International Doctoral Program Environment Water

International Research Training Group nupus

Application of Copulas as a New Geostatistical Tool

Presented by

Jing Li

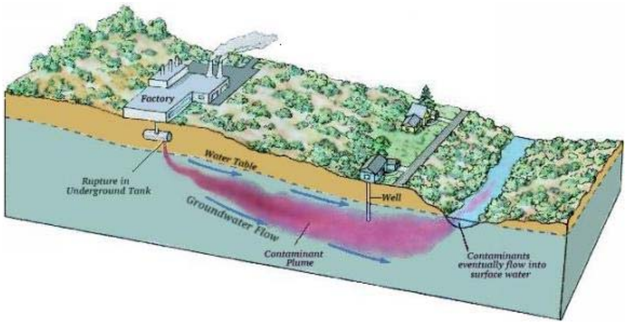
Supervisors

András Bardossy, Sjoerd Van der zee, Insa Neuweiler

Background and Motivations

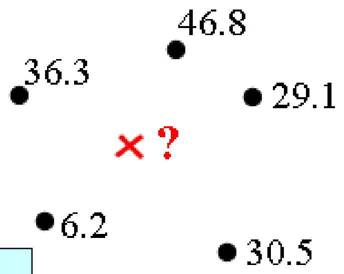
Tasks of environmental engineer or hydrogeologist
- estimation of natural processes

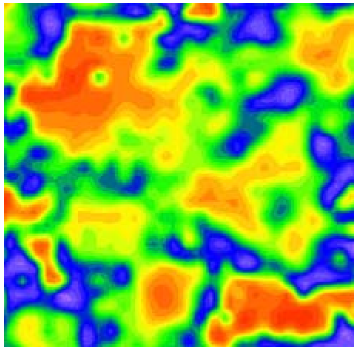
Spatial variability, complexity of natural process
- Geostatistical methods



Available information ? – information needed

Spatial dependence:
$x(\mathbf{s}) = f(\ x(\mathbf{s}_i),\ i=1,...,n\ )$
f()?, f() – spatial configuration





Spatial interpolations, spatial simulations
- decision making

Background and Motivations

Problem of Traditional Geostatistics

Variogram as the sole descriptor of dependence:

- two point statistics, averaged dependence, susceptible to outliers

Interpolation and simulation:

- Gaussianity assumption (symmetrical and minimum spatial continuity for extremes)

Kriging variance for uncertainty analysis:

- measurement density (not value-dependent)

Aim of this PhD work

Develop a strategy of using the concept of copulas as a better alternative to the traditional geostatistics for spatial modeling.

Outline of the Research Work

Model Building

- Using copulas to describe the spatial dependence and apply scale-invariant and higher order dependence measures
- Derive theoretical copulas for spatial modeling
- Develop an appropriate model inference approach

Applications

- Develop Interpolation approach using copulas
- Simulate random fields with non-Gaussian dependence
- Using copulas to guide observation network design for environmental variables

Outline of the Research Work

- Using copulas to describe the spatial dependence and apply scale-invariant and higher order dependence measures
- Derive theoretical copulas for spatial modeling
- Develop an appropriate model inference approach

Model Building

- Develop Interpolation approach using copulas
- Simulate random fields with non-Gaussian dependence
- Using copulas to guide observation network design for environmental variables

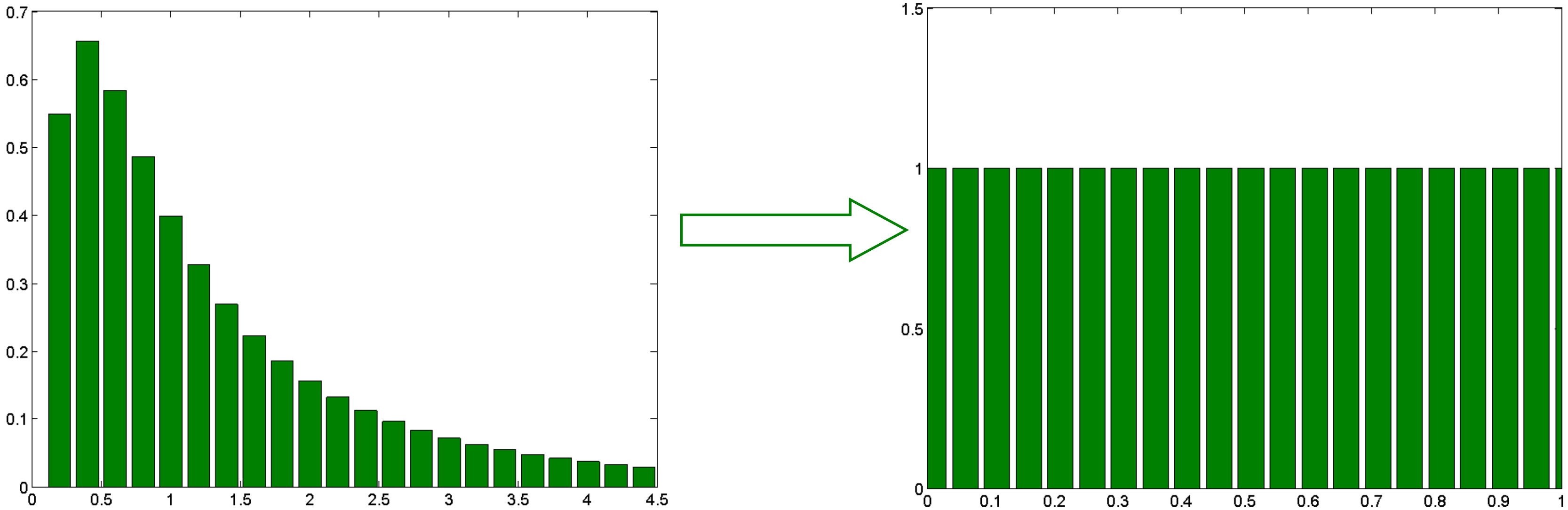
Applications

Copula and Spatial Dependence

Definition of copula

- Copula is a standardized multivariate distribution with all univariate margins being uniformly distributed on [0,1]:

$$C : [0,1]^n \rightarrow [0,1]$$



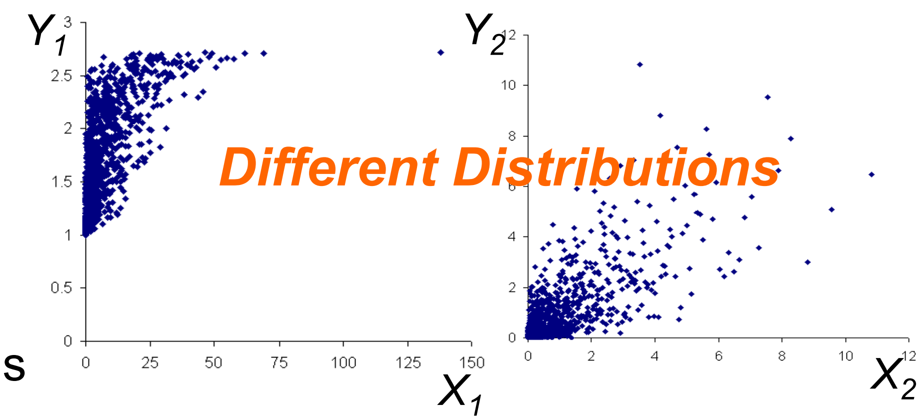
Copula and Spatial Dependence

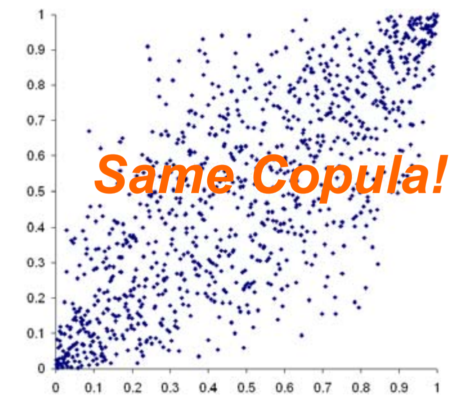
Advantage of using copula

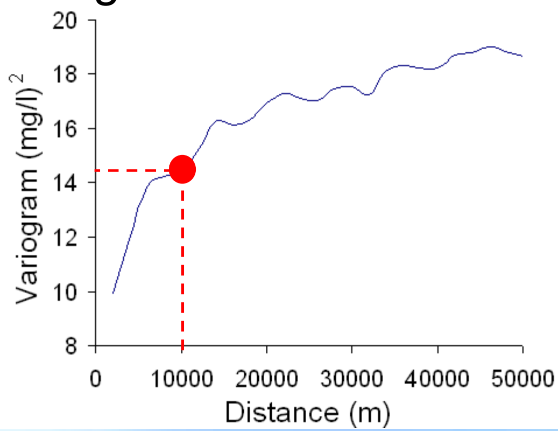
- Captures the pure dependence of RVs without the influence of marginal.

- Scale invariant : no problem for outliers and data transformations

- Full distribution: more informative than variogram







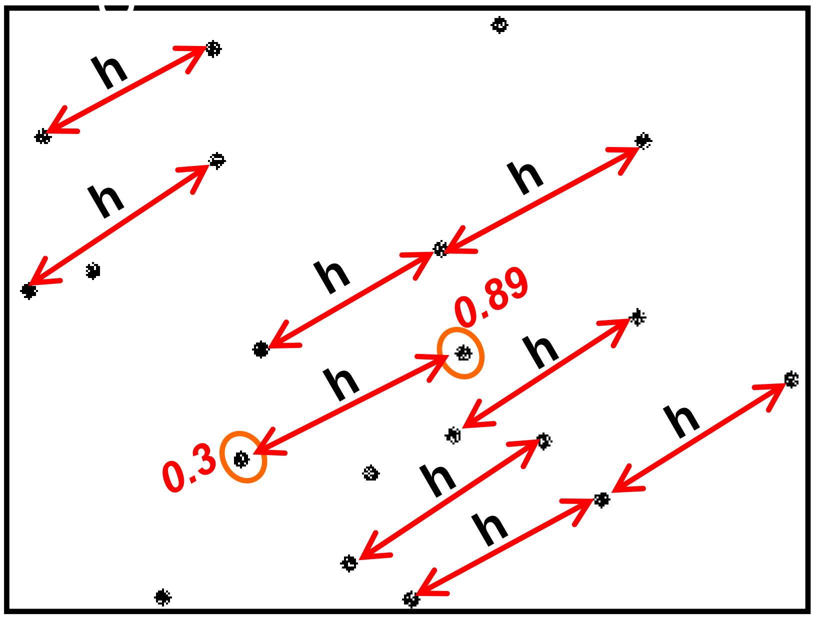
Copula and Spatial Dependence

Empirical bivariate spatial copula

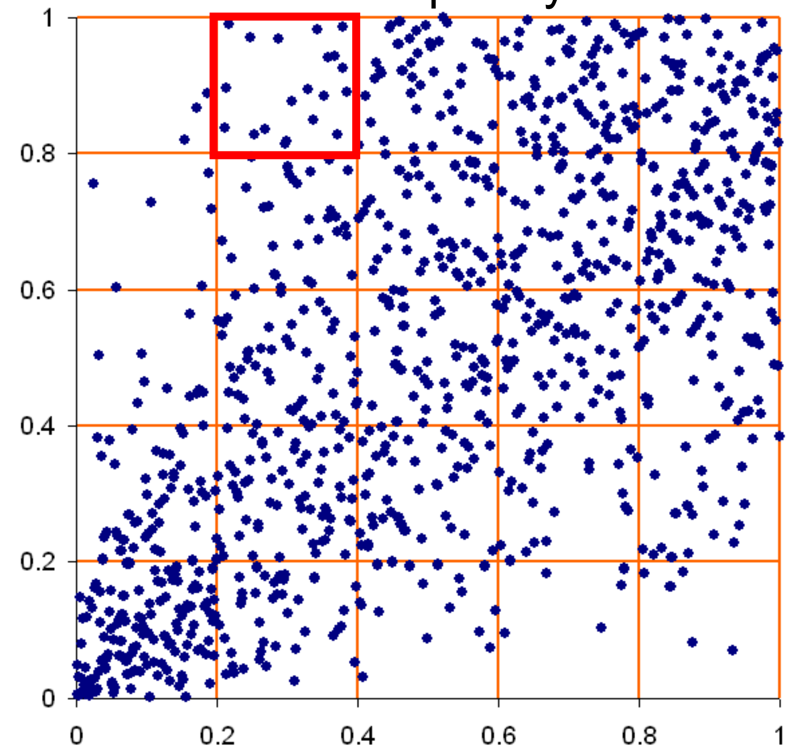
1. For a certain ***h***, select out the pairs.

