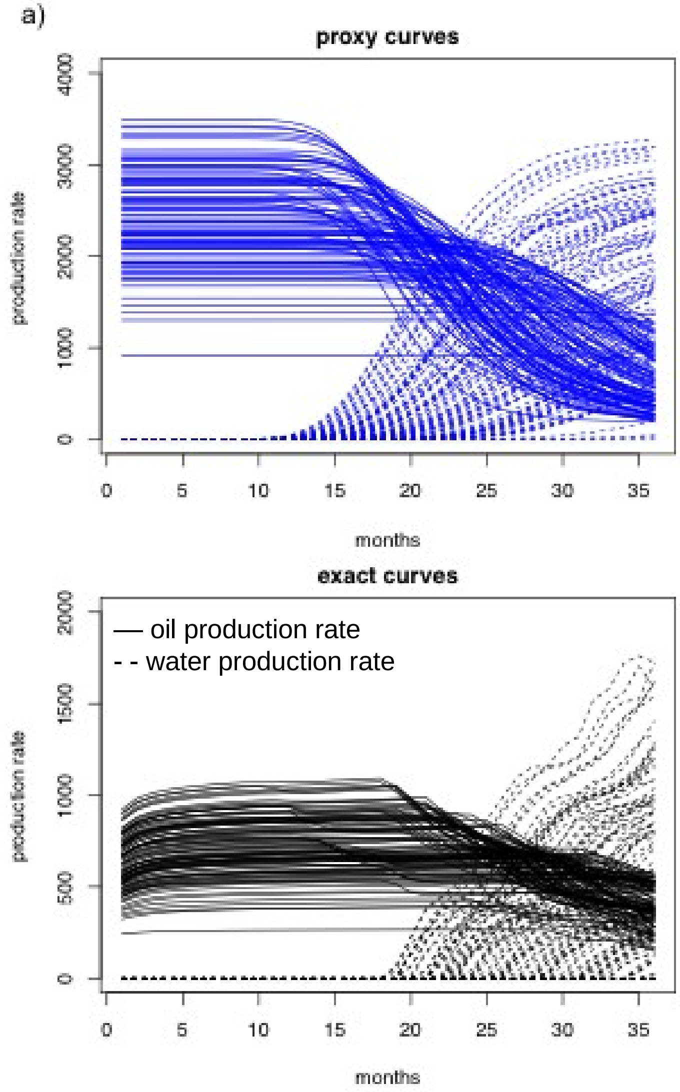
- To sample the posterior probability density function
- Typical application 10^5 iterations
- finite length chains should be able to explore all areas of the prior space
- Increase the step length of the chains?
  - Drastic reduction of the acceptance rate
  - High number of wasted simulations

2-stage MCMC*

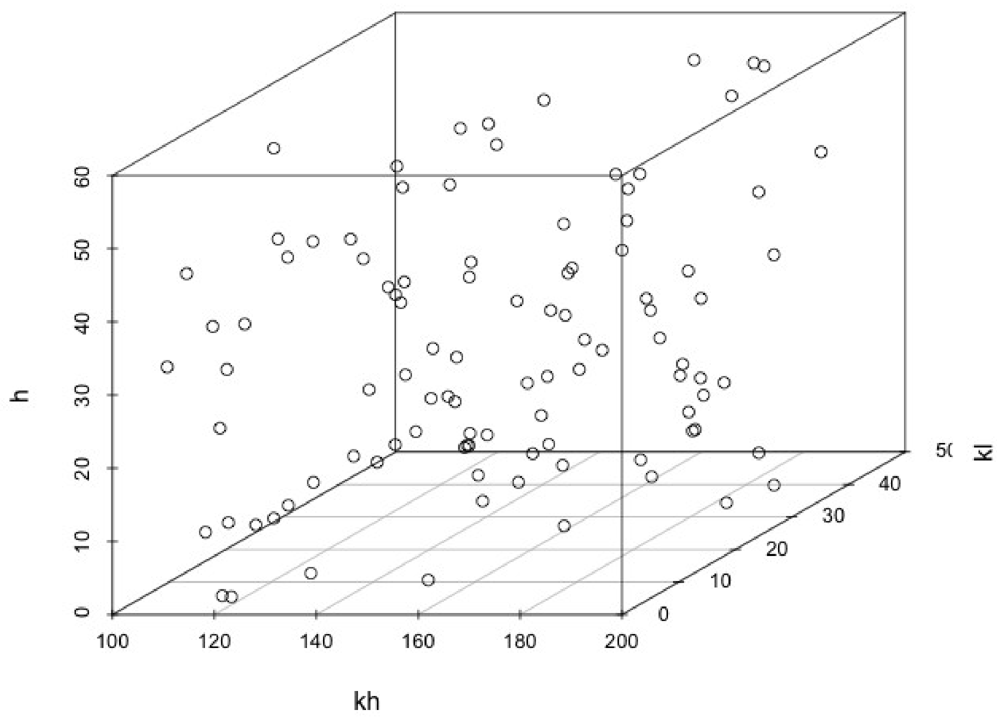
- Avoid unnecessary run of the exact solver
- Reject samples based on the predicted response

