MULTIPLE-POINT STATISTICS TO ASSESS COMPLEX SPATIAL UNCERTAINTY

Philippe Renard

Stochastic Hydrogeology Group
University of Neuchâtel
Switzerland

MascotNum, 25th of April 2014, Zürich

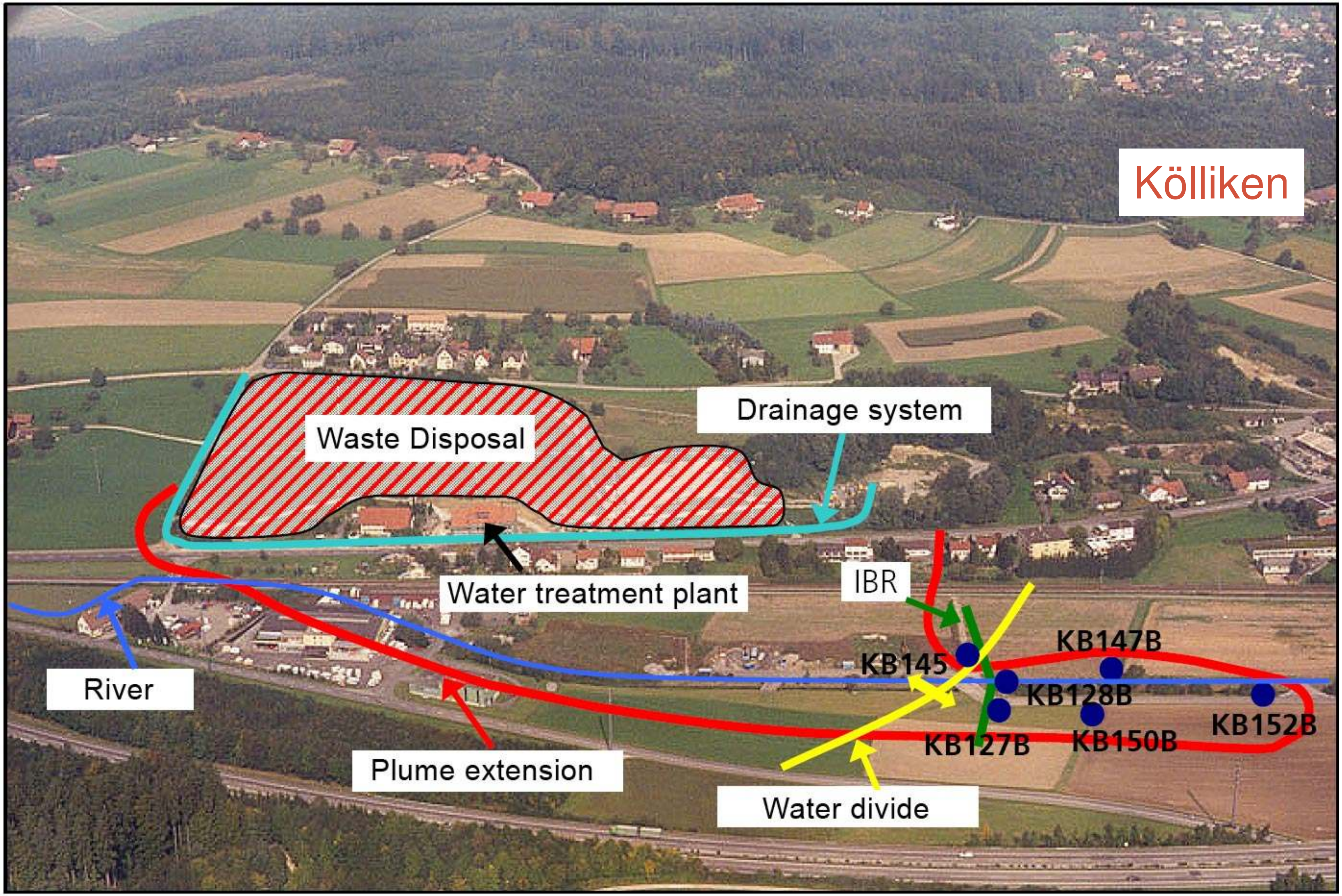
J. Kerrou, G. Mariethoz,
J. Straubhaar, A. Comunian,
G. Pirot, F. Oriani

FNSNF
SWISS NATIONAL SCIENCE FOUNDATION

MOTIVATION FOR MPS

What is our problem / limits of current approaches

Will the contamination reach drinking water supply?

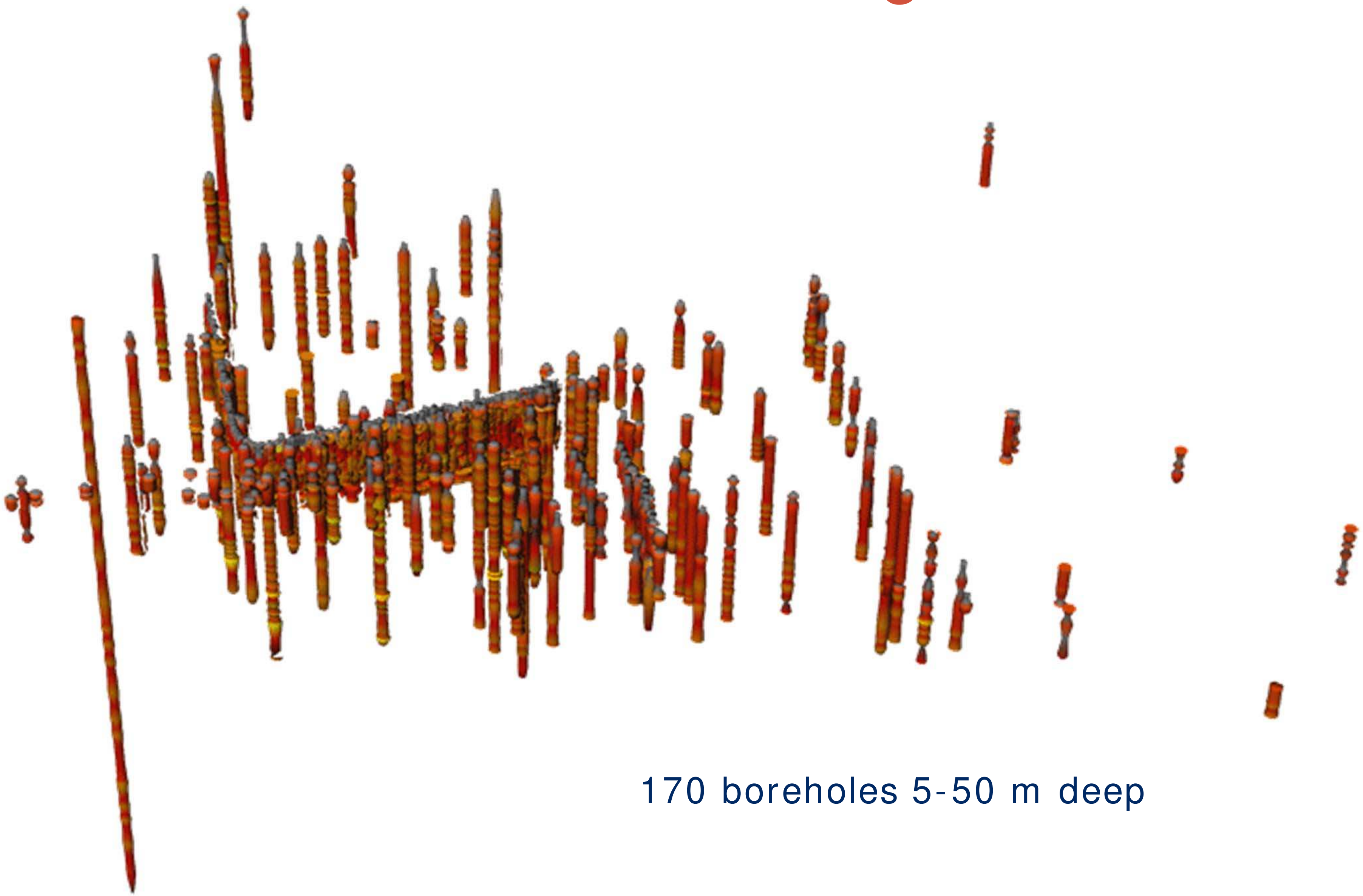


Confining building

Waste excavation

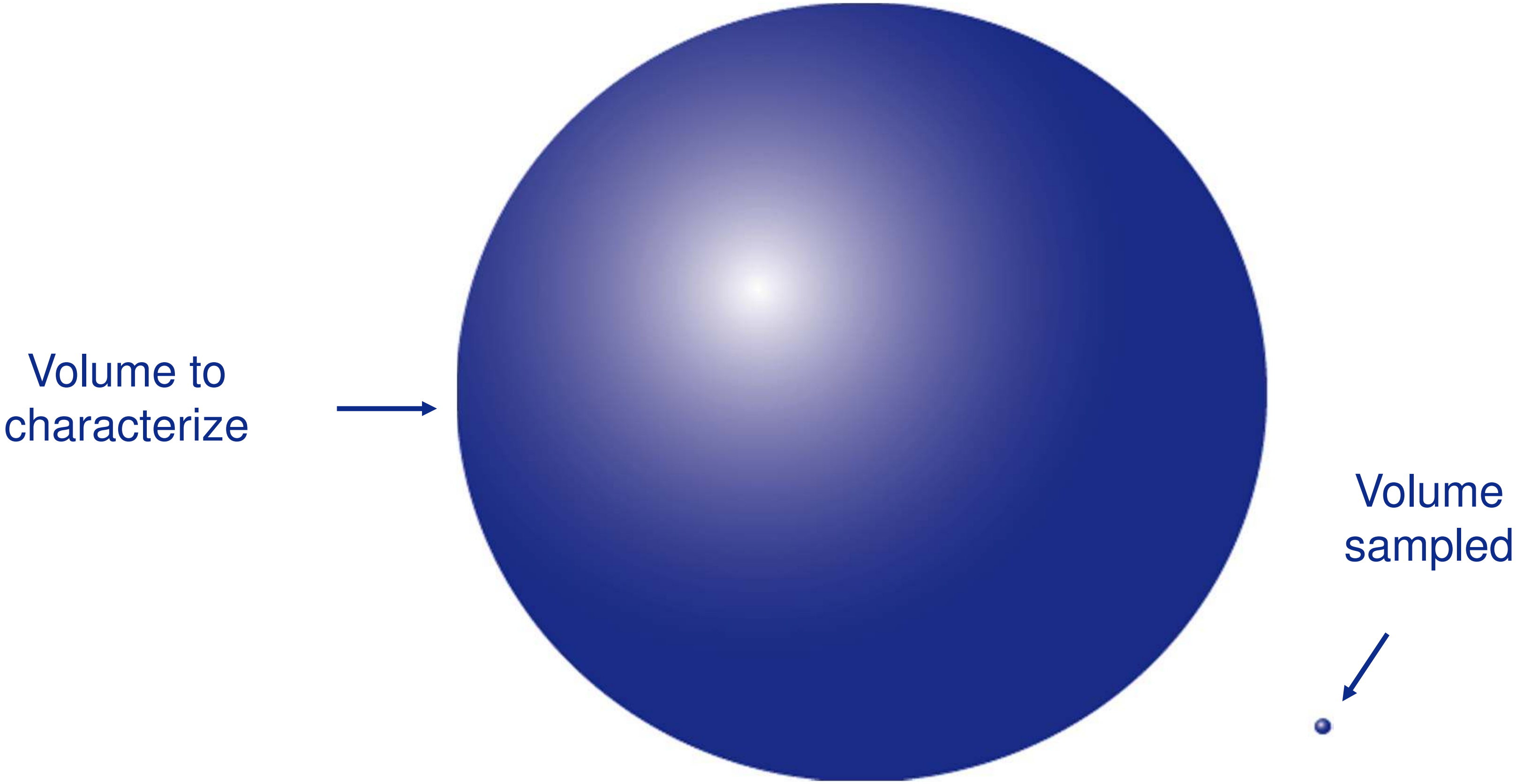
Total cost > 770 M. CHF

Extensive investigations



170 boreholes 5-50 m deep

Characterization issue



Volume to characterize : 850 x 400 x 70 = 24 millions m^3

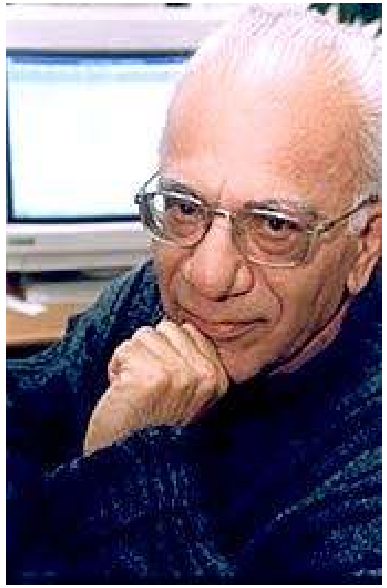
Volume sampled by the boreholes: 6700 x 0.01 x π = 200 m^3

Understanding groundwater flow

- Well established PDE / Numerical models
- Huge uncertainty due to
  - rock heterogeneity + lack of data (field of parameter)
  - badly controlled boundary conditions / source terms
- Long tradition (>30 years) to use Gaussian random fields
  - Interpolating parameters
  - Understanding the physics of heterogeneous materials
  - Estimating uncertainty
- Heavy numerical forward models (e.g.CO_2 sequestration)

- Today → how to build random fields of input parameter

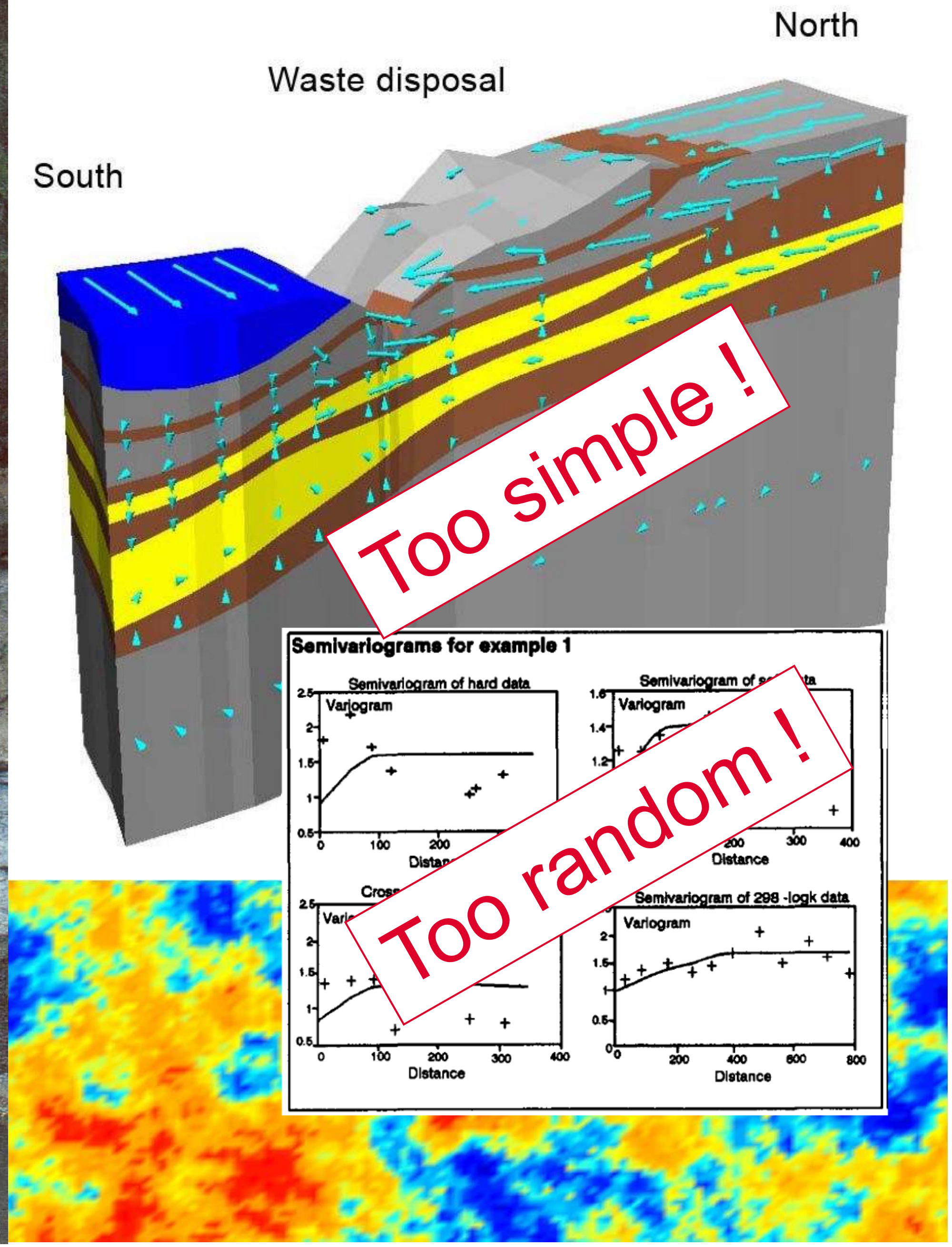
Gedeon Dagan (2002)



- “The stochastic modeling of groundwater has developed considerably ... [but it] hasn’t yet become a routine tool”
- Debate in the community
- Situation is more subtle
- Various issues
  - Education: lack of people
  - Structural issues: consulting market
  - Relevance of the models

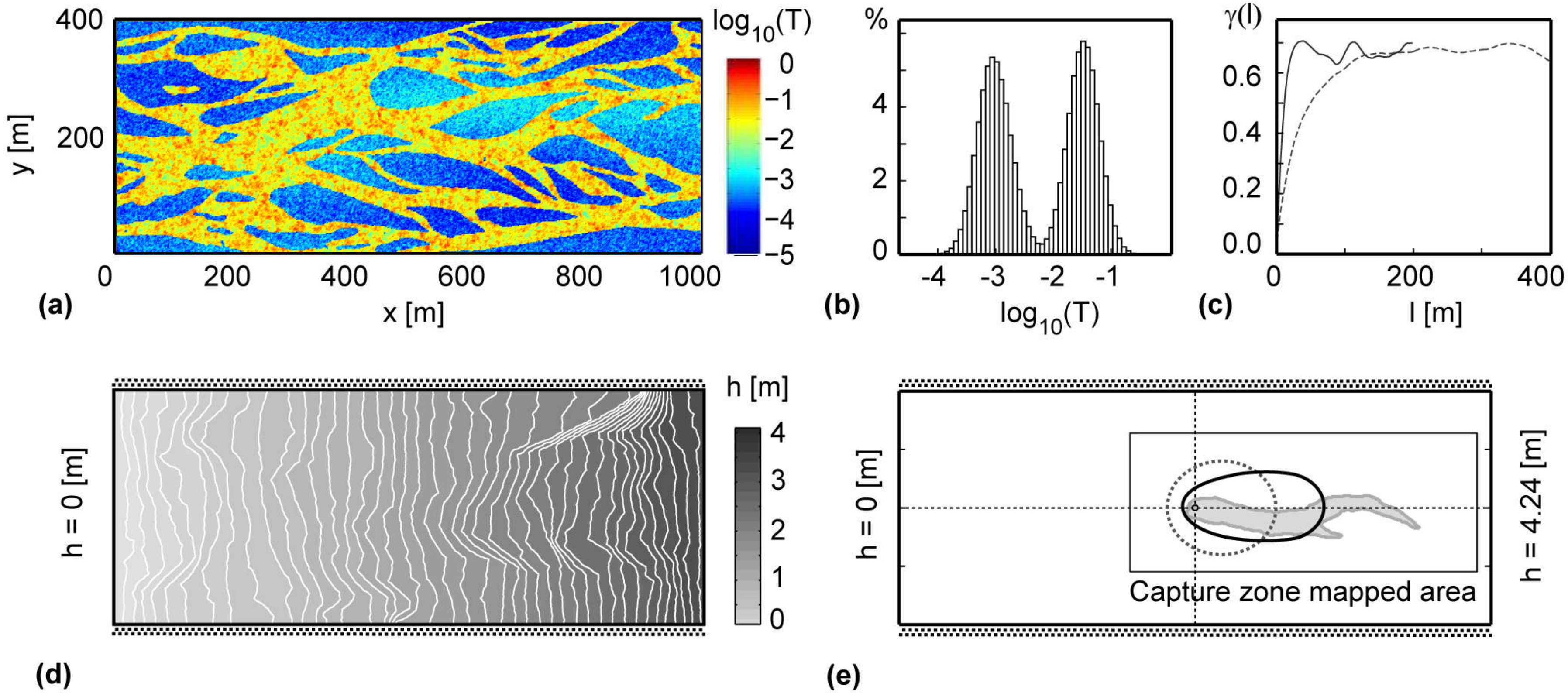
Dagan, 2002, EOS Transactions, 83(53)   Renard, 2007, Groundwater, 45(5): 531-541

One fundamental reasons



A synthetic example

Transmissivity field

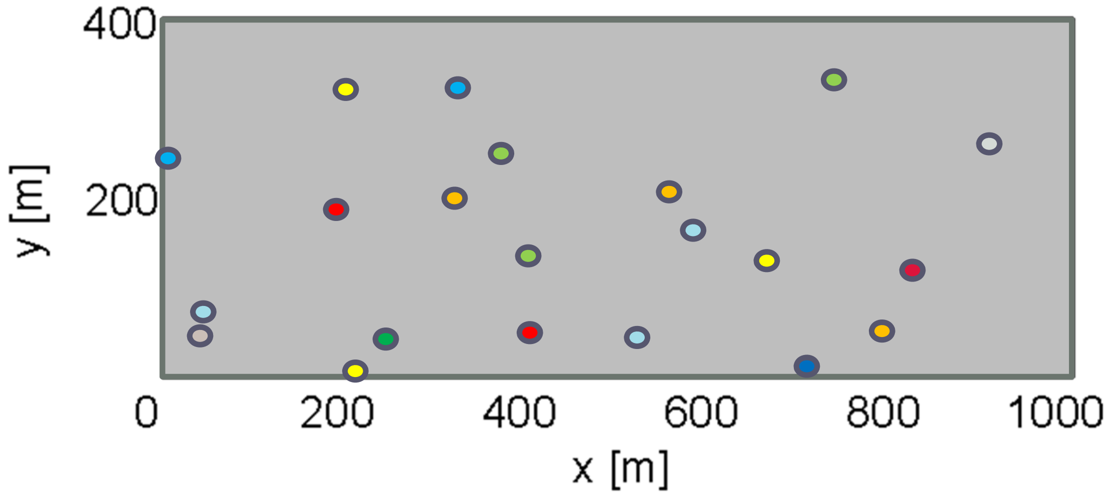


Characterization data set

Forecasts data set

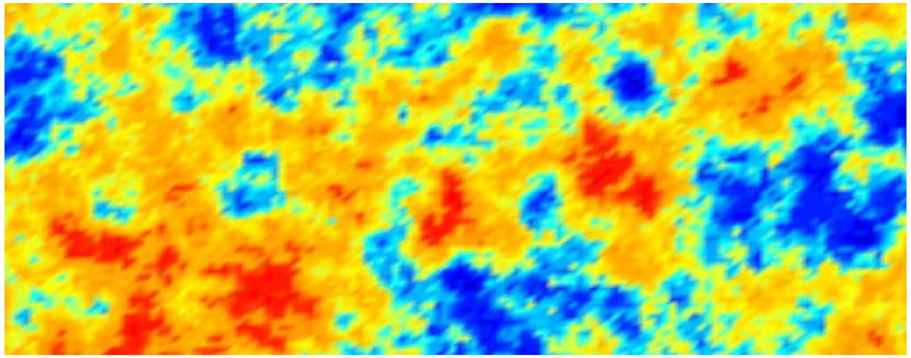
Kerrou, Renard, Hendricks-Franssen, and Lunati (2008) AWR, 31(1): 147-149

Sampling the reference



Heterogeneity characterization

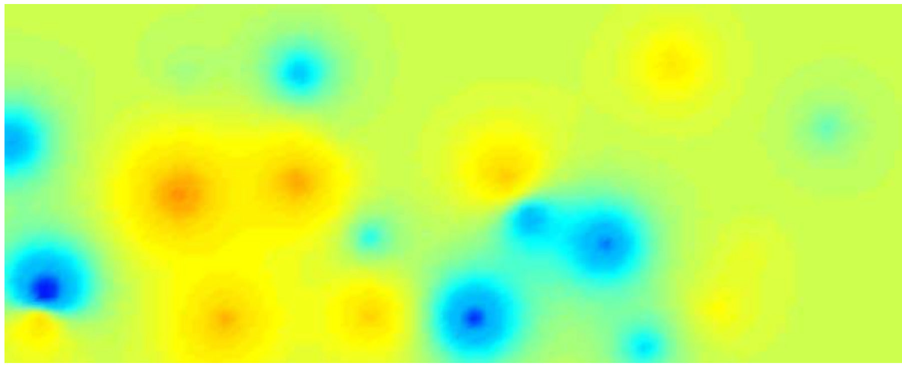
one simulation of $\log_{10}$ (T)



Simulation of N conditional to 21 values

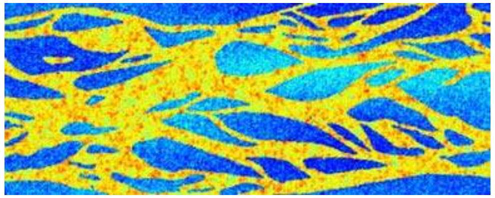
Back transform to get Y = $\log_{10}$ (T)