2. Define a regular grid on the unit square.

3. Count the pair of the cumulative distribution (*cdf)* values in the corresponding section of the grid.

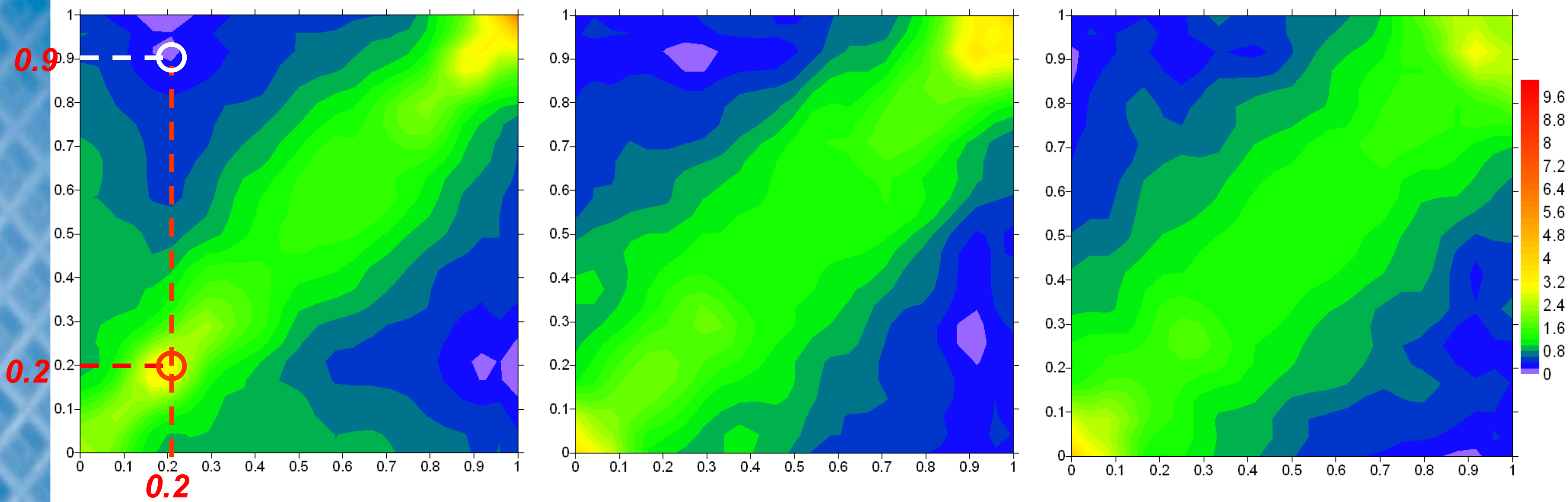


Bivariate frequency = n/N



Copula and Spatial Dependence

Empirical bivariate spatial copula



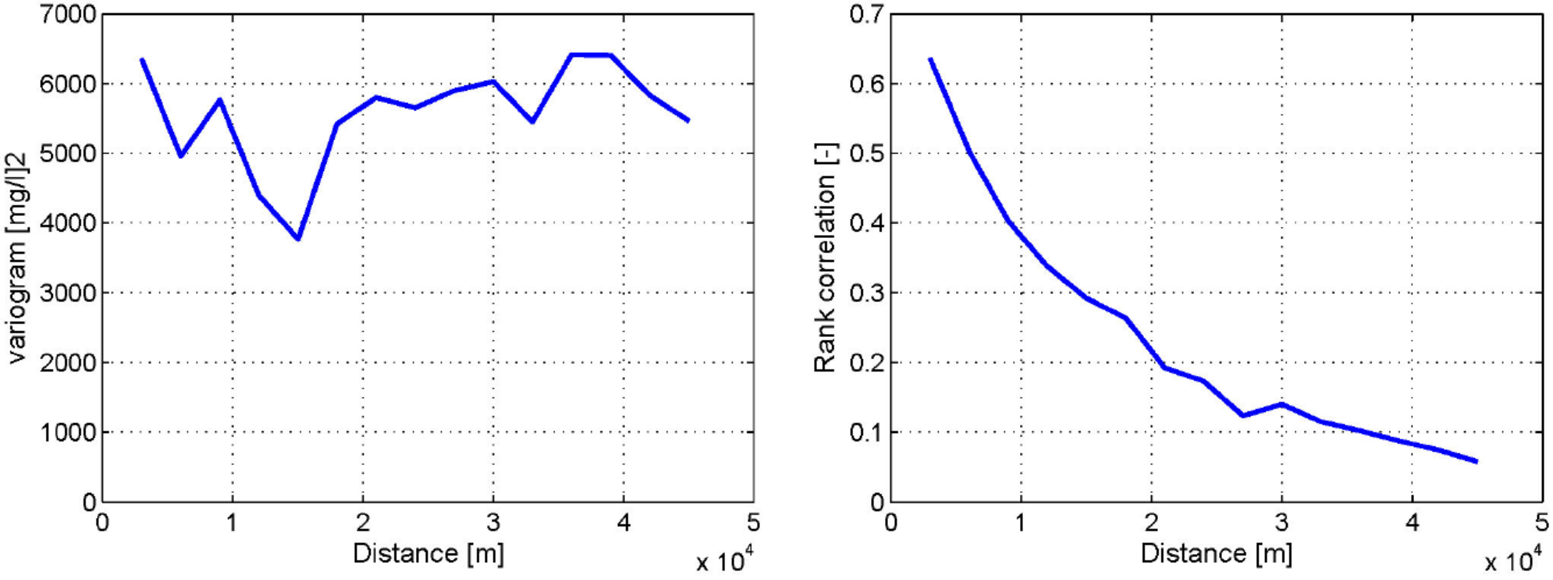
Bivariate copula densities of chloride concentration in groundwater of Baden-Württemberg for separation lengths 3km (left), 6km (middle) and 9km (right)

Copula and Spatial Dependence

Measure of dependence

1. Rank correlation/Spearman‘s rho – scale invariant

$$\rho_s = \frac{E\left[\left(U - E(U)\right)\left(V - E(V)\right)\right]}{\sqrt{Var(U)}\sqrt{Var(V)}} = 12\iint_{\mathbf{I}^2} uv\, dC(u,v) - 3$$



Variogram (left) and rank correlation (right) over distance of chloride

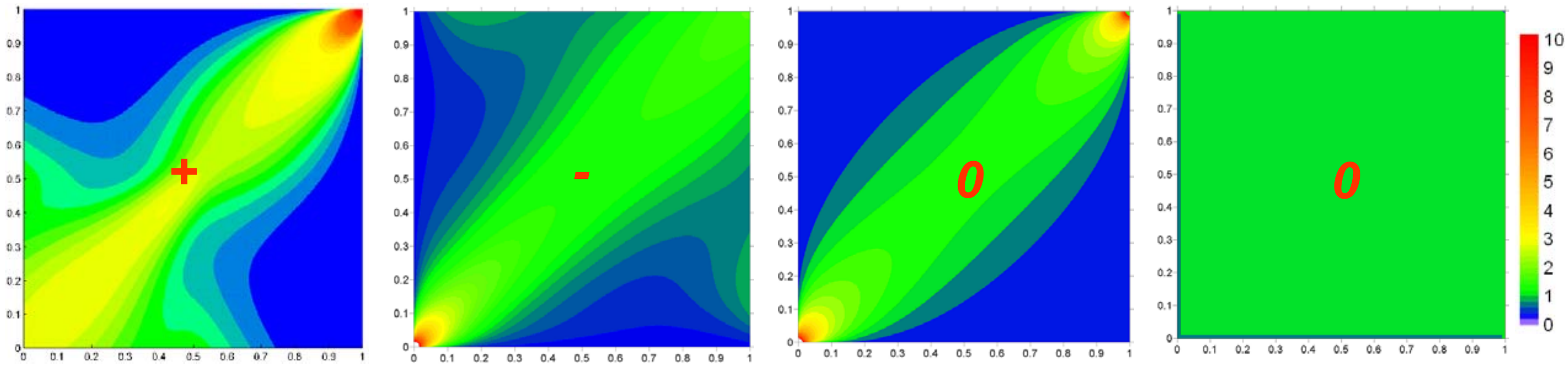
Copula and Spatial Dependence

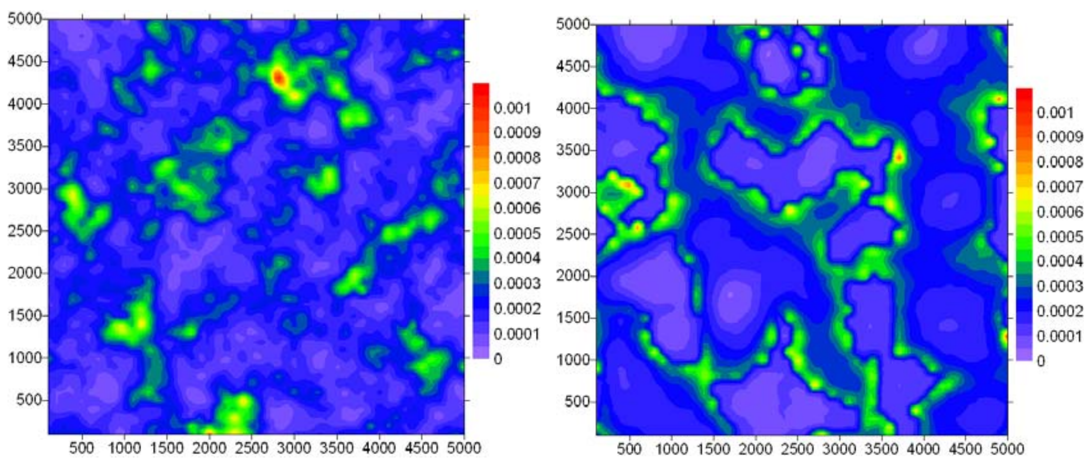
Measure of dependence

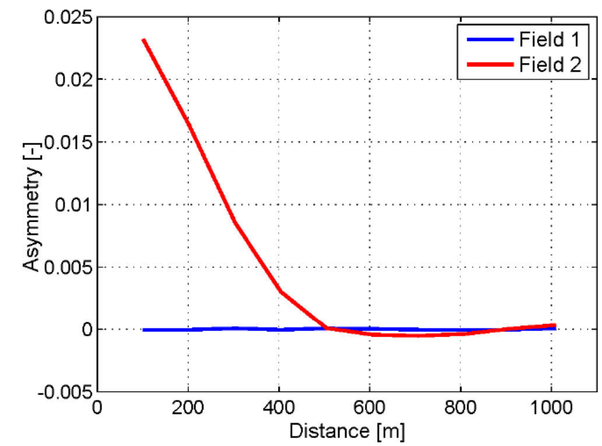
1.Measure of asymmetry – scale invariant and third moment

$$A = E\left[\left(F(Z(\mathbf{x}))-0.5\right)^2\cdot\left(F(Z(\mathbf{x}+\mathbf{h}))-0.5\right)+\left(F(Z(\mathbf{x}))-0.5\right)\cdot\left(F(Z(\mathbf{x}+\mathbf{h}))-0.5\right)^2\right]$$

x**, **h - location and separating vector *F* - marginal distribution of the RV *Z*







Outline of the Research Work

- Using copulas to describe the spatial dependence and apply scale-invariant and higher order dependence measures
- Derive theoretical copulas for spatial modeling
- Develop an appropriate model inference approach

}Model Building

- Develop Interpolation approach using copulas
- Simulate random fields with non-Gaussian dependence
- Using copulas to guide observation network design of environmental variables

}Applications

Theoretical Copulas

Existing copulas for spatial modeling – Gaussian copula

Multivariate Gaussian copula density:

$$c_n(u_1, \ldots, u_n) = \frac{1}{\sqrt{\Gamma}} \left(-\frac{1}{2}\mathbf{x}^T (\Gamma^{-1} - \mathbf{I})\mathbf{x} \right)$$

where ***x*** - the vector whose components are normally distributed variables

Γ - the correaltion matrix

Limitations:

- fully symmetric
- minimum spatical continuity for extremes

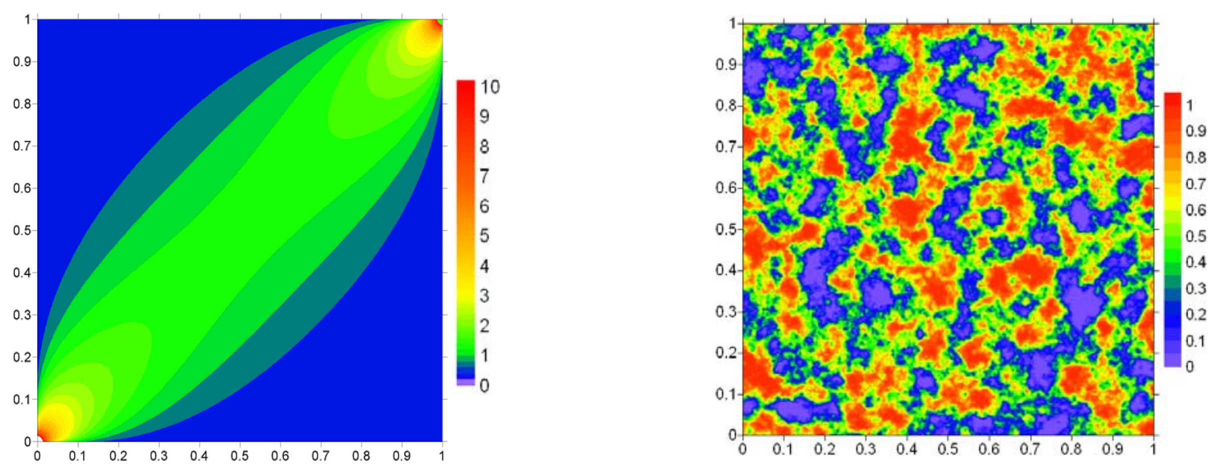


Fig: Bivariate Gaussian copula density (left) and spatial realization of Gaussian copula (right)

Theoretical Copulas

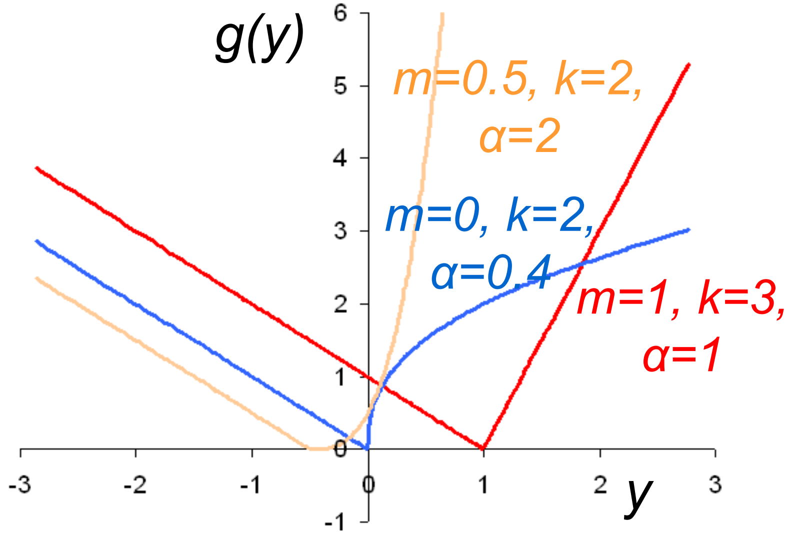
V-transformed normal copula

$g(y) = m-y$ if $y<m$

$g(y) = k(y-m)^{\alpha}$ if $y \geq m$

where $Y \sim N(0,1)$

m, k, α – model parameters



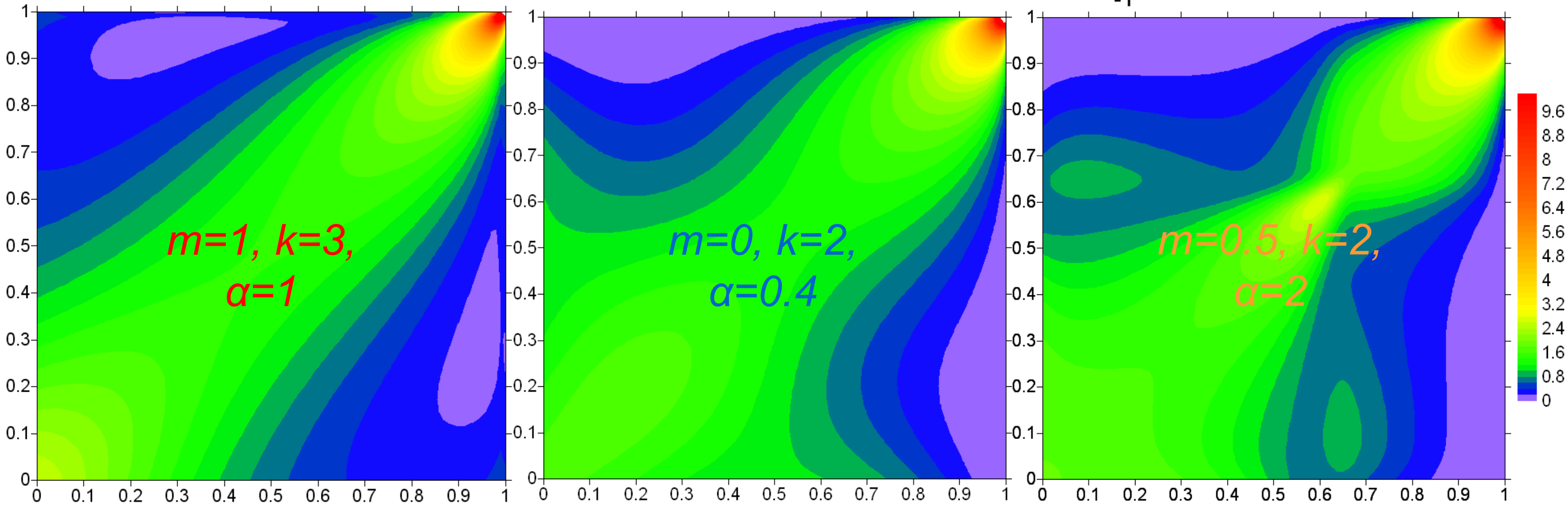


Fig: corresponding bivariate copula densities

Theoretical Copulas

Maximum normal copula

- Maximum of two independent Gaussian processes:

$$\mathbf{Z} = \max(\mathbf{Y}, \mathbf{X})$$

where $\boldsymbol{Y} \sim N(\mathbf{0},\boldsymbol{\Gamma}_1)$, $\boldsymbol{Y}=[Y_1,Y_2, \dots , Y_n]$, $Y_i \sim N(0,1)$
$\boldsymbol{X} \sim N(\mathbf{m},\boldsymbol{\Gamma}_2)$, $\boldsymbol{X}=[X_1,X_2, \dots, X_n]$, $X_i \sim N(m,\sigma^2)$

