Two Basic Algorithms…

for use in stochastic weather generators design

(and some comments on algorithmics)

Etienne Leblois[1], Sheng Chen[1, 2]

etienne.leblois@irstea.fr [1] Irstea Lyon [2] LSCE (Lab. climate CEA & CNRS) Saclay

Norsk hydrologiråd

5th Conference on Modelling Hydrology, Climate and Land Surface Processes

September 17 – 19, 2019, Lillehammer, Norway

Thanks to

- Em. Prof. Lars Gottschalk, a guide and master to many
- Sjur Kolberg, Cristian Lussana, Sara Martino, for joint work around Stochastic Weather Generators
- Nils-Otto Kitterod for inviting me here - and besides, awaking my interest back into bedrock hydrology (once the area of my PhD)

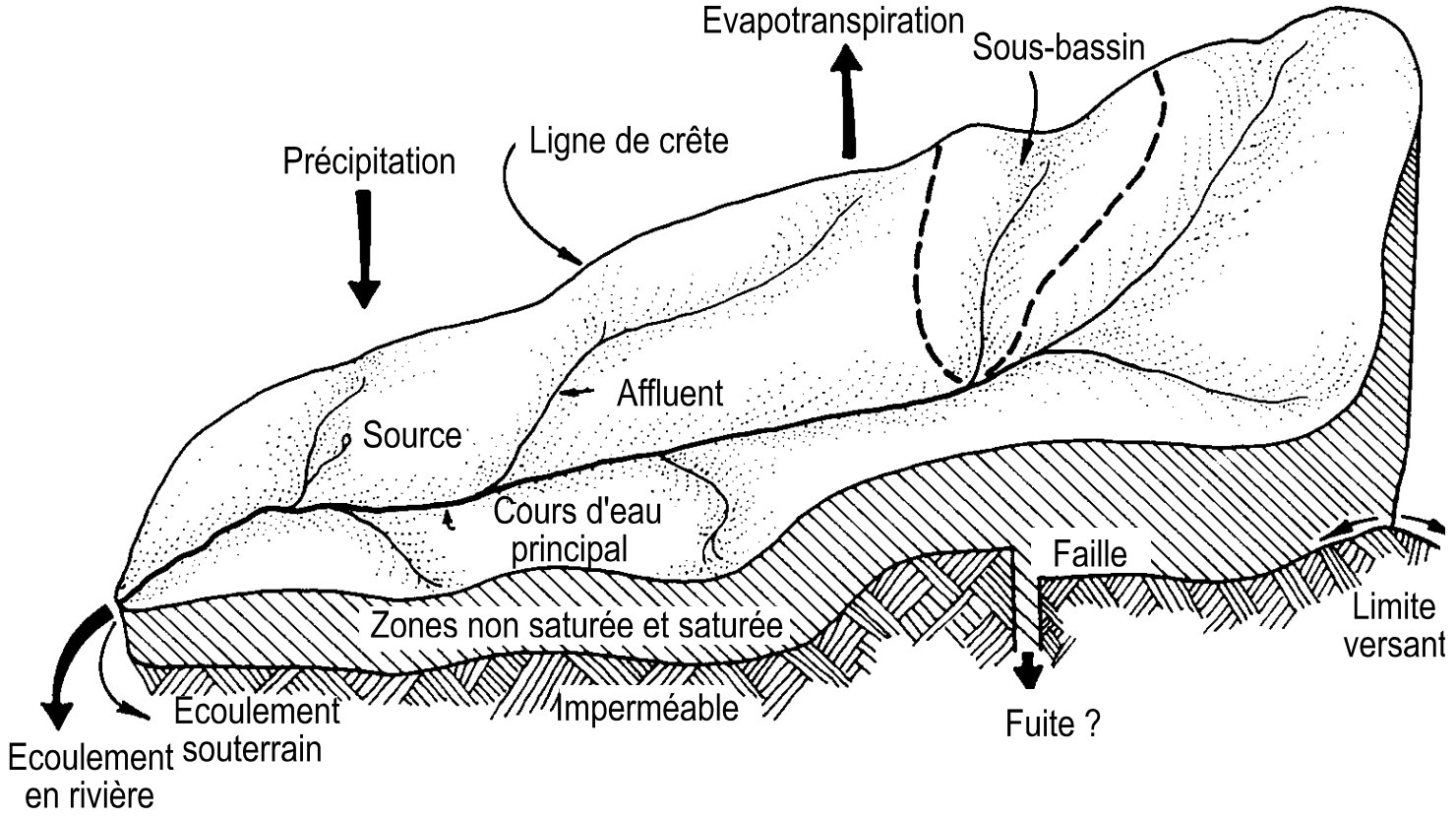
Working with them was much inspiring, giving not only the opportunity to visit Norway

Introduction – Precipitation seen by an hydrologist : spatial structure matters !

- A basin has a given area (space) and characteristic response (time).
- It is mainly sensitive to precipitation averaged over this scale.

- Along the hydrographic network we see small and big basins, selecting quite a range of different scales

- Basins are natural places for aggregation of water fluxes, flow suming up at branches, so that high flow and water resources do combine naturally in a non trivial.
- Man-made actions (targeting water use or water hasard) seems to be quite local, but develop on this background.



Introduction – Precipitation seen by an hydrologist : spatial structure matters !

- Water fluxes in hydrology are not a primary thing. A perspective is to see water fluxes as the transformation of atmospheric conditions by land processes.
- Land processes are largely non-linear in general.

- Most basic example :

  Say a soil is able to absorb 10 mm precipitation.
  What if rain is 10 mm on average ?
  What if rain is spatially 10+10 mm ? 12 +8 mm? 14+6 mm ?

- In hydrology working with averages may not suffice.
- To adress use precipitation we need to consider also the spatial variability.

Introduction – Precipitation seen by an hydrologist : spatial structure matters !

- To design and test management strategies, a present-day must-have playground is to set up a numeric model of the basin.
- This numeric model needs inputs from the atmosphere (precipitation, temperature).
- These inputs can be from the past :

  time-series, specific events, reanalysis…
- They can target the possible beyond the observed :
  - resampled/shuffled historical.
  - stochastic weather generators (SWGs) having a explicit parametric structure
    - these are useful to run long input sequences, offering a quite exhaustive view of how the hydrologic system transforms common climate variability.
- SWGs must respect the assessable structure of inputs intervariates, space and time.

Introduction – Precipitation seen by an hydrologist : spatial structure matters !

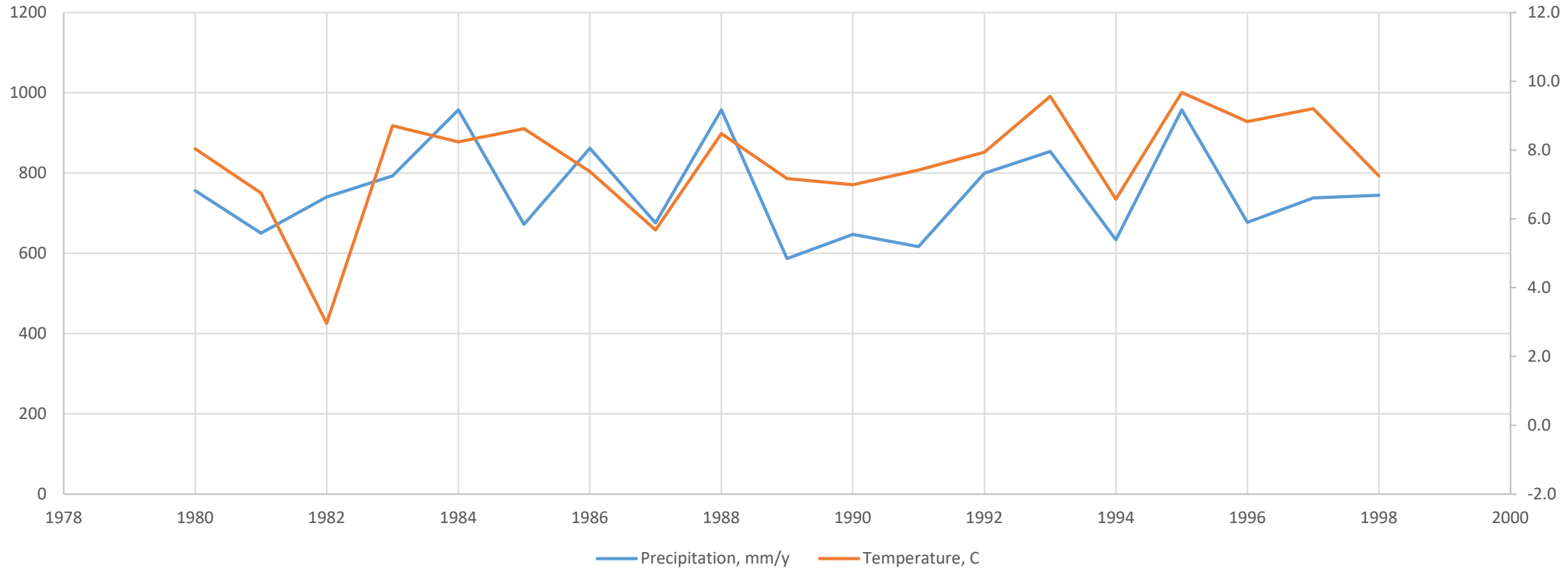
- A precipitation model targeting the needs of hydrology must respect the distribution of rainfall agregates over the said scales.
- One first step is to have expected value and variance correct for a range of spatio-temporal scales.
- For this we need to have the spatio-temporal covariance of precipitation correct, because *covariance governs the variability of sums*

remember $Var(X+Y) = VarX + VarY + 2.Cov(X,Y)$

$\Rightarrow$ This talk is about two algorithms useful in designing SWGs trying to take car of the above.

Algorithm #1
Variations around the multigaussian distribution

Using a minimalistic case : 2 variates sampled, yearly values at one place.