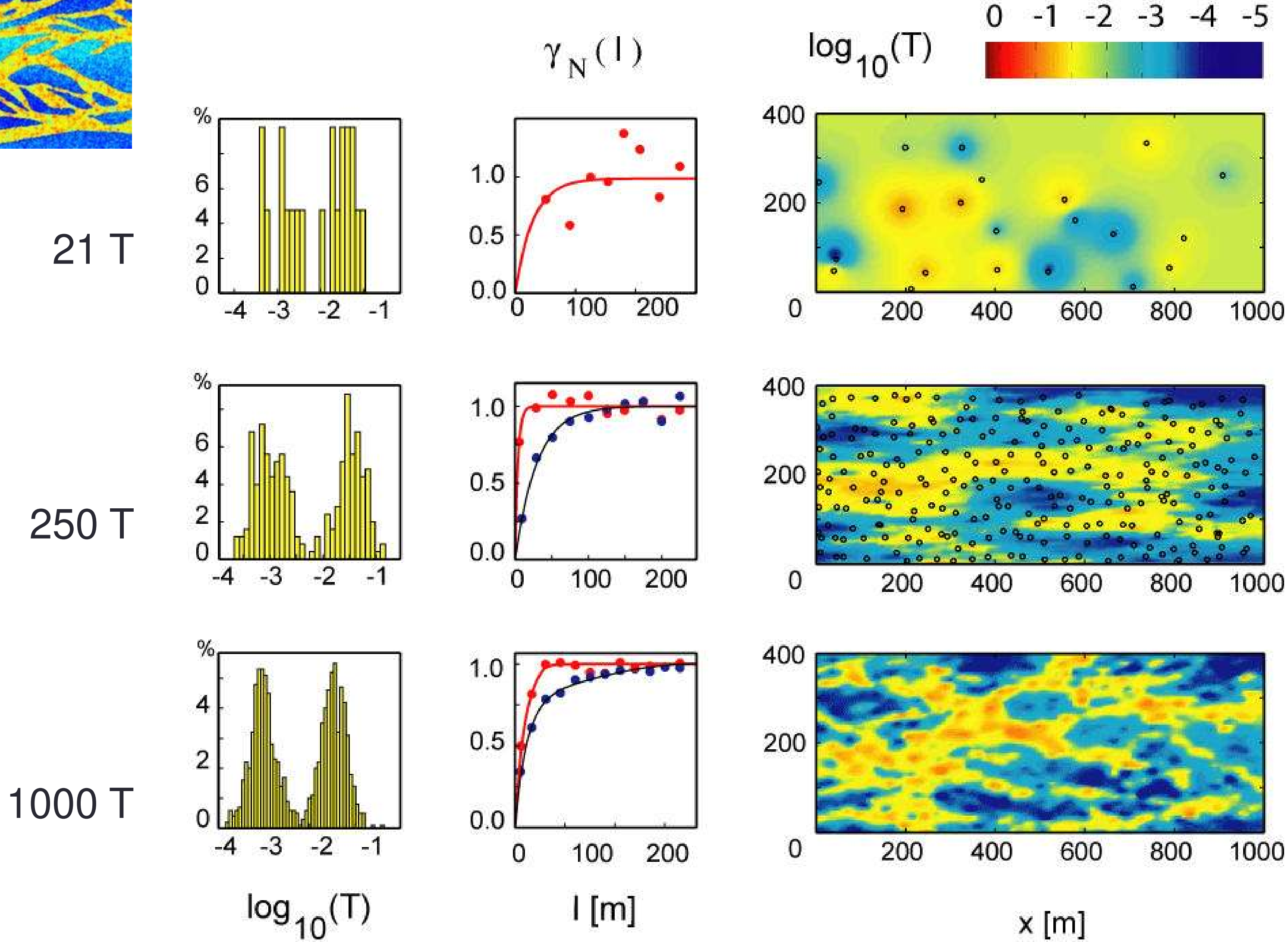
ensemble average of $\log_{10}$ (T)

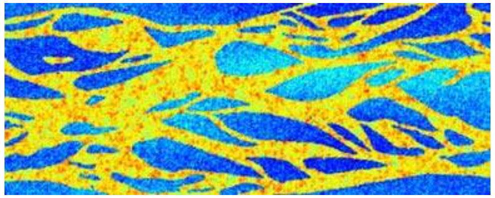


Turning bands method

Adding information T measurements



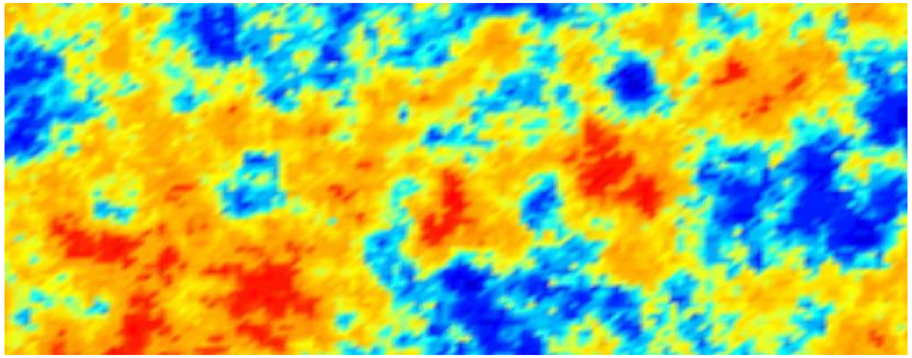




Adding head data

Sequential self calibration method
inverto code (Hendricks-Franssen, 2001)

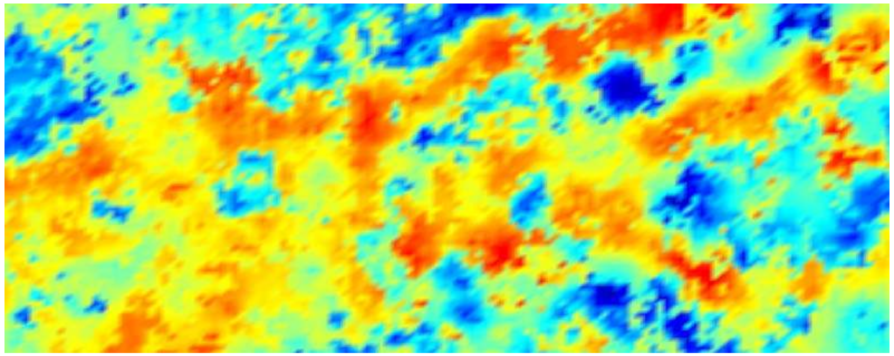
Variogram from the data – 2 master blocks per correlation length



21 T, 0 heads

21 T, 21 heads

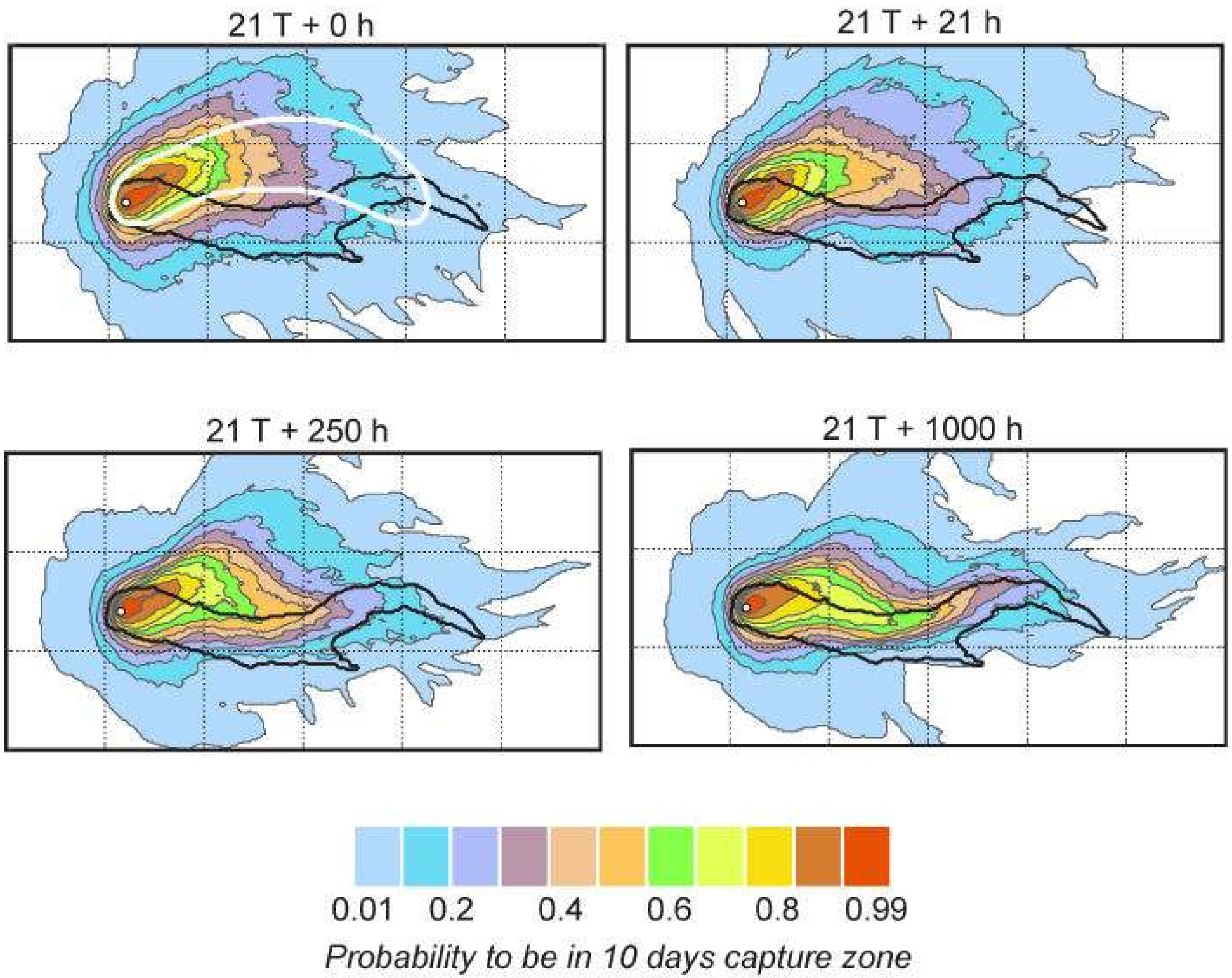
21 T, 250 heads



21 T, 1000 heads

100 simulations

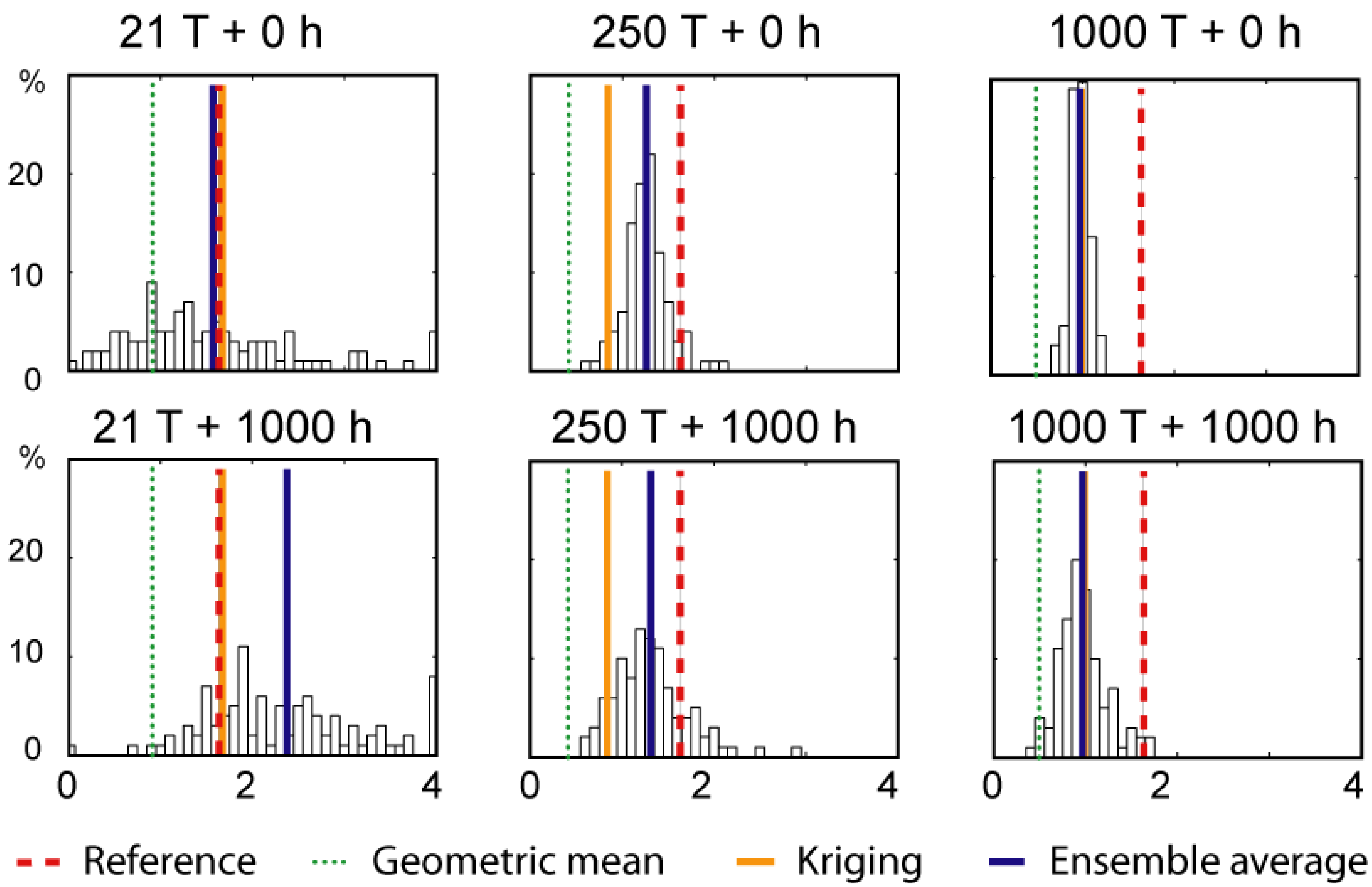
Reliability of transport forecasts



True value belongs to the ensemble forecasts

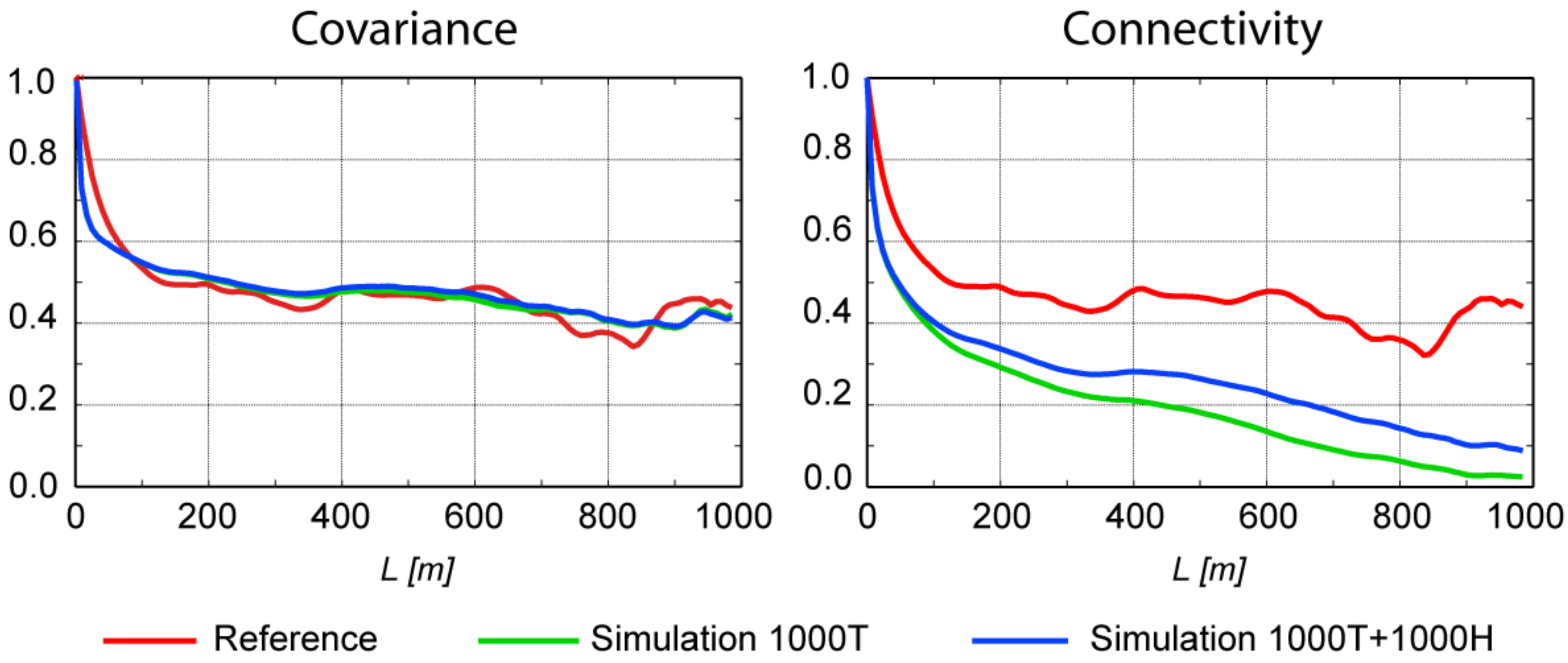
Head conditioning reduces uncertainty and increases accuracy

Limits of the multi-Gaussian approach



Adding transmissivities reduces uncertainty and increases bias
The bias is partly compensated by conditioning to head

Multi-Gaussian approach is insufficient



Covariances are well reproduced by the simulations
Connectivity are not
Conditioning to head improves connectivity

Kerrou, Renard, Hendricks-Franssen, and Lunati (2008) AWR, 31(1): 147-149

Same observations with field data

An Evaluation of Conditioning Data for Solute Transport Prediction

by Timothy D. Scheibe[1,2] and Yi-Ju Chien[1]

The results show that conditioning to a large number of small-scale measurements does not significantly improve model predictions, and may lead to biased or overly confident predictions.

However, conditioning to geophysical interpretations with larger spatial support significantly improves the accuracy and precision of model predictions.

all cases, the effects of model error appear to be significant in relation to parameter uncertainty.

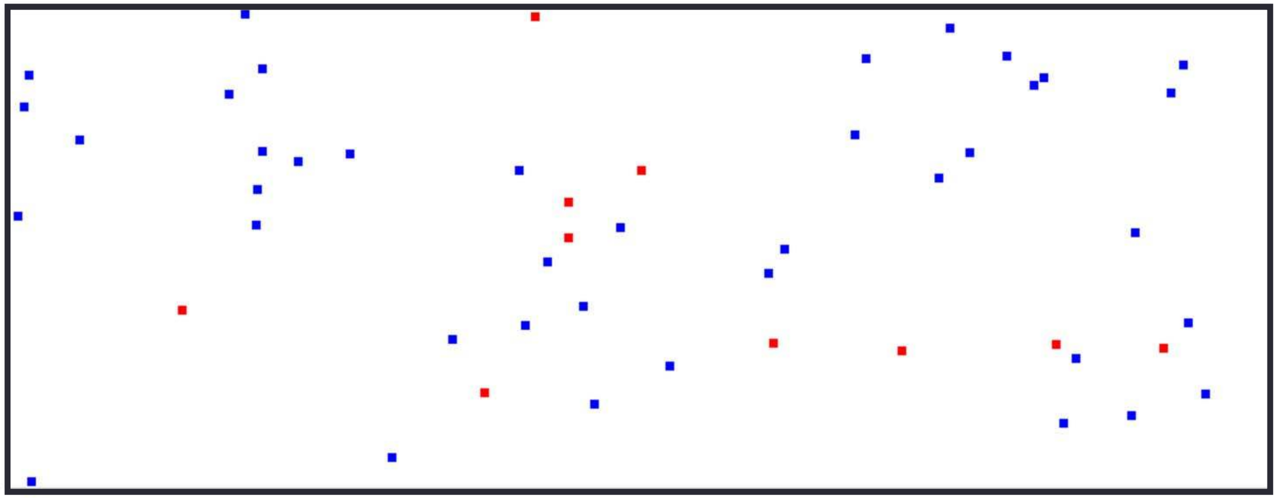
Scheibe and Chien (2003) Groundwater 41(2): 128-241

MULTIPLE POINT STATISTICS

What is it? / Principle of the method

Data set

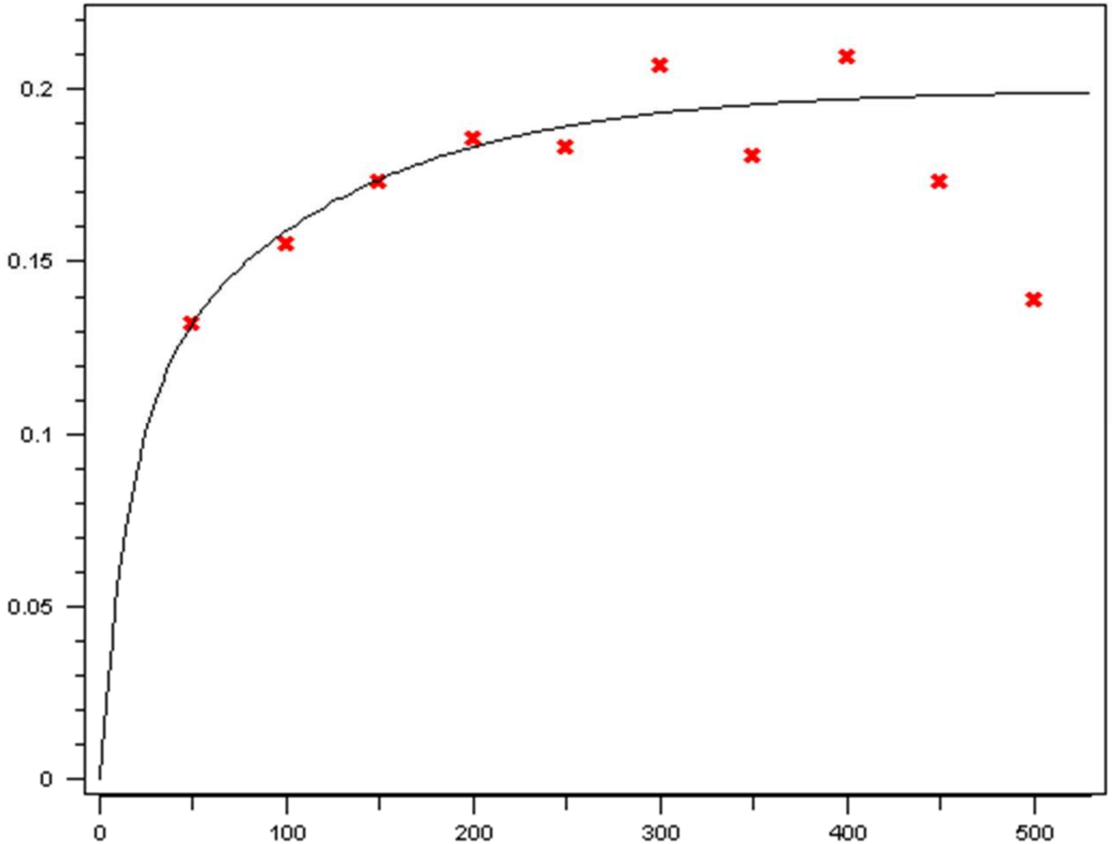
Map of geological samples



■ (blue)	$I(x) = 0$	$\Longleftrightarrow$	$x \in$ Clay
■ (red)	$I(x) = 1$	$\Longleftrightarrow$	$x \in$ Sand