**Christen and Fox (2005), Efendiev et al. (2005, 2006)*

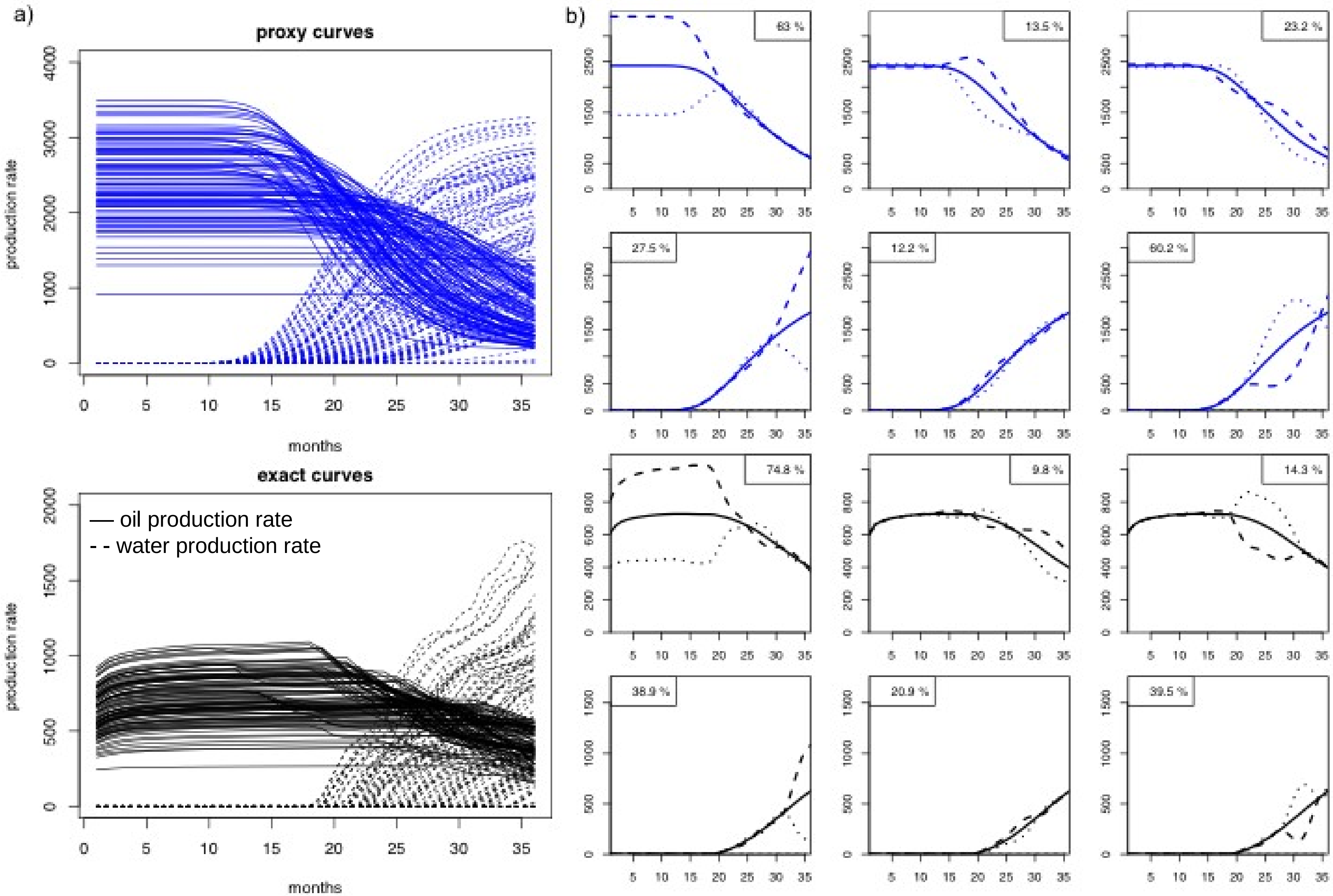
Training set and dimension reduction



Training set: 100 curves sampled from the parameters space