Building a statistical sample out of observed data

data as observed		
year	P	T
1979	**1096**	**11.4**
1980	**756**	**8.0**
1981	**650**	**6.7**
1982	**741**	**3.0**
1983	**793**	**8.7**
1984	**958**	**8.2**
1985	**672**	**8.6**
1986	**861**	**7.4**

data reshaped as 2-year statistical individuals				
	P_1	T_1	P	T
1979-1980	**1096**	**11.4**	**756**	**8.0**
1980-1981	**756**	**8.0**	**650**	**6.7**
1981-1982	**650**	**6.7**	**741**	**3.0**
1982-1983	**741**	**3.0**	**793**	**8.7**
1984-1985	**793**	**8.7**	**958**	**8.2**
1985-1986	**958**	**8.2**	**672**	**8.6**
1986-1987	**672**	**8.6**	**861**	**7.4**

A note on inference

- From this we observe a covariance matrix C

The covariance matrix

	P_1	T_1	P	T
P_1	4	2	1	1
T_1	2	3	1	1
P	1	1	4	2
T	1	1	2	3

same year — successive years

P was divided by 100 (scaling)

- C is positive definite.
  It admits a Cholesky decomposition : it exists one only lower triangular matrix L, so that $C = L.L^T$

L, the Cholesky matrix to C

2.00	0	0	0
1.00	1.41	0	0
0.50	0.35	1.90	0
0.50	0.35	0.85	1.38

Bridge to multivariate autoregressive process perspective

- Standard use of Cholesky matrix L is to multiply it by a vector of independent N(0,1) variates. The resulting vector has covariance C.
- So Cholesky is instrumental in simulating a given dependence.

- Let us consider a new matrix L_1 « first year binding » having the first half lines as in L and the last half lines as in identity matrix I_n.

L1, first-year binding matrix					
2.00	0	0	0		as in L
1.00	1.41	0	0		
0	0	1	0		as in Id
0	0	0	1		

We want to simulate a multivariate time serie

- Let $A=(L_1)^{-1}\,L$
- Given
  - one-year data
  - Independent N(0,1) values for a second year

  Matrix A unbinds the one-year data back into independent N(0,1), then builds a two years samples following C.
- So A is instrumental in extended one year into a next one.
- Looking at A, we find that
  - Down-left in A is a one-year autoregression matrix AR
  - Down right in A is a Cholesky noise matrix N
  - $N.N^T$ is the part of covariance within the 2nd year not accounted by the autoregression (Schur complement).

A, the matrix to extend one year into two

1	0	0	0
0	1	0	0
0.500	0.354	1.904	0
0.500	0.354	0.853	1.377

AR

0.500	0.354
0.500	0.354

N

1.904	0
0.853	1.377

$N.N^T$

3.625	1.625
1.625	2.625

Total variance within one year

4	2
2	3

Link with geostatistics

- In the covariance matrix C is all to make a simulation of the 2nd year conditional to the 1st.
- Now using geostatistical words
  - Having half of n values, you make a simple kriging of one variate (n/2)+1
  - Adding a random deviation based on the kriging variance makes a conditional simulation.
  - Now you go to the next variate (sequential gaussian simulation) until the 2nd year is complete.

- Using statistical words : this is chain rule for multivariate simulation
  - P(X)
  - P(X,Y) = P(Y|X)*P(X)
  - P(X,Y,Z) = P(Z|(X,Y)) *P(X,Y) = P(Z|(X,Y)) *P(Y|X) * P(x)

- Sequential gaussian simulation is essentially the same as multivariate AR based on the same underlying C.
- Then you can shift one year and go for a 3rd year : the AR perspective makes long simulations possible were usually geostatistics do not go.

Link with copula perspective

- Copulas simulate directly vectors with specific link between their marginal cdf.
- Matrix C here is directly related to a n dimensional Gaussian copula.

- We notice that the built-in capacity of copula to simulate some variates knowing the others ones - not the first use of copulas.

- Copula research insists that
  - gaussian copulas is just one of many possibilities.
  - resigning the gaussian copula (were covariance serves as the measure of dependence) is a price to pay ; if we accept it we can independently study marginals distribution and the link between variates.

- The Gaussian copula is the most easy way for a geostatistician to enter the modern world of copulas and understand he/she can survive and possibly contribute there.

Lessons from algorithm #1

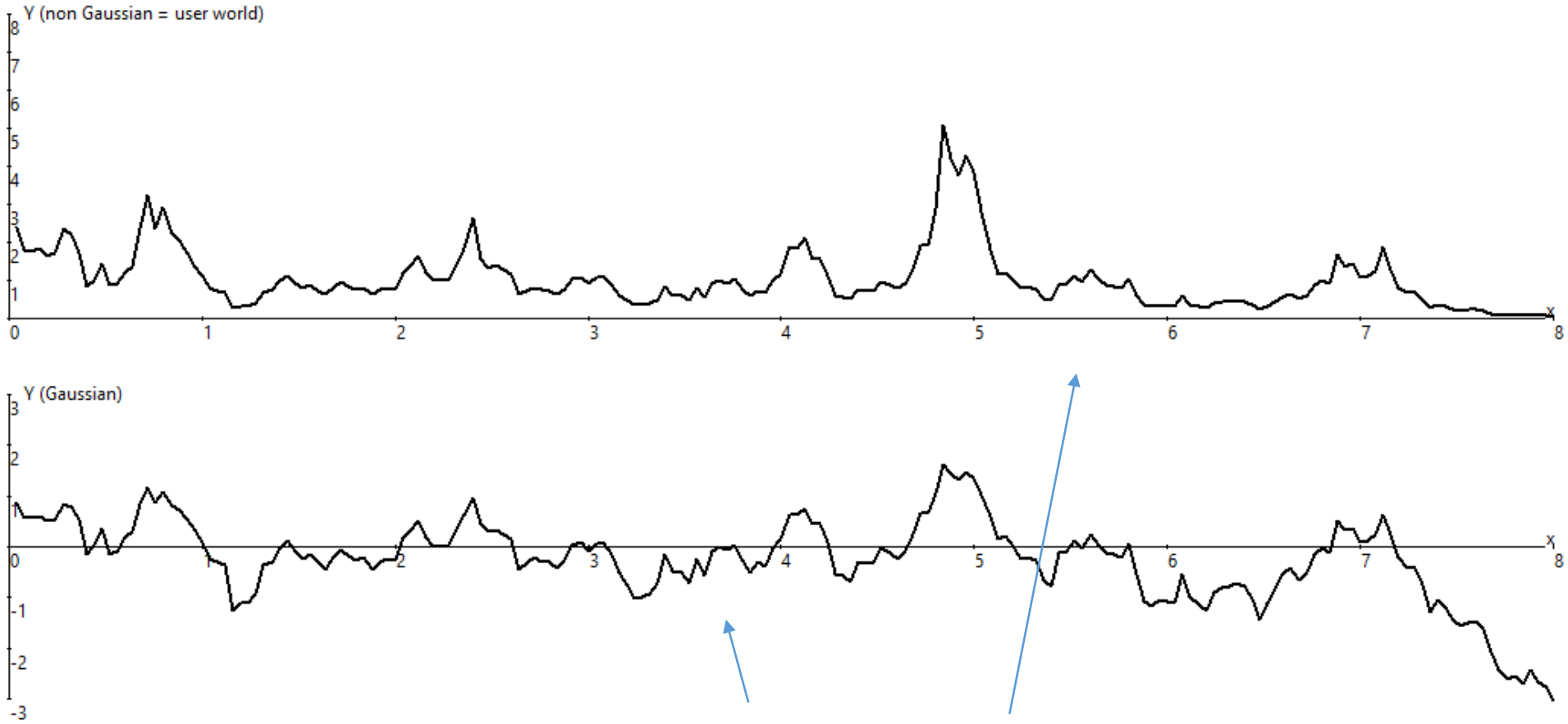
- Having the covariance matrix at their core,

  multivariate AR *geostatistic simulation* *gaussian copula*

  perfectly coincide.
- Combining the various perspectives helps to understand any of these and apply them with flexibility and opportunity.
- Understanding the basic mathematics that makes versatility possible. Using ready packages, certainly powerful and time saving, may restrict the user to predefined « use cases ».

Algorithm #2
Non-deterministic desagregation of precipitation

Preliminary : how to desagregate bloc values in a non gaussian context ?

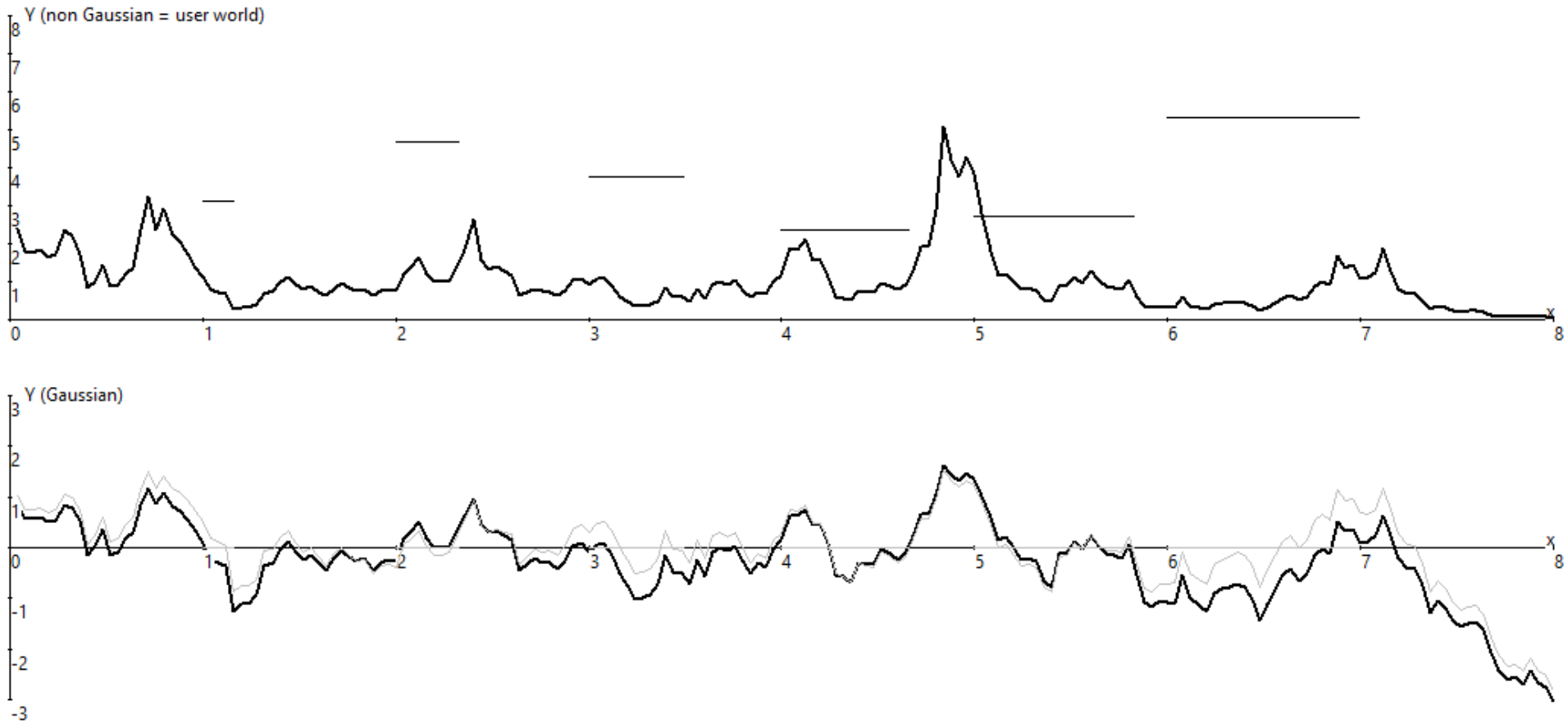


(2)