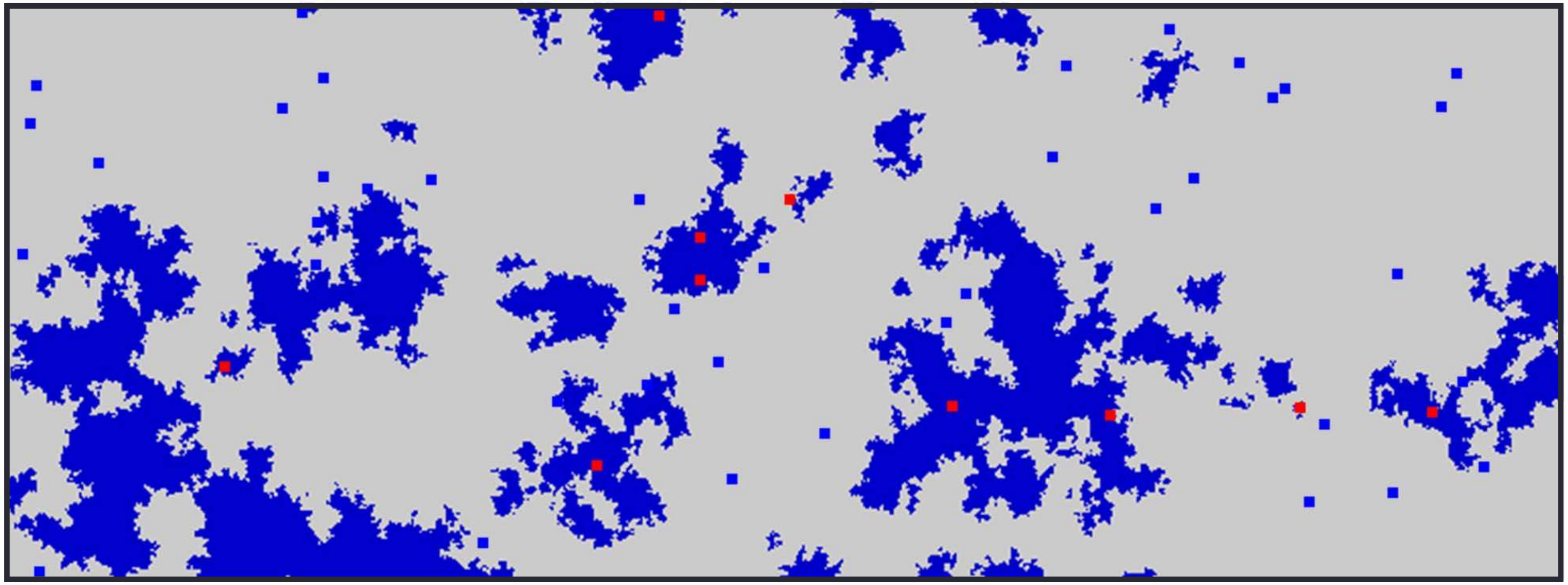
Indicator variogram

$$\gamma_i(h) = \frac{1}{2}E\left[\left(I_i(x+h) - I_i(h)\right)^2\right]$$



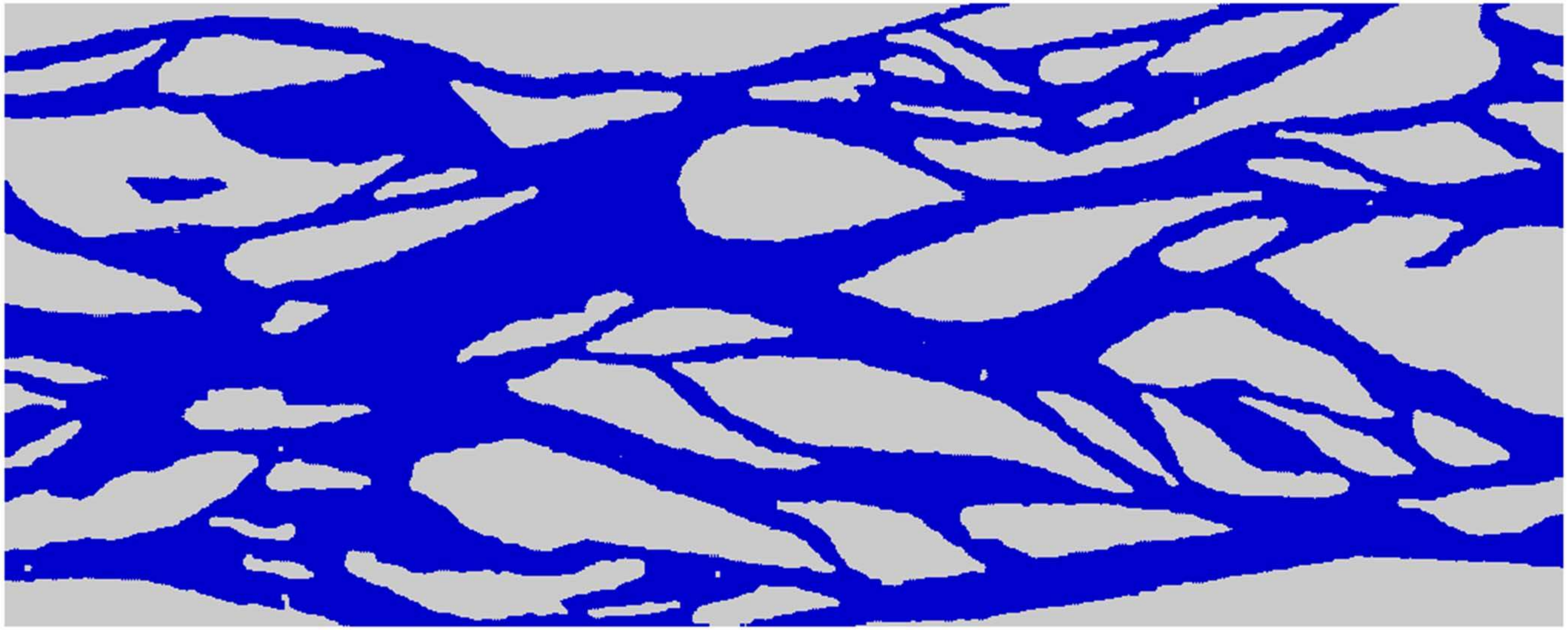
h = lag distance

Sequential indicator simulation



Honors the variogram and proportions of the data

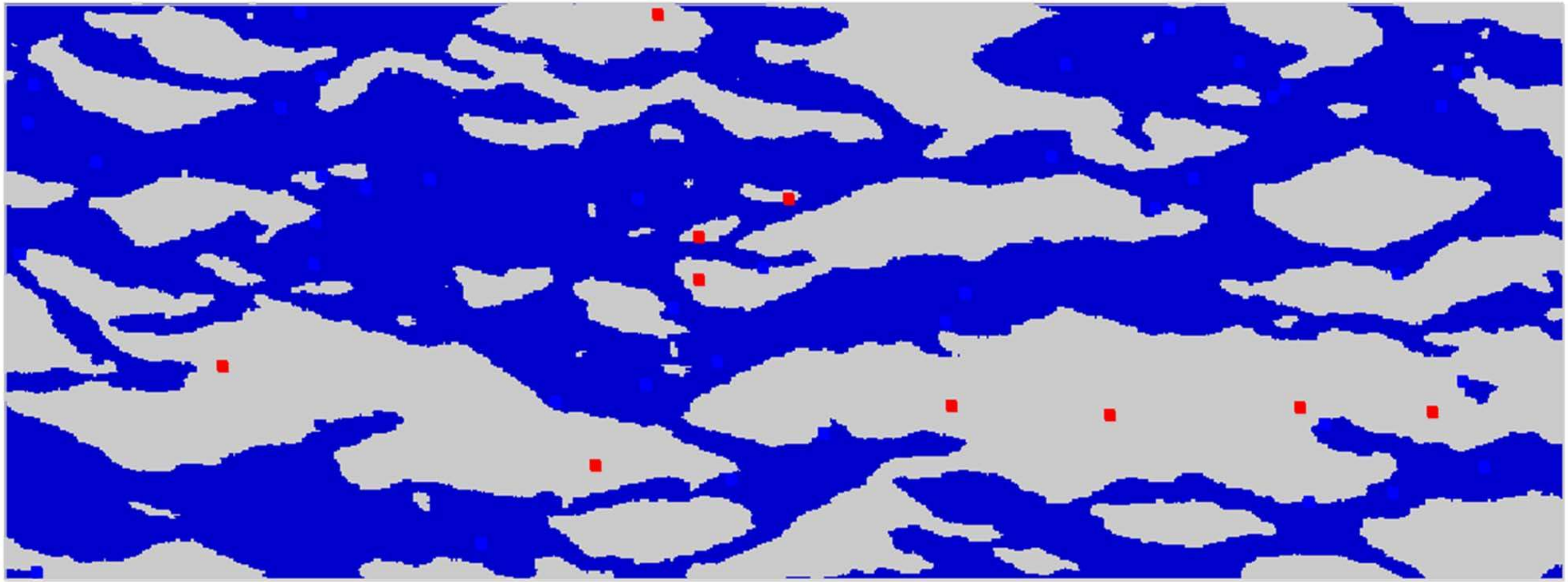
Braided chanel



Ohau river, New Zealand.

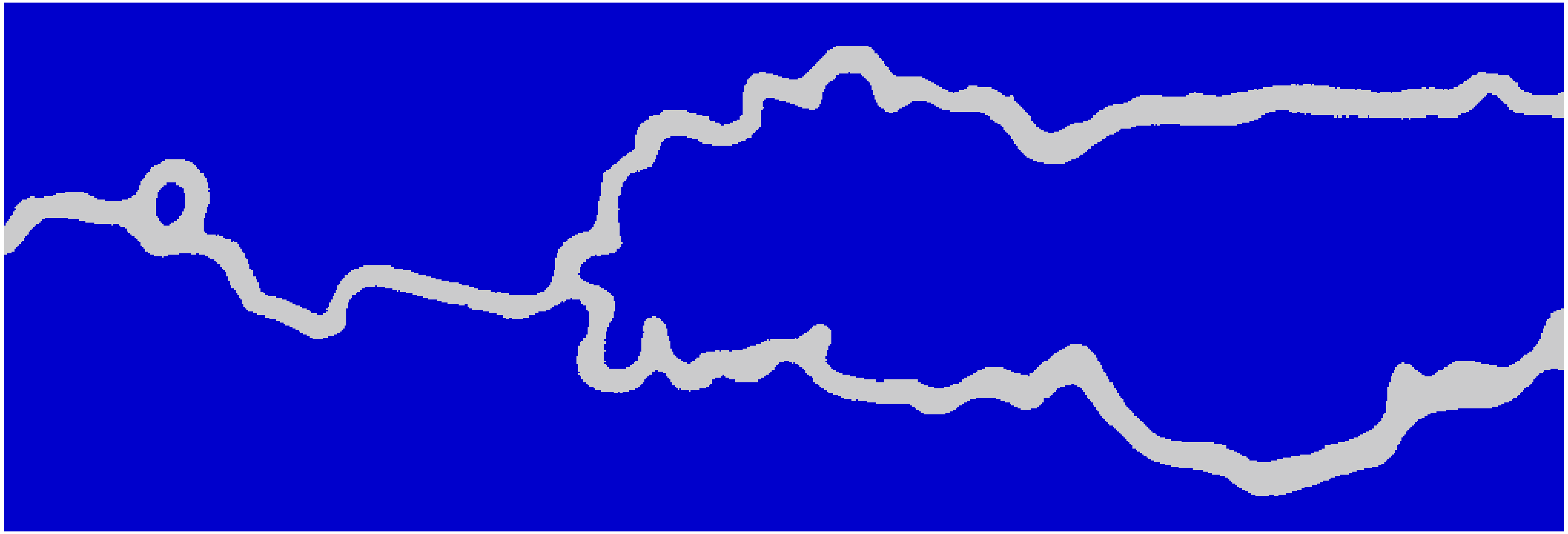
Renard (2007) Groundwater, 45(5): 531-541

Multiple-points simulation



Renard (2007) Groundwater, 45(5): 531-541

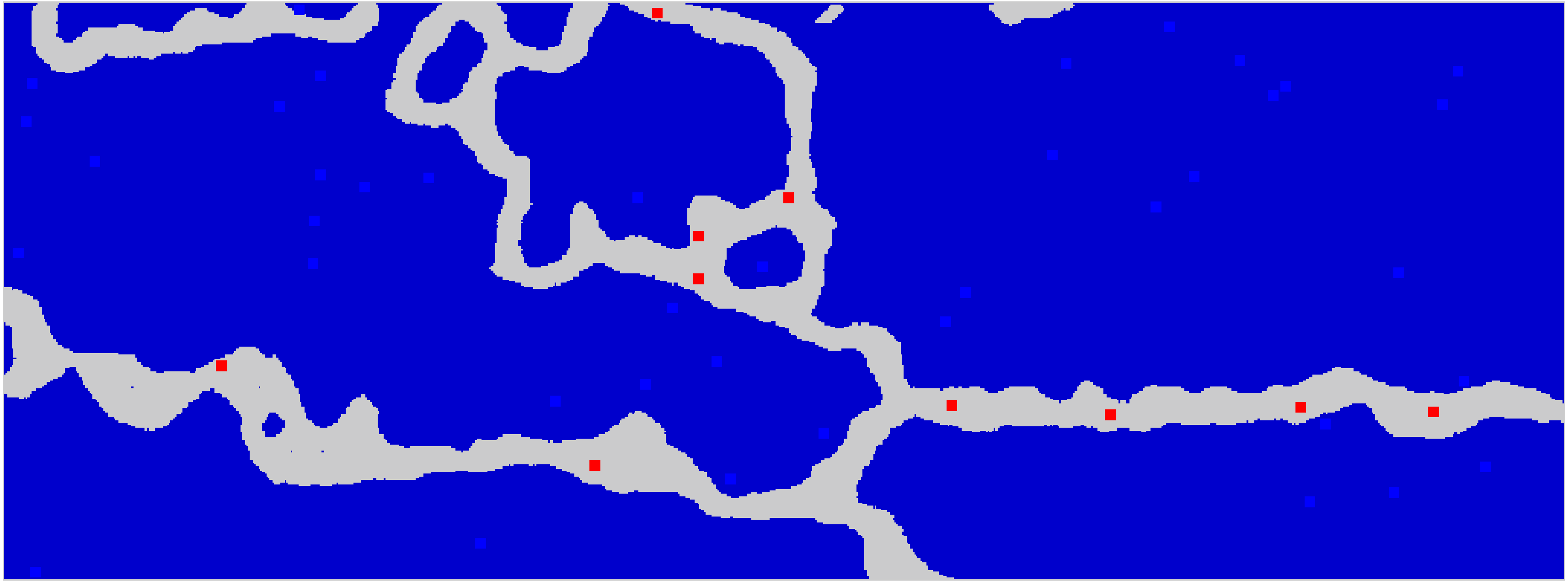
Meanders



Citronelle oil field, Alabama

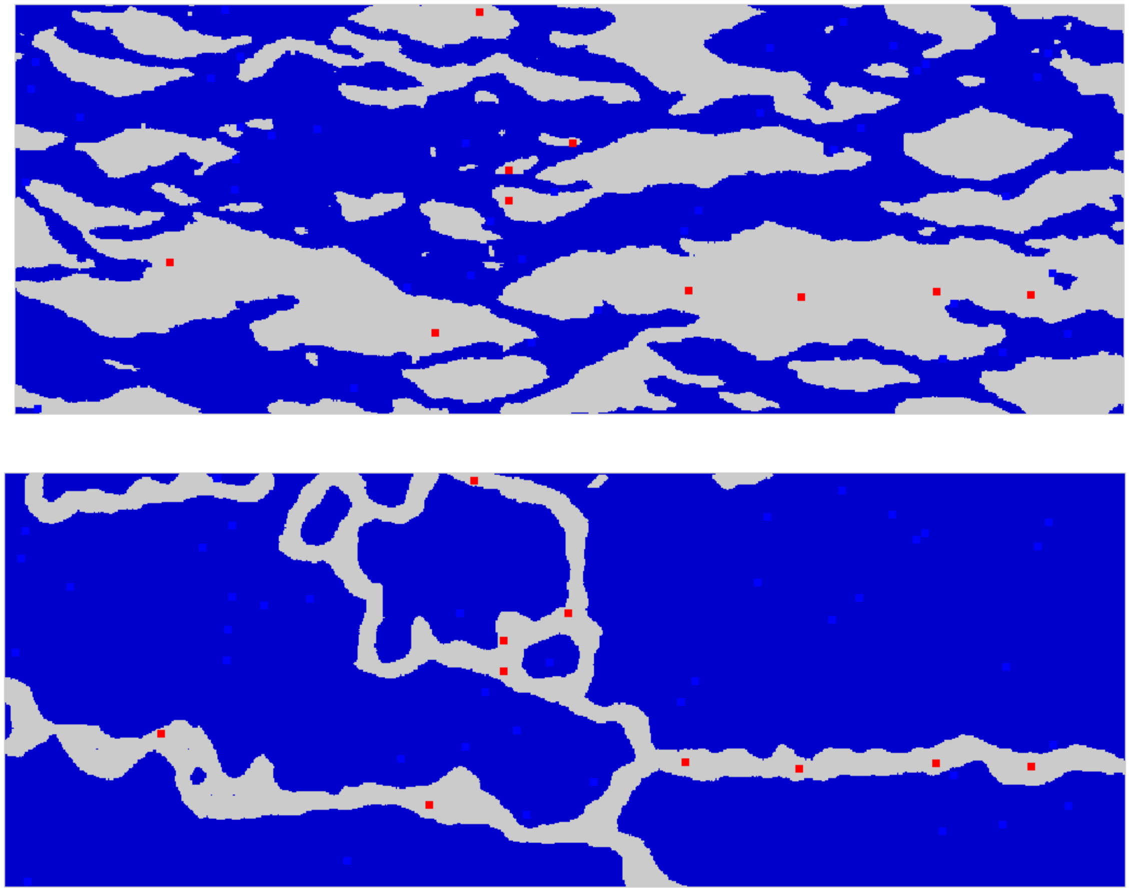
Renard (2007) Groundwater, 45(5): 531-541

Multiple-points simulation



Renard (2007) Groundwater, 45(5): 531-541

Importance of the conceptual model



Renard (2007) Groundwater, 45(5): 531-541

3 innovations

- Field data are not sufficient:

  Training Image (TI)

- Two point statistics are not sufficient:

  Multiple-point statistics (MPS)

- Analytical statistical model not tractable:

  non parametric approach

Guardiano and Srivastava (1993) in Soares (Ed) Geostatistics Troia 1992:133-144.

Principle of the method

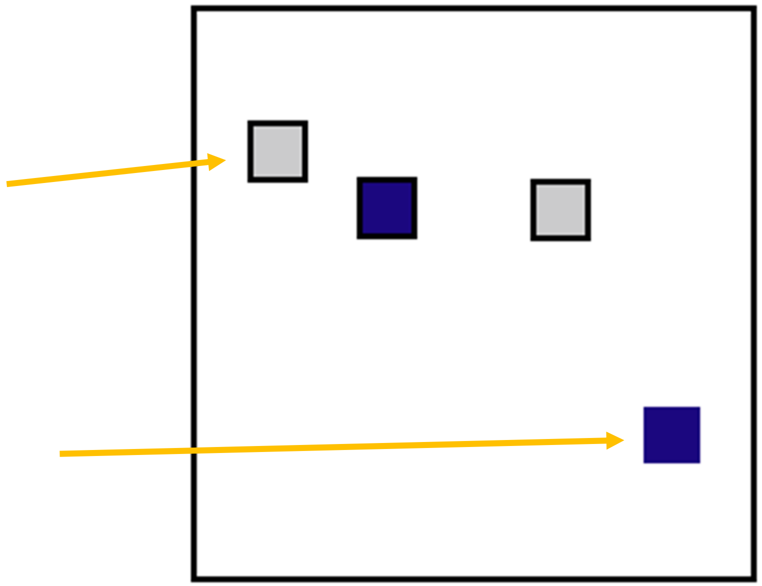
Domain to model

Data: geological observations

Indicator variable $I(x)$

$I(x) = 1$ Sand

$I(x) = 0$ Clay



Principle of the method

Sequential simulation method

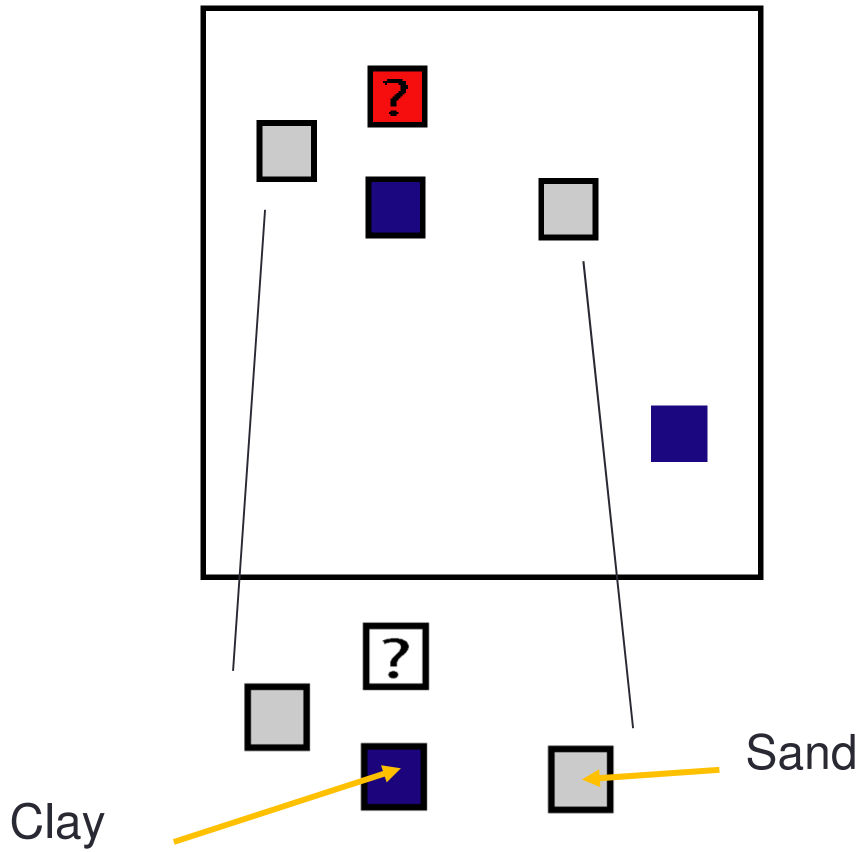
Choice of a location *x*

We want to find I(x)

Find the n closest nodes

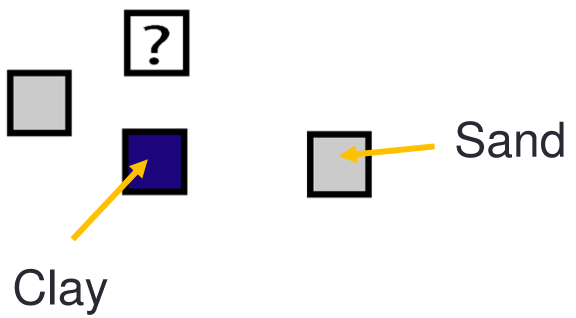
Find the corresponding values of I

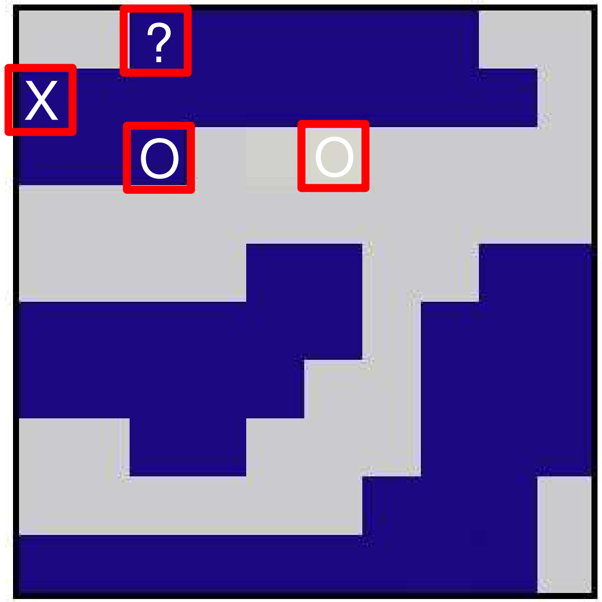
data event

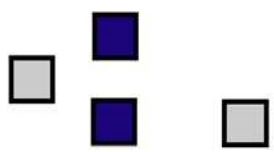
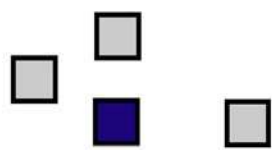


Analysis of the Training image

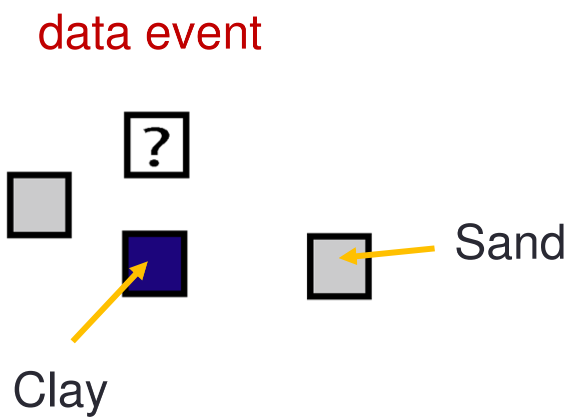
data event

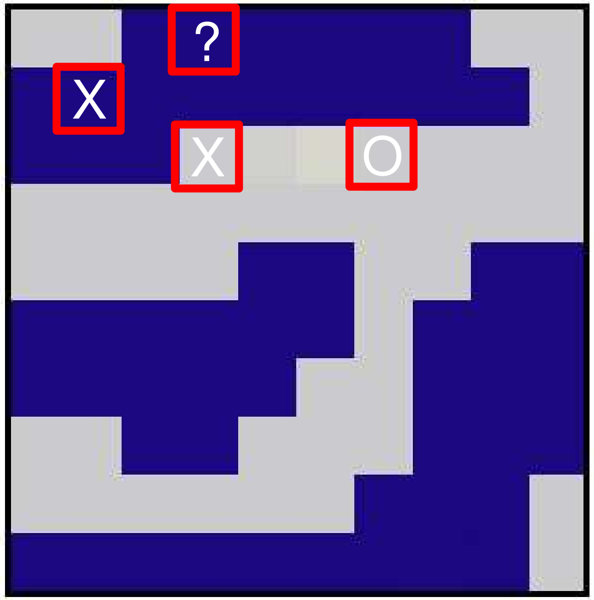


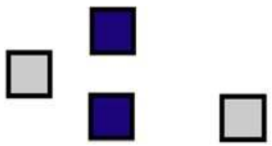
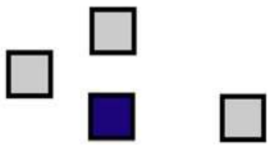


Event		
Counter	0	0

Analysis of the Training image

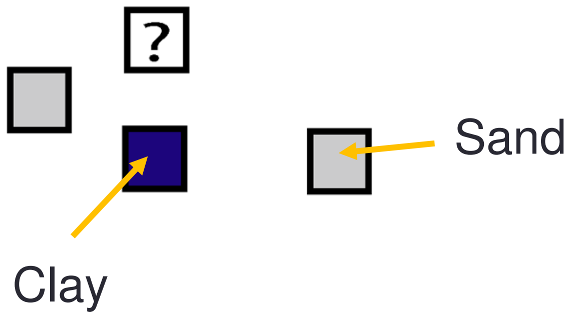


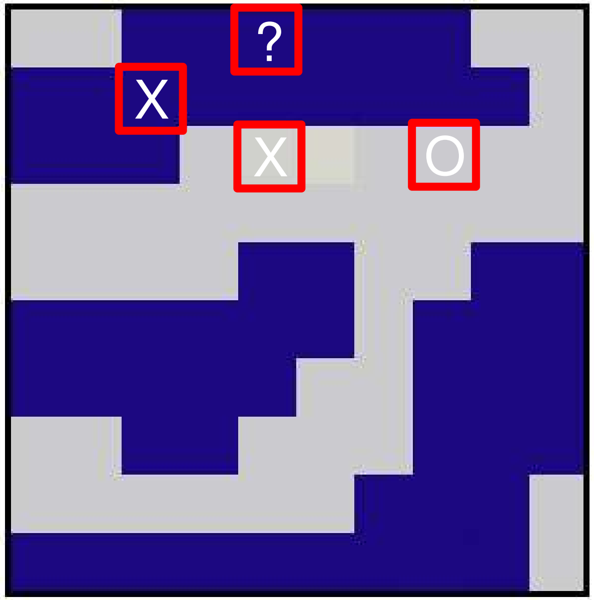


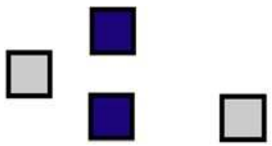
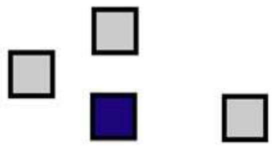
Event		
Counter	0	0