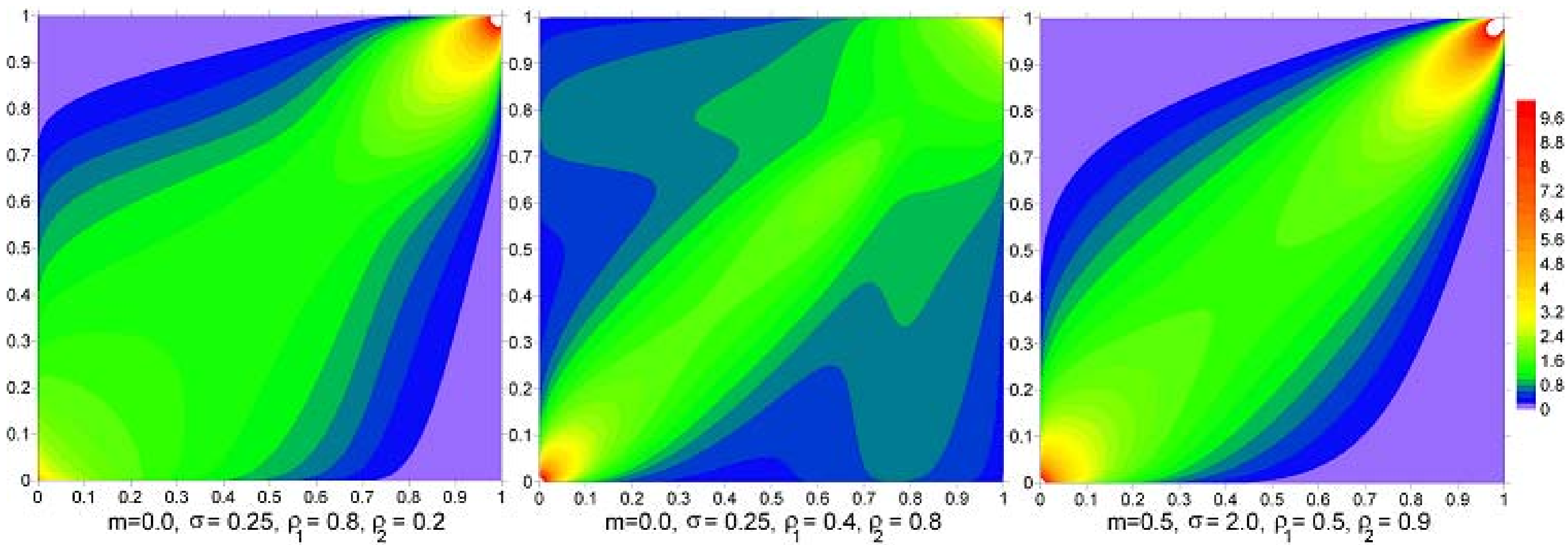
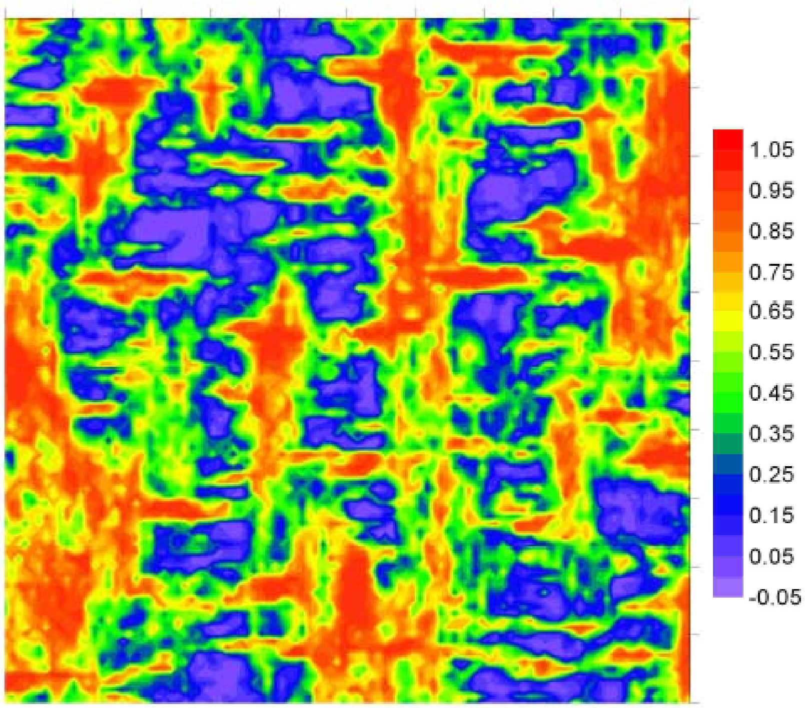


Fig: Examples of bivariate densities of maximum normal copula

Theoretical Copulas

Maximum normal copula

- Effects of two random processes



Outline of the Research Work

- Using copulas to describe the spatial dependence and apply scale-invariant and higher order dependence measures
- Derive theoretical copulas which are suitable for spatial modeling
- Develop an appropriate model inference approach

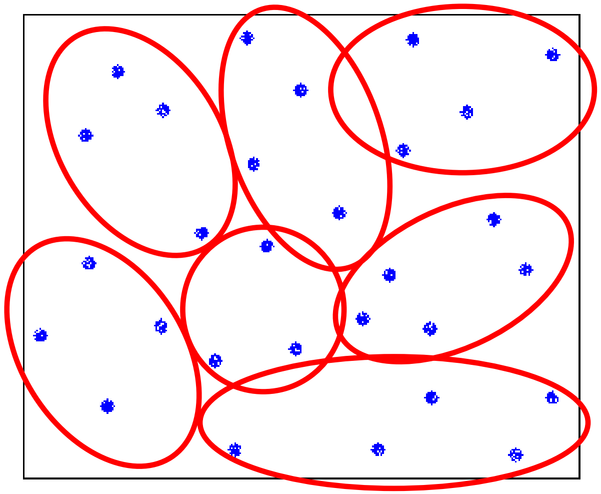
Model Building

- Develop Interpolation approach using copulas
- Simulate random fields with non-Gaussian dependence
- Using copulas to guide observation network design of environmental variables

Applications

Model Inference

1. The observation set is divided into several disjoint subsets



2. For each subset and a given parameterization of the copula, the likelihood is calculated.

$$c(S_k,\theta) = c\left(F_z(Z(\mathbf{u}_1)),\ldots,F_z(Z(\mathbf{u}_{n(k)})),\theta\right)$$

c – denotes the copula density
θ – parameters of the theoretical copula
F_Z – marginal distribution of the random variable Z
u_i – locations of points within the subset S_k

3. Since there are no overlaps between the subsets, the overall likelihood is the product of the individual ones.

$$\text{MAX } L(\theta | Z(\mathbf{u}_1),\ldots,Z(\mathbf{u}_n)) = \prod_{k=1}^{K} c(S_k,\theta)$$

K – total number of the subsets

Outline of the Research Work

- Using copulas to describe the spatial dependence and apply scale-invariant and higher order dependence measures
- Derive theoretical copulas which are suitable for spatial modeling
- Develop an appropriate model inference approach

Model Building

- Develop Interpolation approach using copulas
- Simulate random fields with non-Gaussian dependence
- Using copulas to guide observation network design of environmental variables

Applications

Outline of the Research Work

- Using copulas to describe the spatial dependence and apply scale-invariant and higher order dependence measures
- Derive theoretical copulas which are suitable for spatial modeling
- Develop an appropriate model inference approach

} Model Building

- Develop Interpolation approach using copulas
- Simulate random fields with non-Gaussian dependence
- Using copulas to guide observation network design of environmental variables

} Applications

Interpolation using Copulas

Procedure of interpolation

1. Transform the observed values $z(\mathbf{s}_i)$ to cumulative distribution (*cdf*) values using the empirical distribution *F( )*

$$u_i = F\left(z(\mathbf{s}_i)\right)$$

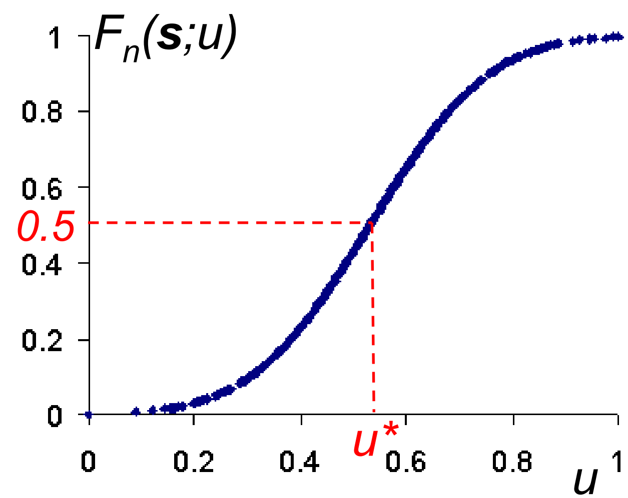
2. Calculate the conditional distribution at the unsampled location ***s*** conditioned on the neighbouring observations with the help of conditional copula:

$$F_n\left(\mathbf{s};u\right) = C_{\mathbf{s},n}\left(u \middle| u_1 = F\left(z(\mathbf{s}_1)\right), \cdots, u_n = F\left(z(\mathbf{s}_1)\right)\right)$$

3. Select one statistics u^* (e.g., median) from the conditional copula as the interpolator

4. Transform the interpolated values back into the original space using the empirical distribution

$$z^* = F(u^*)$$

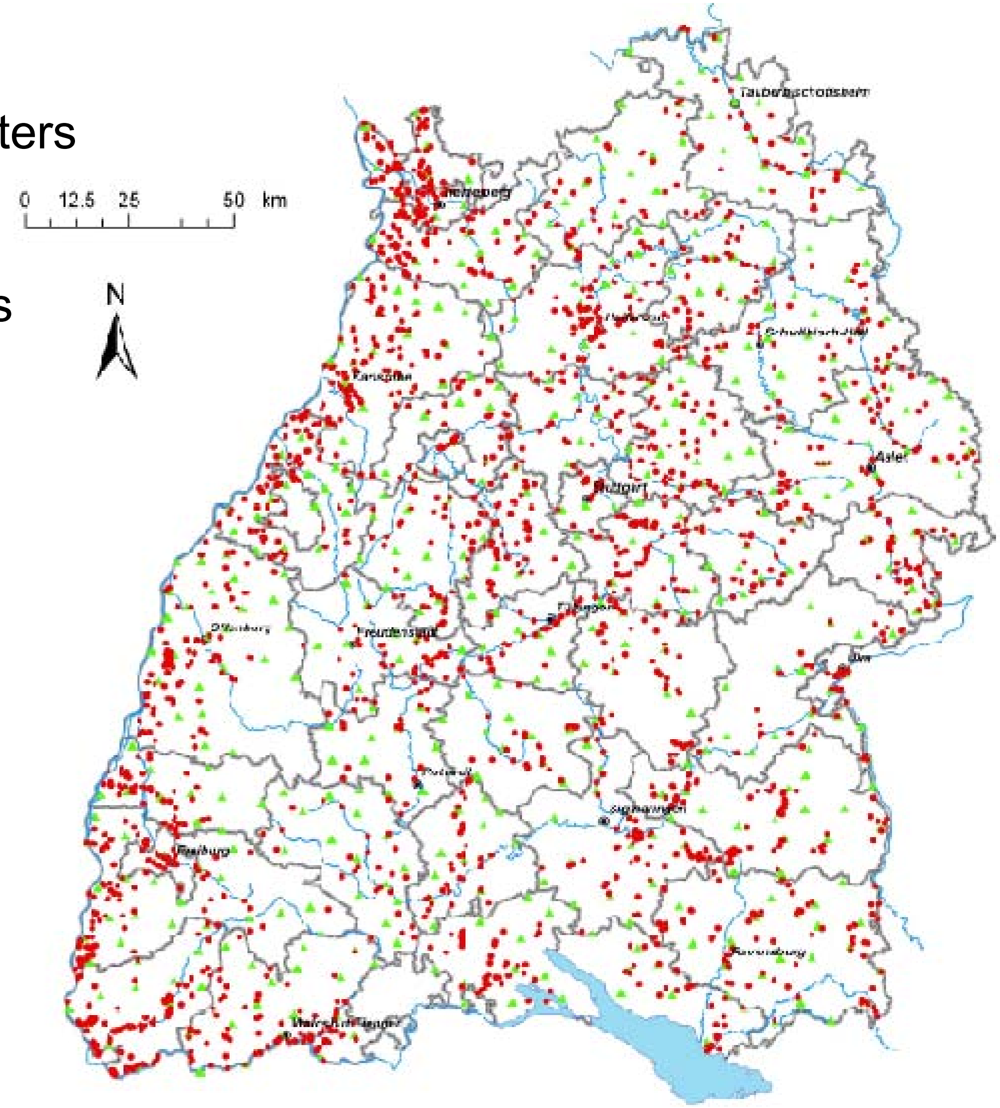


Interpolation using Copulas

Application

Groundwater quality parameters
in Baden-Württemberg:

more than 2000 observations
- chloride
- *pH*
- *nitrate*
- *sulfate*
- *dissolved oxygen*