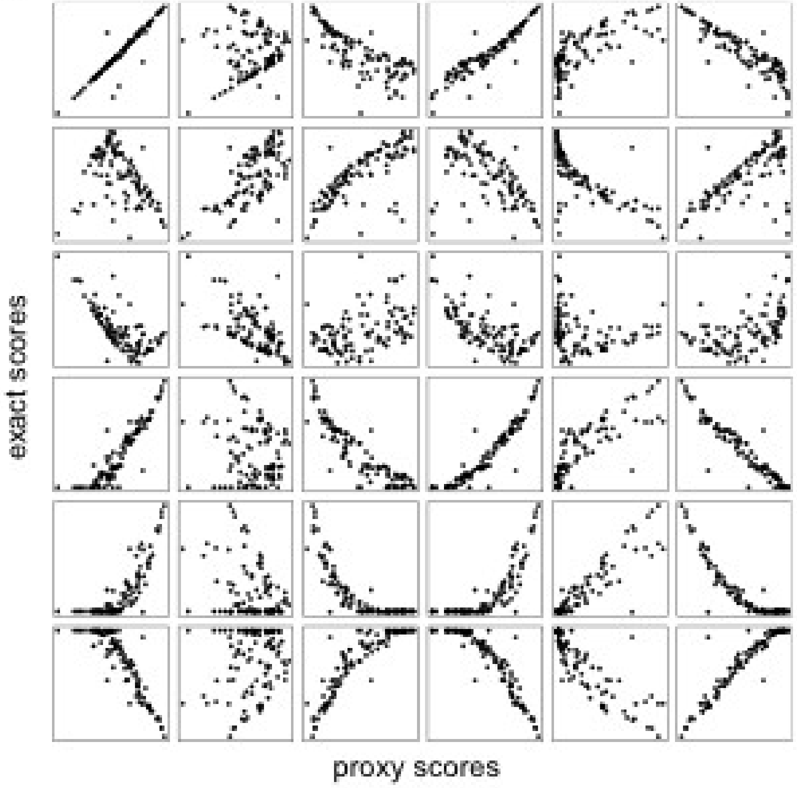


Training set and dimension reduction

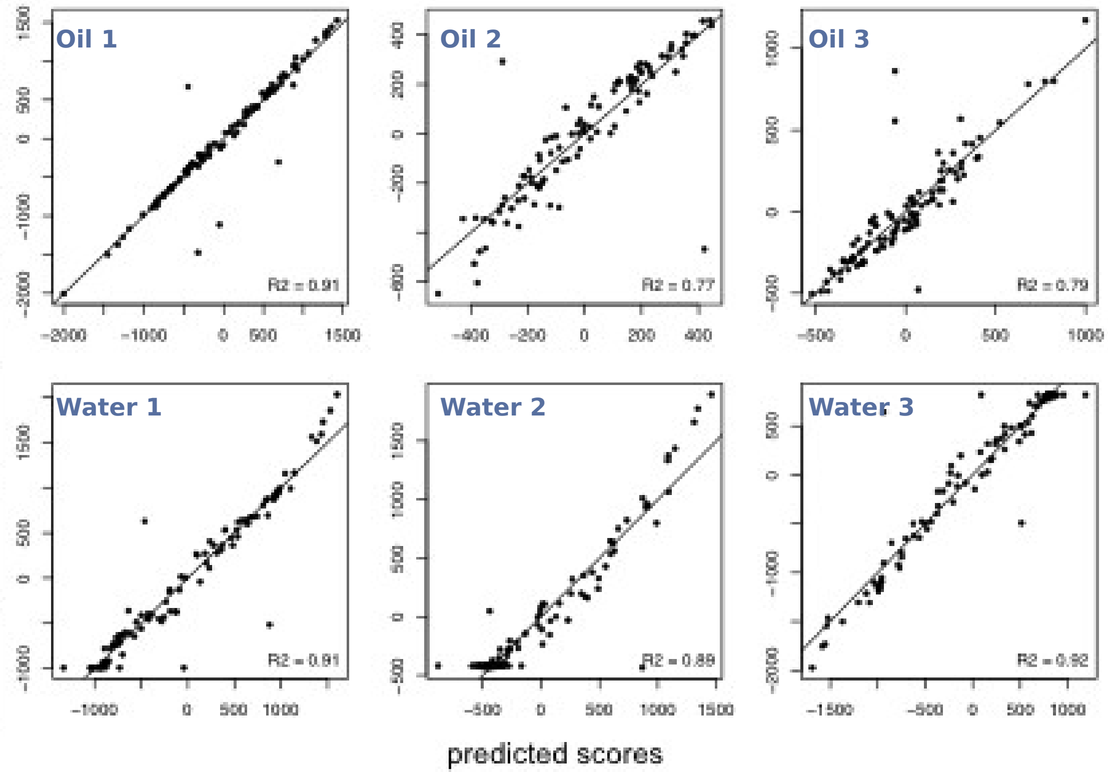


Construction of the regression model