Analysis of the Training image

data event

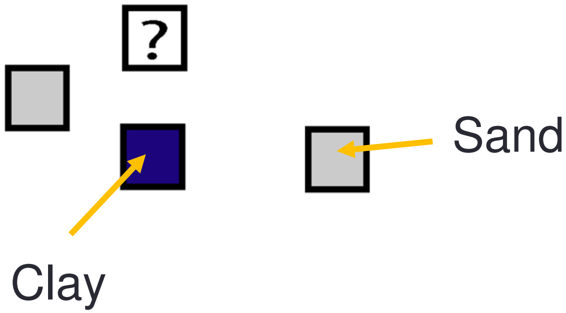


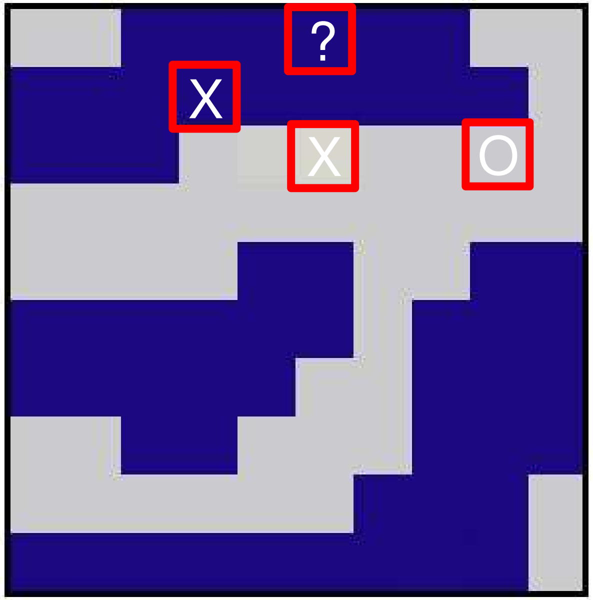


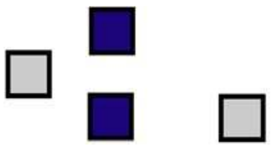
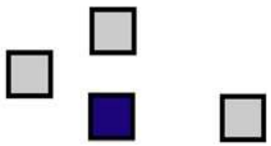
Event		
Counter	0	0

Analysis of the Training image

data event

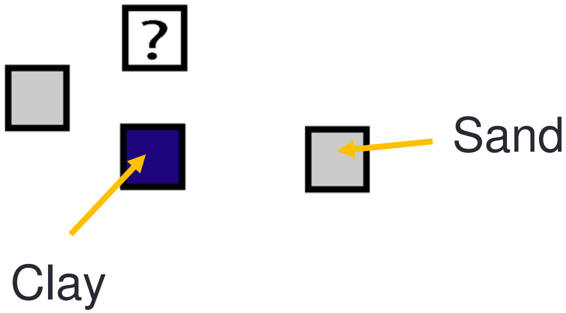


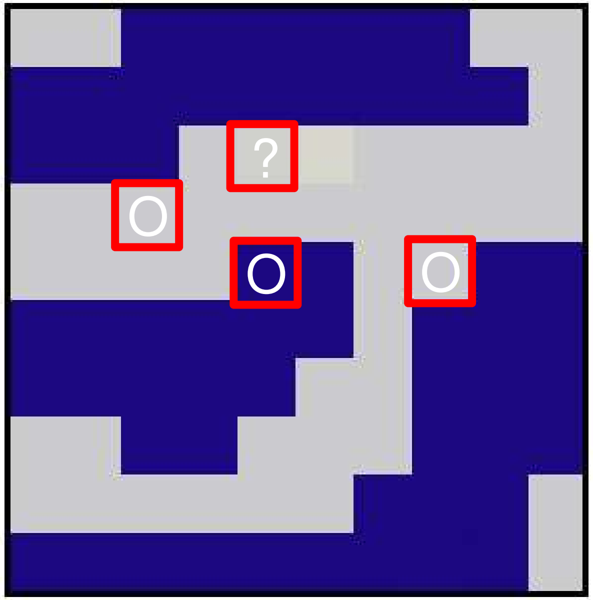


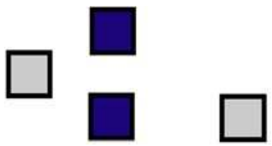
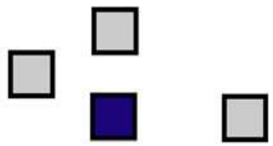
Event		
Counter	0	0

Analysis of the Training image

data event

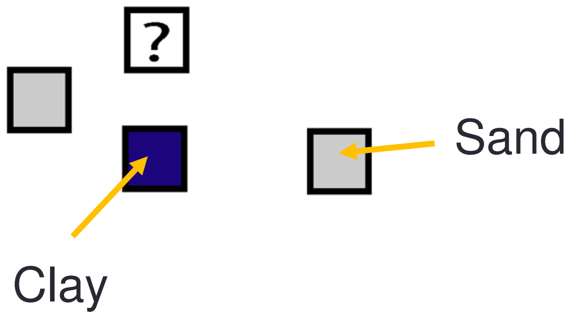


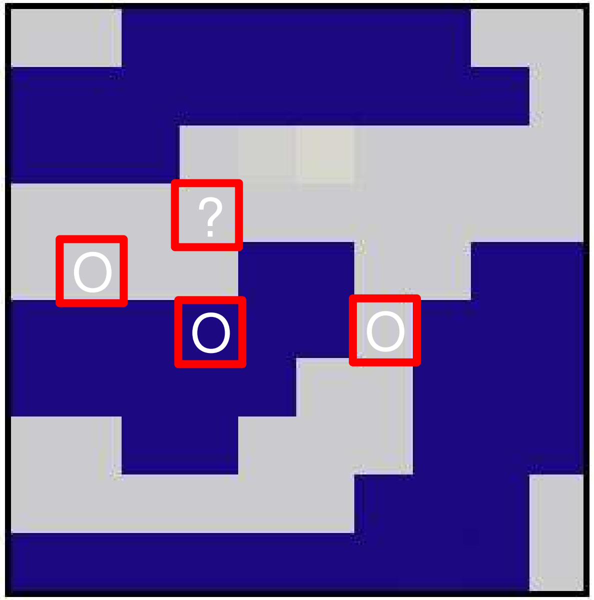


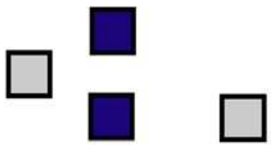
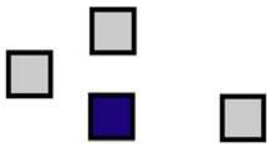
Event		
Counter	0	1

Analysis of the Training image

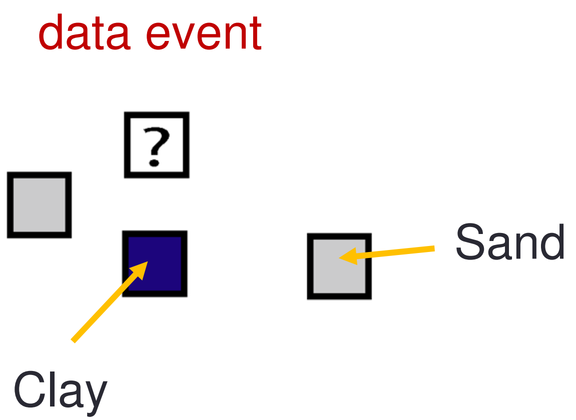
data event

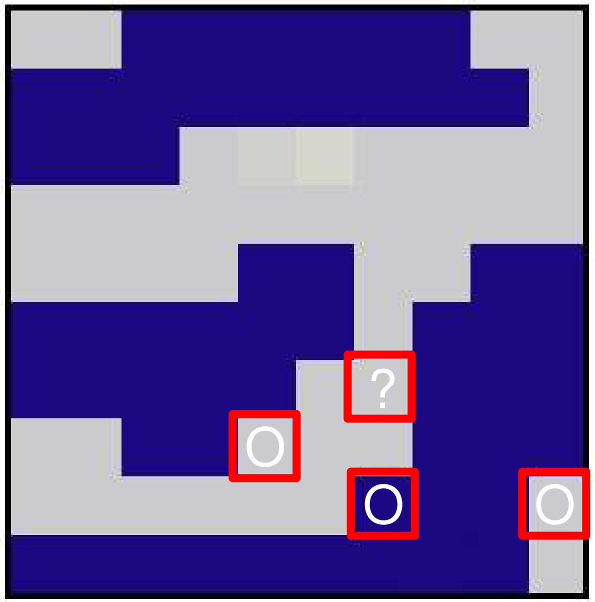


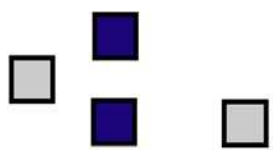
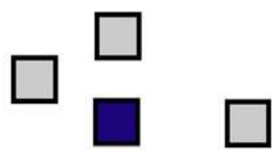


Event		
Counter	0	2

Analysis of the Training image

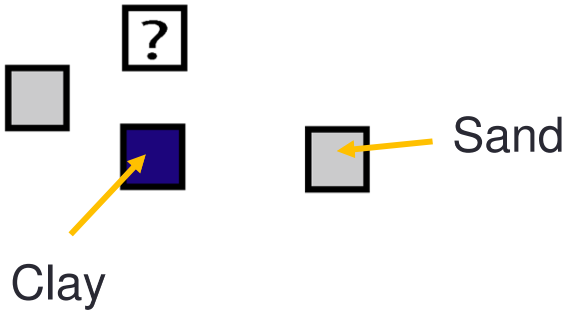


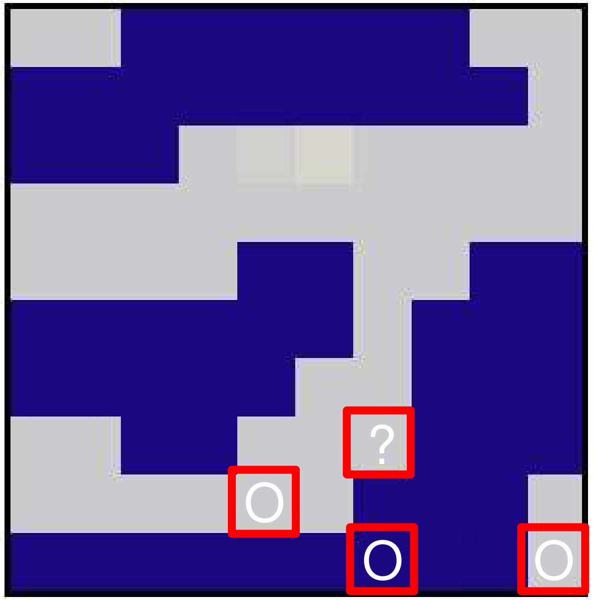


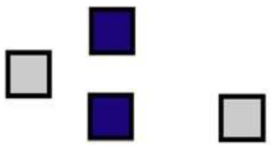
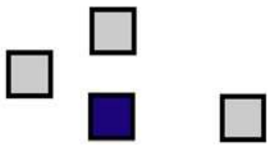
Event		
Counter	0	3

Analysis of the Training image

data event

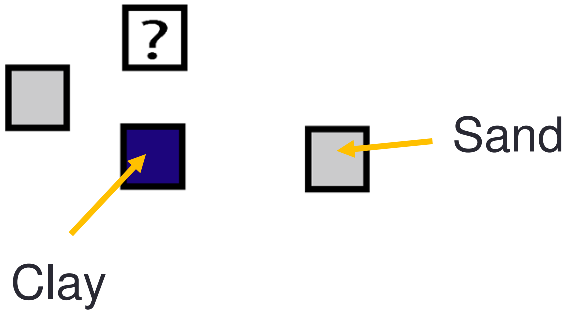


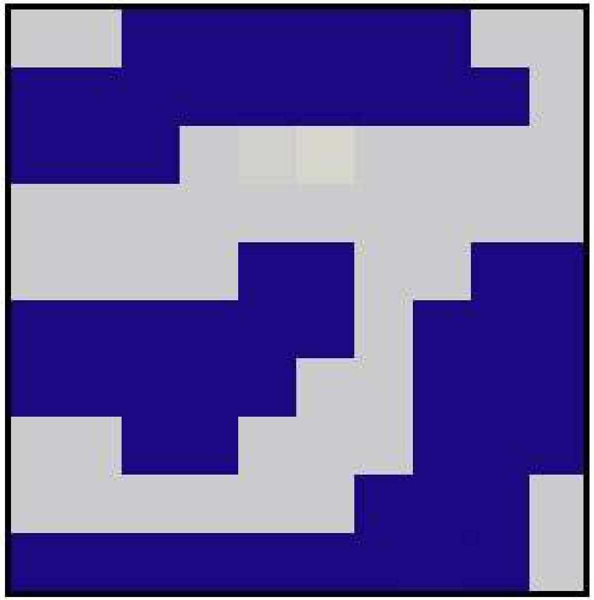


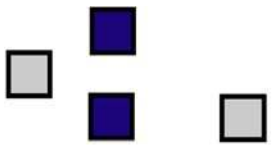
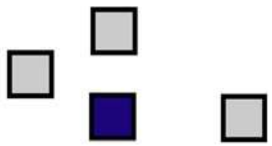
Event		
Counter	0	4

Analysis of the Training image

data event

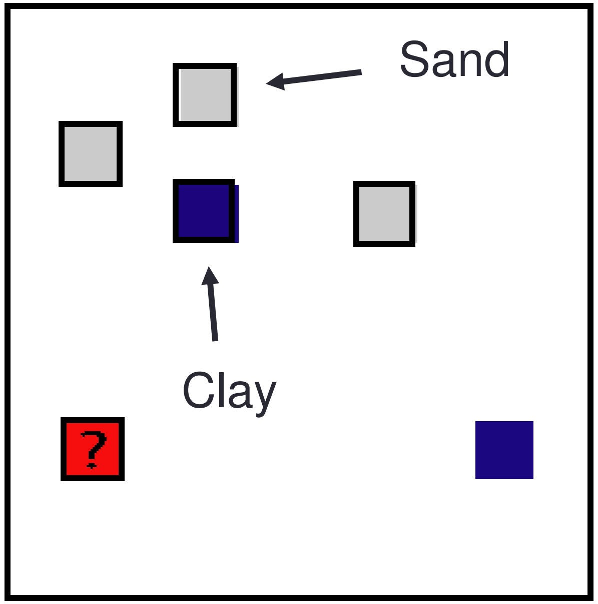




Event			Total
Counter	0	4	4
Probability	0 / 4 = 0	4 / 4 = 1	

Sampling the cpdf

I(x) is drawn from the conditional distribution



Another point is randomly selected, simulated, and so on until the whole domain is filled

Technical difficulties

- Scanning the TI for every pixel is inefficient

- Solutions
  - Analyzing the TI and storing the events within a predefined neighborhood (limited dimension)
    snesim / impala
    Implies additional algorithmic tricks (multigrids, subgrids, data migration, etc)

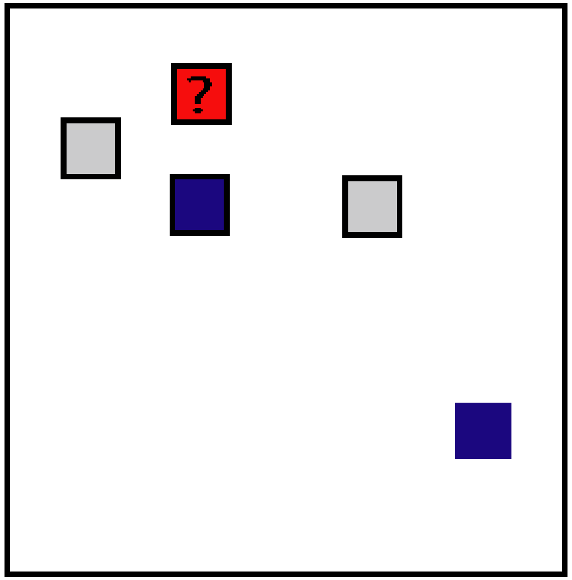
  - Directly sample the training image

Direct sampling (Deesse)

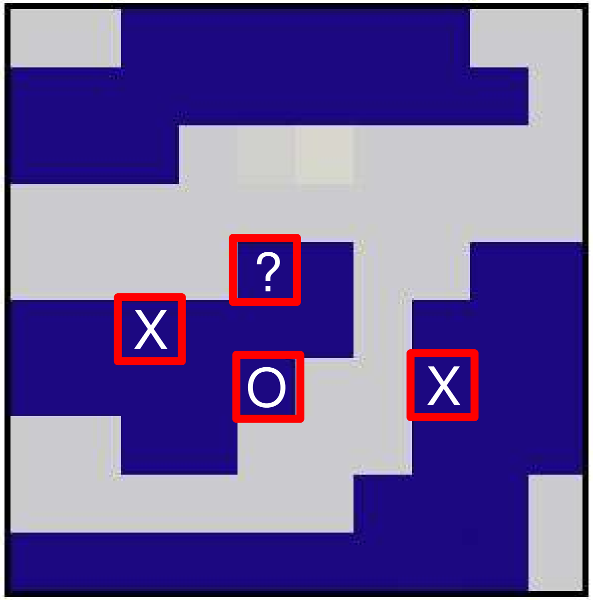
- Does not use a catalog of patterns
- Allows to extend the technique to continuous and multiple variables
- Allows to get rid of the fixed template size and multigrids
- Our main tool today

Direct sampling

Simulation grid



Training image



Conditioning data event

$$\mathbf{d}_n(\mathbf{x}) = \{Z(\mathbf{x}+\mathbf{h}_1), \cdots, Z(\mathbf{x}+\mathbf{h}_n)\}$$

Candidate data event

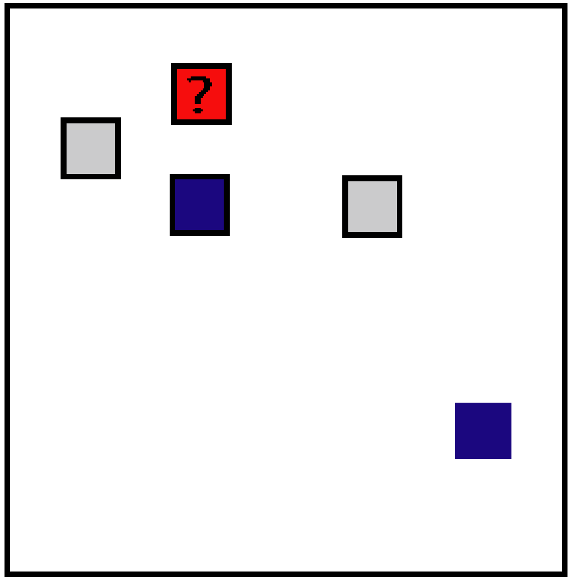
$$\mathbf{d}_n(\mathbf{y}) = \{Z(\mathbf{y}+\mathbf{h}_1), \cdots, Z(\mathbf{y}+\mathbf{h}_n)\}$$

$$d\left\{\mathbf{d}_n(\mathbf{x}), \mathbf{d}_n(\mathbf{y})\right\} = \frac{1}{n}\sum_{i=1}^{n} a_i$$

$$a_i = \begin{cases} 0 & if\ Z(\mathbf{x}+\mathbf{h}_i) = Z(\mathbf{y}+\mathbf{h}_i) \\ 1 & if\ Z(\mathbf{x}+\mathbf{h}_i) \neq Z(\mathbf{y}+\mathbf{h}_i) \end{cases}$$

Direct sampling

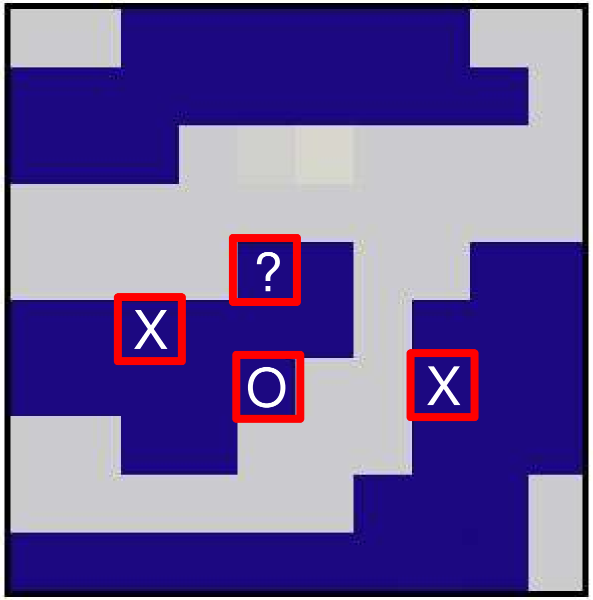
Simulation grid



Conditioning data event

$$\mathbf{d}_n(\mathbf{x}) = \{Z(\mathbf{x}+\mathbf{h}_1), \cdots, Z(\mathbf{x}+\mathbf{h}_n)\}$$

Training image

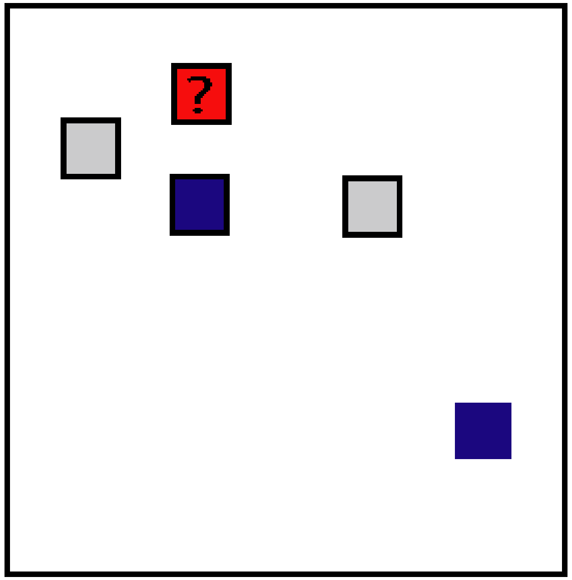


Candidate data event

$$\mathbf{d}_n(\mathbf{y}) = \{Z(\mathbf{y}+\mathbf{h}_1), \cdots, Z(\mathbf{y}+\mathbf{h}_n)\}$$

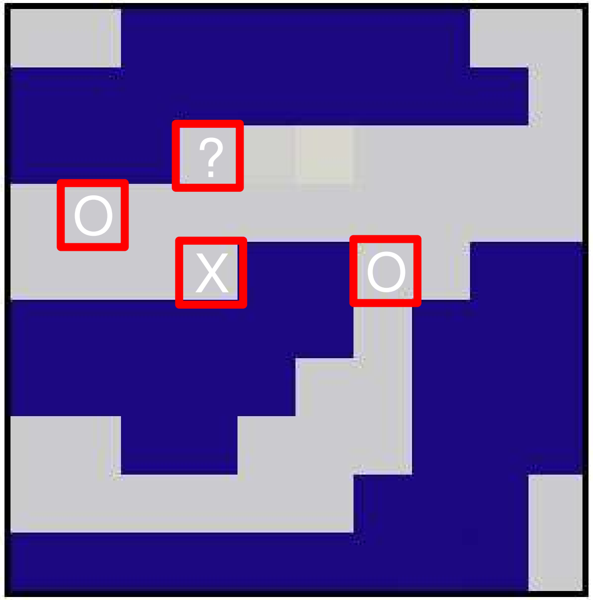
$$d\{\mathbf{d}_n(\mathbf{x}), \mathbf{d}_n(\mathbf{y})\} = \frac{2}{3}$$

Direct sampling



Conditioning data event

$$\mathbf{d}_n(\mathbf{x}) = \{Z(\mathbf{x}+\mathbf{h}_1), \cdots, Z(\mathbf{x}+\mathbf{h}_n)\}$$



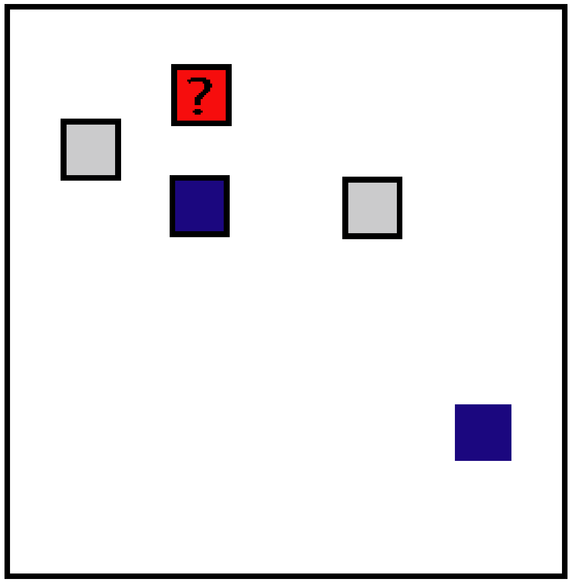
Candidate data event

$$\mathbf{d}_n(\mathbf{y}) = \{Z(\mathbf{y}+\mathbf{h}_1), \cdots, Z(\mathbf{y}+\mathbf{h}_n)\}$$

$$d\{\mathbf{d}_n(\mathbf{x}), \mathbf{d}_n(\mathbf{y})\} = \frac{1}{3}$$

Direct sampling

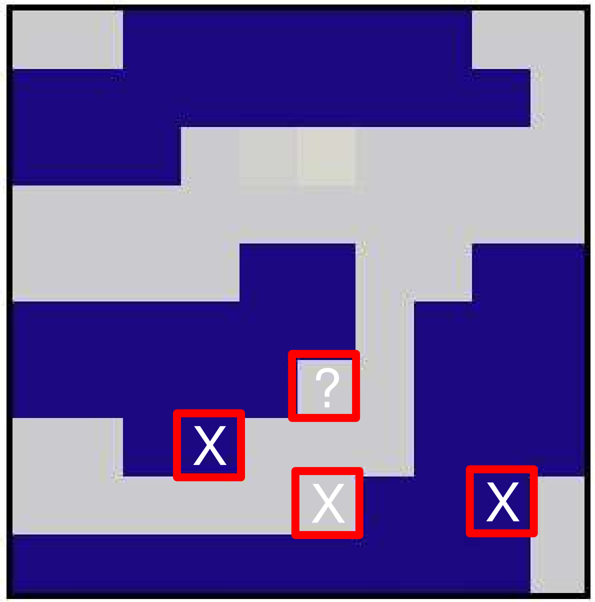
Simulation grid



Conditioning data event

$$\mathbf{d}_n(\mathbf{x}) = \{Z(\mathbf{x}+\mathbf{h}_1), \cdots, Z(\mathbf{x}+\mathbf{h}_n)\}$$