Interpolation using Copulas

Empirical copulas

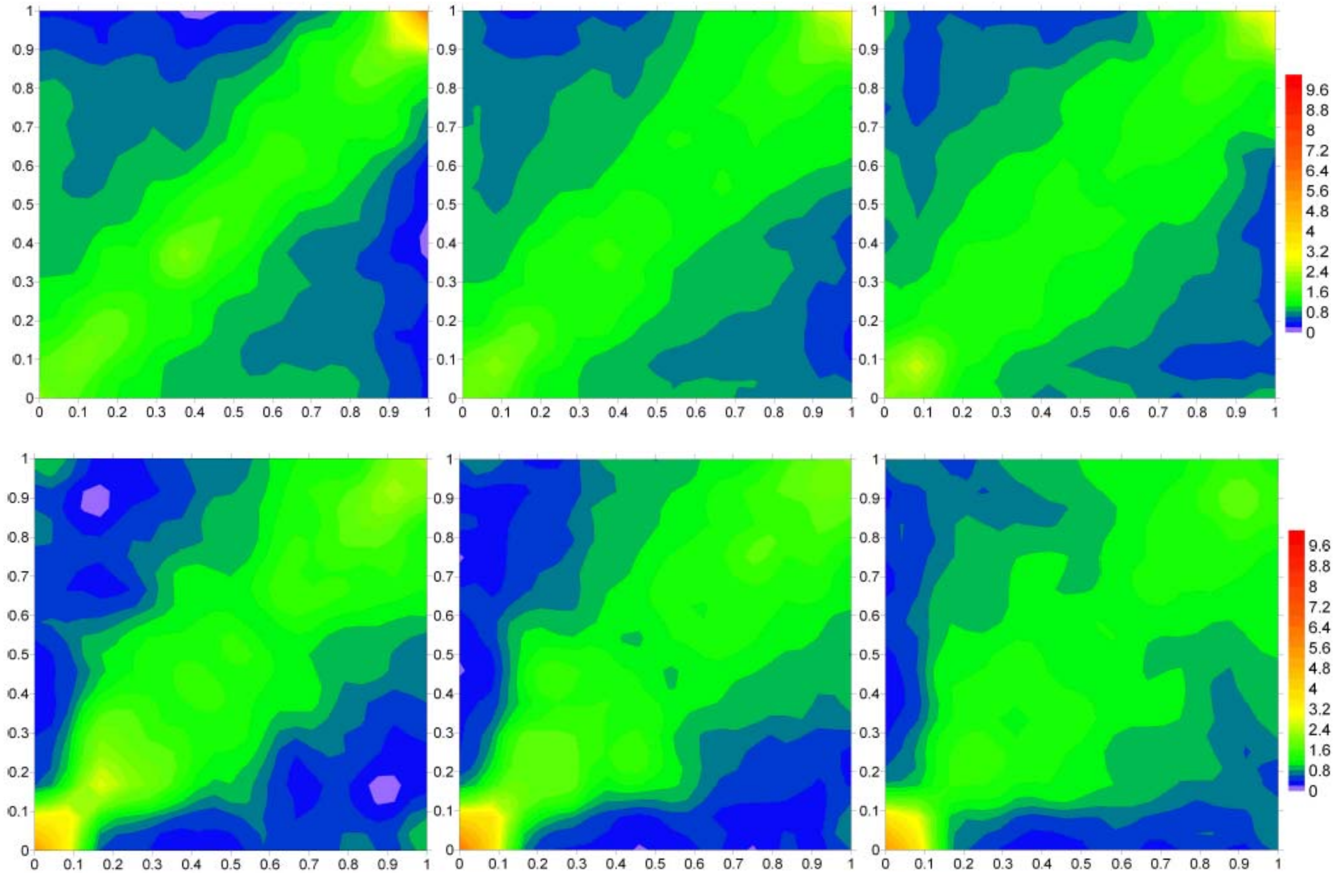


Fig: Empirical copulas of nitrate (upper line) and pH (lower line) for the separation lengths of 3km, 6km and 9km.

Interpolation using Copulas

Interpolation Methods

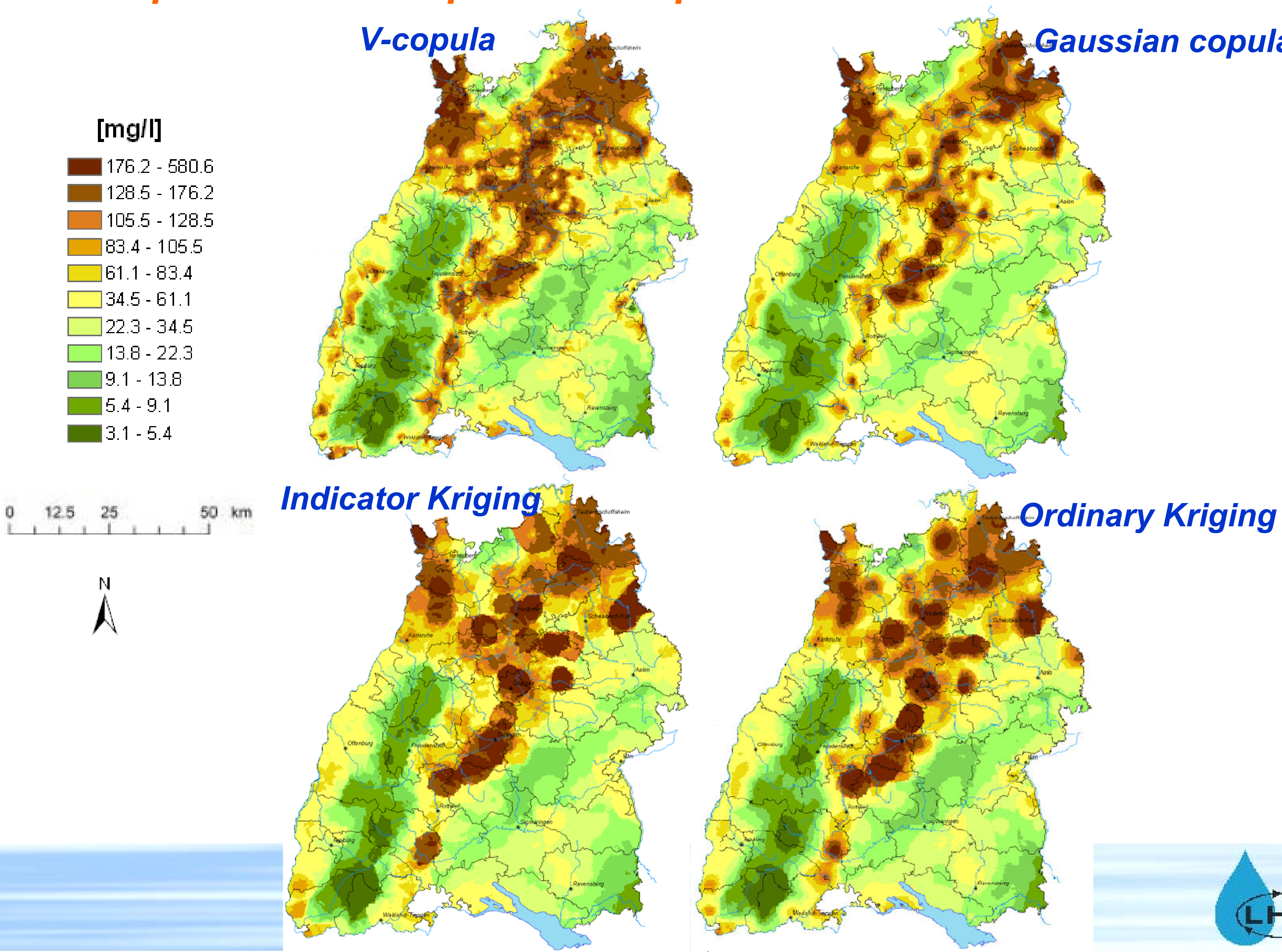
- V-transformed normal copula

For comparison:

- Gaussian copula
- Ordinary Kriging
- Indicator Kriging

Interpolation using Copulas

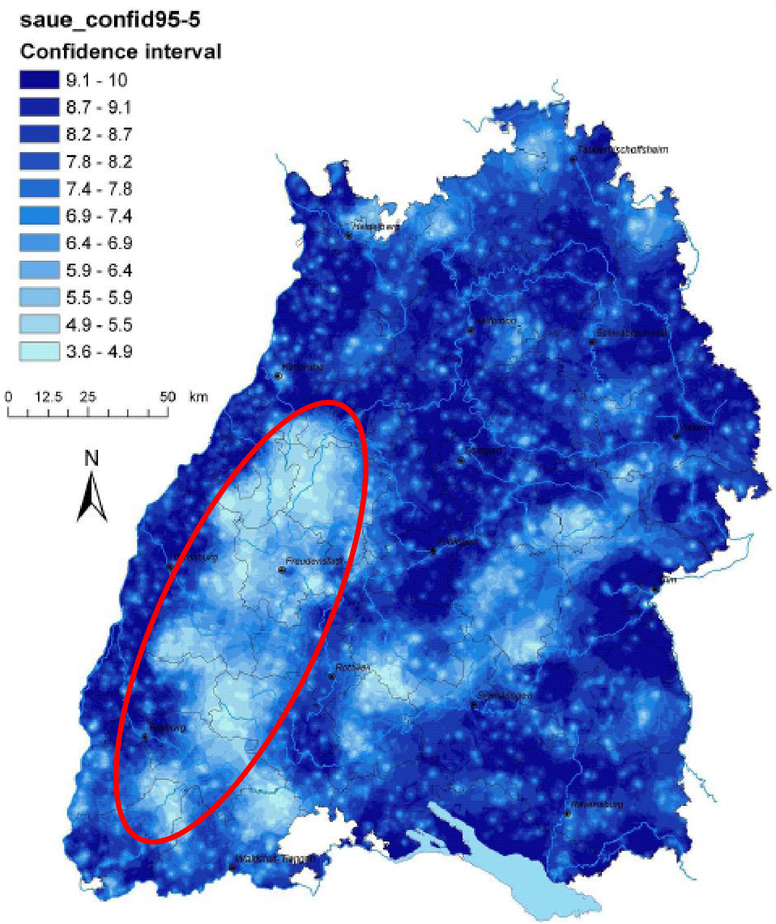
Comparison of interpolation maps - sulfate