a) **Scatterplot of the exact and proxy scores**

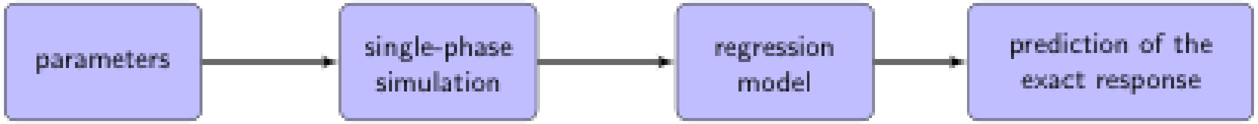


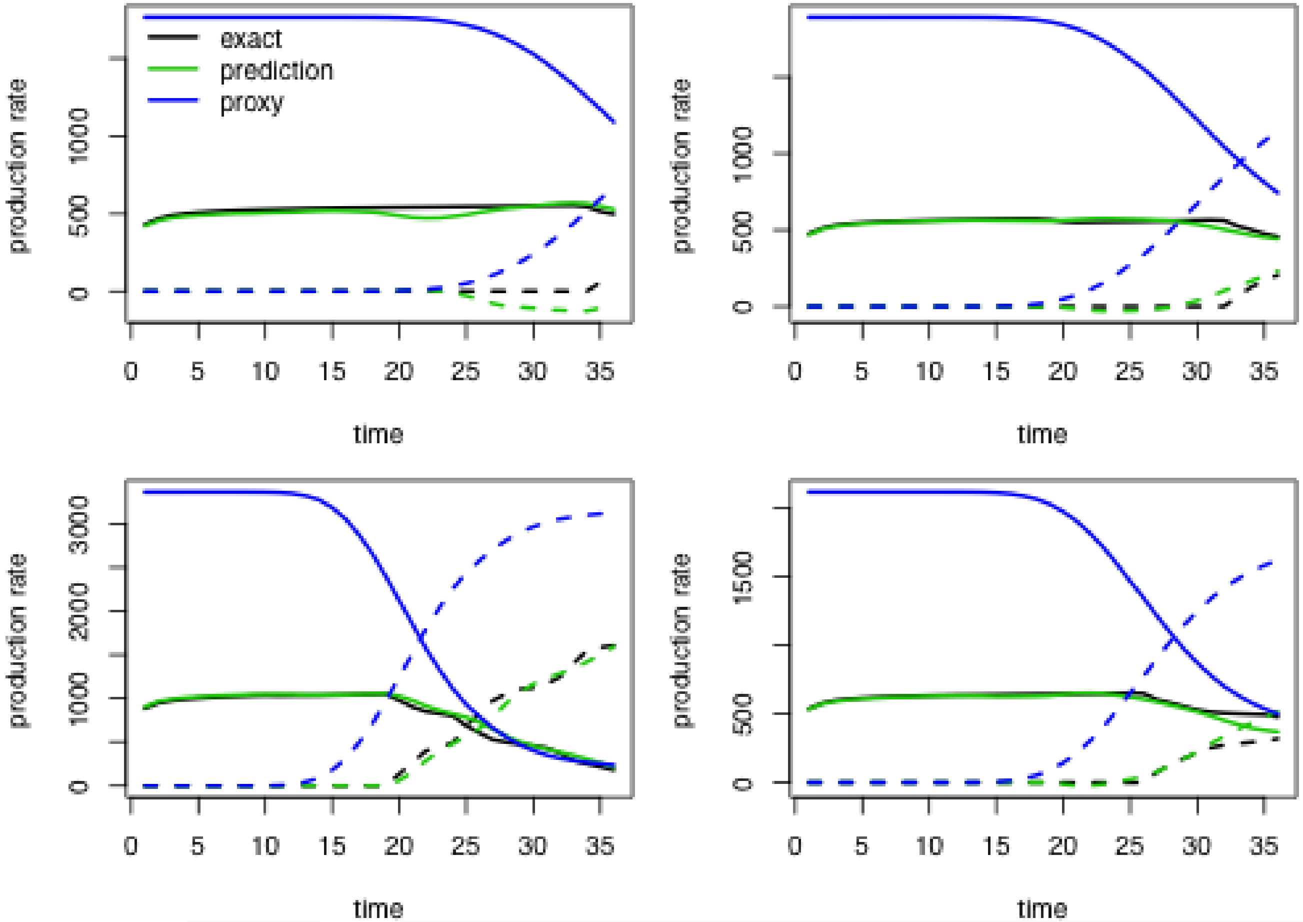
b) **Plot of the exact VS predicted scores**