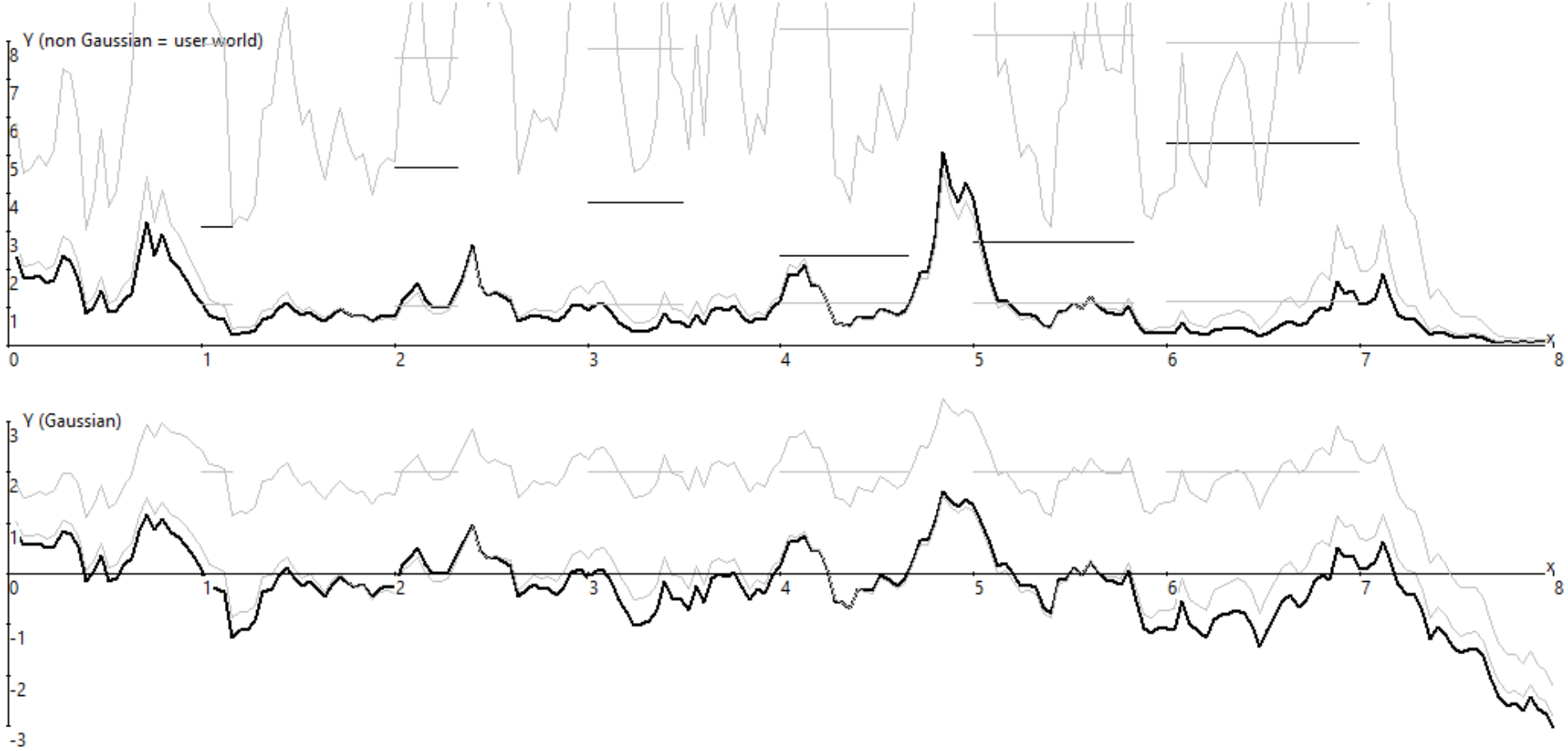
(1)

We start from a Gaussian free simulation (1) we translate into real-world engineering values (2)

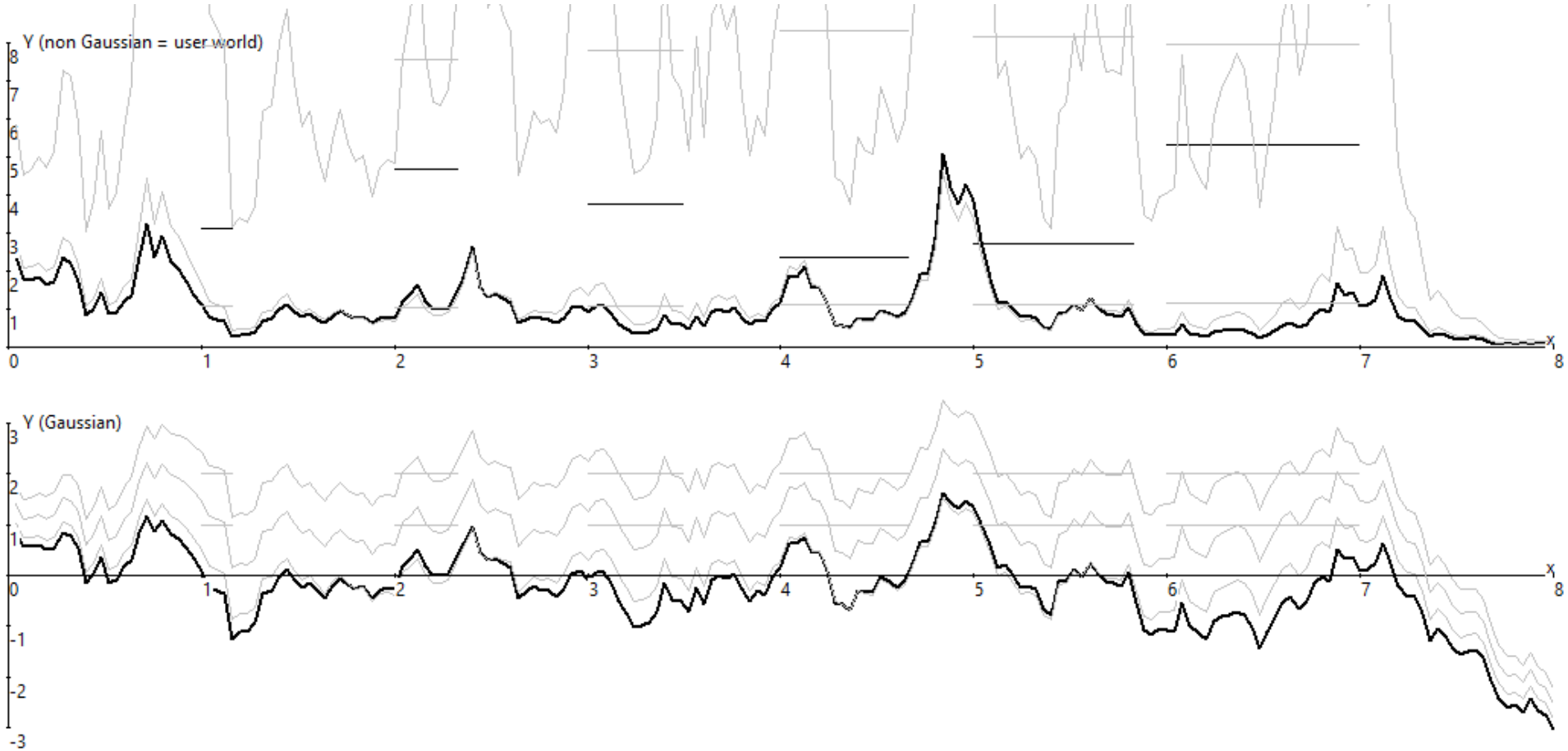
We want to respect bloc values in user world.