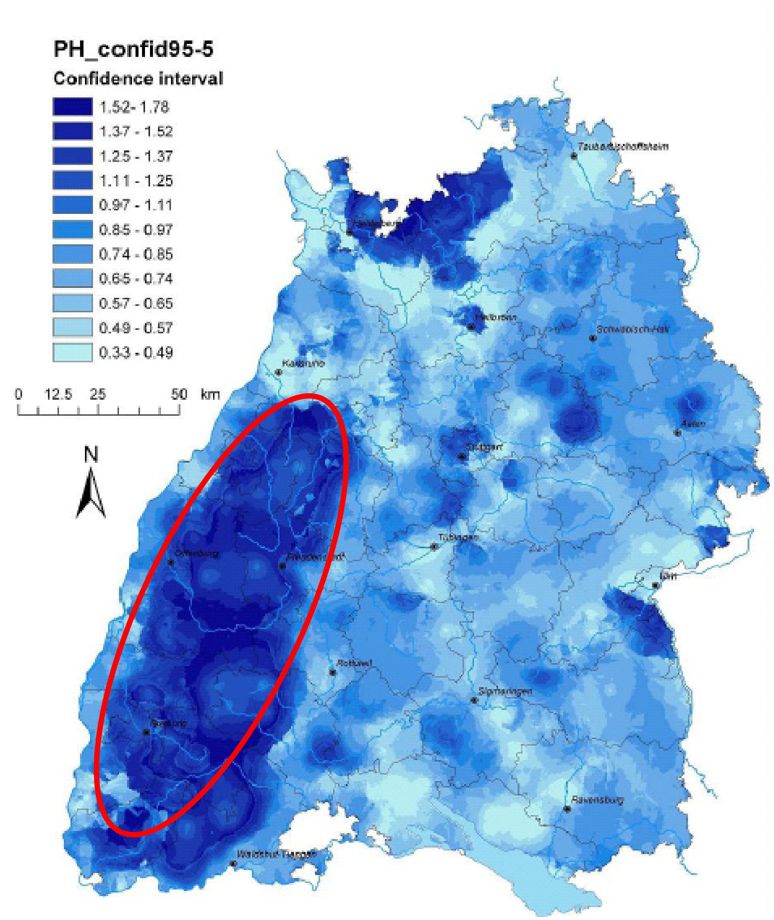
Interpolation using Copulas

Confidence intervals – from V-copula

90% confidence interval = *F(0.95)-F(0.05)*



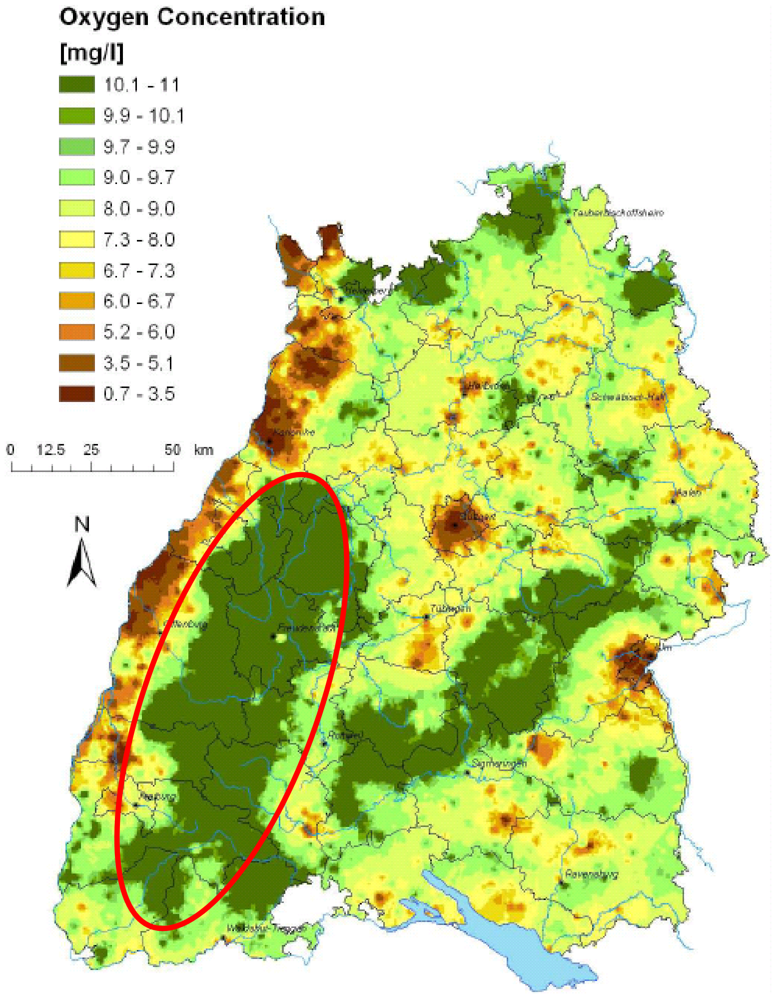
dissovled oxygen



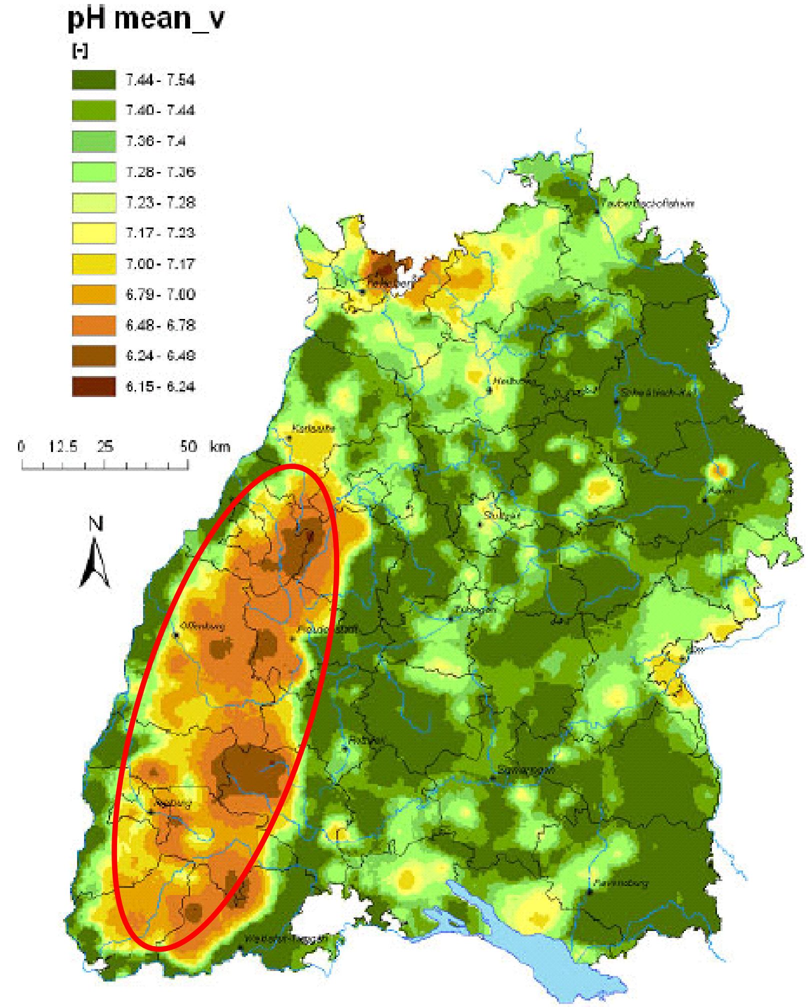
pH value

Interpolation using Copulas

Interpolation maps



dissovled oxygen



pH value

Interpolation using Copulas

Crossvalidation results

Mean absolute error

	Chloride [mg/l]	Nitrate [mg/l]	pH [-]	Dissolved oxygen [mg/l]	Sulfate [mg/l]
V-copula	14.861	13.689	0.192	1.876	34.992
G-copula	15.380	13.938	0.194	2.049	38.128
O.Kriging	16.817	13.853	0.198	1.911	42.365
I.Kriging	16.561	15.501	0.200	1.989	43.979

Outline of the Research Work

- Using copulas to describe the spatial dependence and apply scale-invariant and higher order dependence measures
- Derive theoretical copulas which are suitable for spatial modeling
- Develop an appropriate model inference approach

Model Building

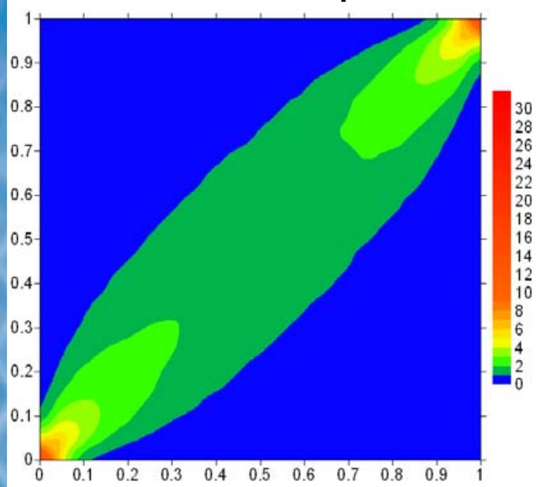
- Develop Interpolation approach using copulas
- Simulate random fields with non-Gaussian dependence
- Using copulas to guide observation network design of environmental variables

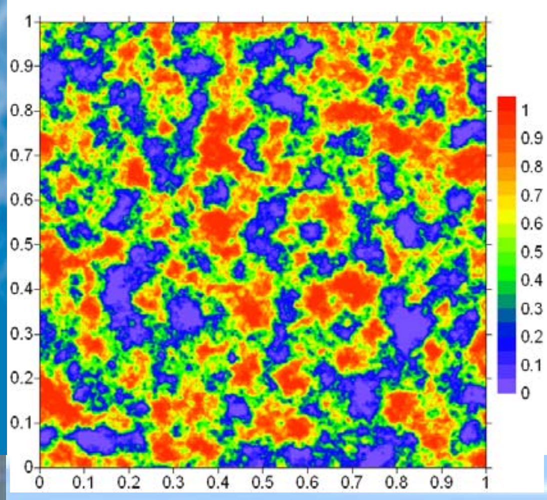
Applications

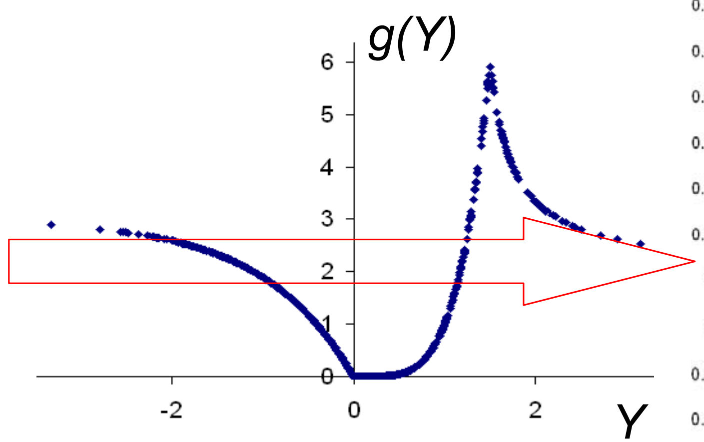
Simulation of non-Gaussian Fields

Unconditional simulation

Apply non-monotonic transformation (e.g. V-shaped transformation) to a Gaussian process – non-Gaussian process

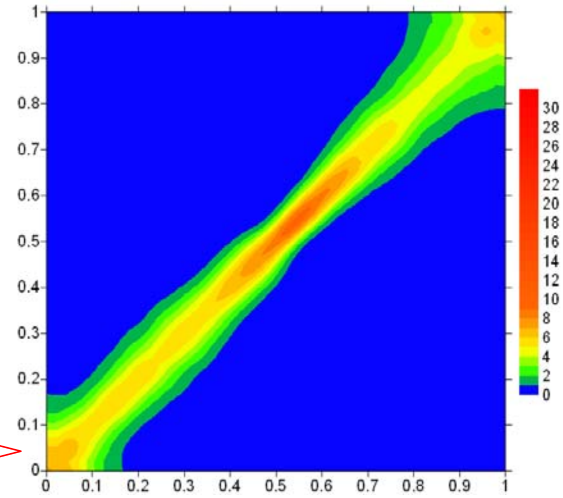


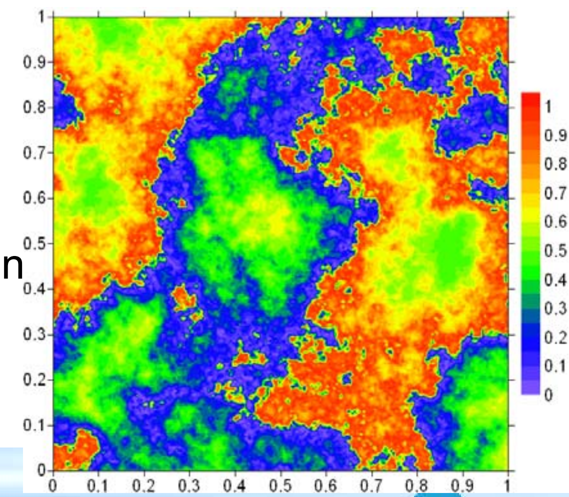




Y – Gaussian variable

g( ) – non-monotonic transformation





Simulation of non-Gaussian Fields

Unconditional simulation

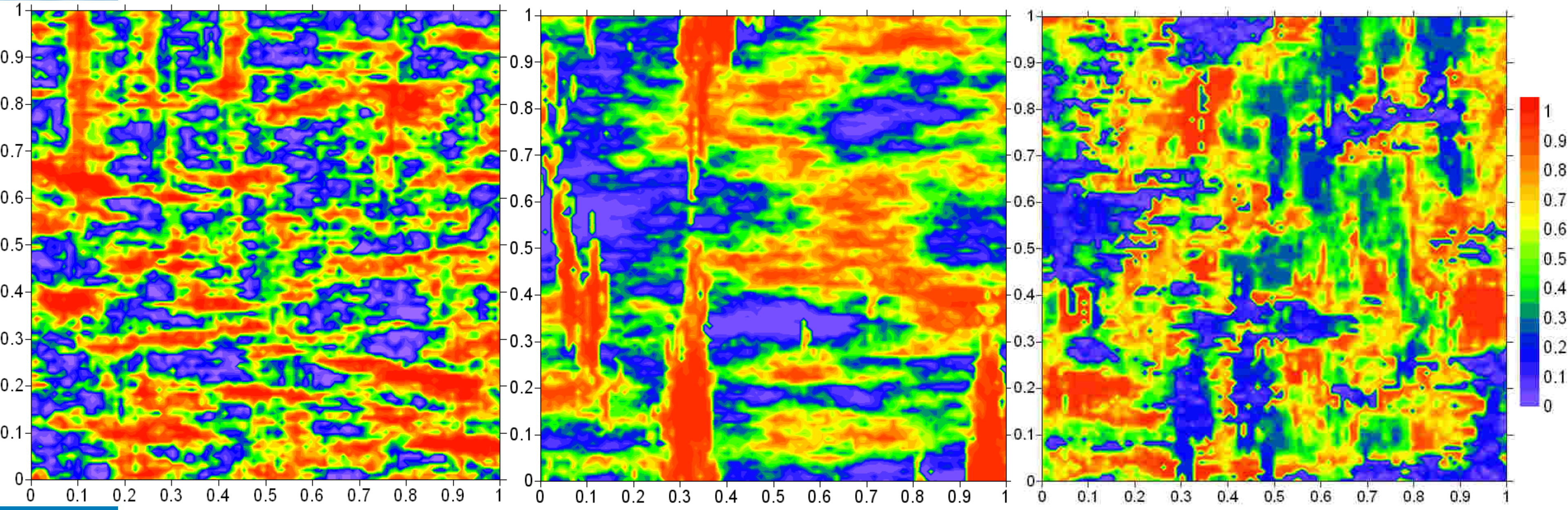
Combination of two processes:

$$Z = f(Y_1, Y_2)$$

f – combination function (e.g. *f* = *max*)

Y_1, Y_2 – independent Gaussian processes

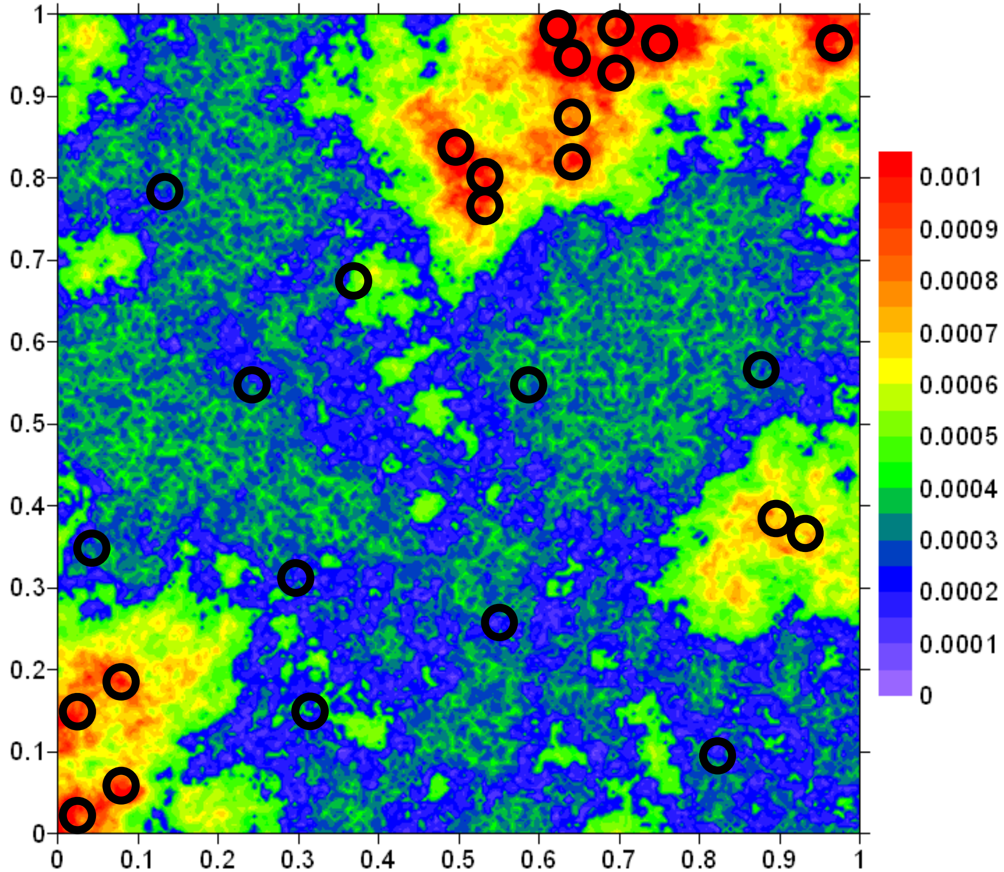
(for this case with orthogonal anisotropies to model layering and macropores simultaneously)