The proxy is useful to predict the exact response

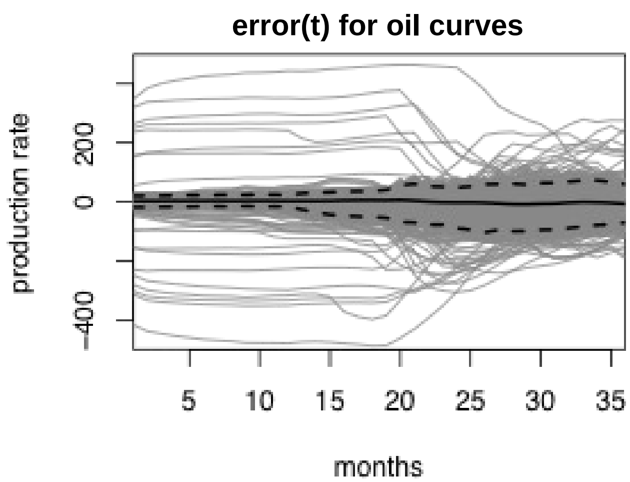
Four examples of predictions

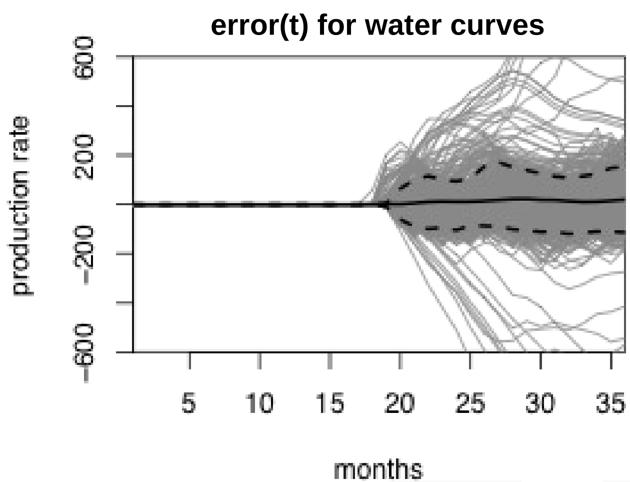




Evaluation of the performance of the error model

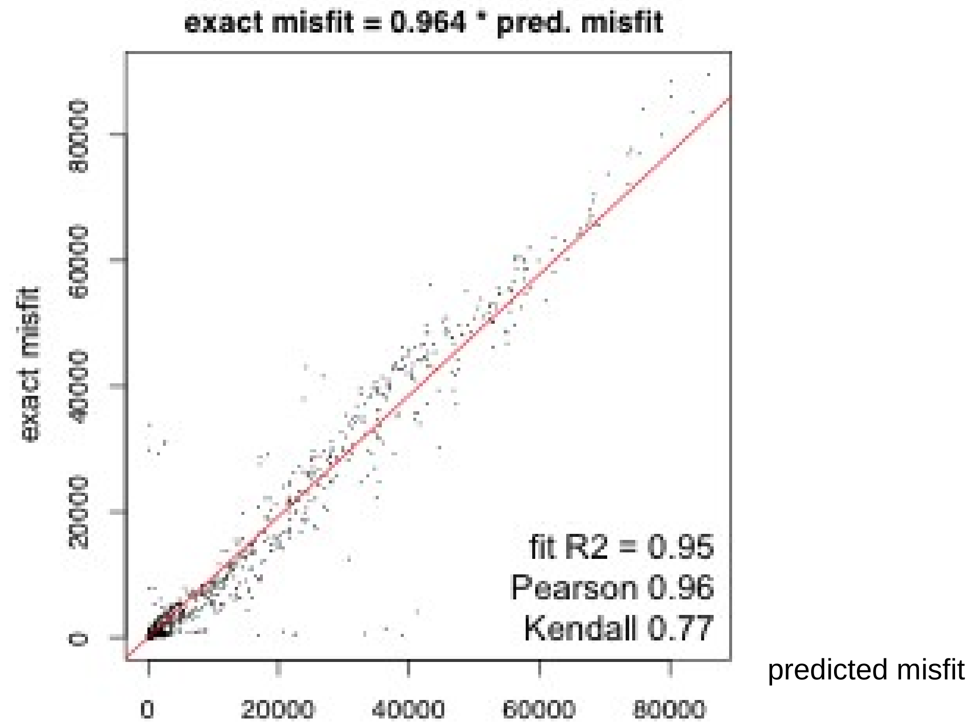
Test set of 1000 realizations



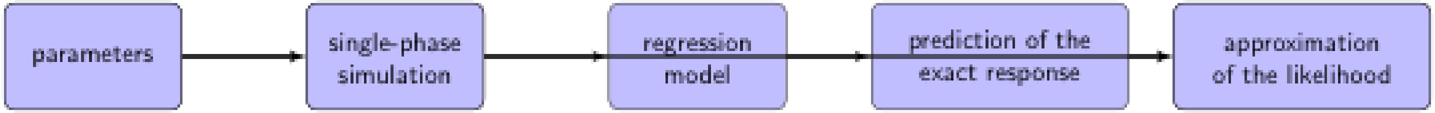


Predicted curves → predict the misfit:

$$M = \frac{1}{36}\sum_{t=1}^{36}\frac{(C_{ref}^{oil}(t) - C^{oil}(t))^2}{2\sigma^2} + \frac{1}{7}\sum_{t=30}^{36}\frac{(C_{ref}^{water}(t) - C^{water}(t))^2}{2\sigma^2}$$

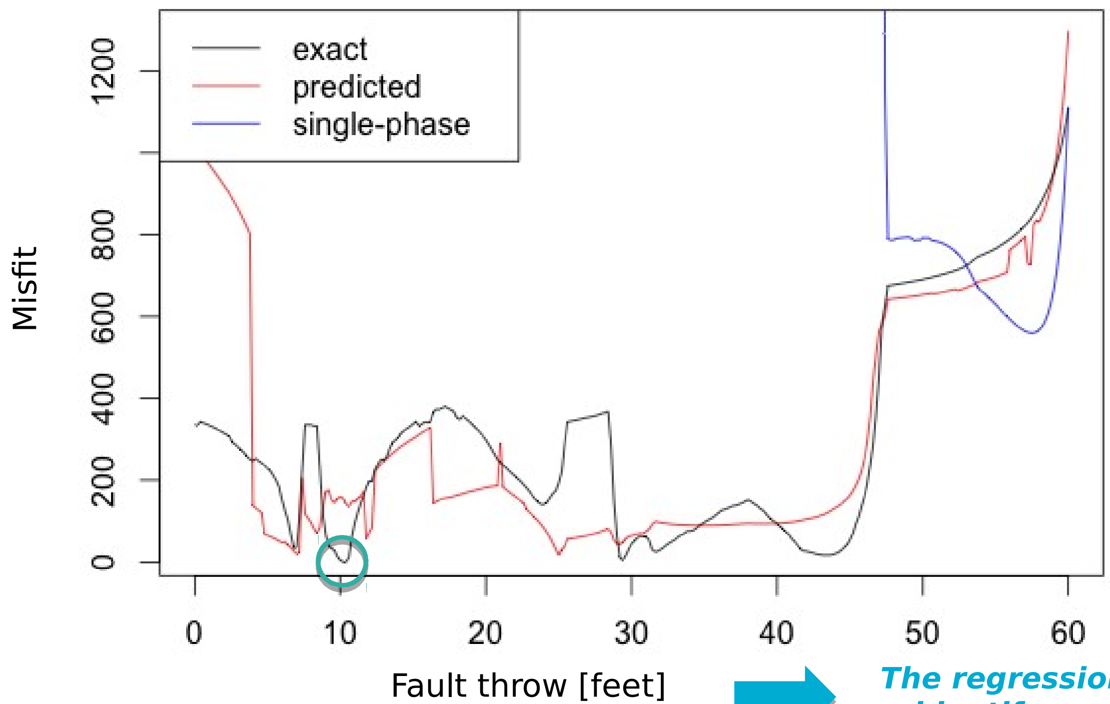


Is the error model necessary?