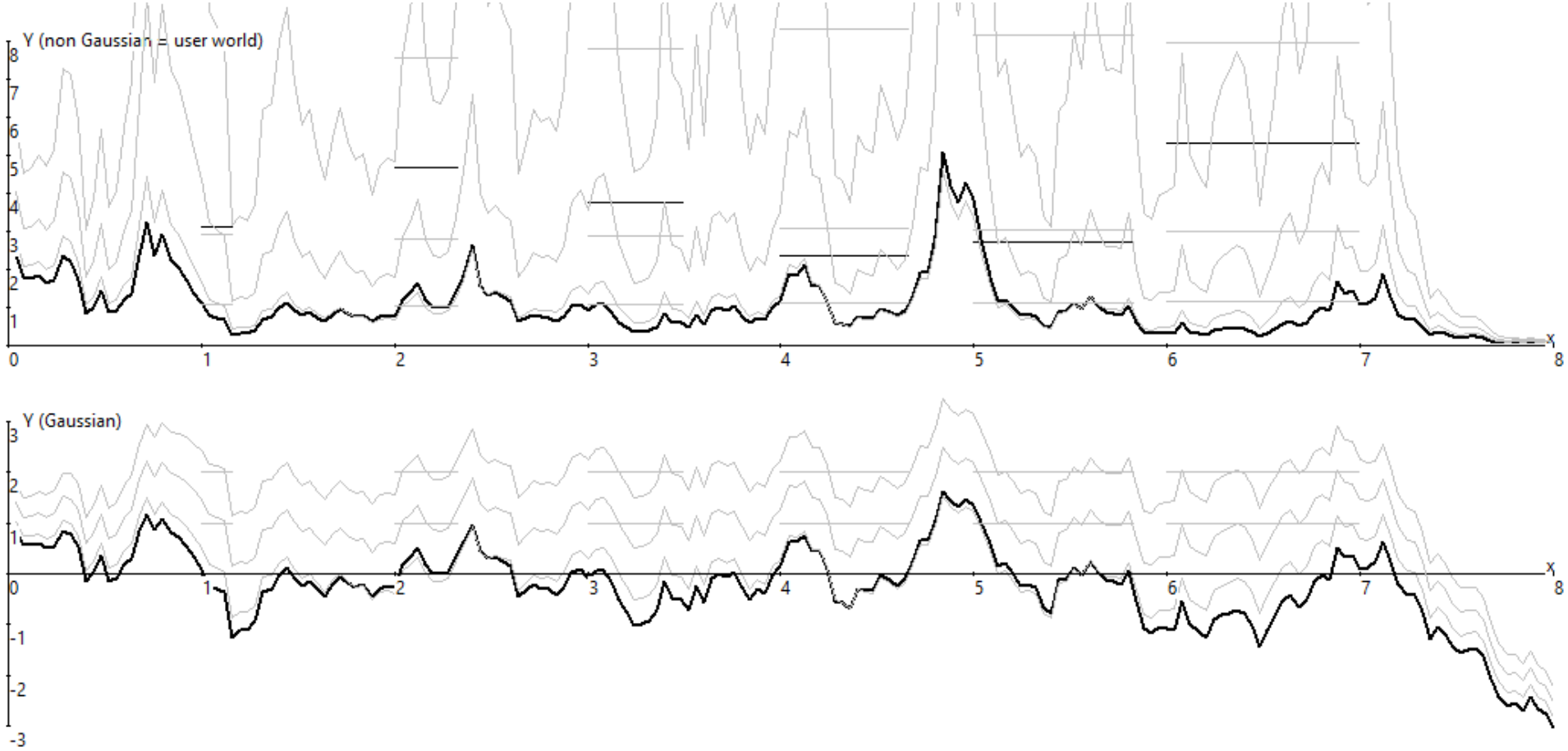
We know (using block kriging) how to enforce Gaussian block averages but they are not in a 1-1 relation to real world averages.



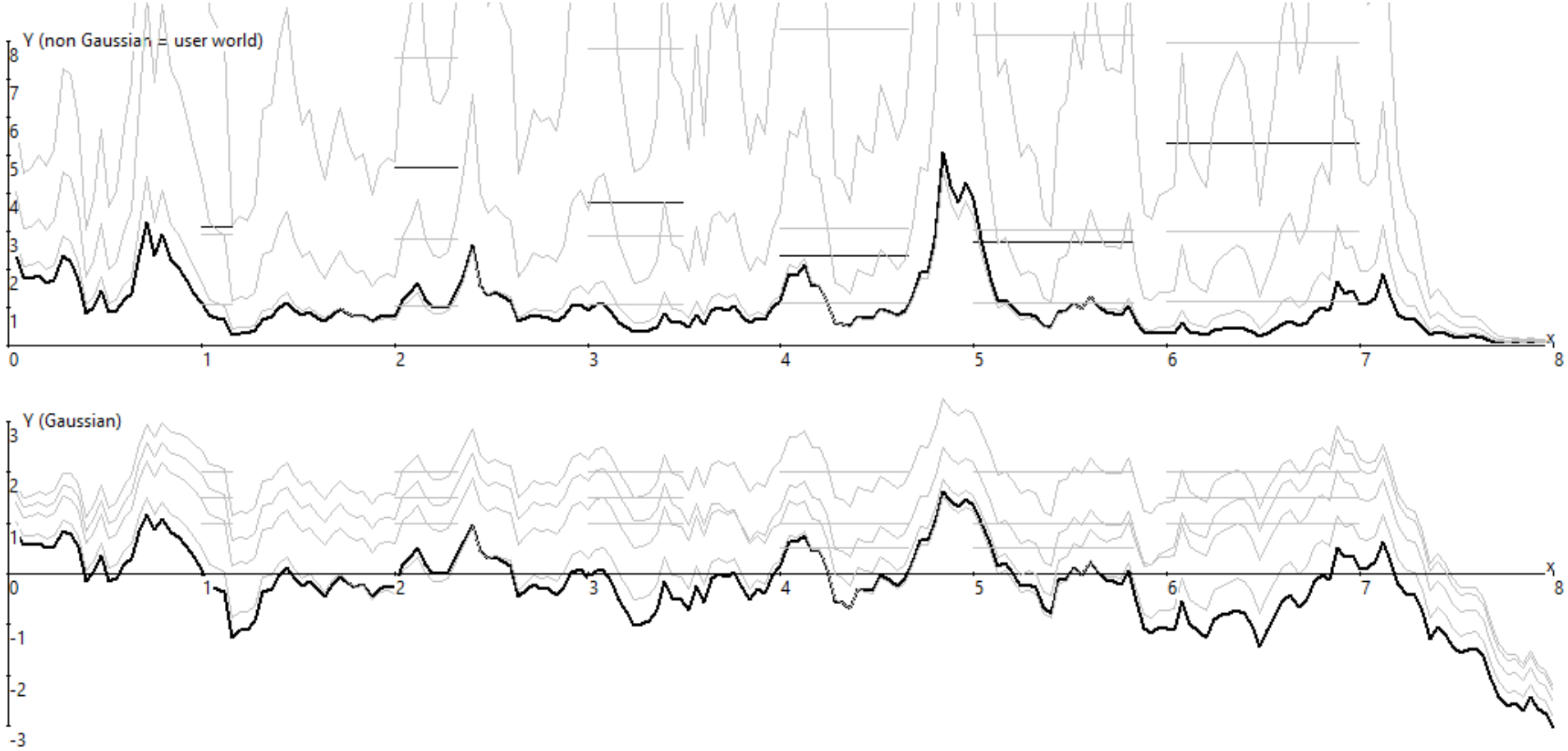
Arbitrary block values in Gaussian world will translate too high or too low in the user world.



Fortunately there is a positive relation between gaussian world averages and real world averages

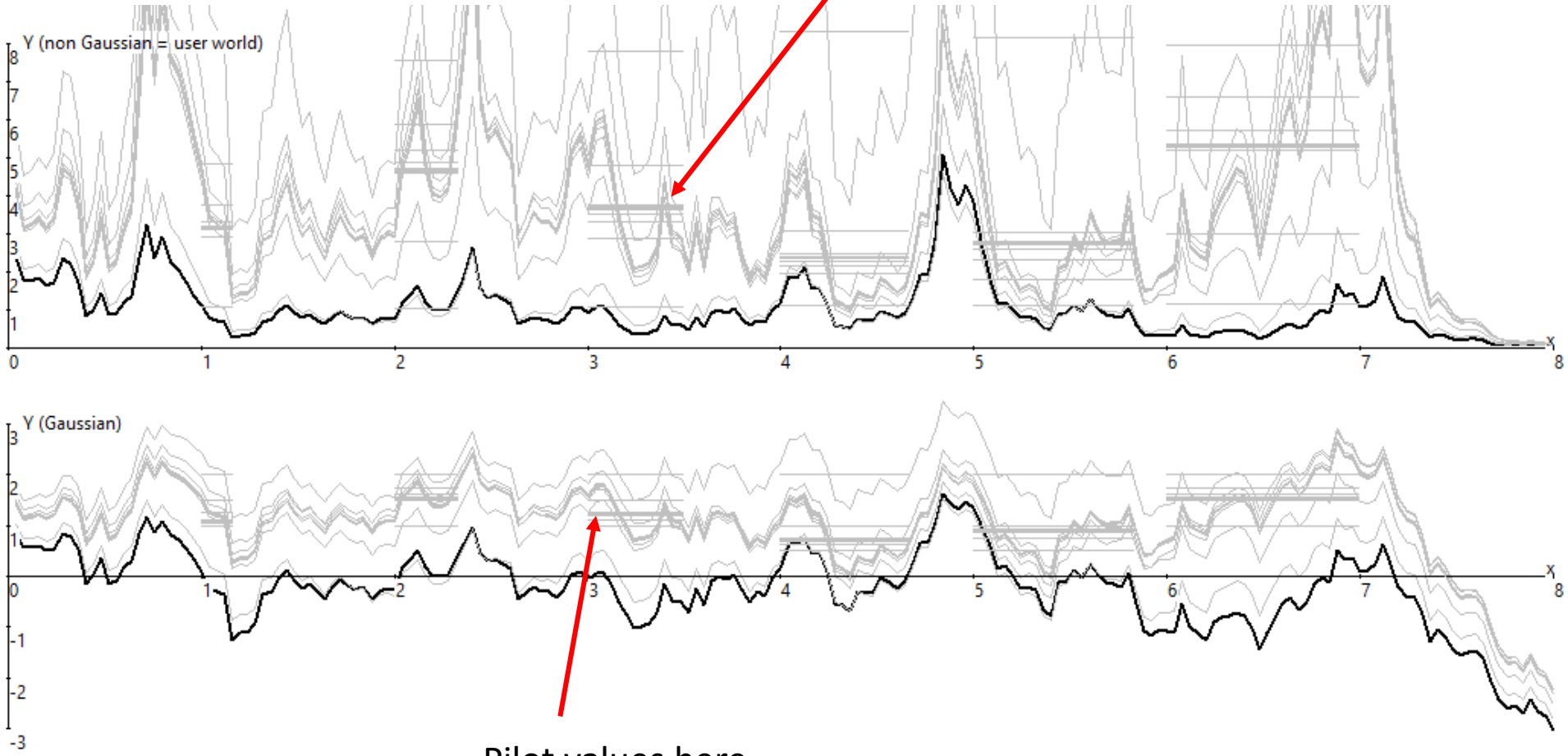


So we can use them as *pilot values* to guide the simulation to the target (De Marsily, 1986)