Training image



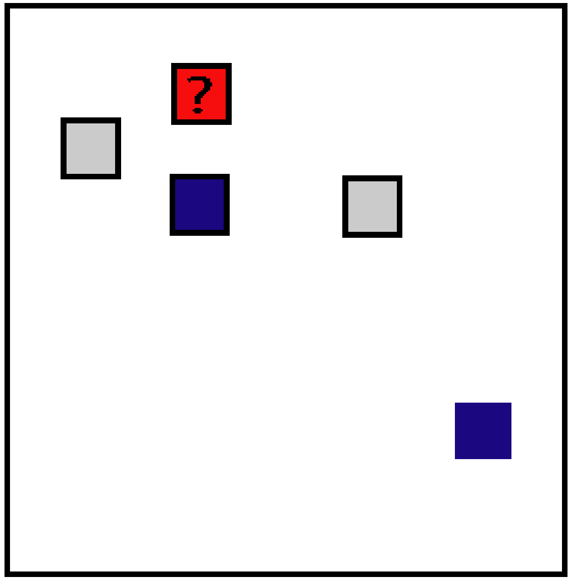
Candidate data event

$$\mathbf{d}_n(\mathbf{y}) = \{Z(\mathbf{y}+\mathbf{h}_1), \cdots, Z(\mathbf{y}+\mathbf{h}_n)\}$$

$$d\{\mathbf{d}_n(\mathbf{x}), \mathbf{d}_n(\mathbf{y})\} = \frac{3}{3}$$

Direct sampling

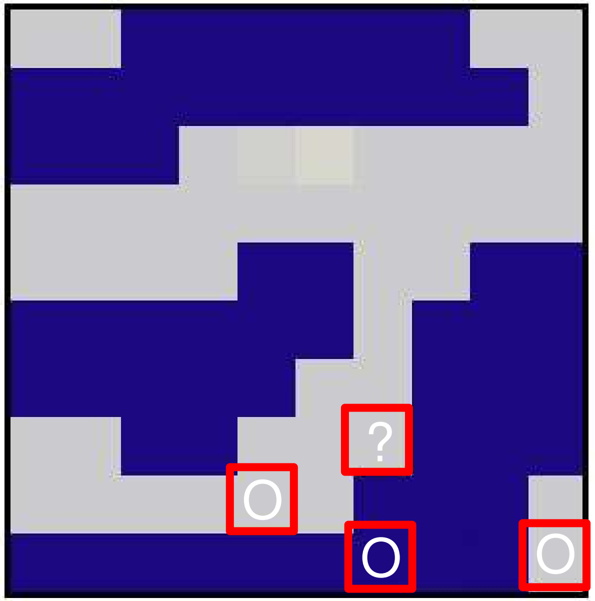
Simulation grid



Conditioning data event

$$\mathbf{d}_n(\mathbf{x}) = \{Z(\mathbf{x}+\mathbf{h}_1),\cdots,Z(\mathbf{x}+\mathbf{h}_n)\}$$

Training image



Candidate data event

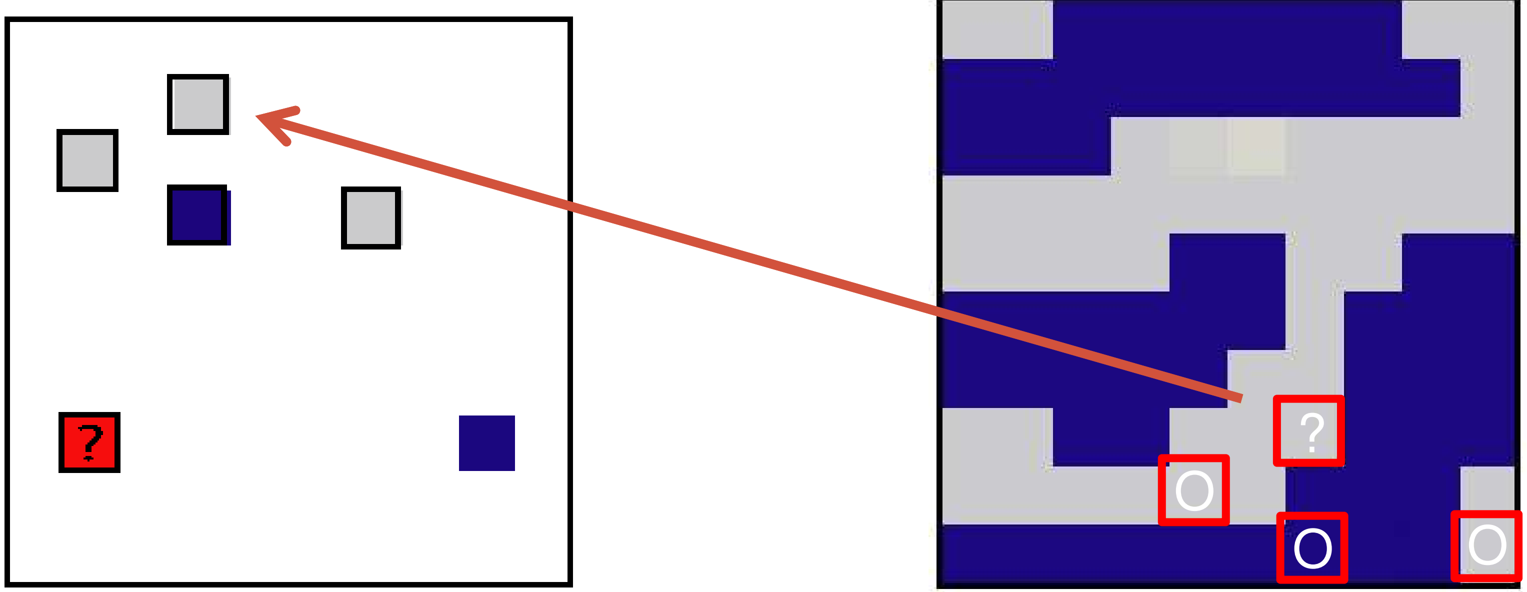
$$\mathbf{d}_n(\mathbf{y}) = \{Z(\mathbf{y}+\mathbf{h}_1),\cdots,Z(\mathbf{y}+\mathbf{h}_n)\}$$

$$d\{\mathbf{d}_n(\mathbf{x}),\mathbf{d}_n(\mathbf{y})\} = \frac{0}{3}$$

Direct sampling

Simulation grid

Training image



Conditioning data event

$$\mathbf{d}_n(\mathbf{x}) = \{Z(\mathbf{x}+\mathbf{h}_1),\cdots,Z(\mathbf{x}+\mathbf{h}_n)\}$$

Candidate data event

$$\mathbf{d}_n(\mathbf{y}) = \{Z(\mathbf{y}+\mathbf{h}_1),\cdots,Z(\mathbf{y}+\mathbf{h}_n)\}$$

$$d\{\mathbf{d}_n(\mathbf{x}),\mathbf{d}_n(\mathbf{y})\} = \frac{0}{3}$$

Basic Direct Sampling algorithm

- Distance :

$$d\{\mathbf{d}_n(\mathbf{x}),\mathbf{d}_n(\mathbf{y})\} = \frac{1}{n}\sum_{i=1}^{n} a_i \qquad a_i = \begin{cases} 0 & if\ Z(\mathbf{x}+\mathbf{h}_i) = Z(\mathbf{y}+\mathbf{h}_i) \\ 1 & if\ Z(\mathbf{x}+\mathbf{h}_i) \neq Z(\mathbf{y}+\mathbf{h}_i) \end{cases}$$

- DS algorithm consists in scanning the Training Image
  - Until we find the first data event such that

$$d\{\mathbf{d}_n(\mathbf{x}),\mathbf{d}_n(\mathbf{y})\} < t$$

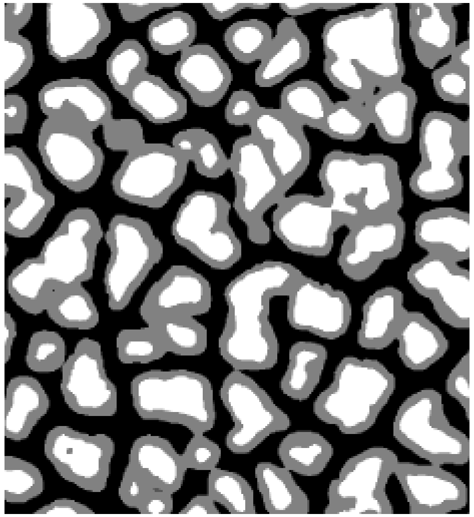
  - Or until a certain fraction **f** of the training image has been scanned, then the best data event is selected

Mariethoz, Renard, Straubhaar (2010) WRR, doi:10.1029/2008WR007621

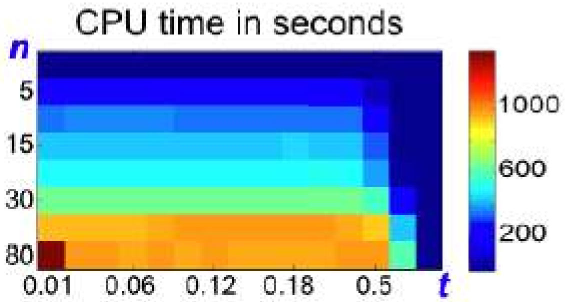
Direct Sampling parameters

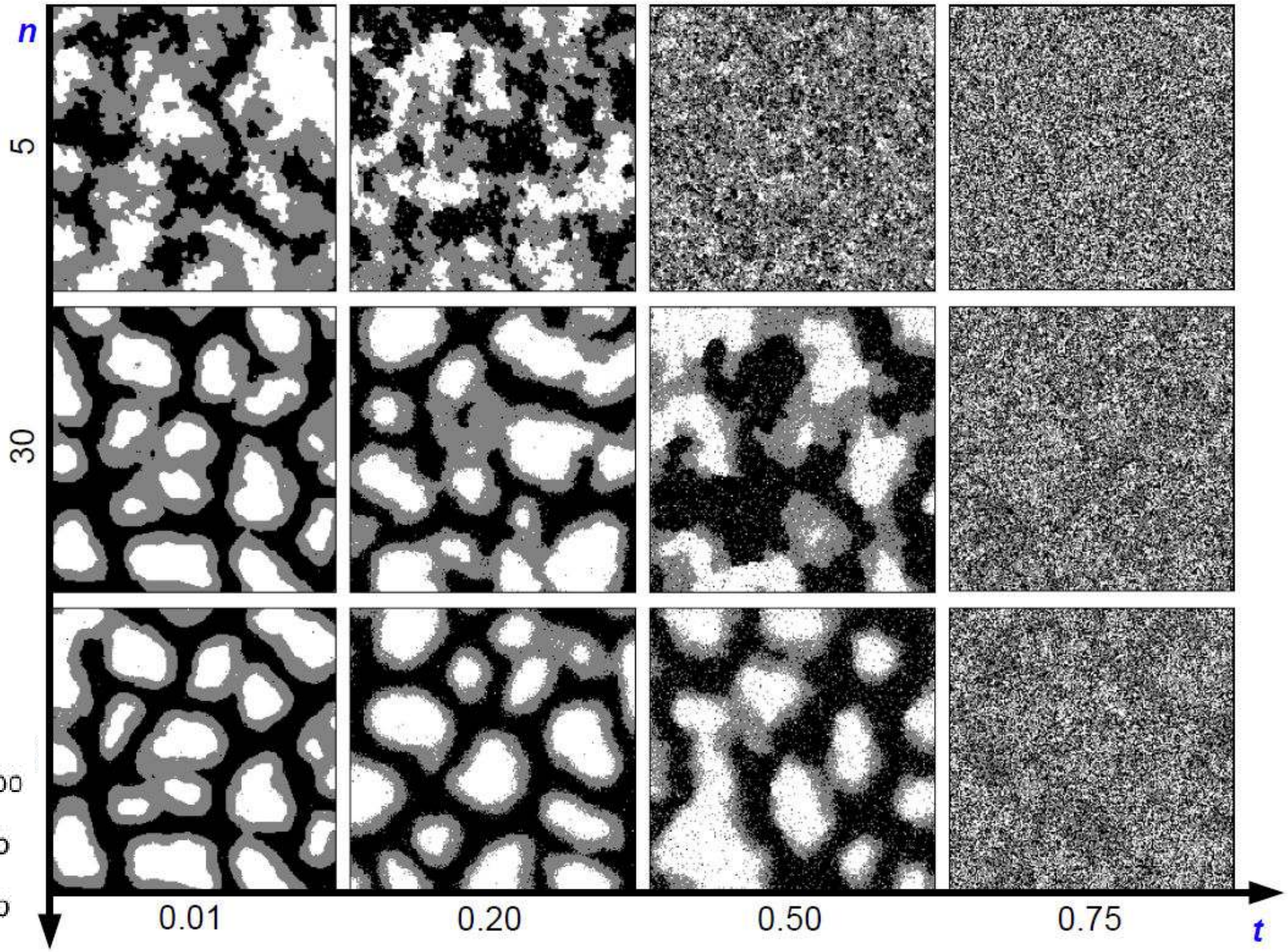
- Size of the neighborhood: **n**
  - Depends on space dimension, 10 to 100
- Acceptance threshold for the distance: **t**
  - A value between 0 and 1
- Maximum scan fraction of the Training Image: **f**
  - Between 0.1 and 0.5

Parameter sensitivity



Training image



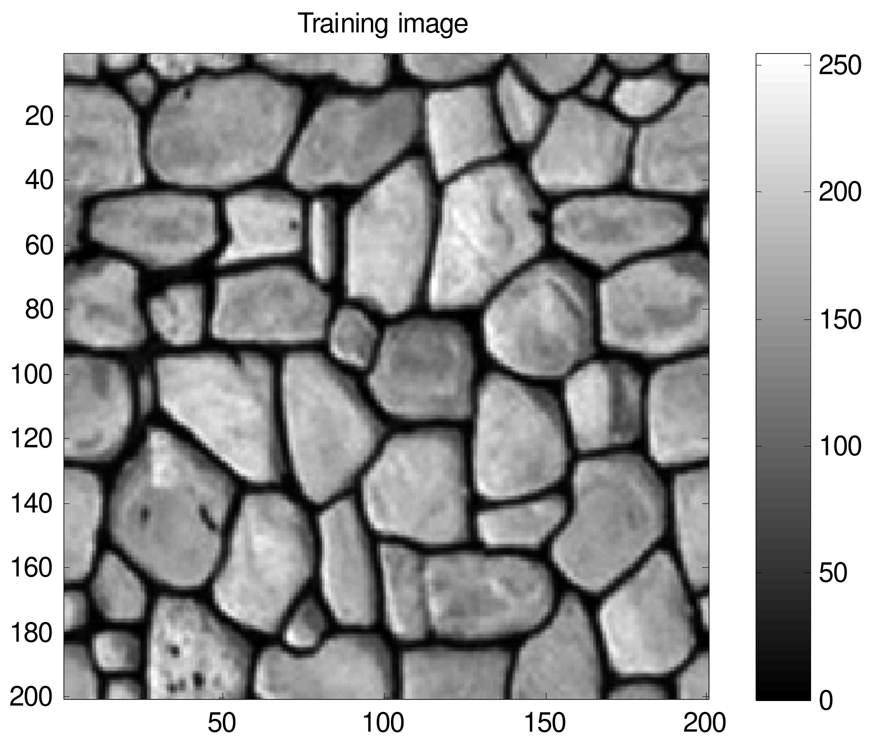


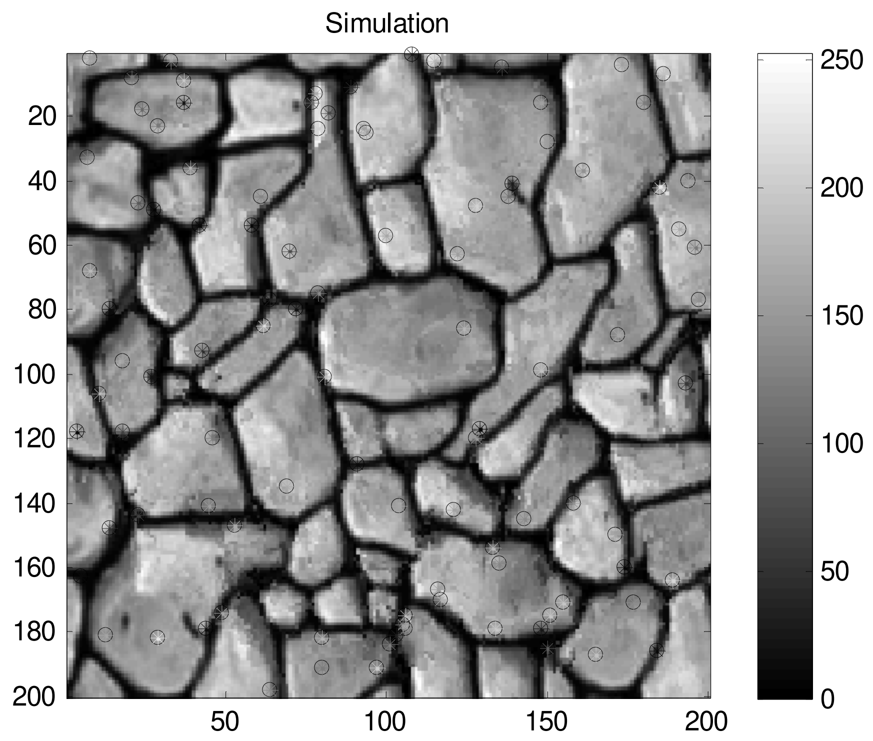
Simulations with f=0.3

Meershman, Pirot, et al. (2013) Computers and Geosciences, 52: 307-324

Continuous variable

$$d\{\mathbf{d}_n(\mathbf{x}), \mathbf{d}_n(\mathbf{y})\} \propto \sum_{i=1}^{n} \left[Z(\mathbf{x} + \mathbf{h}_i) - Z(\mathbf{y} + \mathbf{h}_i) \right]^2$$

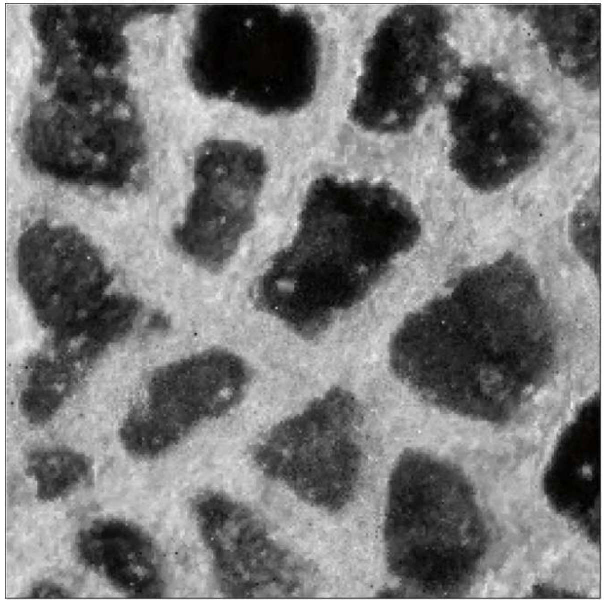




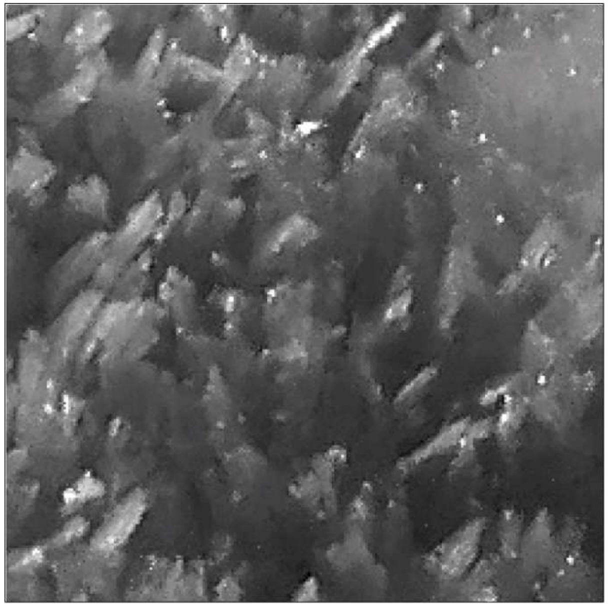
Mariethoz, Renard, Straubhaar (2010) WRR, doi:10.1029/2008WR007621

Examples of simulations

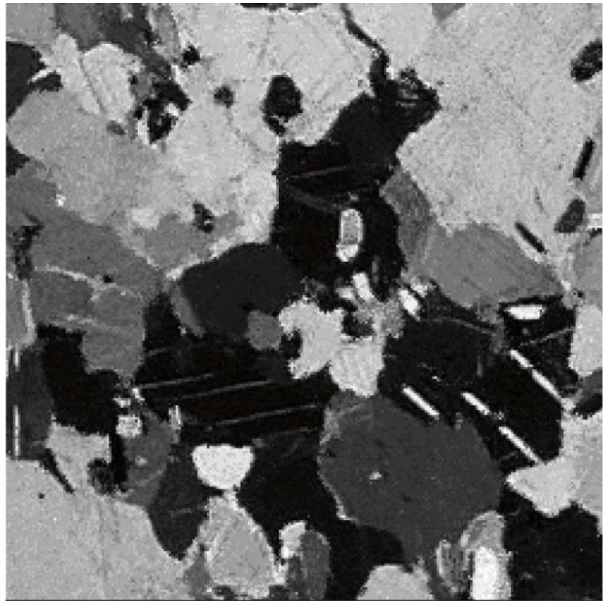
Ice wedge polygonal soil



Snow



Thin slice marble

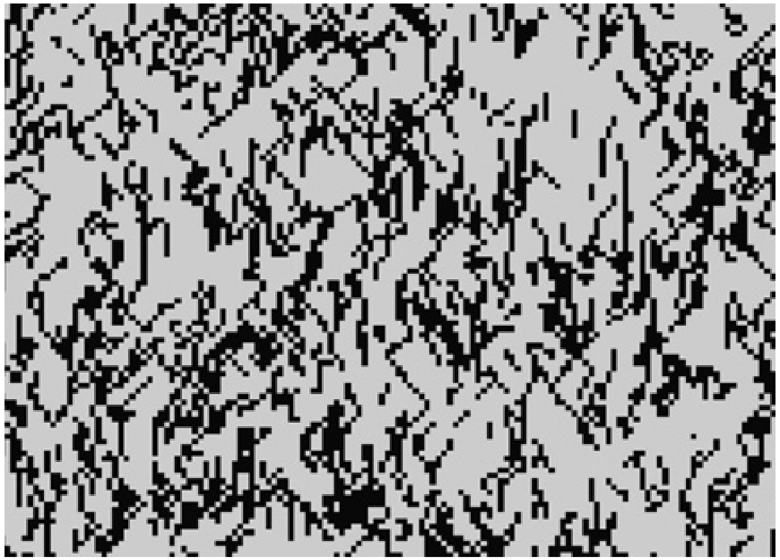


Meershman, Pirot, et al. (2013) Computers and Geosciences, 52: 307-324

Training image stationarity

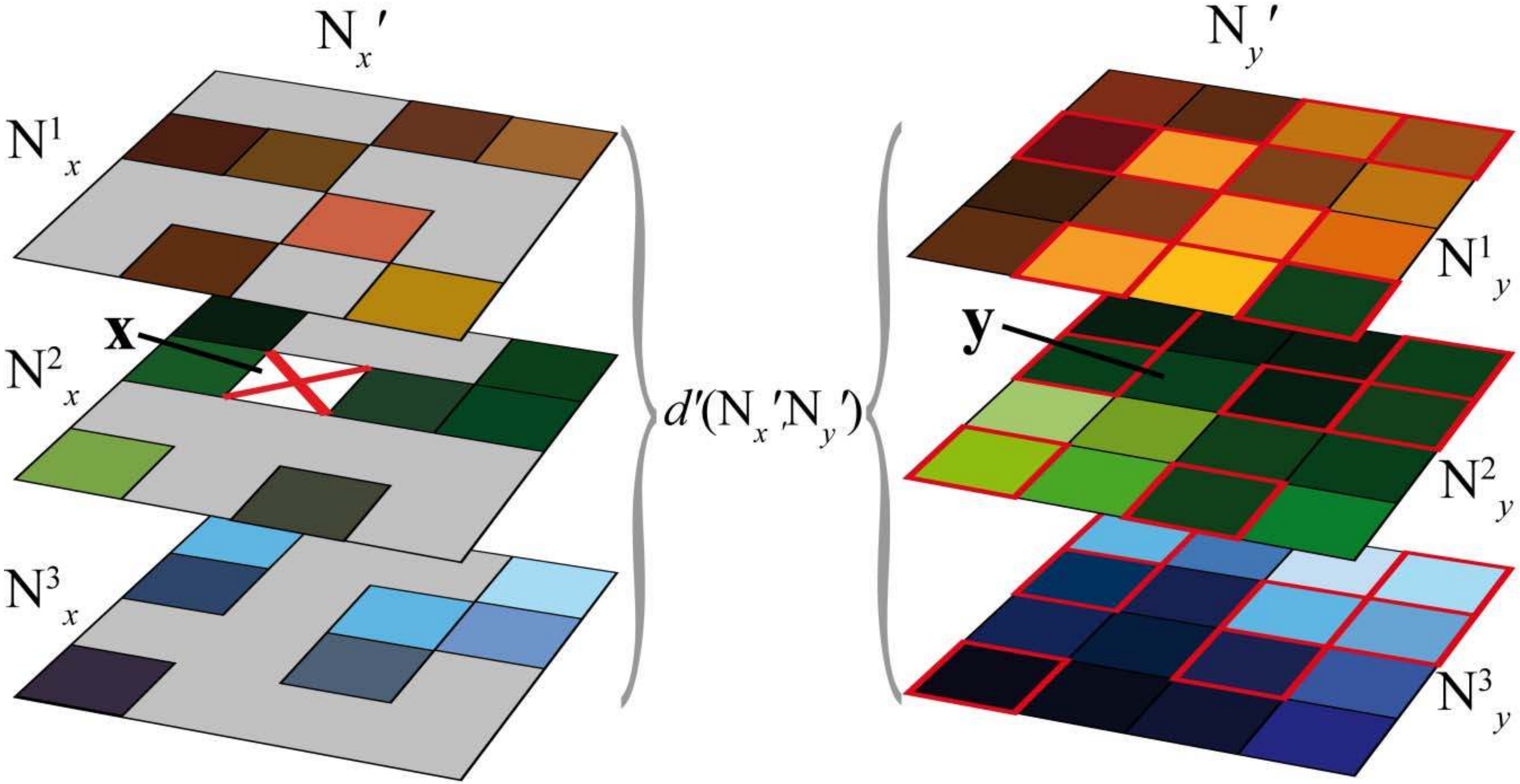
Training image (TI)