$$\mathcal{L} = e^{-M} \qquad M = \frac{1}{36}\sum_{t=1}^{36}\frac{(C_{ref}^{oil}(t) - C^{oil}(t))^2}{2\sigma^2} + \frac{1}{7}\sum_{t=30}^{36}\frac{(C_{ref}^{water}(t) - C^{water}(t))^2}{2\sigma^2}$$

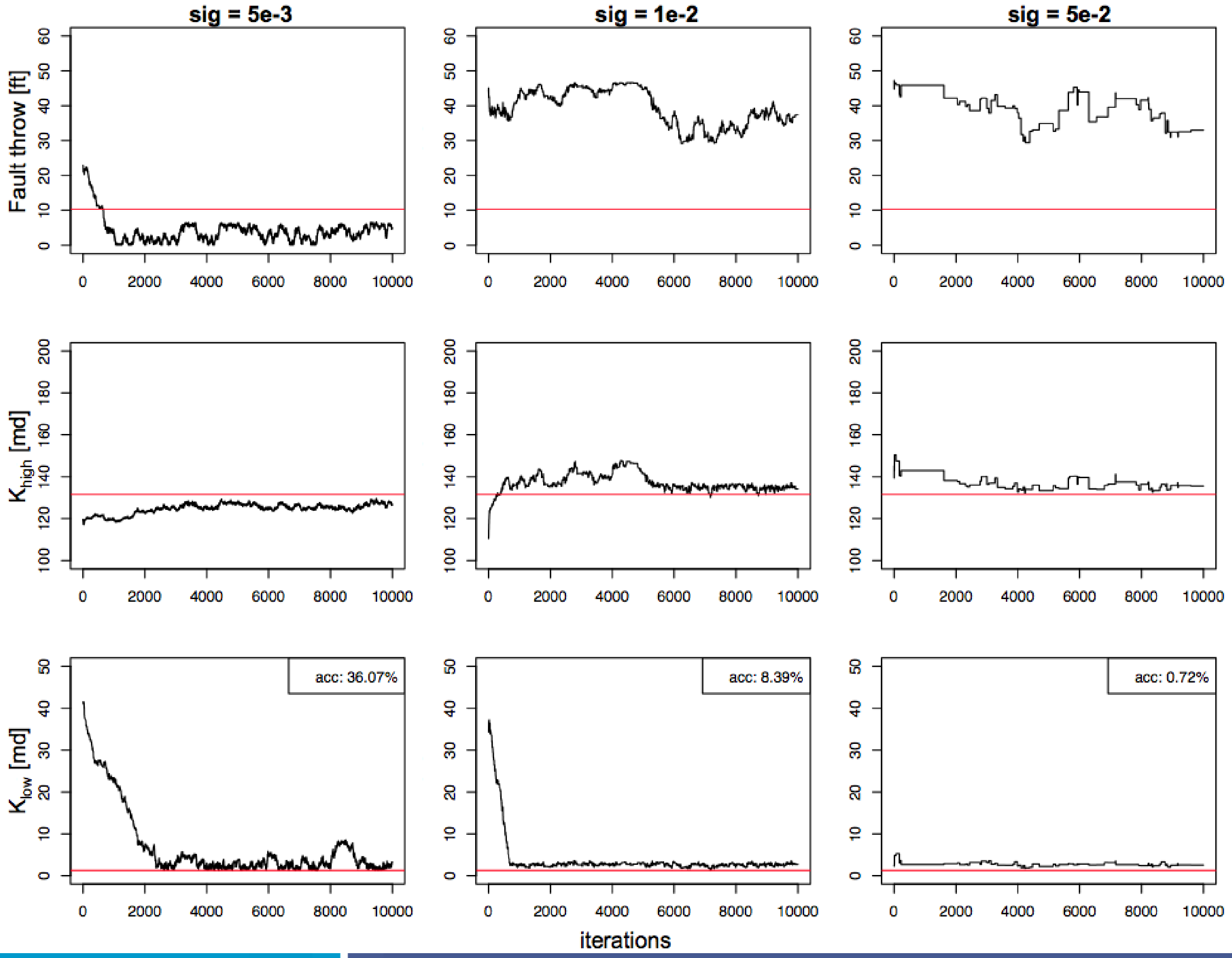
K_{high} = 131.6
K_{low} = 1.3
true parameter