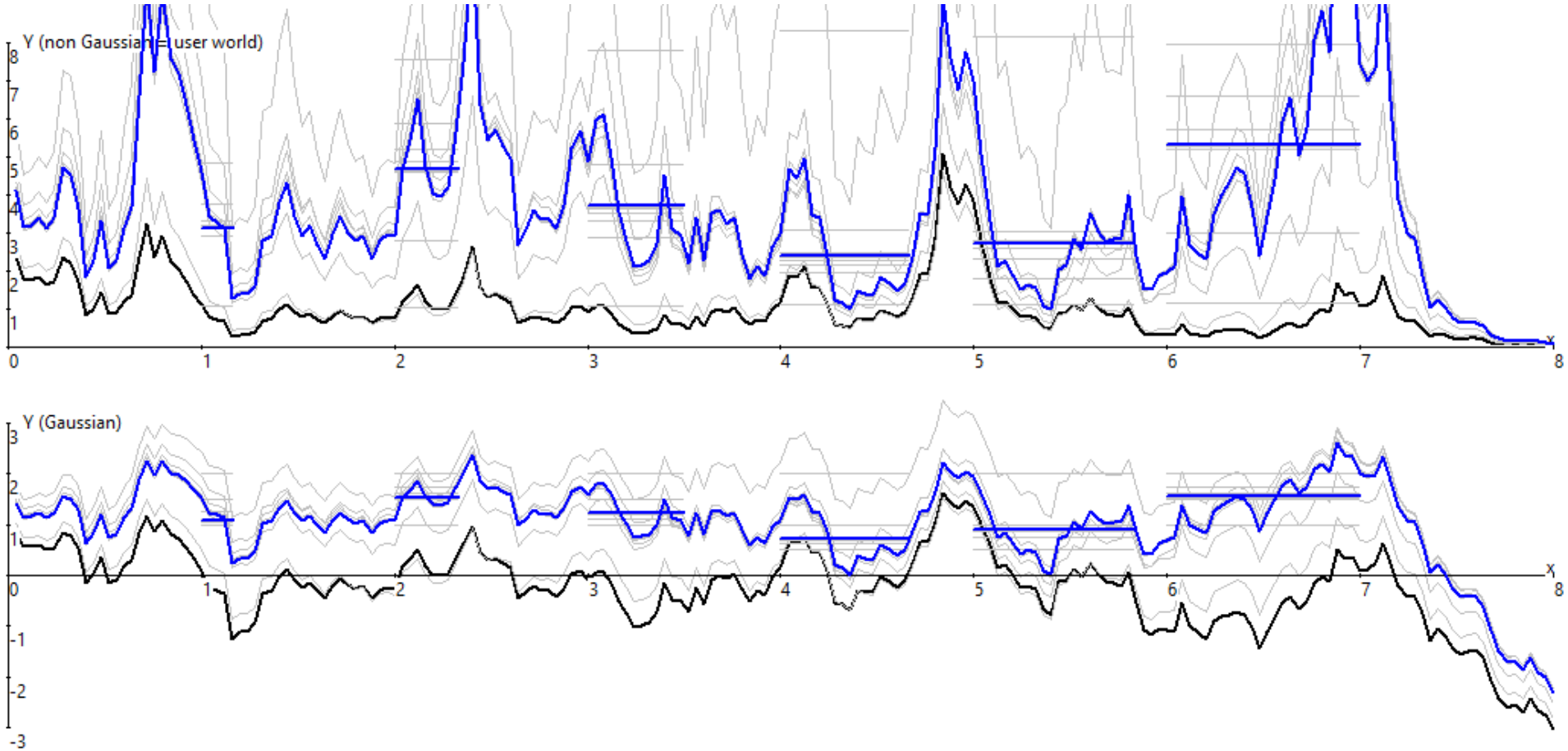
..this mean optimization…

Assessment here

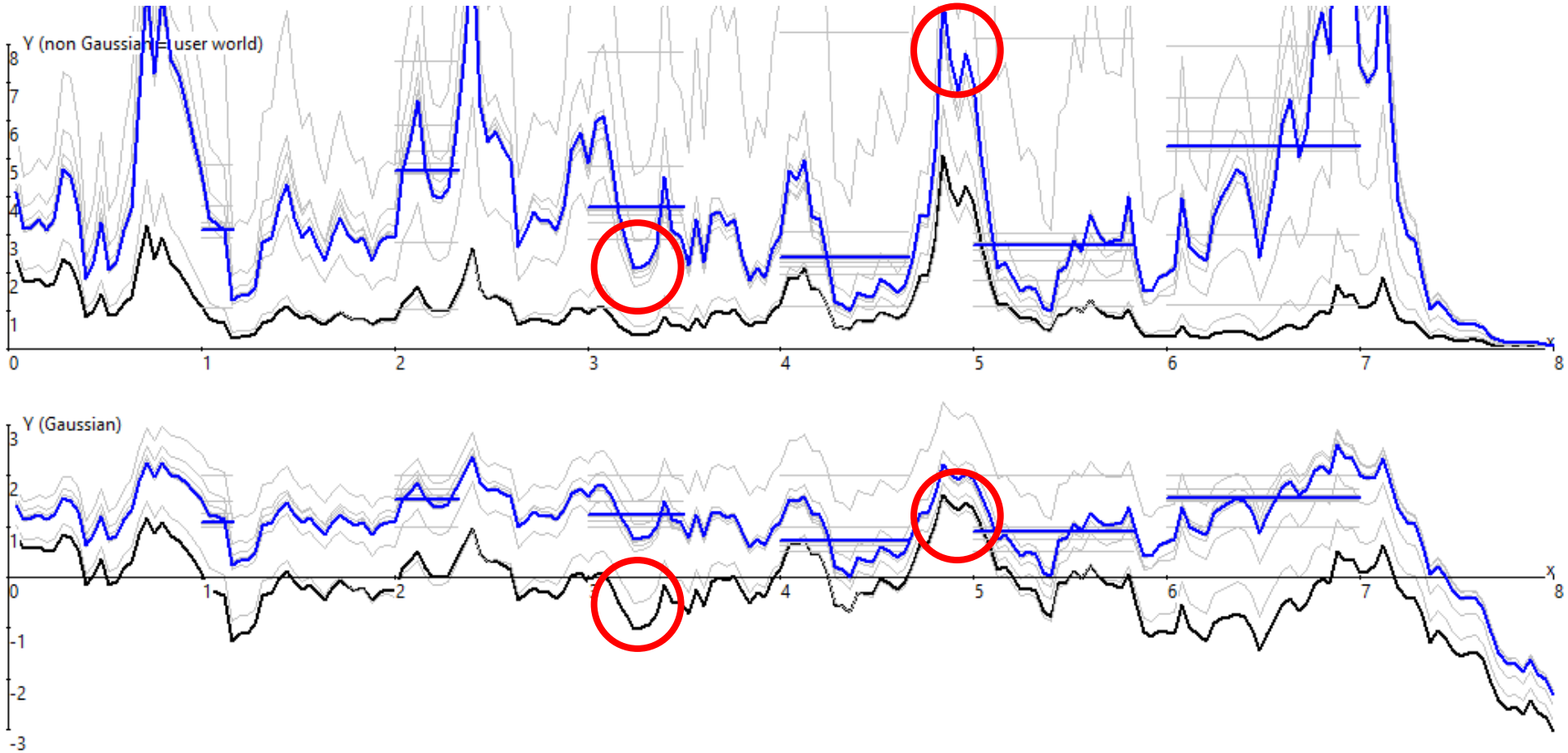


Pilot values here

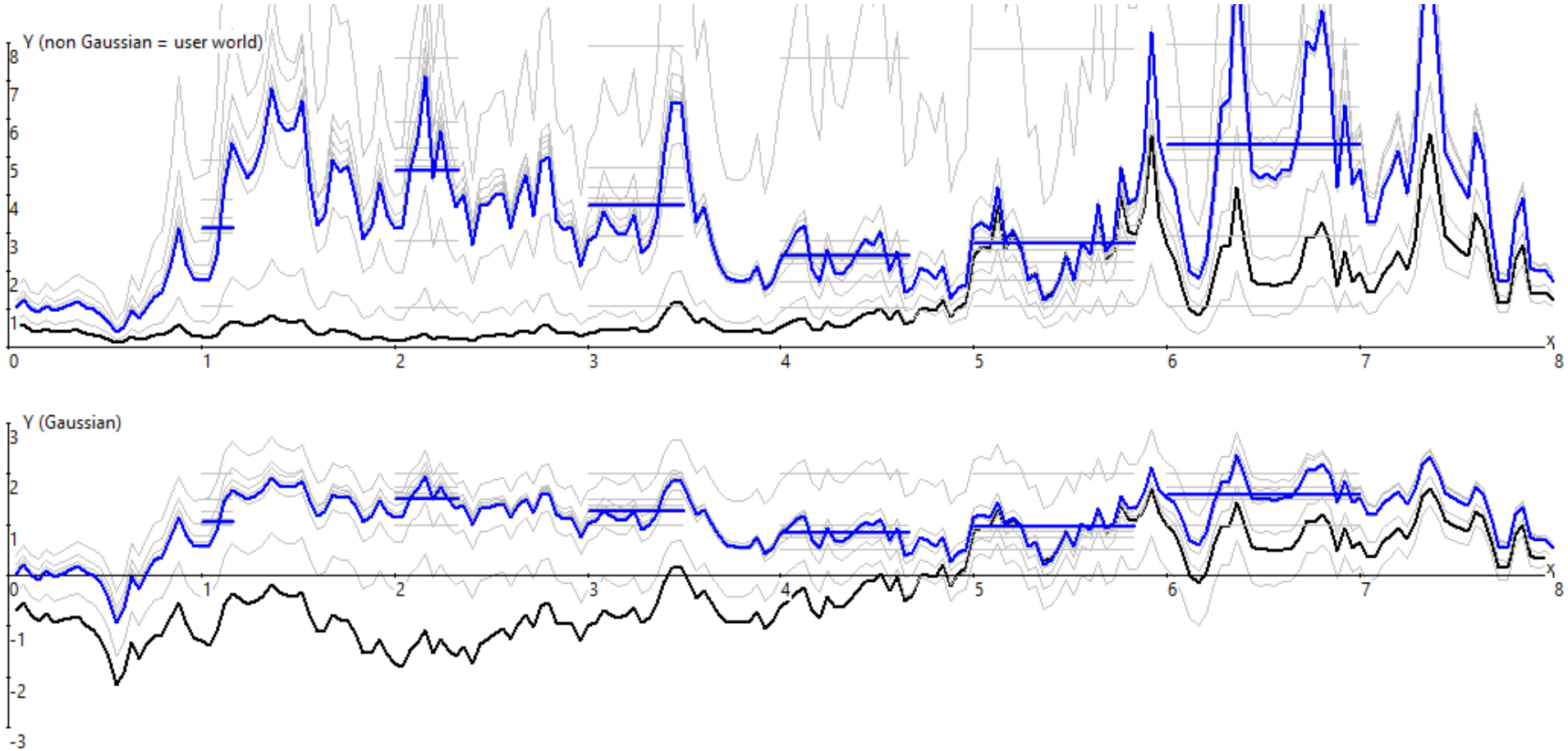
Finally the job is done (and we use blue ink)