Simulation



Boucher (2009) Computers and Geosciences, 35: 1151–1158

Multiple variables

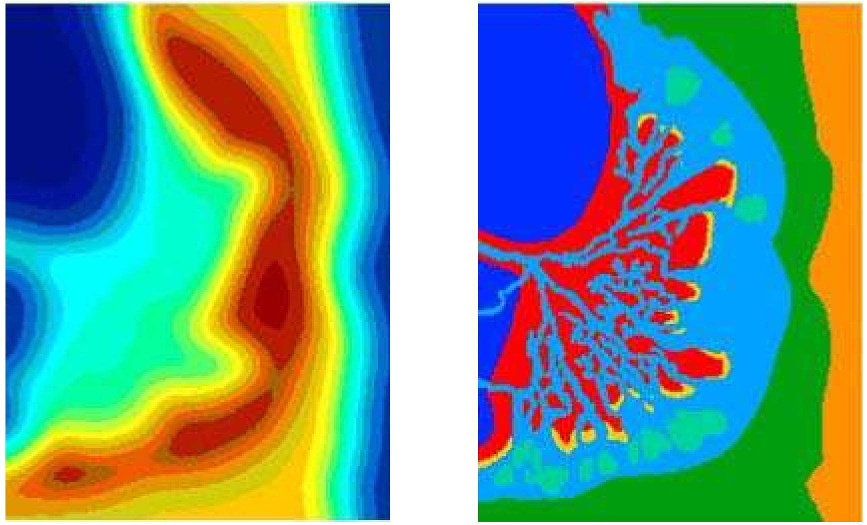


$$d^i\left\{\mathbf{d}_n(\mathbf{x}), \mathbf{d}_n(\mathbf{y})\right\} < t_i$$

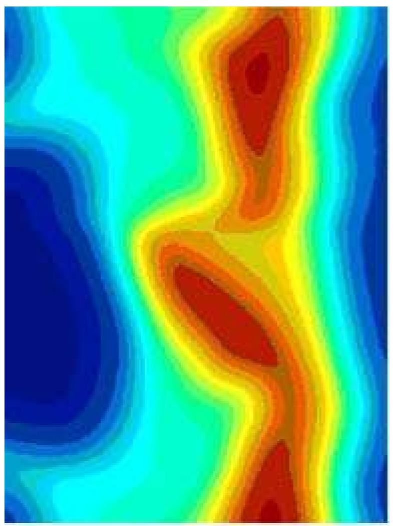
Multivariate simulation

Training data set

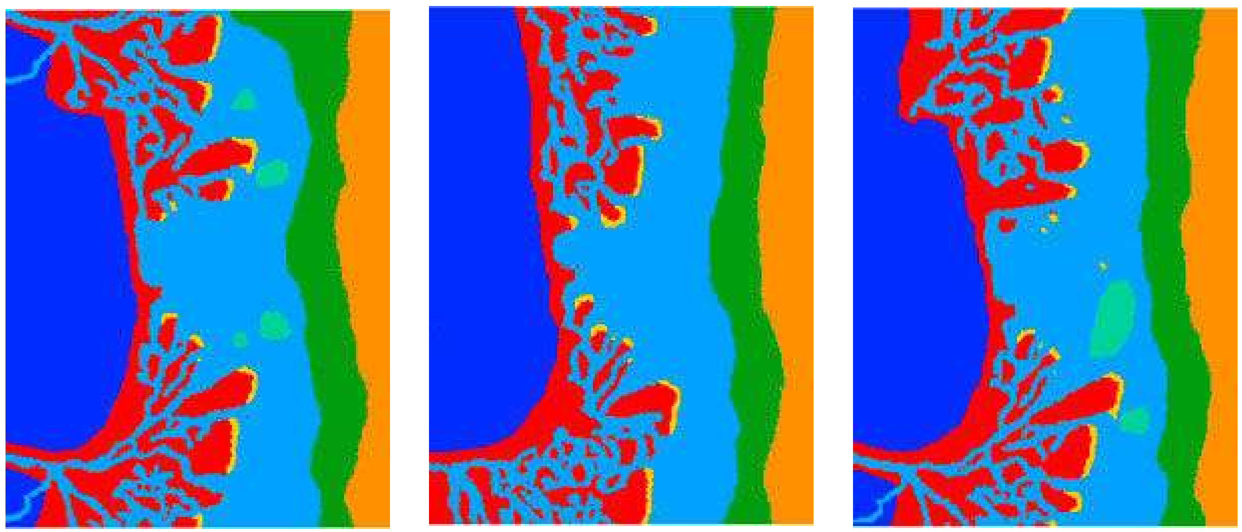
Variable 1 Variable 2



Conditioning variable
Variable 1



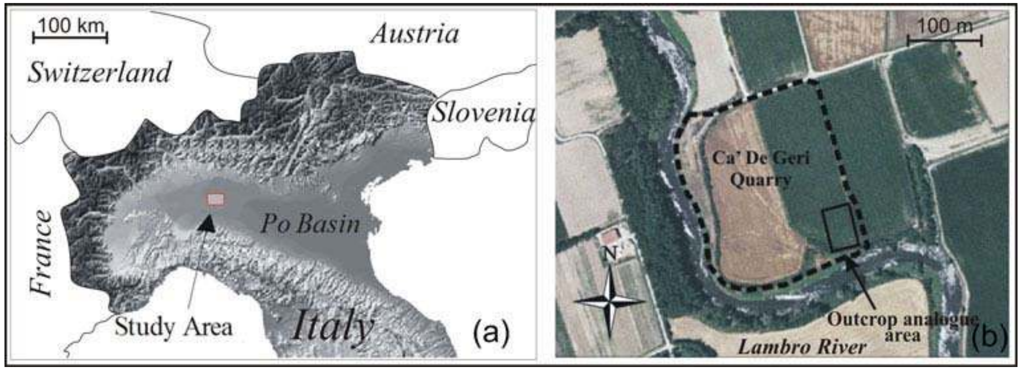
3 Simulations
Variable 2

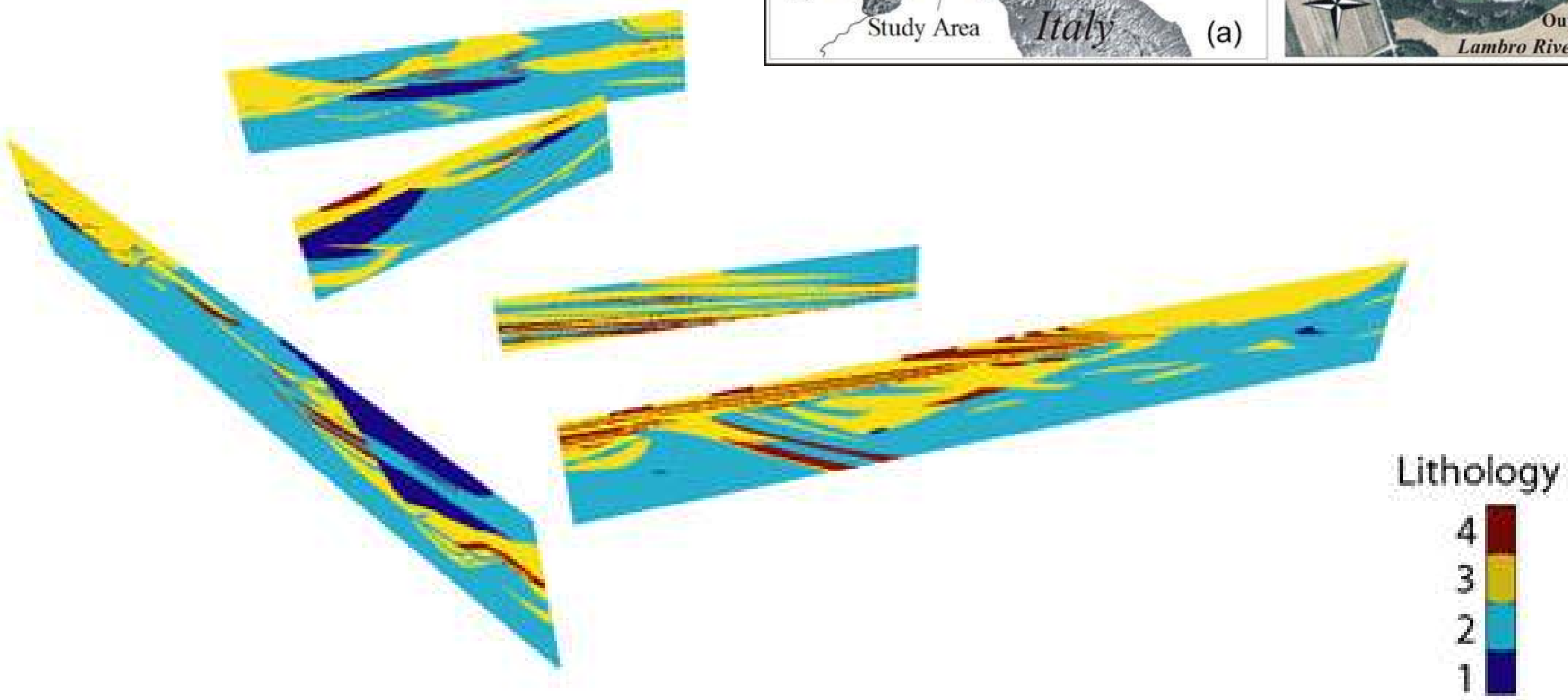


A FEW EXAMPLES OF APPLICATIONS

3D geology / Rainfall simulation / Reconstruction of missing data / etc.

Reconstructions from sections

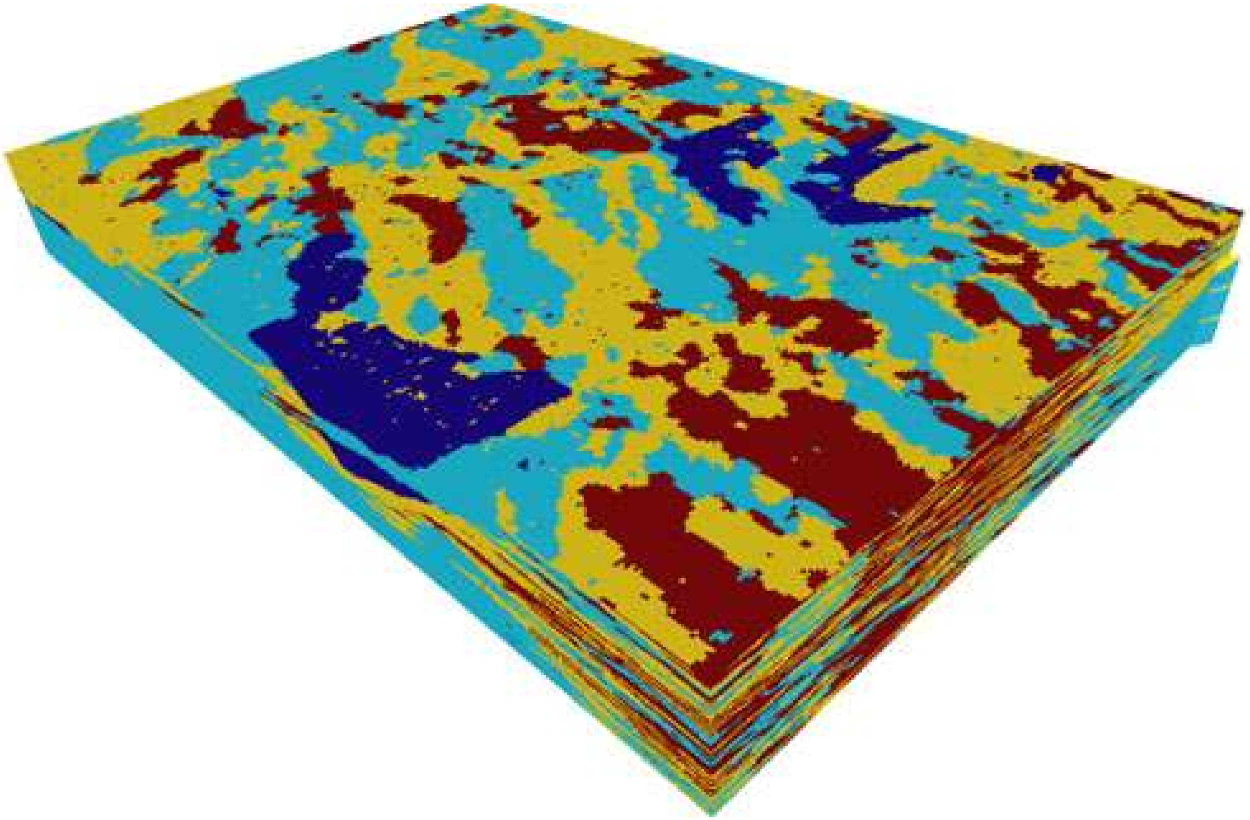




Data provided by Dell'Arciprete, Felleti, Bersezio

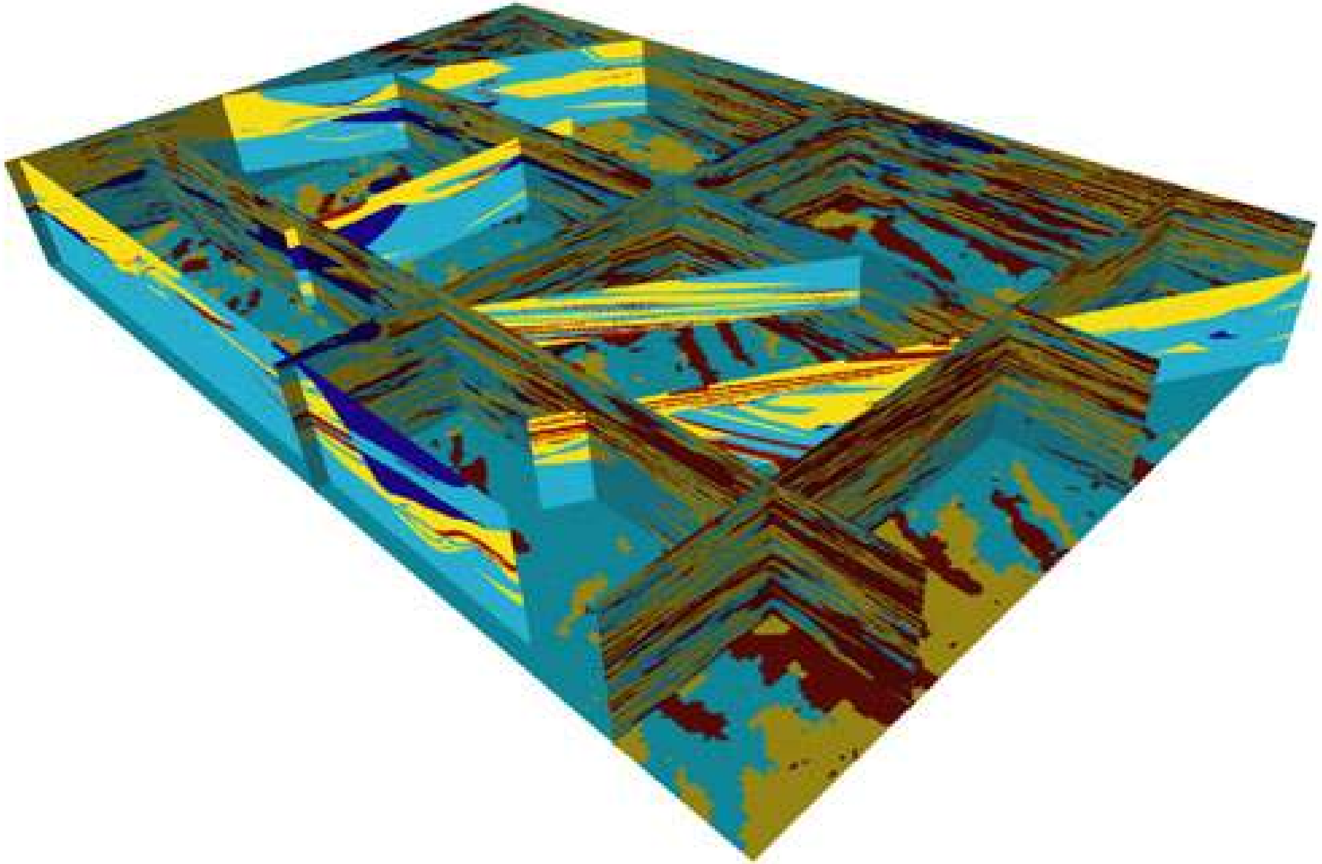
Mariethoz, Renard (2009) Mathematical Geosciences. 42(3): 245–268

One simulation



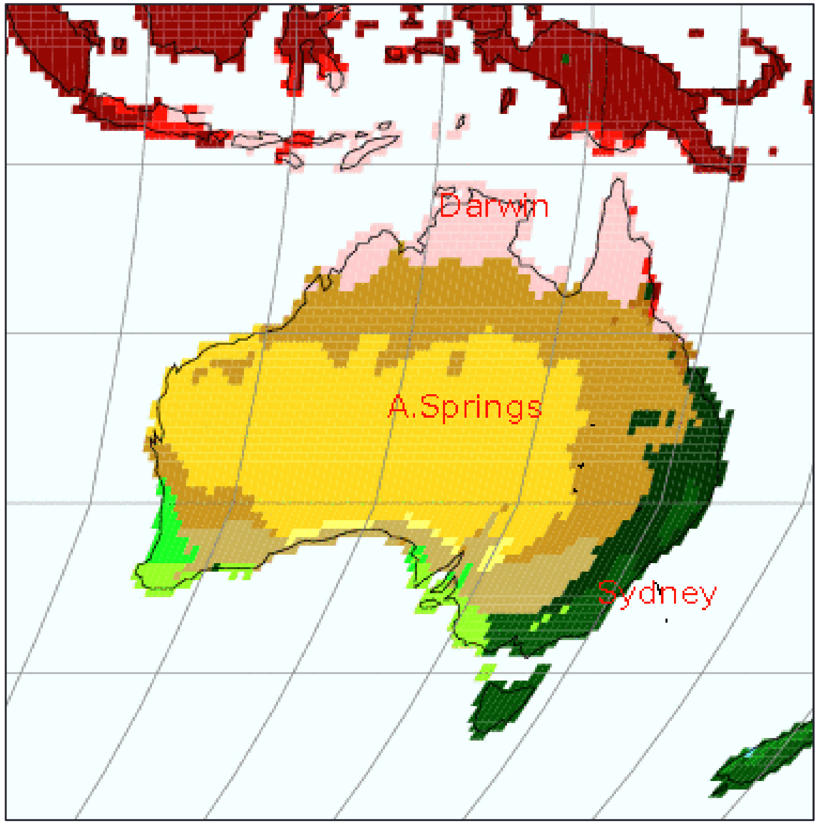
Mariethoz, Renard (2009) Mathematical Geosciences. 42(3): 245–268

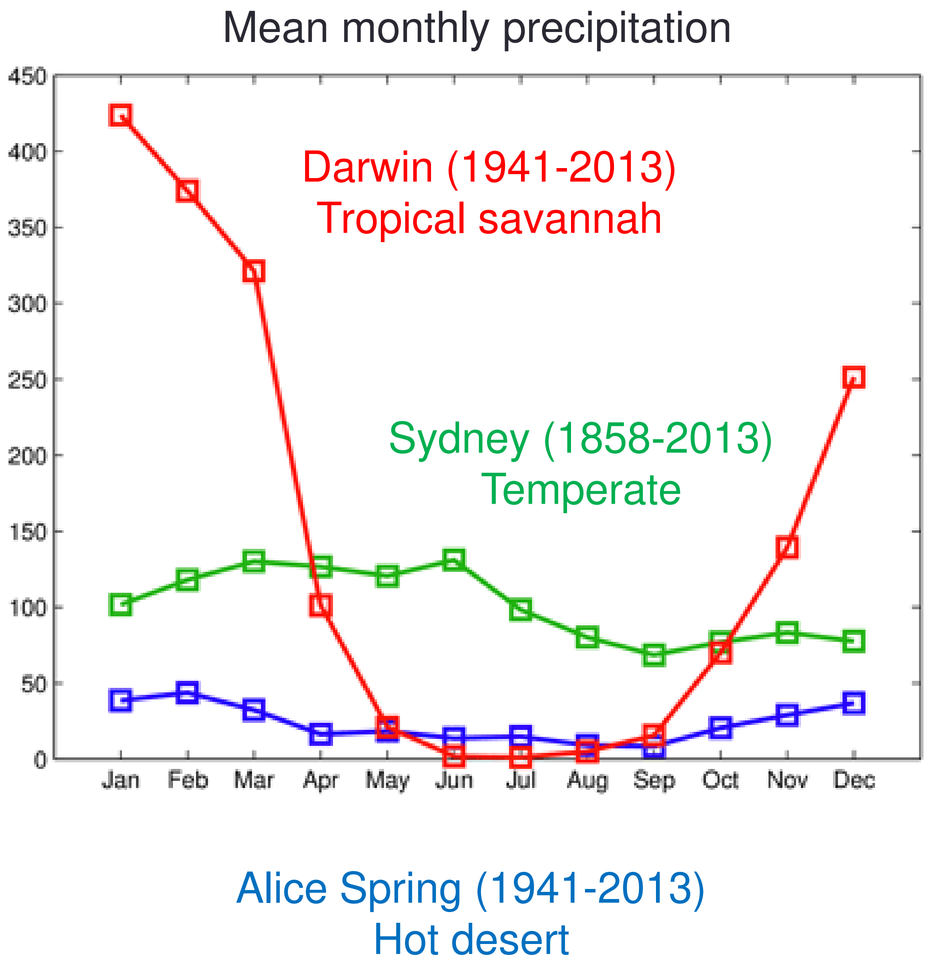
Sections in the simulation



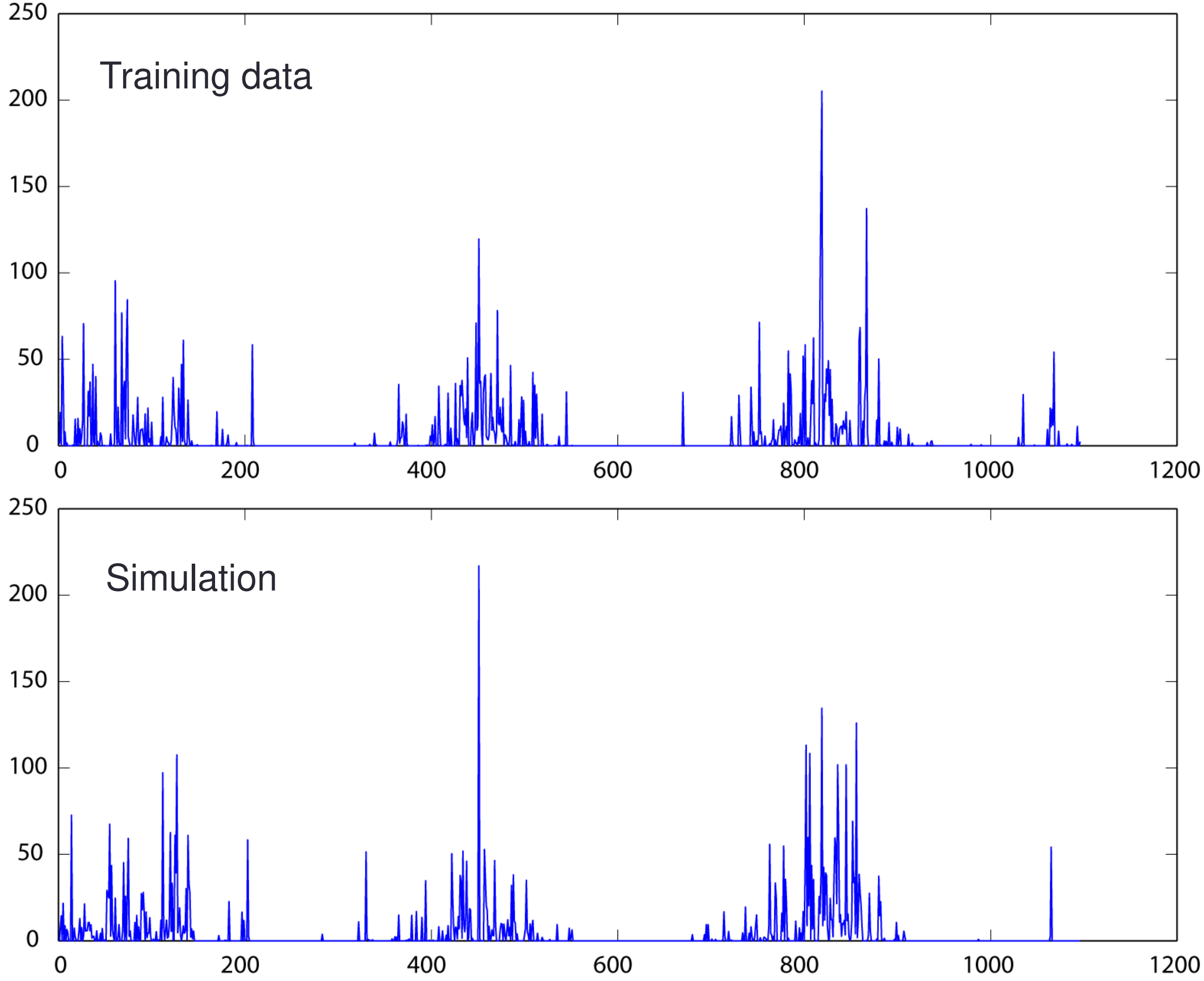
Mariethoz, Renard (2009) Mathematical Geosciences. 42(3): 245–268