The regression model is necessary to identify regions in the parameter space

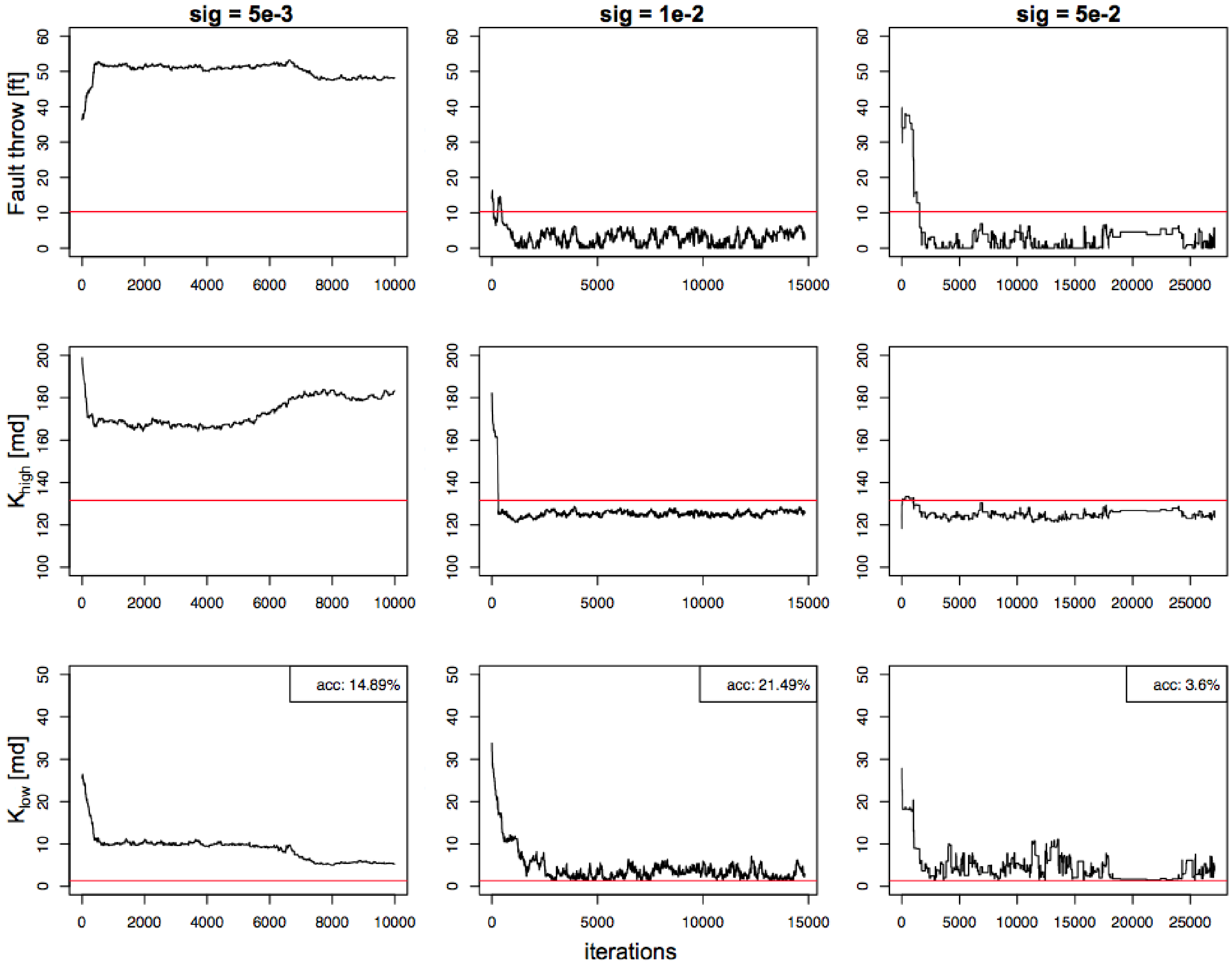
Metropolis-Hastings results

3 chains for different step size
Length: 10'000 evaluations



2-stage MCMC results

3 chains for different step size
Length: equivalent MH



Comparison of the results

random walk	nb of it.			nb of acc. 1stage sim			nb of acc. 2stage sim			acc. rate			
σ	C1	C2	C3	C1	C2	C3	C1	C2	C3	C1	C2	C3	mean
	Metropolis-Hasting												
$5 \cdot 10^{-3}$	10'000	10'000	10'000				1'631	3'247	1'291	18.1%	36.1%	14.3%	22.8%
$1 \cdot 10^{-2}$	10'000	10'000	10'000				1'683	755	628	18.7%	8.4%	7.0%	11.4%
$5 \cdot 10^{-2}$	10'000	10'000	10'000				179	65	48	2.0%	0.7%	0.5%	1.1%
	Two-stage MCMC												
$5 \cdot 10^{-3}$	10'000	10'000	10'000	4'760	5'299	176	367	789	41	7.7%	14.9%	23.3%	15.3%
$1 \cdot 10^{-2}$	14'372	14'815	31'738	9'666	9'656	7'820	2'060	2'075	331	23.3%	21.5%	4.2%	16.3%
$5 \cdot 10^{-2}$	28'337	31'777	27'108	9'341	9'261	9'370	393	518	337	4.2%	5.6%	3.6%	4.5%

2-stage MCMC with the error model

- Higher acceptance rate
- Longer chains can be run for the same computational cost

However

- Nowhere near convergence
- ICF still a very challenging problem
- As the Swiss say: “ça va pas mieux mais plus longtemps !”