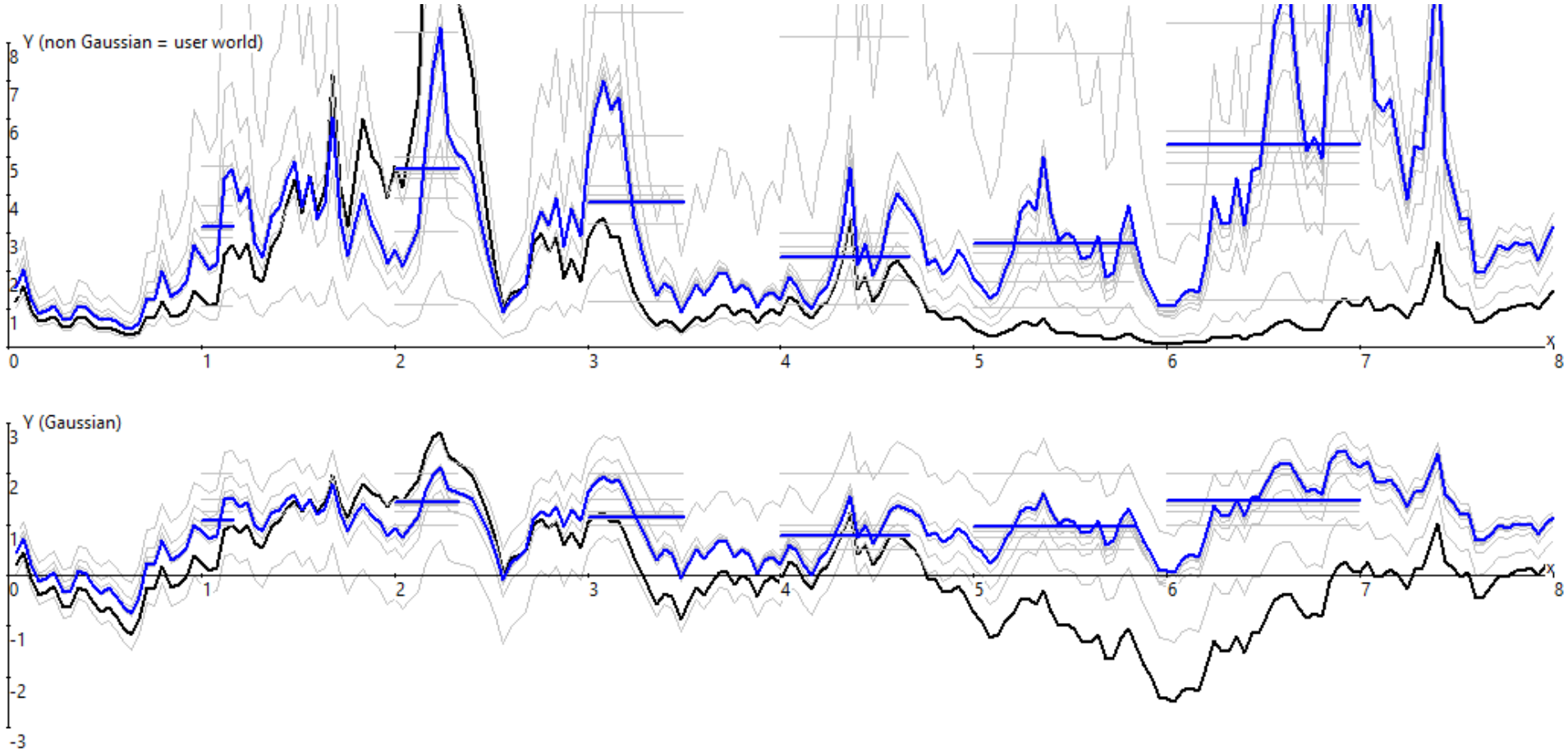
Still, the build was initiated from a free simulation that reflects in the results ?



Then other solutions are possible…



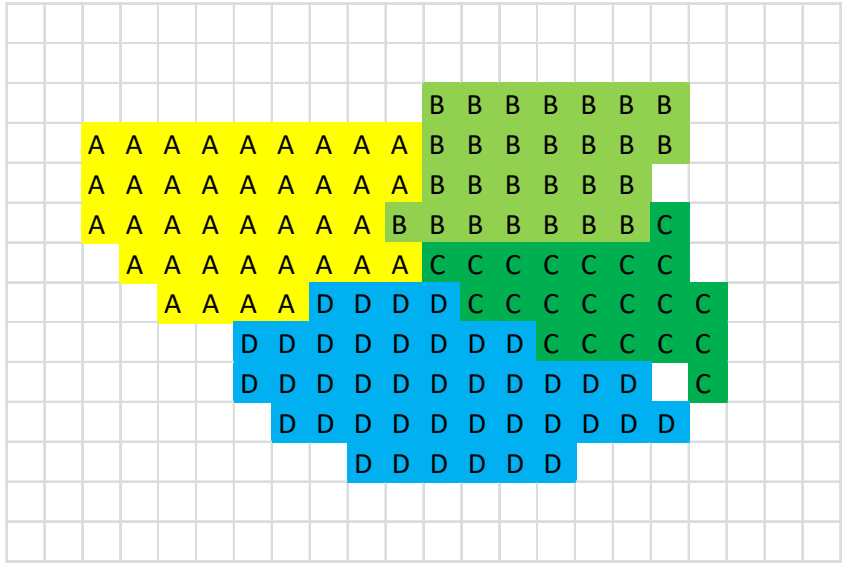
… infinite many in fact, sharing the parental variability as filtered by the constraints



Now we adapt to precipitation disagregation

- Large scale averages (satellite, GCM, scenarii) known for precipitation (mm) and wetness (0-1)

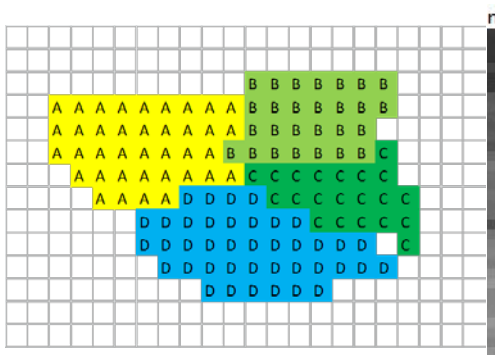
	Area => Time step	A	B	C	D
Wetness [0-1]	1	0.4	0.6	0.7	0.8
	2	0.3	0.8	0.6	0.9
	…				
precipitation (mm)	1	1	1.2	2.3	3.2
	2	1.5	1.3	2.1	2.8
	…				

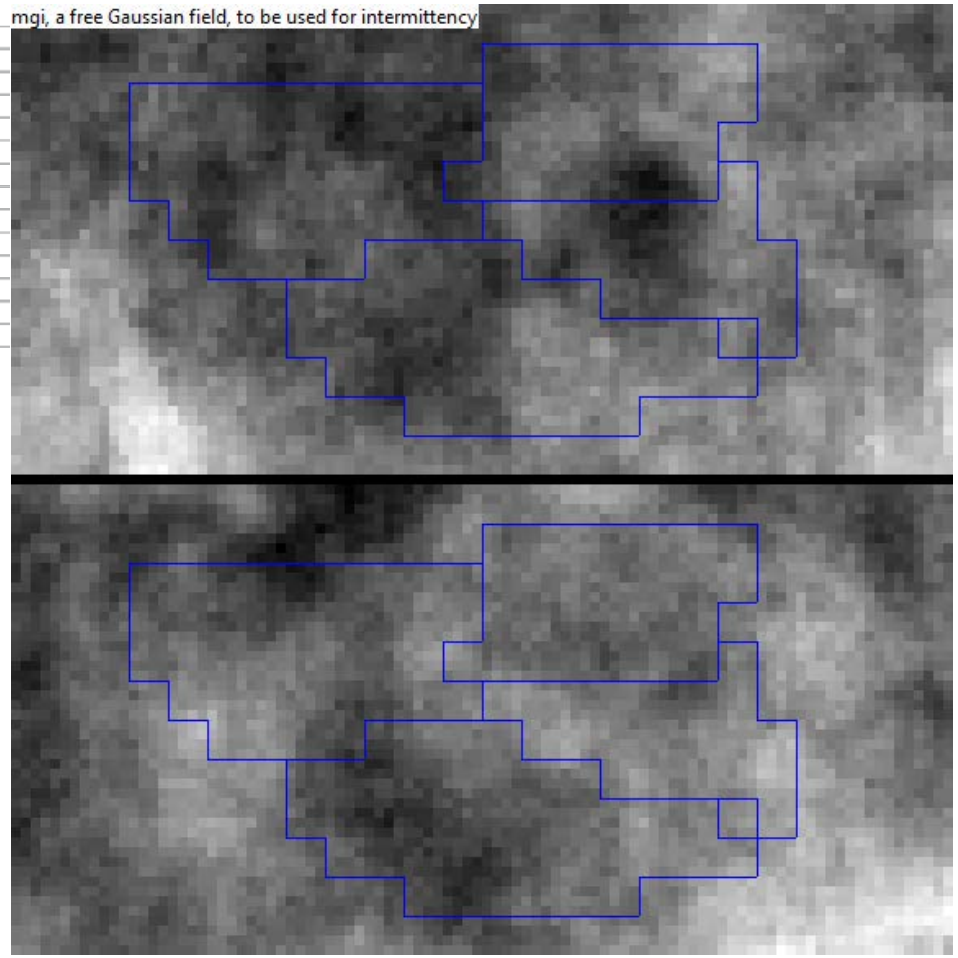


- We want to simulate a suitable small scale
- Variability is assumed known at small scale (say, from local raingauges)

- We use the basic algorithm twice
  - Once to get a wet-dry pattern respecting given areal wetness
  - Once to get an intensity pattern respecting given areal precipitation.