Simulation of non-Gaussian Fields

Conditional simulation

- Generation of random fields with prescribed variability honoring the measurements at the sampling locations



Simulation of non-Gaussian Fields

Conditional simulation – sequential simulation

1. Transform the observed values into *cdf* values

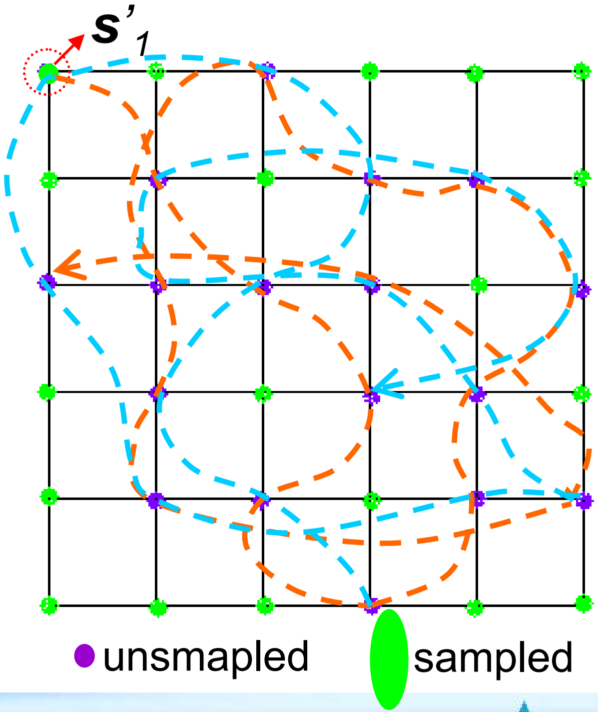
2. Define a random path through all the unsampled points. At the first point **s**'$_1$, the cumulative conditional distribution (*ccdf*) is calculated conditioned on the *m* original observations

$$F\left(\mathbf{s}_1';u_1'\middle|(m)\right)=C_{\mathbf{s}_1',m}\left(u_1'\middle|u_1=F\left(z(\mathbf{s}_1)\right),\cdots,u_m=F\left(z(\mathbf{s}_m)\right)\right)$$

3. Draw from this *ccdf* an estimate, $z^1(\mathbf{s}'_1)$ (Monte Carlo simulation), and add this point to conditioning data for all the subsequent simulations.

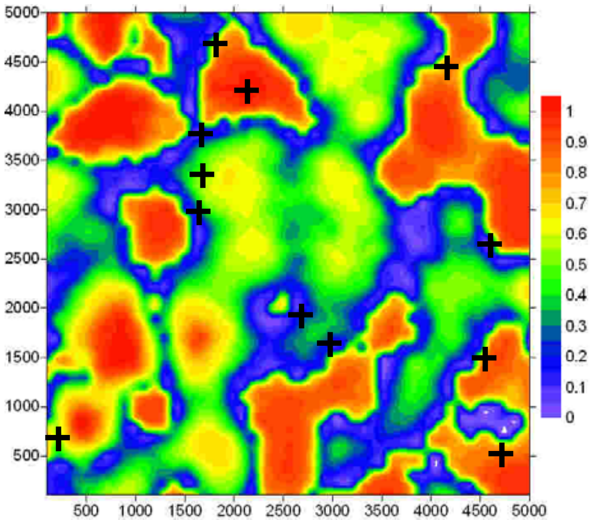
4. Repeat until all of the unsampled points have a simulated value.

5. A second realization would start with the original conditioning data and visiting the unsampled points in a different sequence.

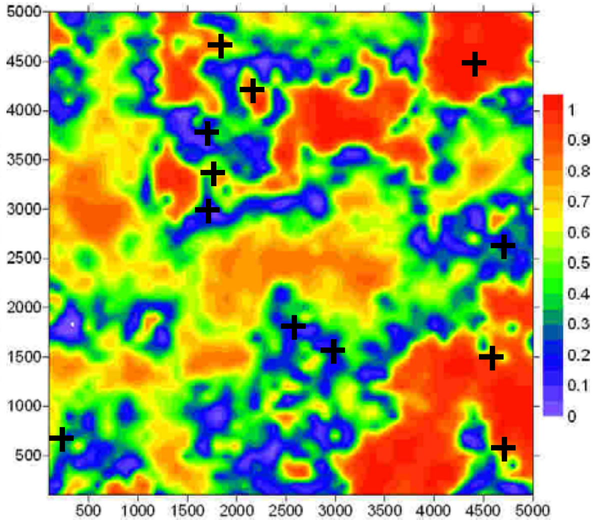


Simulation of non-Gaussian Fields

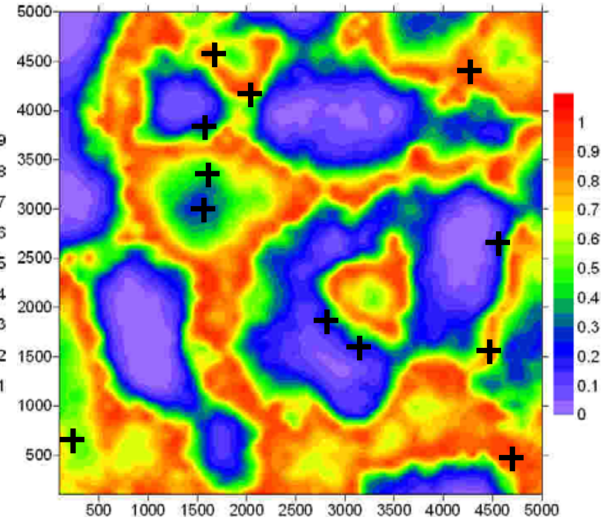
Conditional simulation



α=0.1, m=0.5, k=4.0



α=1.0, m=0.5, k=6.0



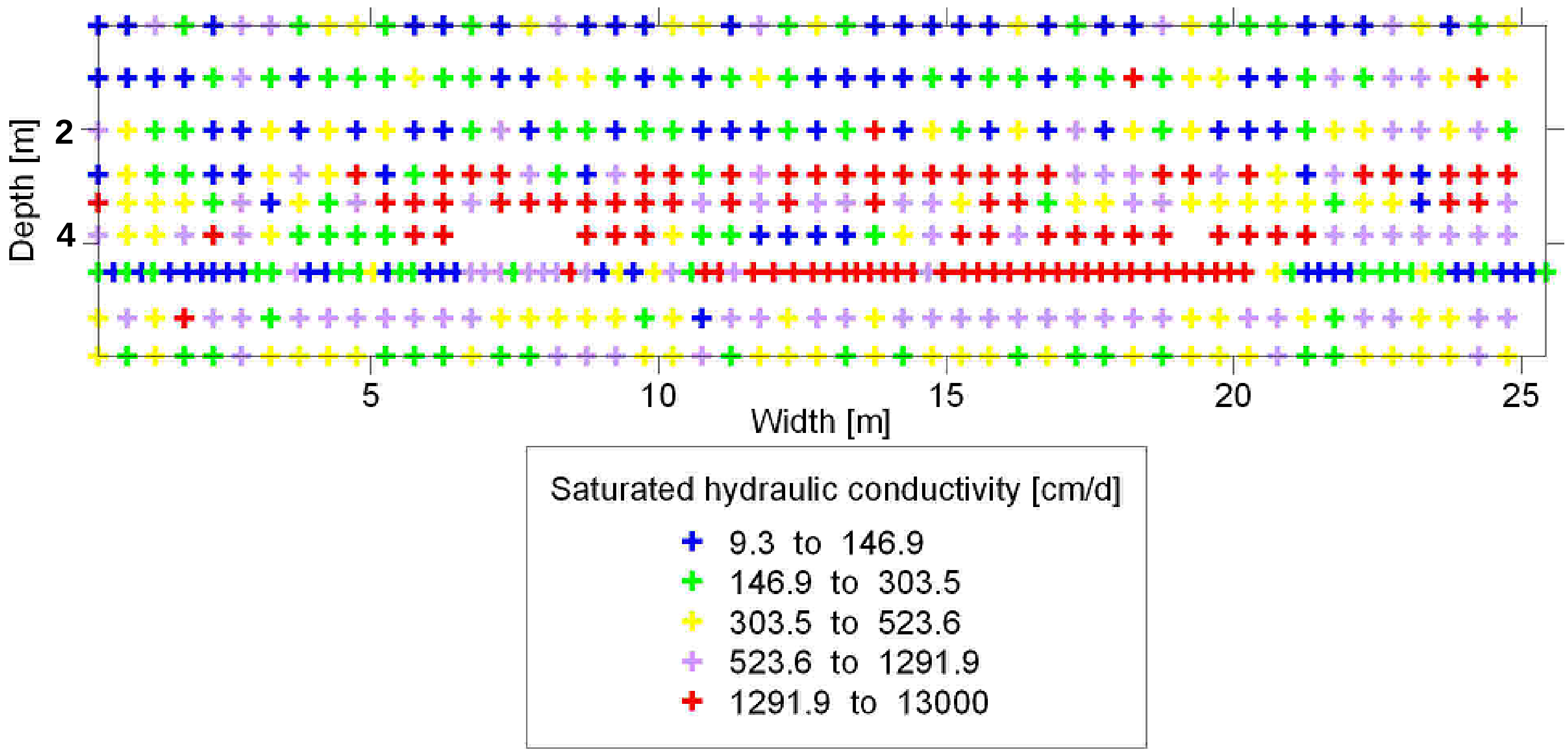
α=0.5, m=0.0, k=2.5

Simulation of non-Gaussian Fields

Application

Las Cruces Trench Site (northeast of Las Cruces, New Mexico)

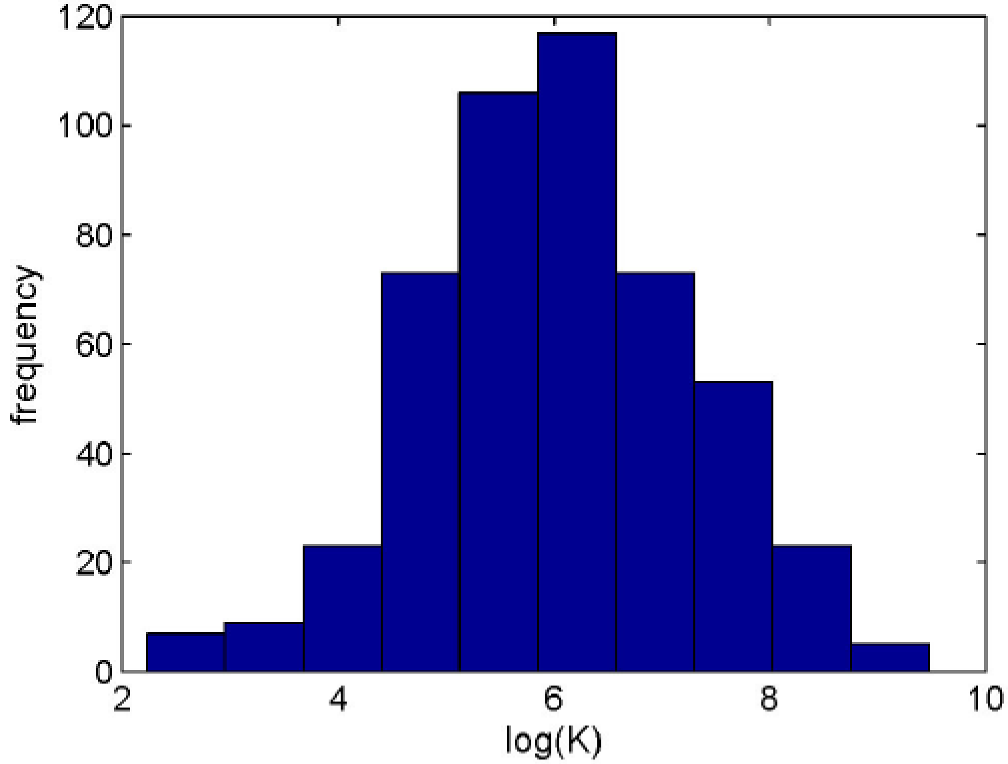
- saturated hydraulic conductivity
- *25 m* wide and by *6 m* deep
- sampling space about *50 cm*