Conclusion
Key ideas

Prediction model

= proxy + error model

= single-phase + FPCA regression

- Why single-phase flow simulations:
  - Connectivity is what varies between realisations
  - Cheap: pressure is solved only once
- Why error modelling:
  - Missing physics has to be taken in account

Advantages

- Strong reduction of computational costs
- Allows the evaluation of the relevance of the proxy for the specific problem

Outlook

- On going work: sensitivity analysis
- Application to seawater intrusion in coastal aquifer
- Evolve to more complex regression model -> Kernel methods

Acknowledgements

David Ginsbourger, University of Bern

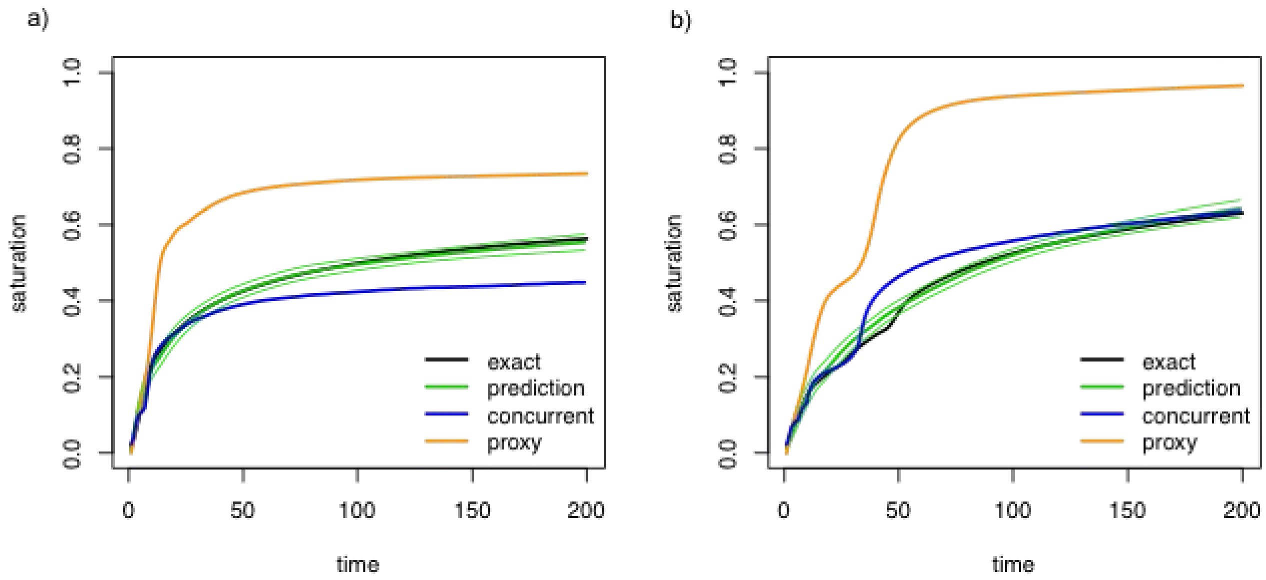
Ahmed H. Elsheikh and Vasily Demyanov, University of Heriot-Watt

References

- L. Josset, D. Ginsbourger and I. Lunati, “Functional error modeling for uncertainty quantification in hydrogeology”, Water Resources Research (2015)
- L. Josset, V. Demyanov, A.H. Elsheikh and I. Lunati, “Accelerating Monte Carlo Markov chains with proxy and error models”, Computer and Geosciences (in revision)
- P. Bayer et al., “Three-dimensional high resolution fluvio-glacial aquifer analog”, J. Hydro 405 (2011) 19
- G. Mariethoz, P. Renard, and J. Straubhaar “The Direct Sampling method to perform multiple-point geostatistical simulations”, Water Resour. Res., 46 (2010)
- J. Ramsay, G. Hooker and S. Graves, “Functional data analysis with R and MATLAB’’, Springer (2009)
- P. Tavassoli et al., “Errors in history matching”, SPE 86883 (2004)

THANK YOU FOR YOUR ATTENTION

Simultaneous confidence bands



$$Pr\Big(\tilde{y}(t) \in [\hat{y}(t) - w_\alpha(t), \hat{y}(t) + w_\alpha(t)] \text{ for all } t\Big) = 1 - \alpha$$

$$w_\alpha(t) = \sqrt{\left(\frac{D_{ex}(N_l - D_{app} - 1)}{N_l - D_{ex} - D_{app}}\right) F_{D_{ex}, N_l - D_{ex} - D_{app}}(\alpha)} \times \sqrt{(1 + \mathbf{b}'(\mathbf{B}'\mathbf{B})^{-1}\mathbf{b})\left(\frac{N_l}{N_l - D_{app} - 1}\right)\boldsymbol{\eta}'(t)\hat{\boldsymbol{\Sigma}}\boldsymbol{\eta}(\boldsymbol{t})},$$

with $\boldsymbol{\eta}(t)$ the values of the exact harmonics

$\hat{\boldsymbol{\Sigma}}$ the covariance matrix of errors

$F(\alpha)$ Fisher's α quantile