A - Simulated 3d gaussian field #1

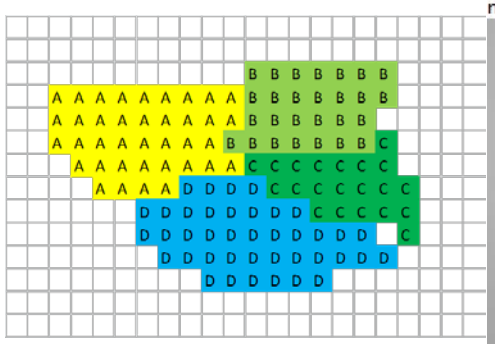


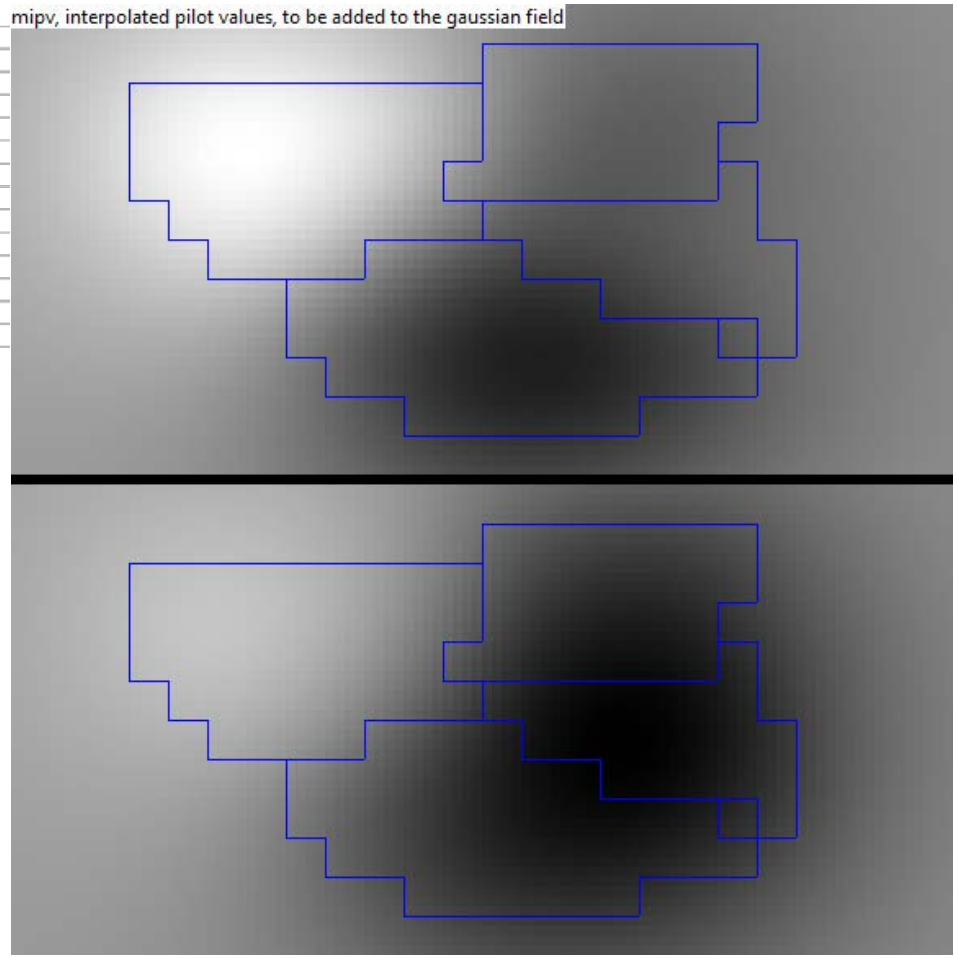


Time 1

Time 2

B – optimised large scale « pilot values » here interpolated as a grid of local deviation to add…

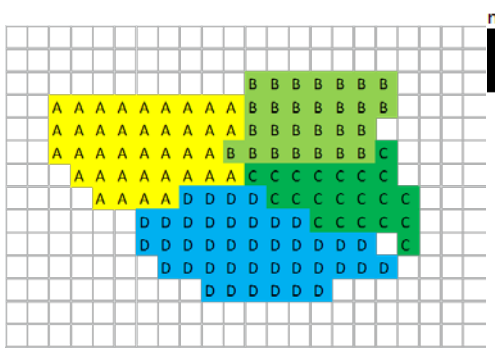


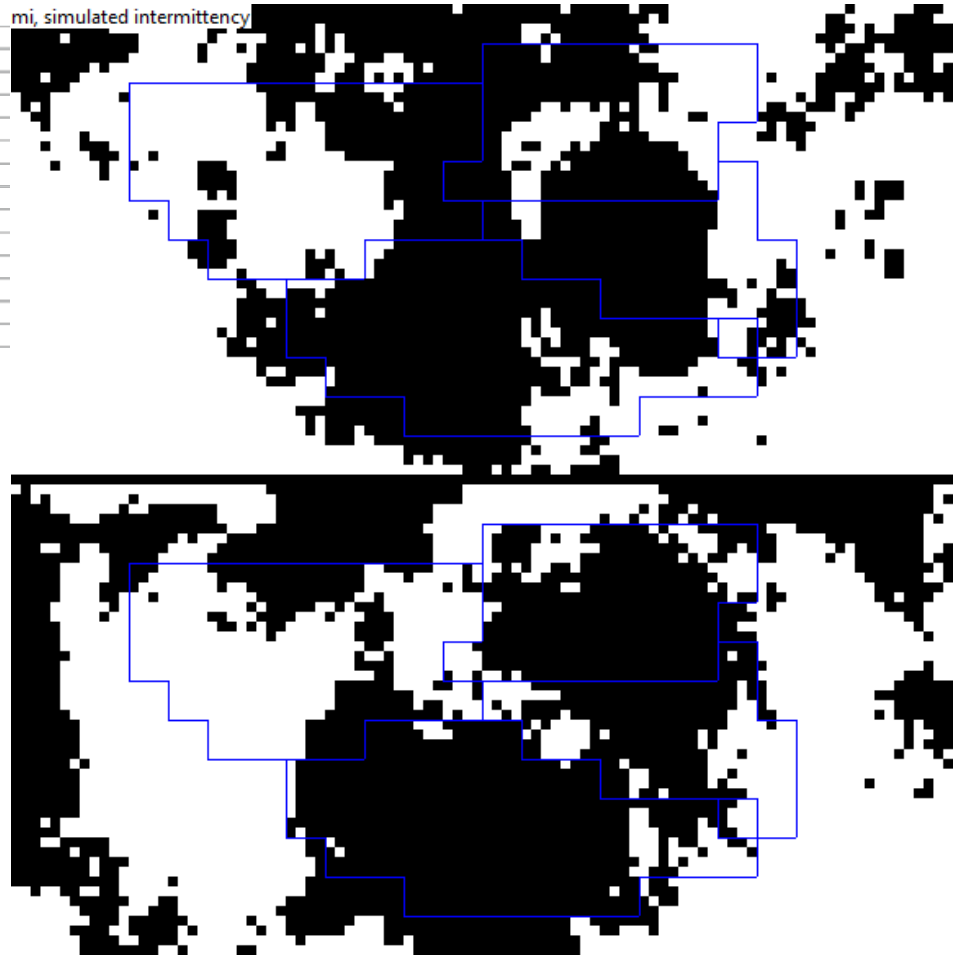


Time 1

Time 2

C – so that the rainy area mask we get (tresholding the sum) respects the expected wetness