Simulation of non-Gaussian Fields

Application – marginal distribution

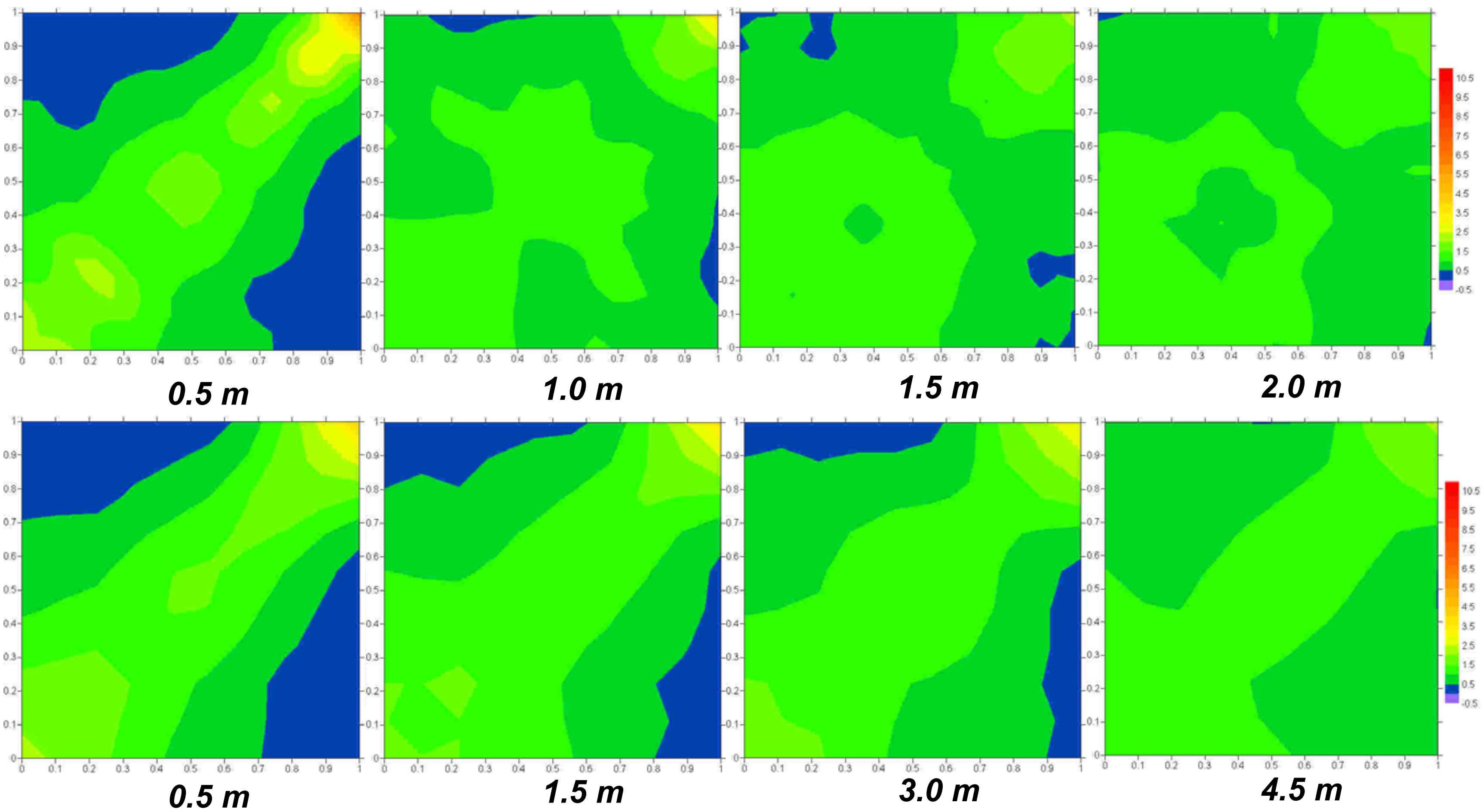
Histogram of log saturated hydraulic conductivity – normal distribution



Simulation of non-Gaussian Fields

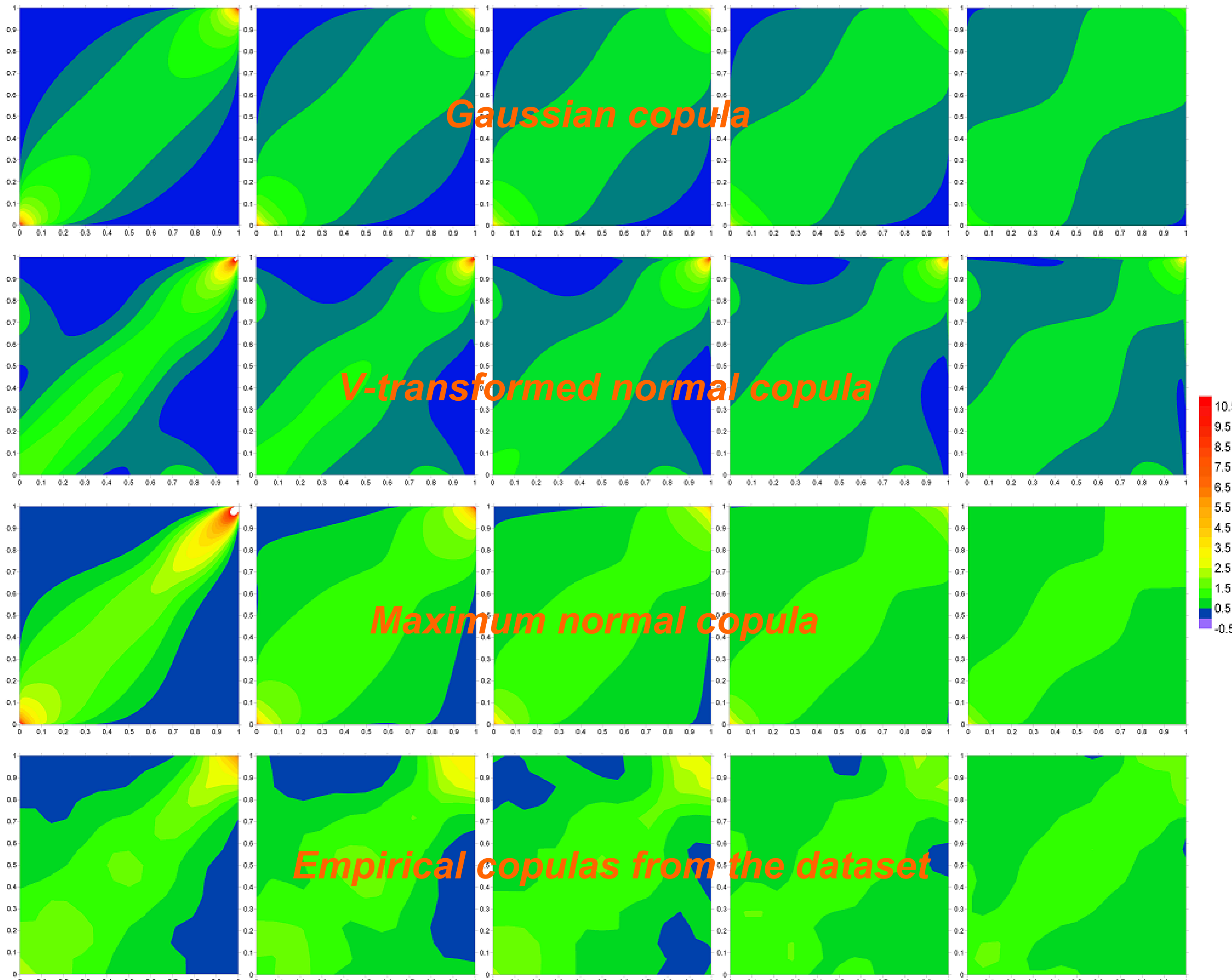
Application – empirical copulas

Empirical copulas along the omnidirection (upper) and horizontal directions (lower) – non-Gaussian behavior



Simulation of non-Gaussian Fields

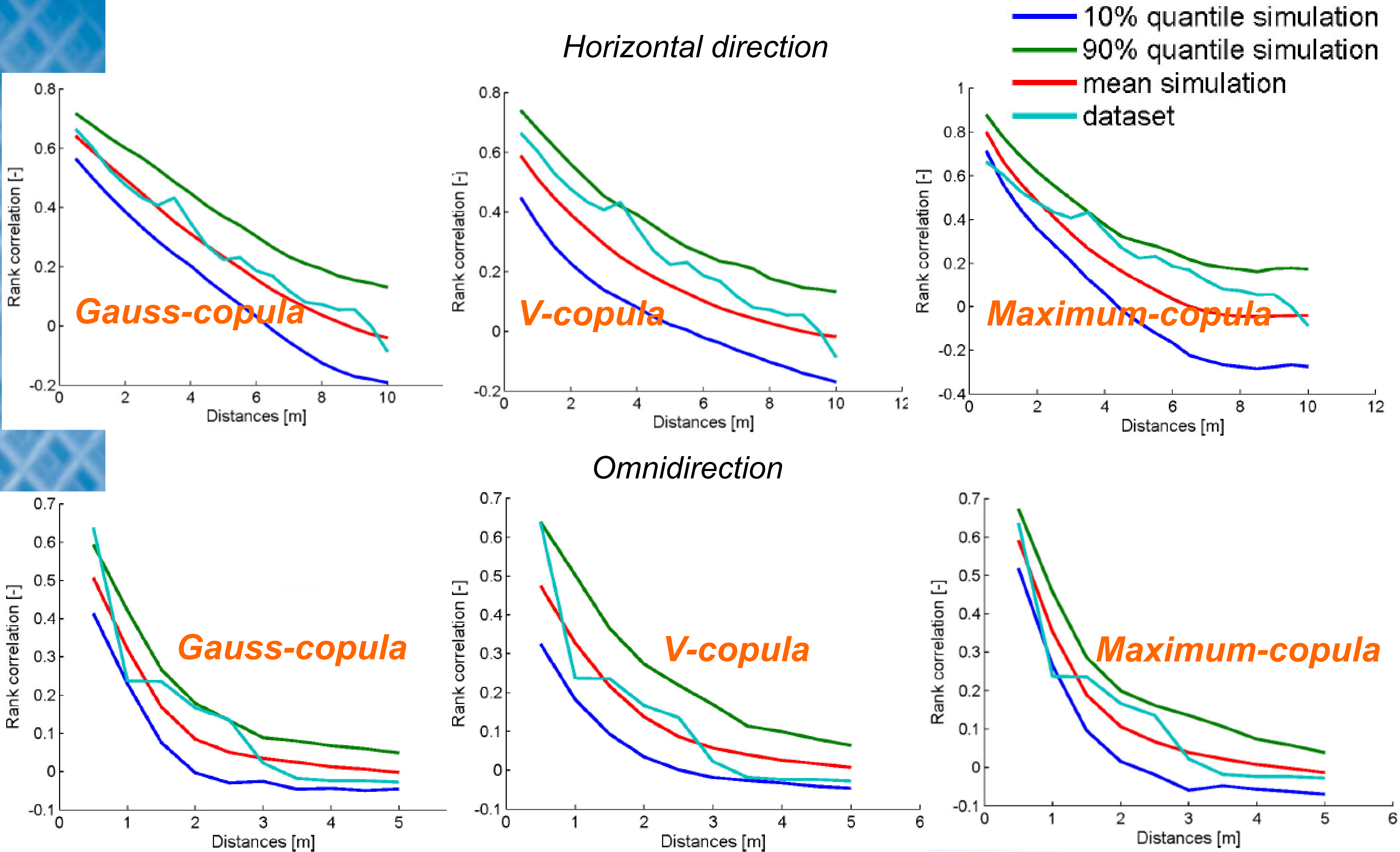
Application – parameterized theoretical copulas



Simulation of non-Gaussian Fields

Application – goodness of fit test

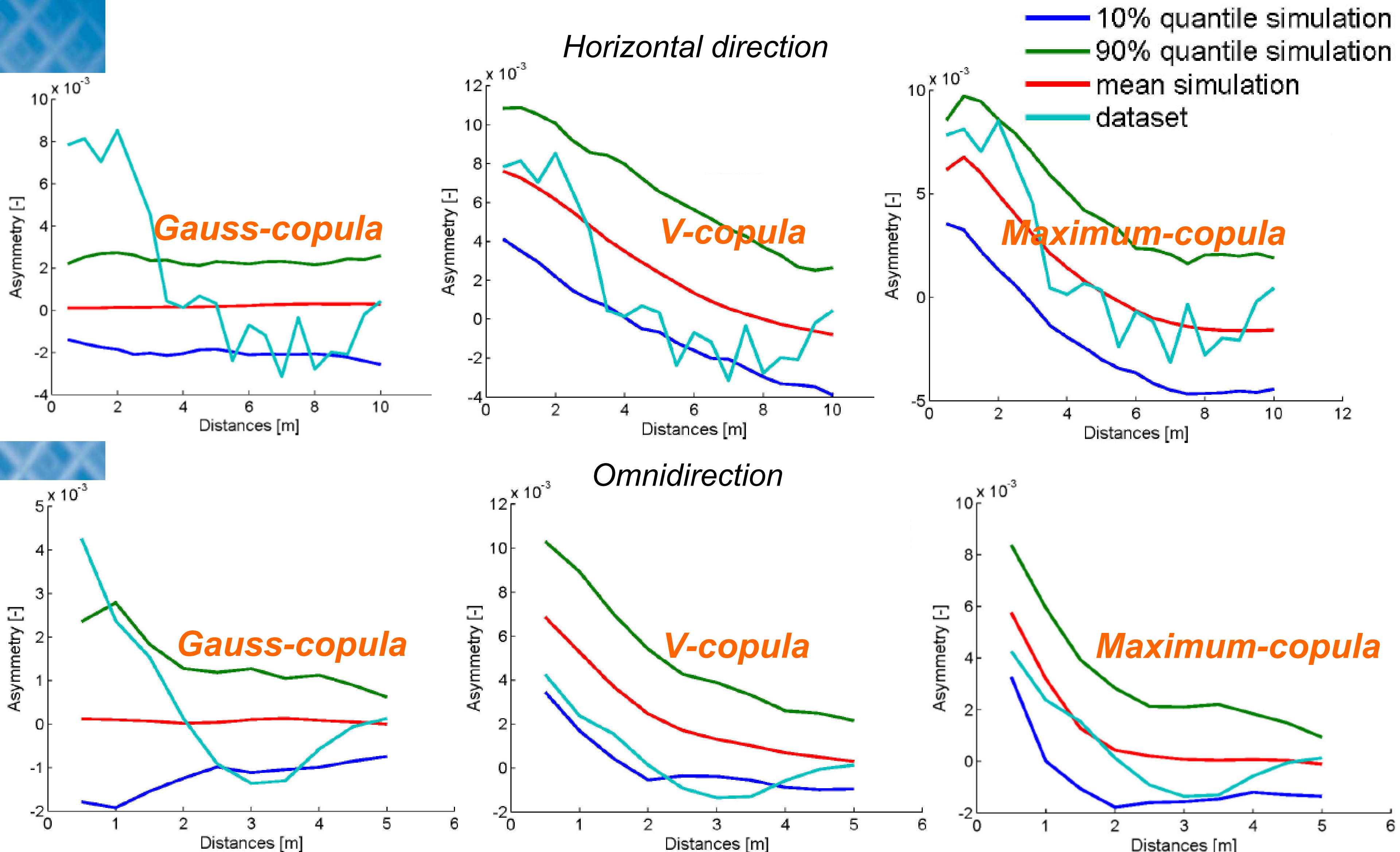
Statistical test over 100 realizations for rank correlation structure



Simulation of non-Gaussian Fields

Application – goodness of fit test

Statistical test over 100 realizations for asymmetry over distance structure



Outline of the Research Work

- Using copulas to describe the spatial dependence and apply scale-invariant and higher order dependence measures
- Derive theoretical copulas which are suitable for spatial modeling
- Develop an appropriate model inference approach

Model Building

- Develop Interpolation approach using copulas
- Simulate random fields with non-Gaussian dependence
- Using copulas to guide observation network design for environmental variables

Applications

Observation Network Design

Purpose oriented network design

Where to collect additional measurements so that the ***objectives of monitoring*** are met in the most cost-effective way?