Rainfall simulation

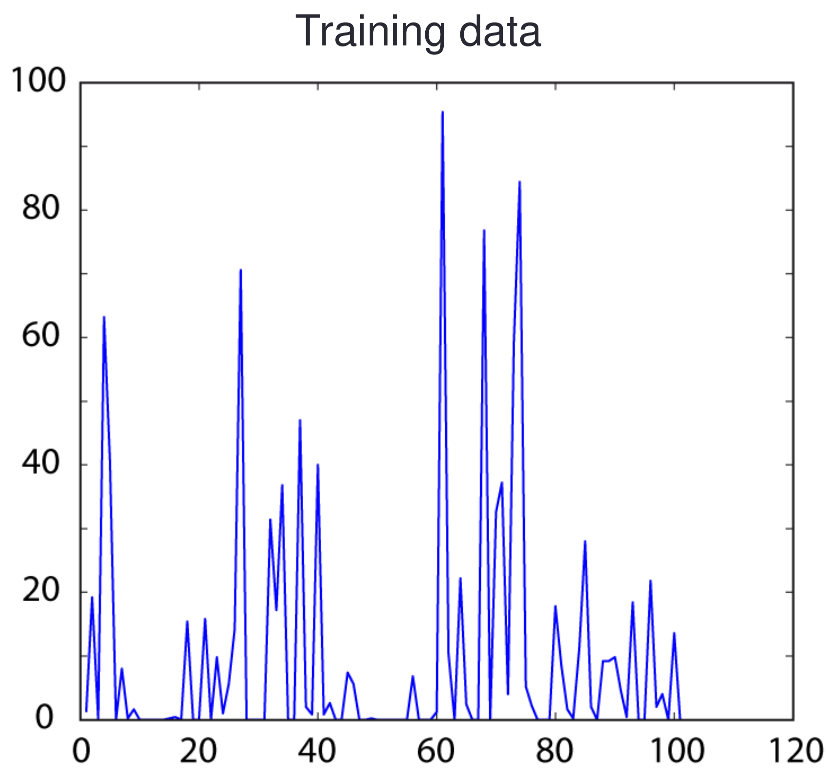


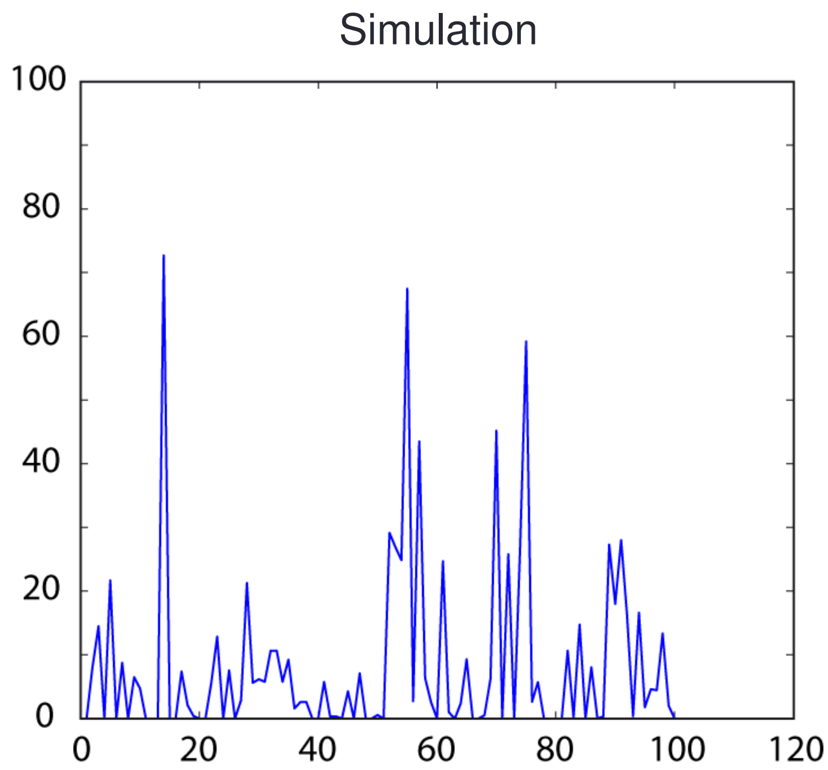


Rainfall simulation (Darwin)

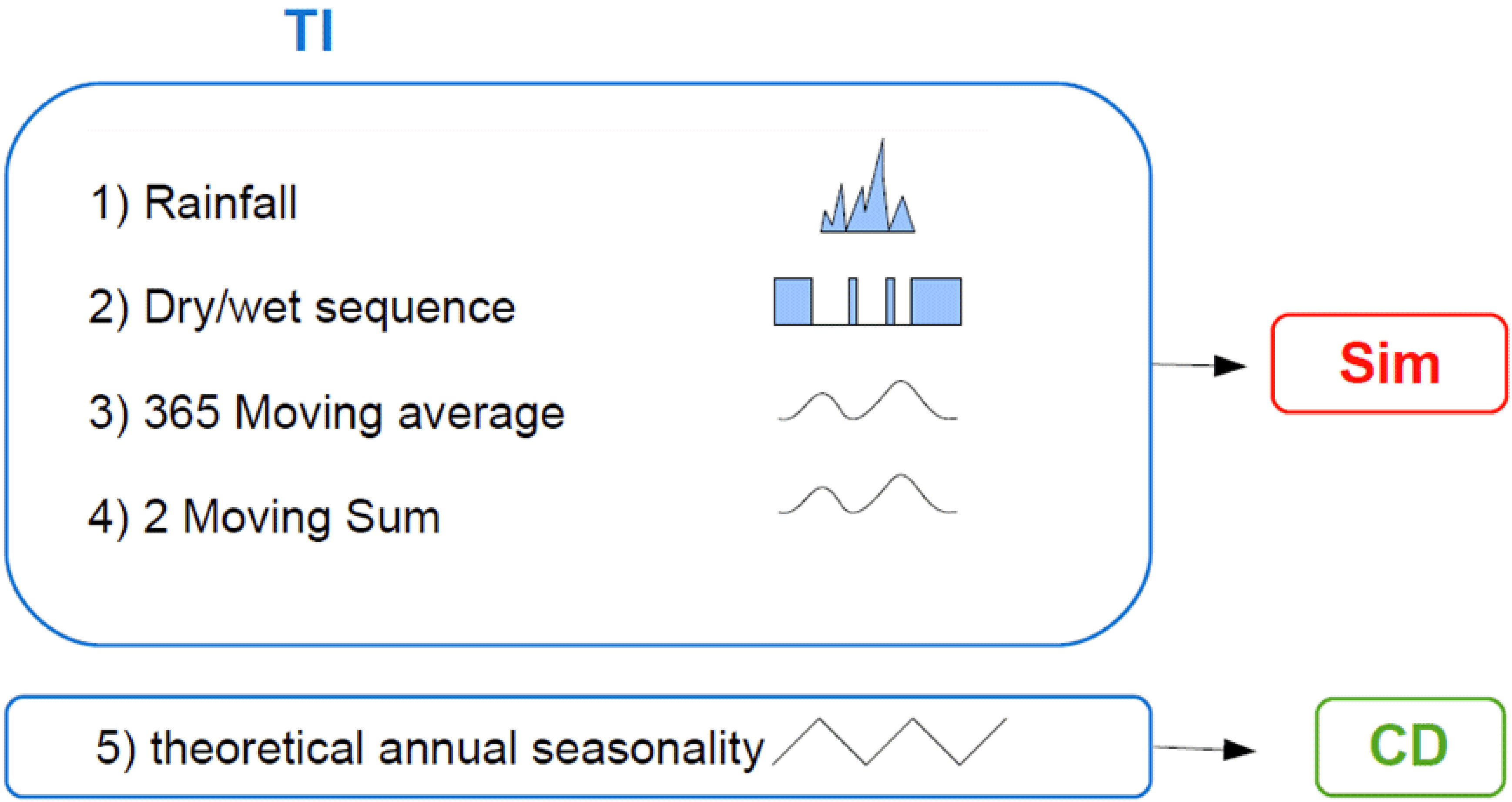


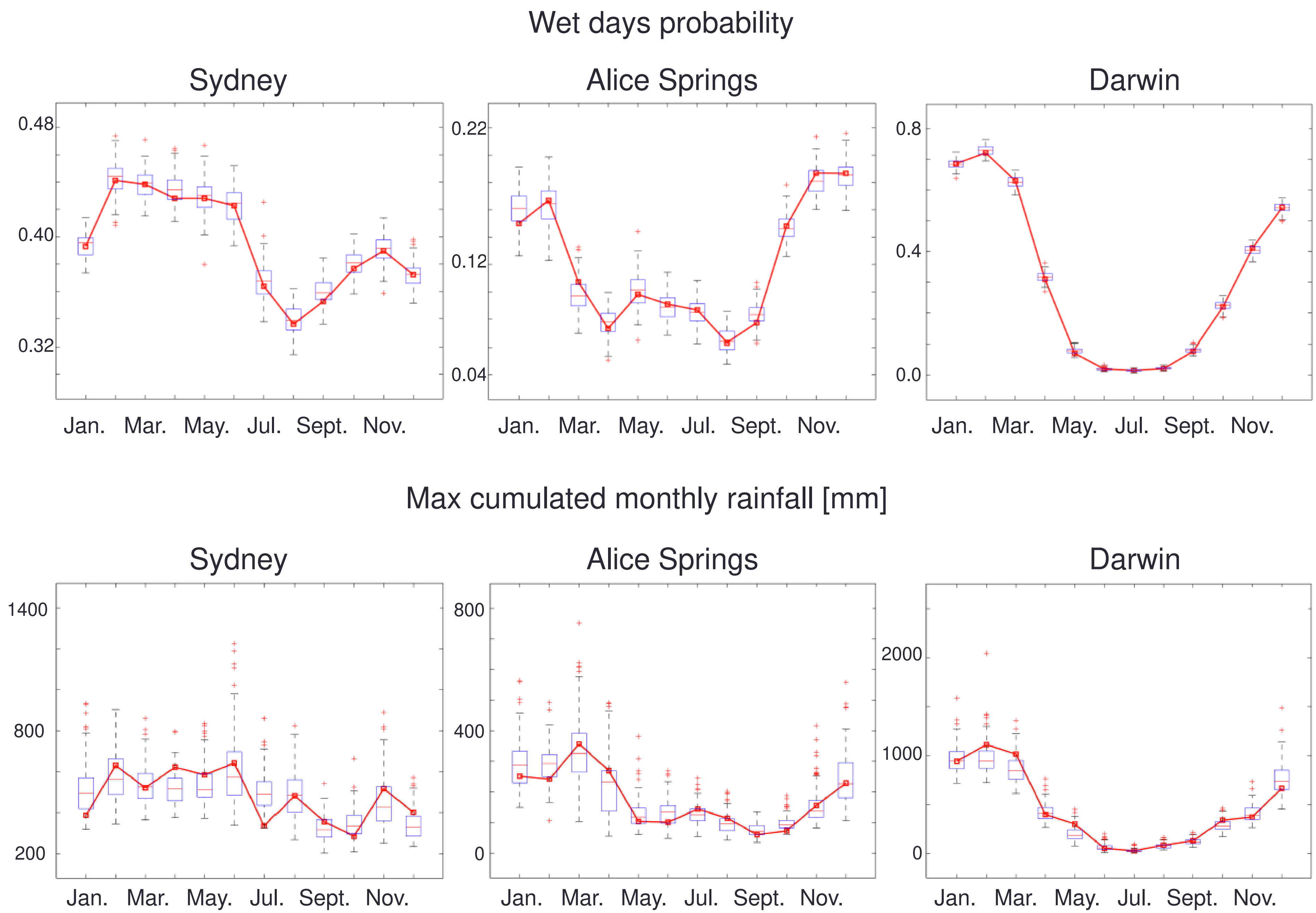
Rainfall simulation (Darwin)



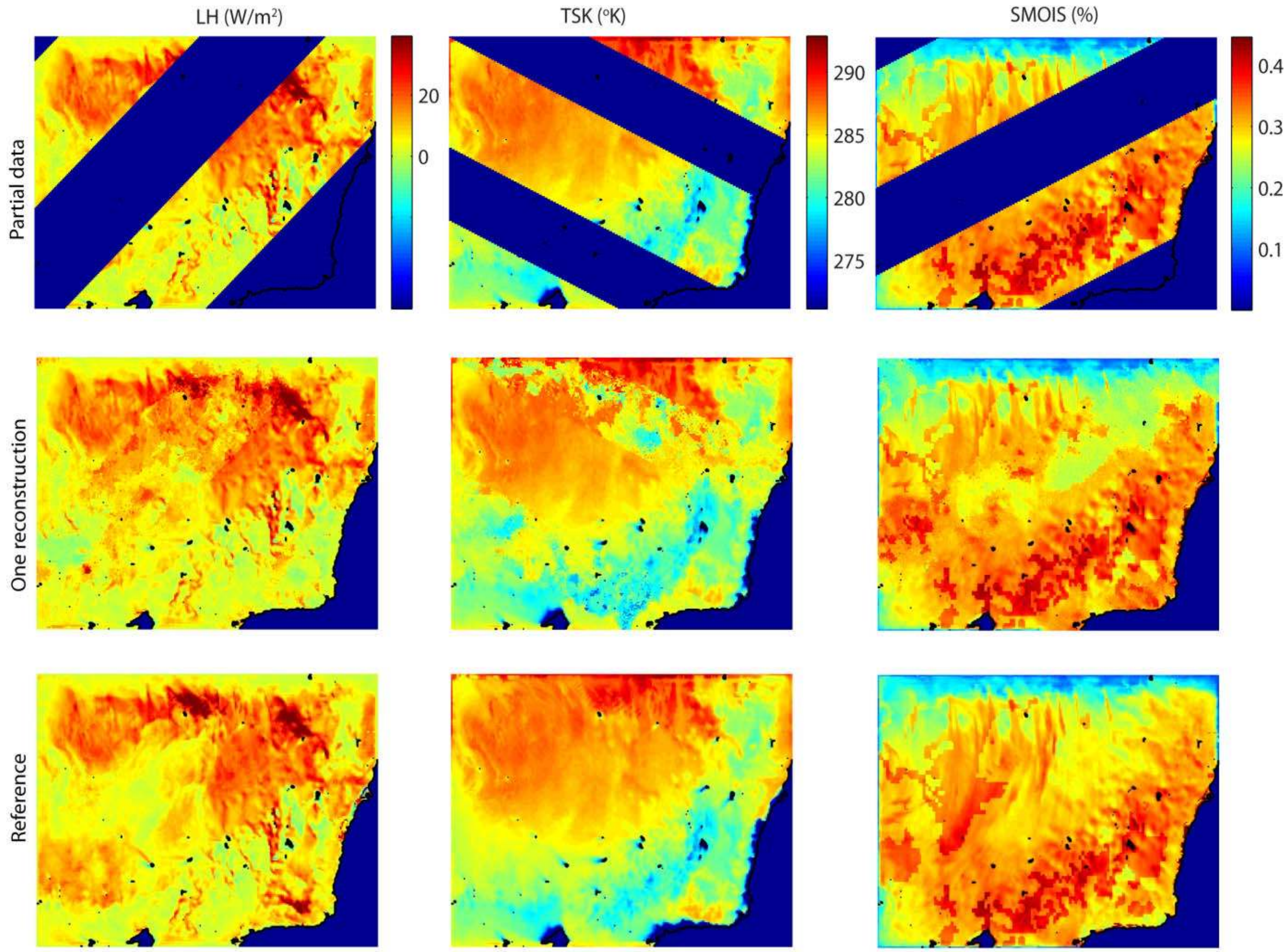


Rainfall simulation procedure



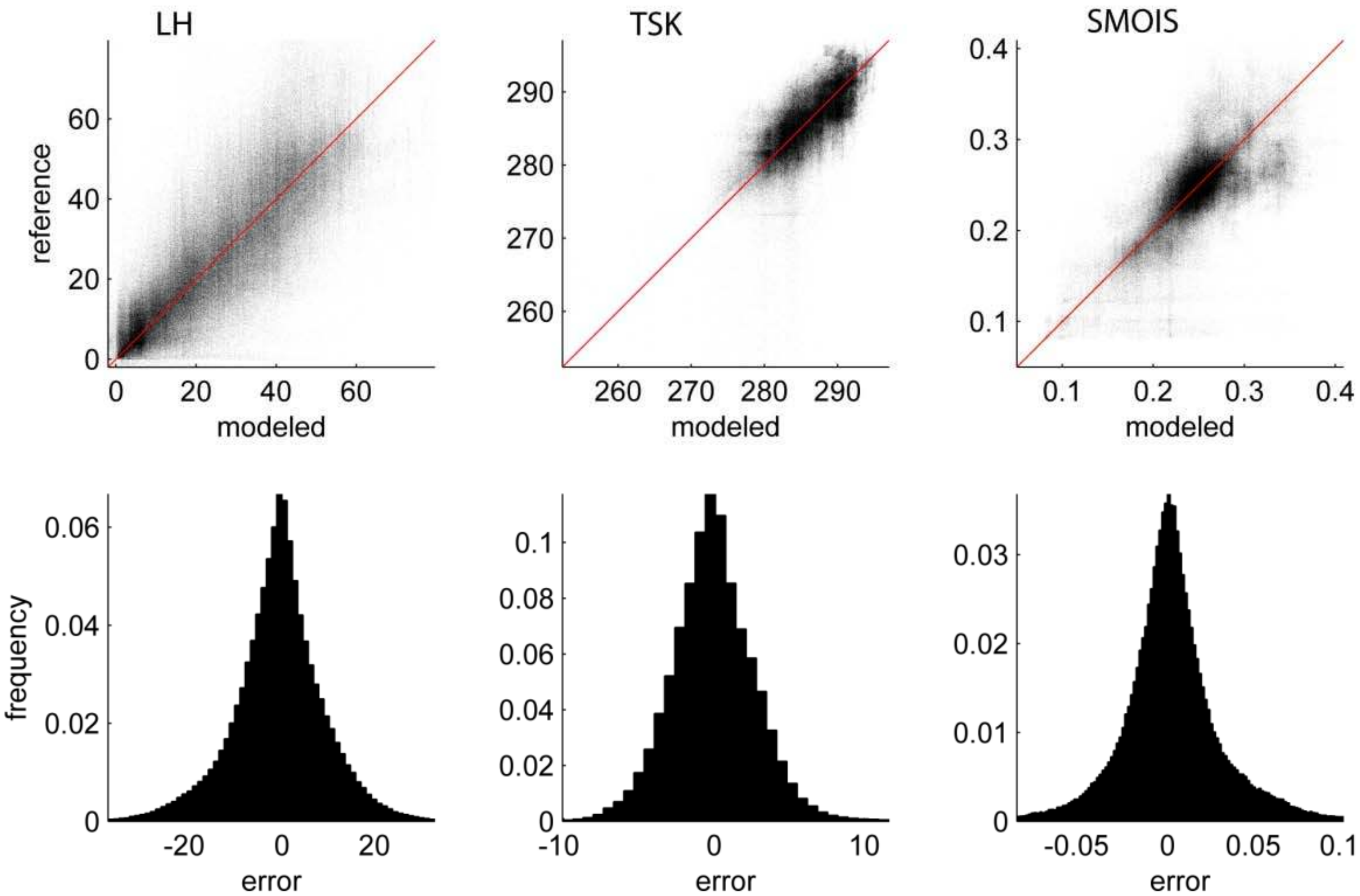

Wet days probability
Sydney
0.48
0.40
0.32
Jan. Mar. May. Jul. Sept. Nov.
Alice Springs
0.22
0.12
0.04
Jan. Mar. May. Jul. Sept. Nov.
Darwin
0.8
0.4
0.0
Jan. Mar. May. Jul. Sept. Nov.
Max cumulated monthly rainfall [mm]
Sydney
1400
800
200
Jan. Mar. May. Jul. Sept. Nov.
Alice Springs
800
400
0
Jan. Mar. May. Jul. Sept. Nov.
Darwin
2000
1000
0
Jan. Mar. May. Jul. Sept. Nov.