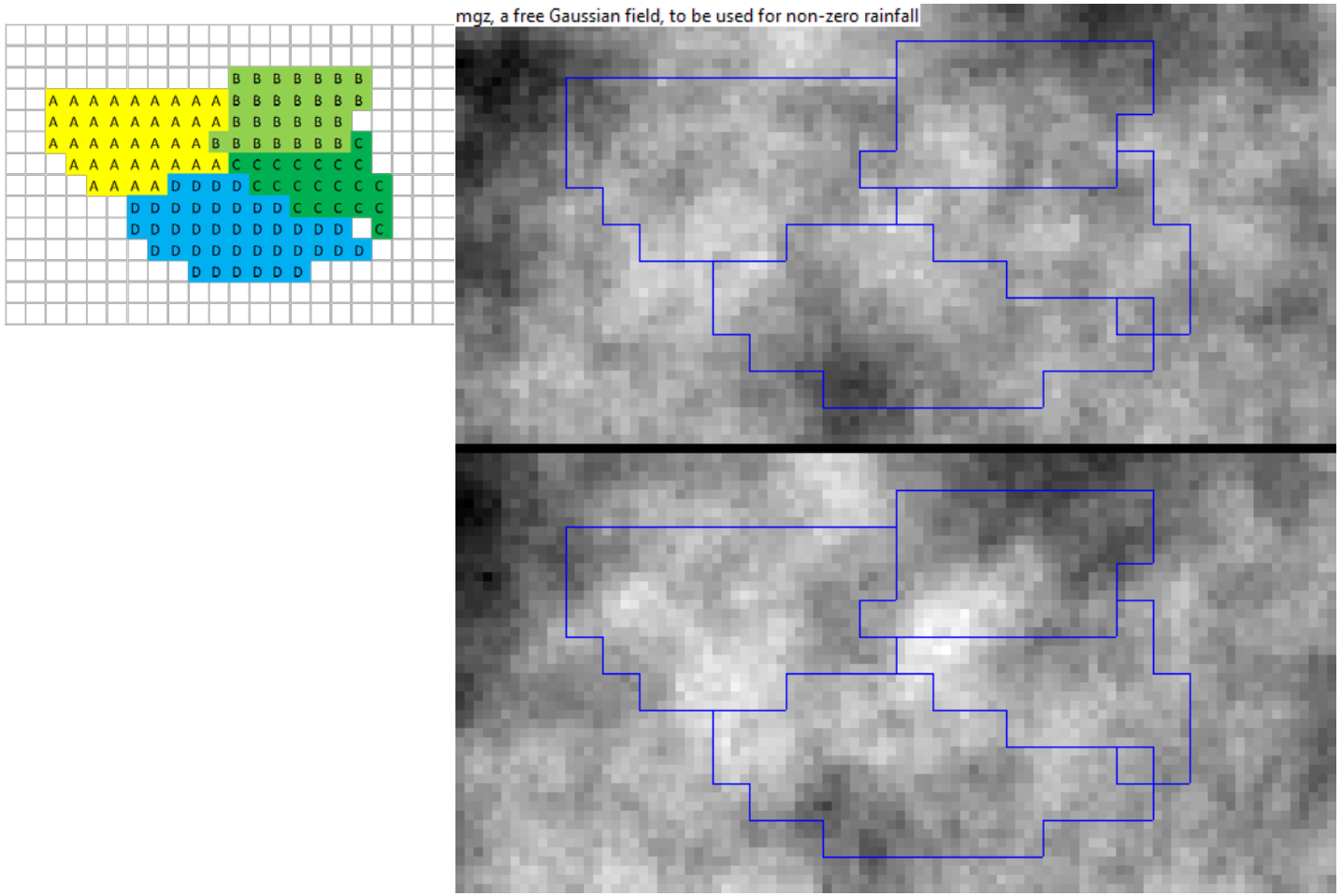




Time 1

Time 2

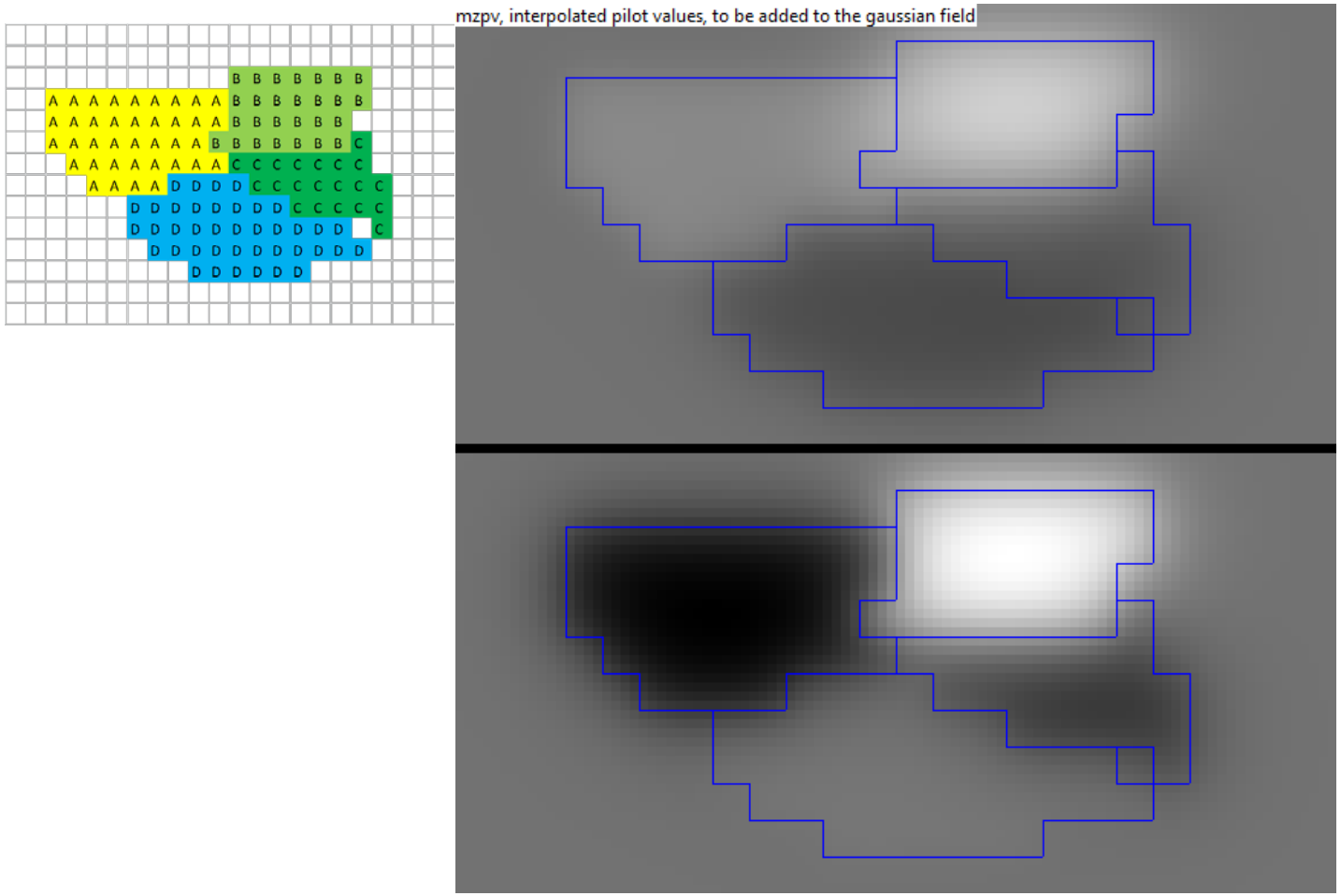
D – Simulated 3d Gaussian field #2



Time 1

Time 2

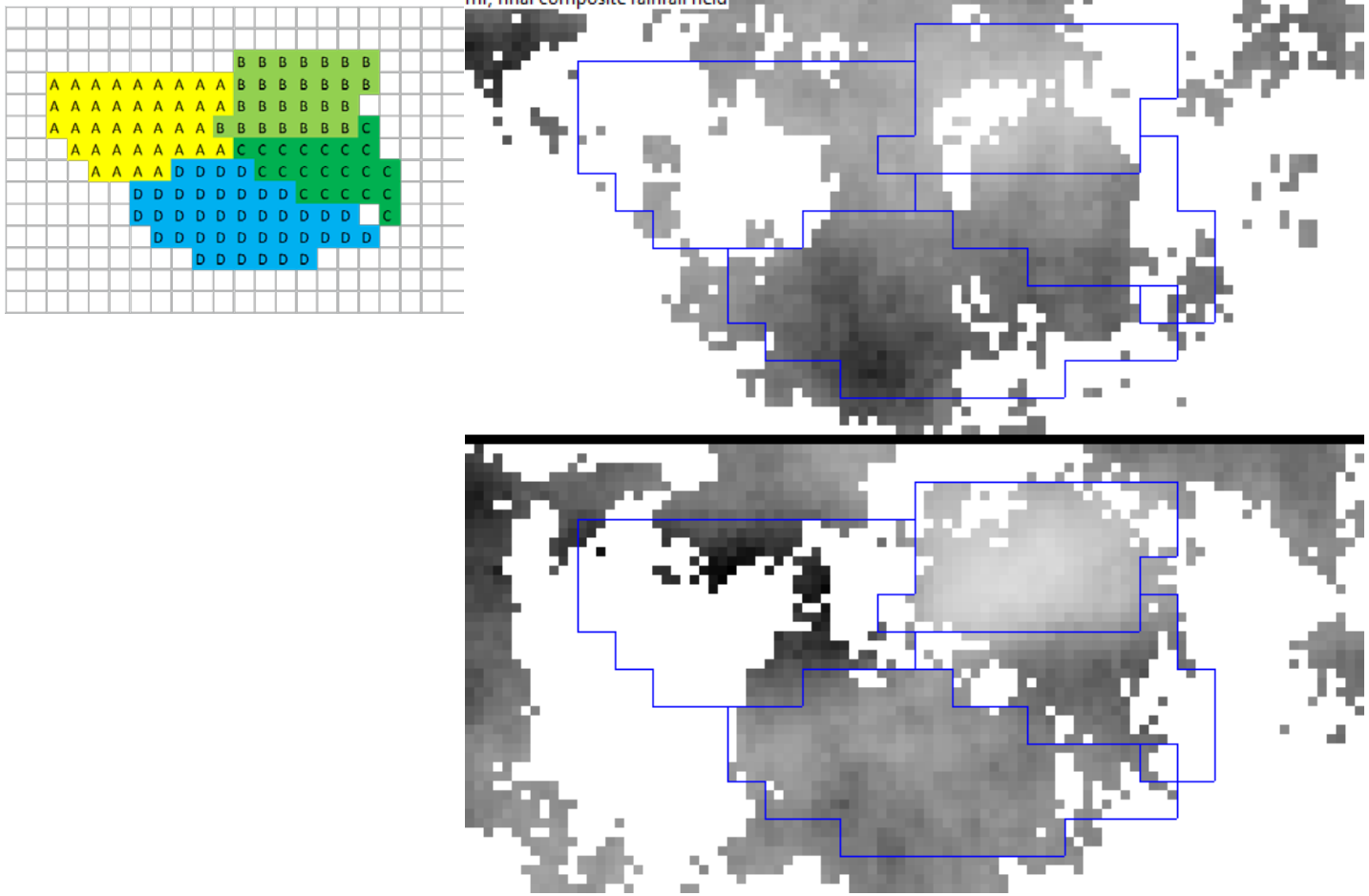
E – and interpolated bloc deviations



Time 1

Time 2

F – …so that rainfall quantities (as sum over rainy areas) match the given values



Time 1

Time 2

Lessons from algorithm#2

- Disagregation needs to supplement large scale information with a model of small-scale variability.
- The disagregation is non-deterministic, in practice uncertainty is expressed by an ensemble of equiprobable solutions.
- Each member is like a possible reality. We can process it across a deterministic model to study consequences and their uncertainties.
- Such a model can be as physically based and as non-linear as the real world is.

Wrap up

Start	Nice hydrologic models are not enough to understand / optimize / secure the behaviour of the water flow in a catchment, if uncontrolled input interacts with non-linear features at ground.
Perspective	Full detail of real world features is intrinsicaly beyond measurement. We shall always need to cope with unsolved variability.
Suggestion	To mimick any unsolved variability shall give at least a probabilistic control on it
Assessment	Techniques are available, of course approximate, yet give pictures of the real world retracing what we know and what we do not know.
	Uncertainty is replaced by ensembles that are convenient to process.
	Developpement, calculation and storage burden are not negligible. It may be acceptable in a team work, if colleagues accept preprocessing of input data.
Hydrologists	Should contribute because of their understanding of the part of variability significant to them.
	Can participate to build an interface between climatology/meteorology and hydrology/water resource/water hasard management
	It would be unfair and inefficient to let meteorologists (or mathematicians) make all the job alone and thereafter complain the result is inadequate to our needs.

Note on the large size of matrices

Large matrix may arise in alg #1

- Let some variables of interest

  (say 2, Precipitation, Temperature)
- Monthly values taken as basic decriptor to include seasonnality

  (say 12 values per year, 24 for 2 years so as to include autoregression)
- Over several areas

  (say 10 sub-areas to a major basin)
- Then n=2 x 24 x 10 = 480

- As C is assessed using a limited observation record, $C\tilde{}$ will have rank << n.
- To get rid of eigenvalues zero, we need to load the diagonal.
  - Adding a small constant to the diagonal (Tikhonof regularization)
  - Equivalent to : using the same observation again and again adding independent noise to the data.
- This makes the covariance matrix C definite positive, so the Choleski decomposition possible