Uncertainty estimation of predictions at the unsampled locations
- extremes may behave differently from the average

Kriging variance:
only reflects the measurement density

Confidence intervals based on copulas:
considers both the data geometry and the data values

Observation Network Design

Methodology

State of nature θ of the variable Z being below or above the threshold β at a *sampled* location ***s*** determines the decision

$$\theta(\mathbf{s}) = \begin{cases} \theta_0(\mathbf{s}) & if\ Z(\mathbf{s}) < \beta \\ \theta_1(\mathbf{s}) & if\ Z(\mathbf{s}) \geq \beta \end{cases}$$

positive decision d_0 (allow to use water)

negative decision d_1 (forbid to use water)

Utility matrix weighs the gain or loss of a certain decision

$U_s(\theta_i,d_i)$	θ_0	θ_1
d_0	k_{00}	k_{01}
d_1	k_{10}	k_{11}

gains

loss

Observation Network Design

Methodology

Expected utility at an *unsampled* location ***s'*** for a decision d_i :

$$E\left(U_{\mathbf{s}} \middle| d_i\right) = k_{i0} \cdot p\left(\theta(\mathbf{s}') = \theta_0\right) + k_{i1} \cdot p\left(\theta(\mathbf{s}') = \theta_1\right) \qquad i = 0{,}1$$

If probability of $\theta=\theta_0$ (Z<β) at the *unsampled* location ***s'*** exceeds a certain limit p_l then d_0 is taken, else d_1 is taken

$$p_l = \frac{k_{11} - k_{01}}{k_{00} - k_{01} - k_{10} + k_{11}}$$

The probability $p(\theta(\mathbf{s}')=\theta_0)=p(Z(\mathbf{s}')<\beta)$ is calculated as the conditional copula:

$$P\left(Z(\mathbf{s}') < \beta\right) = F_n\left(\mathbf{s}', \beta\right) = C_{\mathbf{s}^*,n}\left(F_Z(\beta) \middle| u_1 = F_Z(z_1), \cdots, u_n = F_Z(Z_n)\right)$$

s' *:* unsampled location

u_i *:* quantile values at the existing observation points

Observation Network Design

Methodology

If a new measurement location is added, the conditional copula at the unsampled location can be re-estimated:

$$P\big(Z(\mathbf{s}') < \beta\big) = C_{\mathbf{s}',n+1}\big(F_Z(\beta)\,\big|\,u_1 = F_Z(z_1),\cdots,u_n = F_Z(Z_n), u_{n+1} = F_Z(Z_{n+1})\big)$$

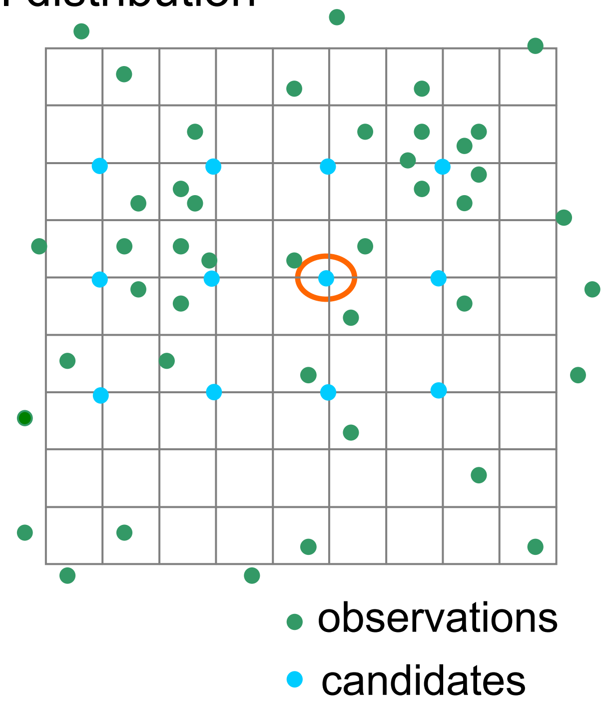
The value u_{n+1} at the new candidate ***s**** should also be estimated from the old observations using conditional copula *C** – full distribution

The expected utility at an unsampled location ***s'***:

$$\int_0^1 E\big[U_{\mathbf{s}'}\,\big|\,u_{n+1}\big]\,dC^*$$

The candidate which produces the highest total utility of the entire estimation grid will be selected

$$\text{MAX}\sum_{i=1}^{m}\int_0^1 E\big[U_{\mathbf{s}'_i}\,\big|\,u_{n+1}\big]\,dC^*$$



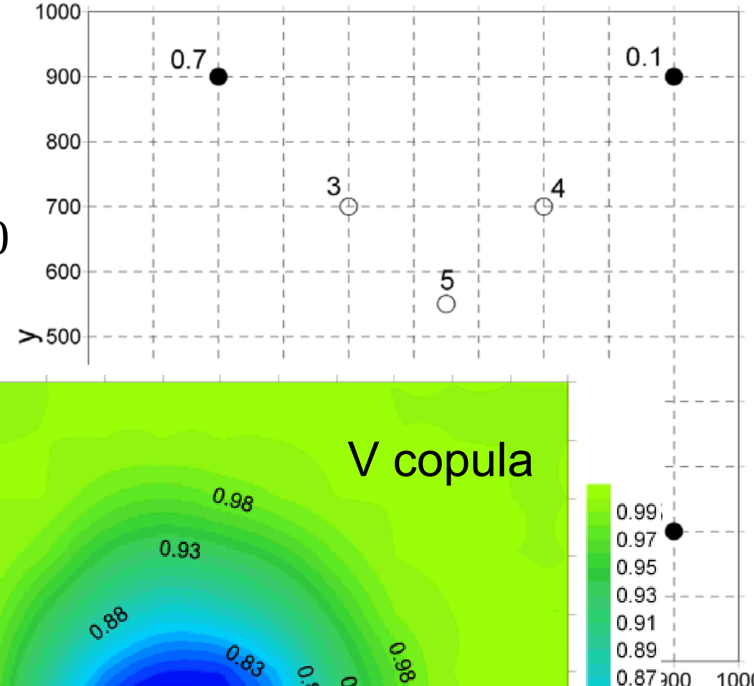
Observation Network Design

Synthetic example

- threshold probability $P(Z(\mathbf{s}) < \beta) = 0.8$
- entry values of the utility matrix:

$$k_{00} = 0.0, \quad k_{01} = -2.0, \quad k_{10} = -1.0, \quad k_{11} = 0.0$$

- V-copula and Gaussian copula



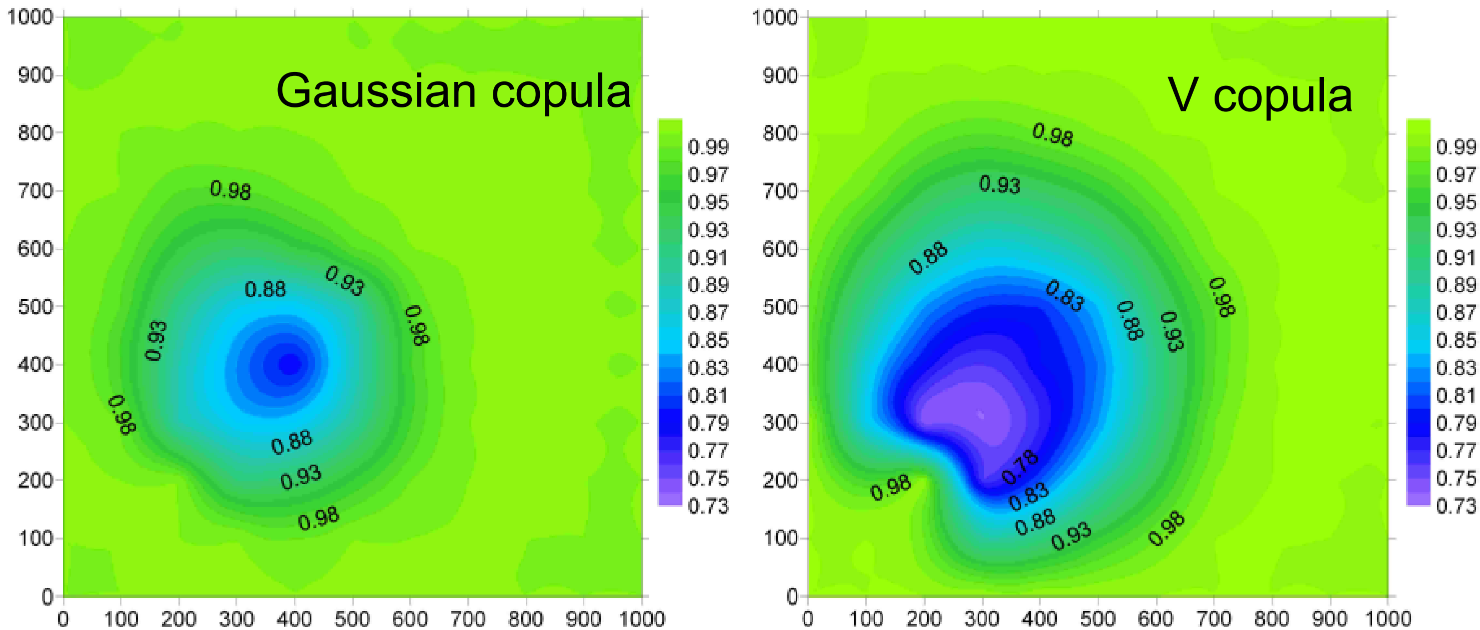


Fig: Contour maps of percentage of positive decisions resulting from Gaussian copula (left) and V-copula (right).

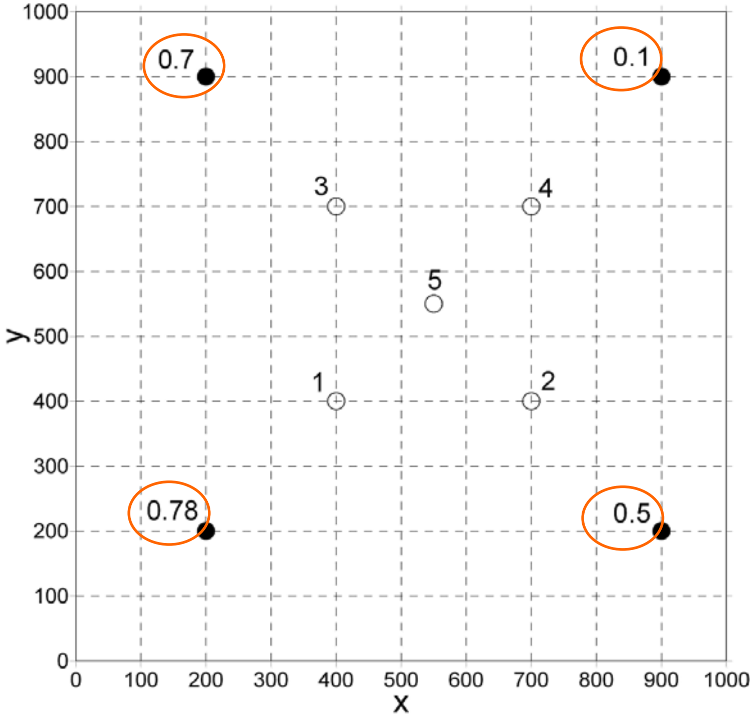
Observation Network Design

Synthetic example

- threshold probability $P\left(Z(\mathbf{s}) < \beta\right) = 0.8$
- entry values of the utility matrix:

$$k_{00} = 0.0, \quad k_{01} = -2.0, \quad k_{10} = -1.0, \quad k_{11} = 0.0$$

- V-copula and Gaussian copula



How about using Indicator Kriging?

- All the observations are below the threshold, IK gives no information on where to measure

Summary and Outlook

Summary

- Empirical copulas and scale-invariant measures are applied to investigate spatial dependence.
- Theoretical non-Gaussian copulas are derived for spatial modeling.
- A model inference approach is developed to parameterize theoretical copulas.
- Methodology of interpolation using copulas is developed and the crossvalidation results of an application to the groundwater quality parameters show that the copula approach gets better performance than Kriging.
- Simulation algorithms of generating realizations with non-Gaussian dependence are developed for both unconditional and conditional cases and statistical tests of simulations of a hydraulic conductivity dataset demonstrate that the non-Gaussian copula is more suitable than the Gaussian copula.
- Conditional copula is embedded into the utility function to guide the observation network design and the synthetic exmaple shows its potential.

Summary and Outlook

Outlook

- Copula models which considers effects of more processes can be developed to model more complex structures.

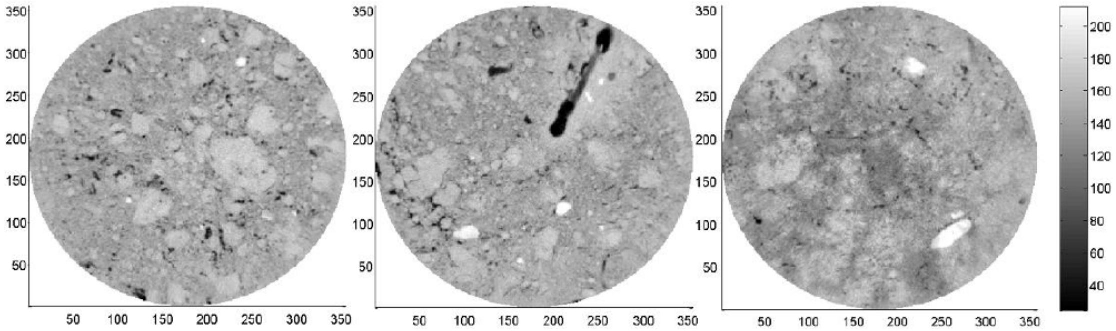


Fig: Horizontal planes from the X-ray tomography of the bulk density of a soil column (A. Bayer, H.-J. Vogel and K. Roth, 2004)

- The application of the concept of copula can be further extended to categorical spatial variables.

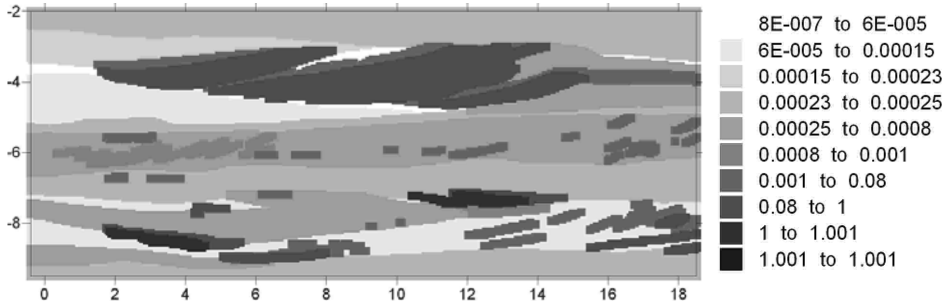


Fig: Surface ground-penetrating radar (GPR) profiling of sediment in the upper Rhine valley. (J. Tronicke, P. Dietrich, U. Wahlig and E. Appel, 2001)

Thank you