Reconstruction of gaps caused by orbital passages on remote sensing images

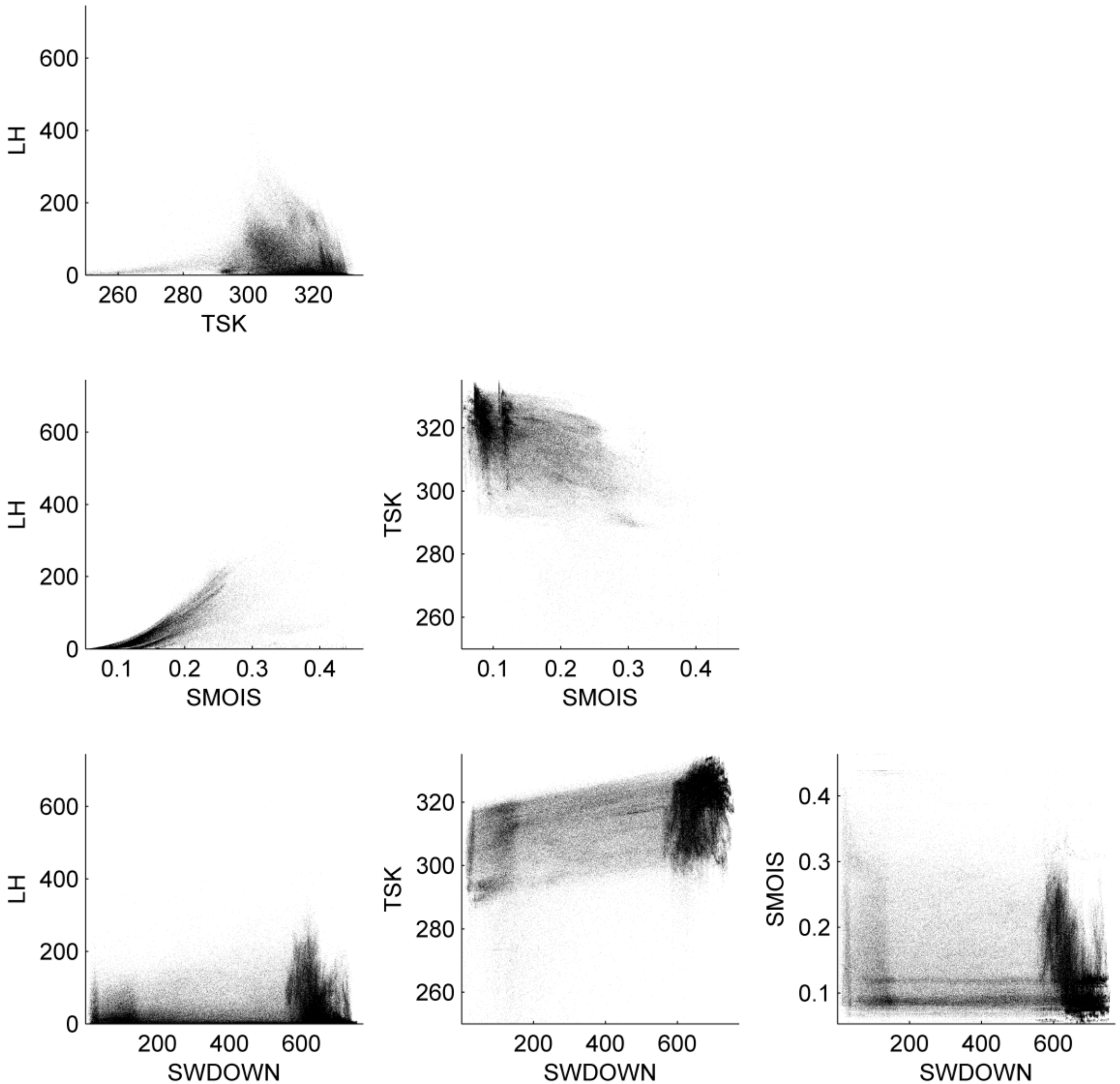


Mariethoz, Mc Cabe and Renard (2013) WRR, 48: doi:10.1029/2012WR012115

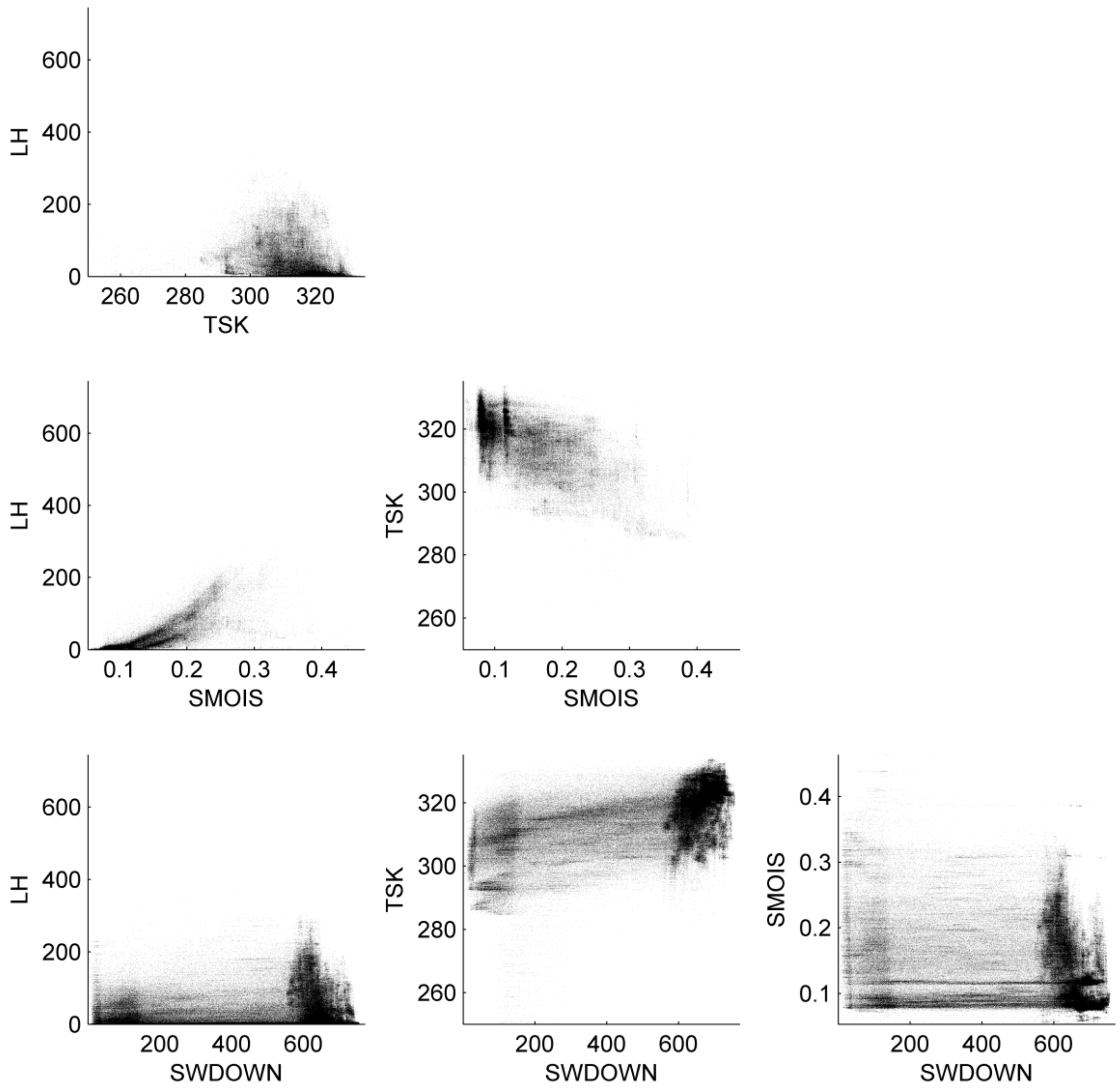
Error analysis



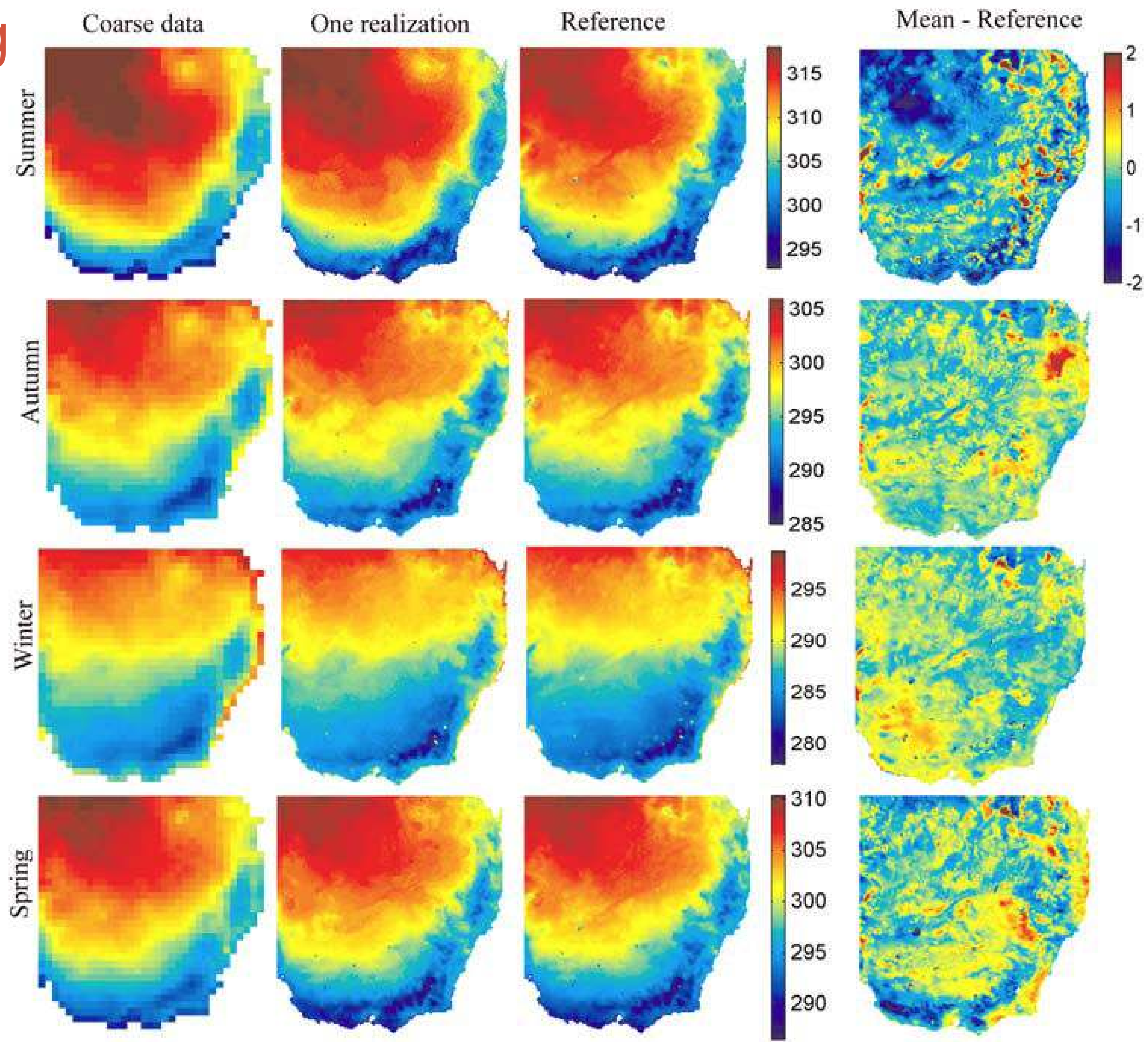
Dependence between reference values



Dependence between simulated values



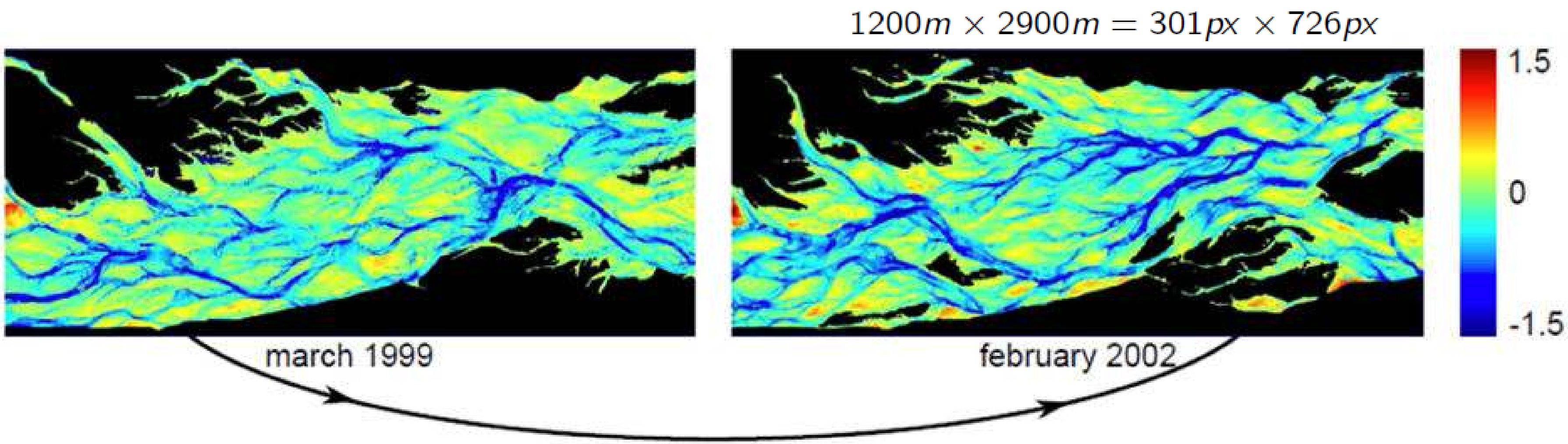
Downscaling



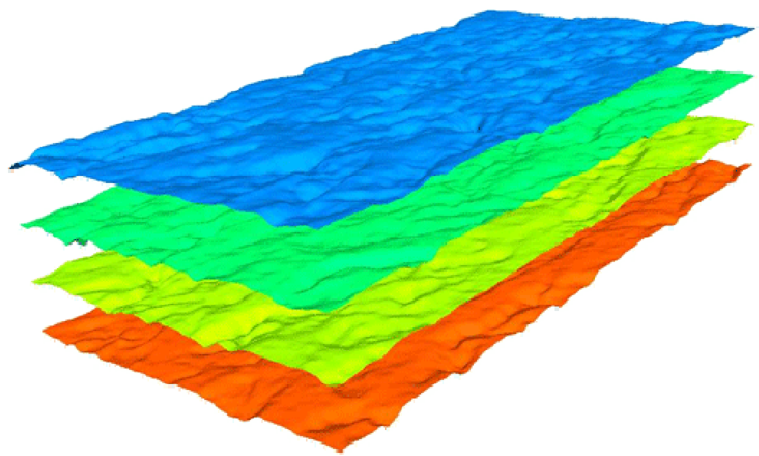
Surface Temperature

Jha, et al (2013) WRR, 49: doi: 10.1029/2012WR012602

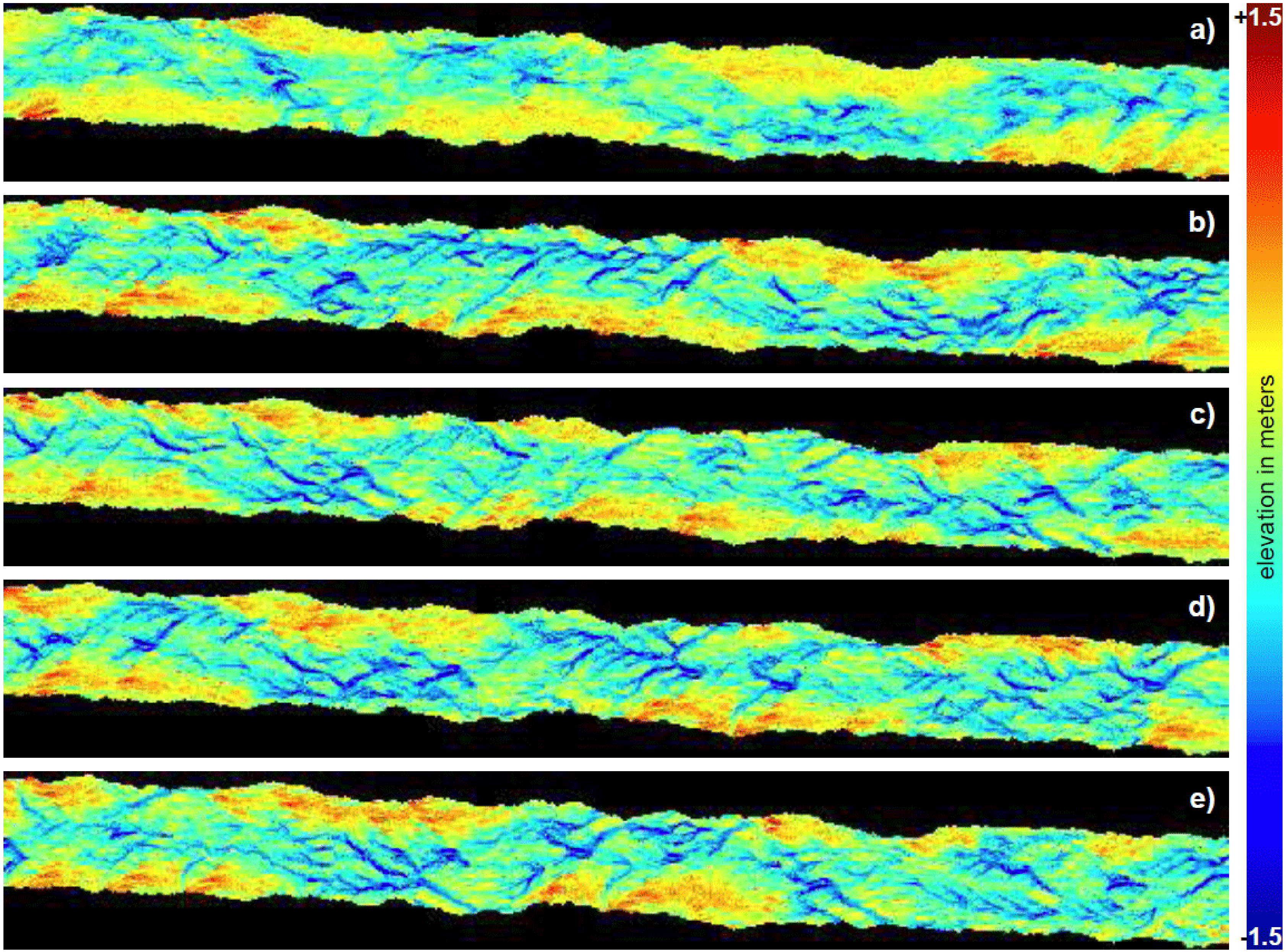
Topography simulations



- High resolution DEM
- Multiple-point statistics to model successive topographies
- Stack them
- Fill the volumes with sediments

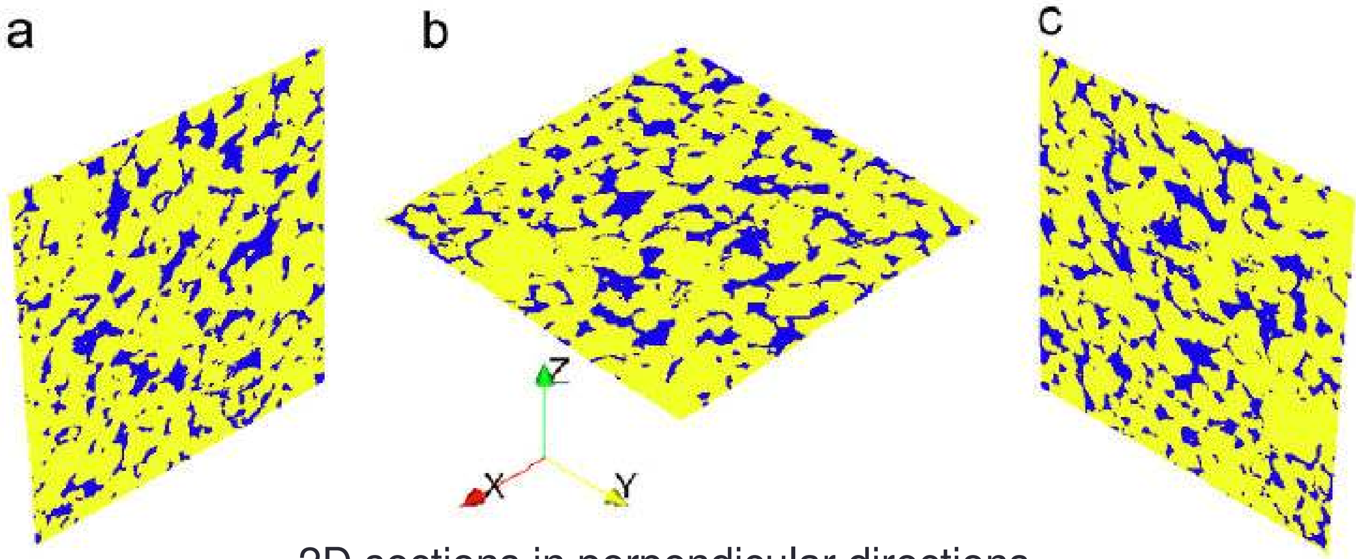


Pirot et al (2014) Geomorphology, 214: doi: doi:10.1016/j.geomorph.2014.01.022



Pirot et al (2014) Geomorphology, 214: doi: doi:10.1016/j.geomorph.2014.01.022

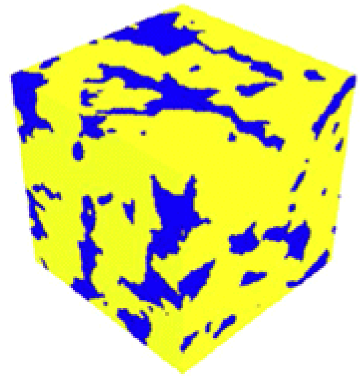
Porous media



2D sections in perpendicular directions

Berea Sandstone

Example of a 3D volume simulated from 2D data sets:

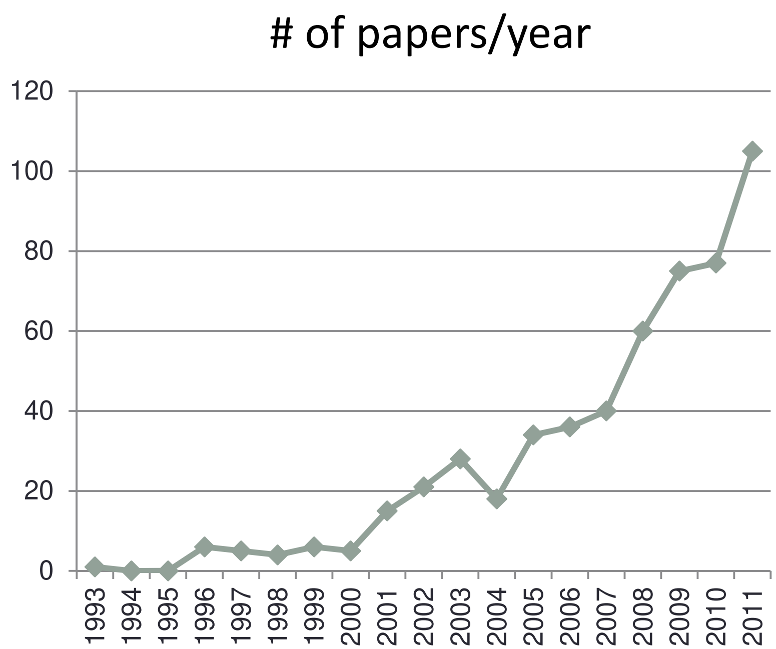


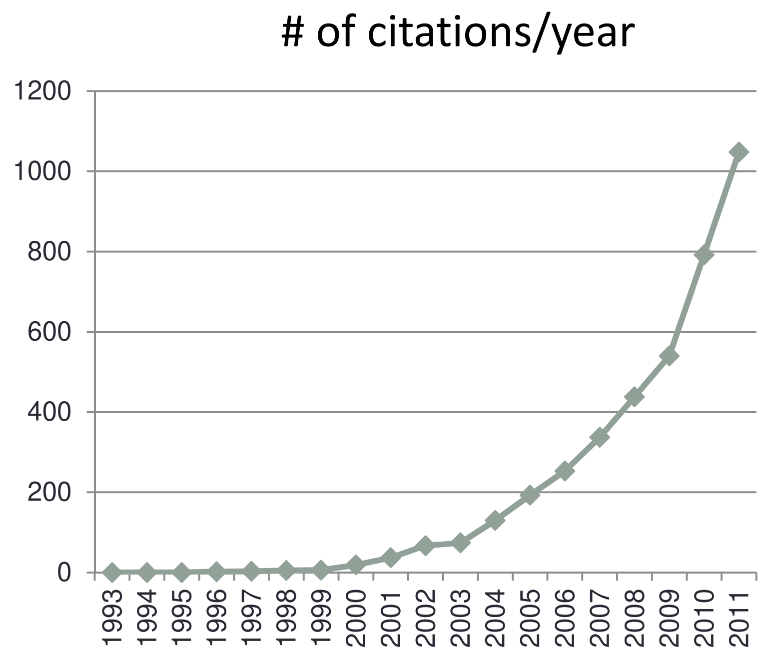
Comunian, Straubhaar and Renard (2013) Computers and Geosciences, 40: 49-65

CONCLUSION

Active field of research

- Guardiano and Srivastava (1993)
- First efficient implementation:
  - Strebelle (2002), probability tree, multi-grid, etc.
- Since then:





MPS Pros / Cons

- Well suited to model complex structures

- General (same code for different structures)
- Easy conditioning
- Integration of secondary data

- Not a well defined Random Function model
- CPU time is longer than other methods
- Where to get the training image?

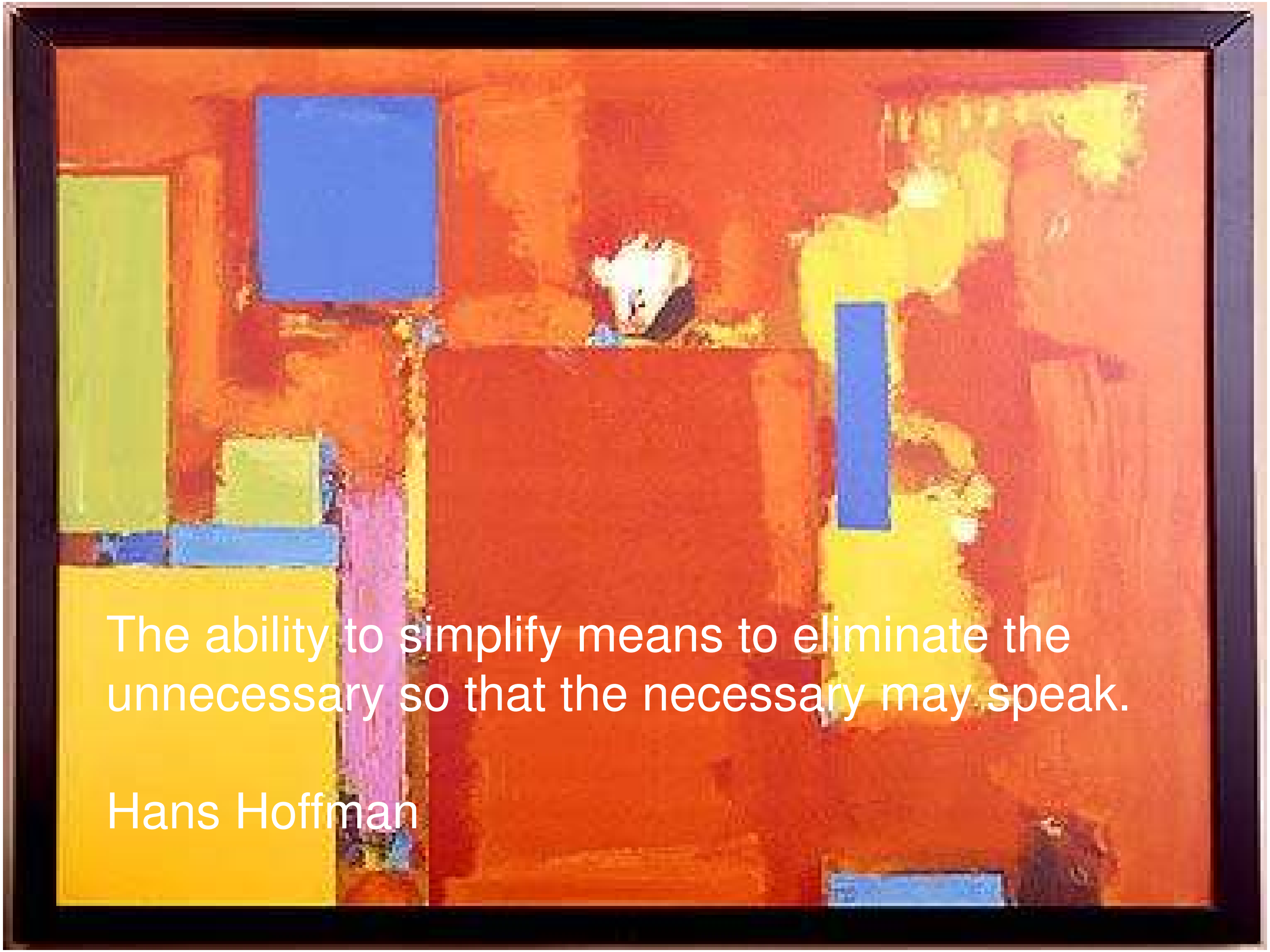
Current research directions

- Applications / demonstrations
- Braided river systems
- Spatio-temporal fields (Precipitation)

- Algorithmic improvements / acceleration

- Multi-scale
- Inverse problem

http://members.unine.ch/philippe.renard/biblio.html


The ability to simplify means to eliminate the unnecessary so that the necessary may speak.
Hans Hoffman