Data simulation in ecological research: Spatial autocorrelation

Severin Hauenstein

Department of Biometry and Environmental System Analysis, University of Freiburg, 79106 Freiburg, Germany

29 November 2015

1 Data simulation with simSAC

The data simulated with simSAC can be varied in predictor landscape, distribution of the response variable and cause of spatial autocorrelation (SAC). The three predictor landscapes are linear and non-linear gradients without noise ("smooth"), unconditional Gaussian random fields from an exponential covariance model ("realistic"), and "real" bio-climatic data from `http://www.worldclim.org` (see fig. 1). The response distribution can be Gaussian, Bernoulli or zero-inflated Poisson. Spatial autocorrelation was caused by adding SAC onto the response variable, omitting an important predictor in the model, the wrong functional form of the model (i.e. the model must miss a quadratic term or interaction) or dispersal (i.e. smoothing out values/ probabilities). The full number of datasets is 45.

Here we choose a minimal set of 12 = 2 landscapes * 2 distributions * 3 SAC causes (see table 1). For the data with a smooth **landscape** the seven predictors are:

- $x_1 = lon$
- $x_2 = lat$
- $x_3 = (lon - \overline{lon})^2$
- $x_4 = (lat - \overline{lat})^2$
- $x_5 = x_3^{x_4} x_4^{x_3}$
- $x_6 = x_1^{x_1} x_3^{x_4}$
- $x_7 = x_2^{x_1} x_4^{x_3} log(x_5 + 1))$

where lon and lat are (simulated) longitude and latitude. As for the data with real landscapes seven real bio-climatic predictors are cropped to the extent of 5N24E to 7S37E. The bio-climatic variables are:

- $x_1 =$ annual mean temperature
- $x_2 =$ precipitation of coldest quarter
- $x_3 =$ mean diurnal range
- $x_4 =$ annual precipitation
- $x_5 =$ temperature seasonality
- $x_6 =$ precipitation of warmest quarter
- $x_7 =$ isothermality

Regardless of the landscape, all predictors are rescaled to [-1,1].

The Gaussian **response distribution** is simulated as $y \sim \mathcal{N}(\beta_0 + \beta_1 x_1 + \beta_2 x_4 + \beta_3 x_4^2 + \beta_4 x_3 * x_4 + \beta_5 x_3, \sigma = 0.2)$ with β-values of 0.8, 0.2, -0.9, 0.8, -0.6 and 0.5, respectively. As for the Bernoulli distribution $y \sim$ Bern(logit$^{-1}(\beta_0 + \beta_1 x_1 + \beta_2 x_4 + \beta_3 x_4^2 + \beta_4 x_3 * x_4 + \beta_5 x_3)$), with β-values of 0.2, 4.5, -1.2, -1.2, -1.1 and 0.9, respectively.

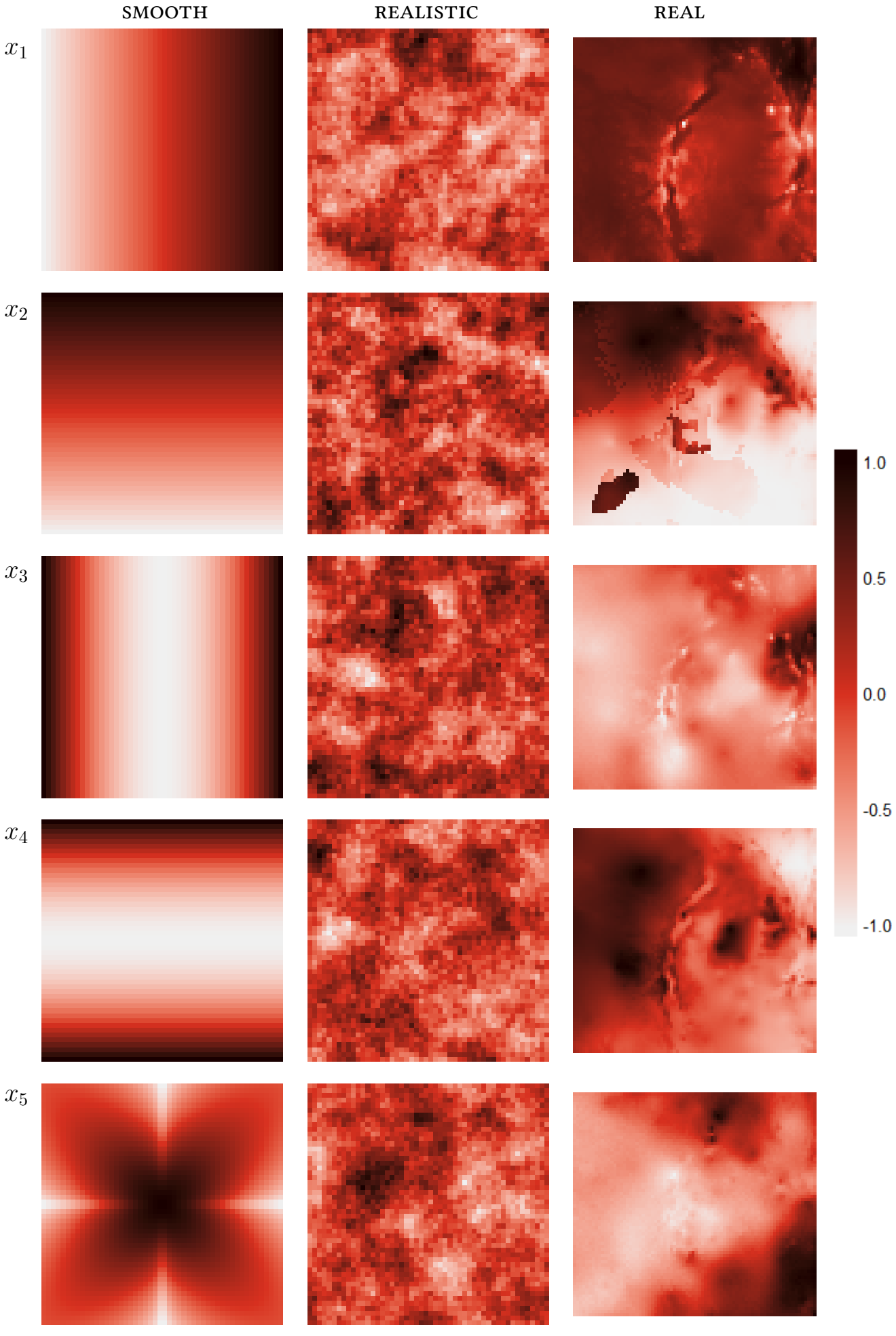
We impose five different **autocorrelation regimes**. of which we use three in the minimal set:

SAC cause 0 The response variable without spatial autocorrelation is used as reference scenario.

SAC cause 1 Adding spatial autocorrelation onto the response variable. We first compute the correlation structure as $\Omega = e^{-0.3D}$, where D is the euclidean distance matrix. By computing the inverse of the Choleski decomposition of Ω we obtain the correlation weights matrix ($w = Chol(\Omega)^{-1}$). We then preform a matrix multiplication of w^{-1} with n values drawn from $\mathcal{N}(\mu = 0, \sigma = 0)$, where $n =$ is the number of rows of D.

SAC cause 2 Omitting x_1 as important predictor

Fig. 2 shows the response and residuals (from a linear model) maps and correlograms for the five different spatial autocorrelation scenarios. For the latter the spatial dependence measure is Moran's I.


SMOOTH
REALISTIC
REAL
x_1
x_2
x_3
x_4
x_5
1.0
0.5
0.0
-0.5
-1.0

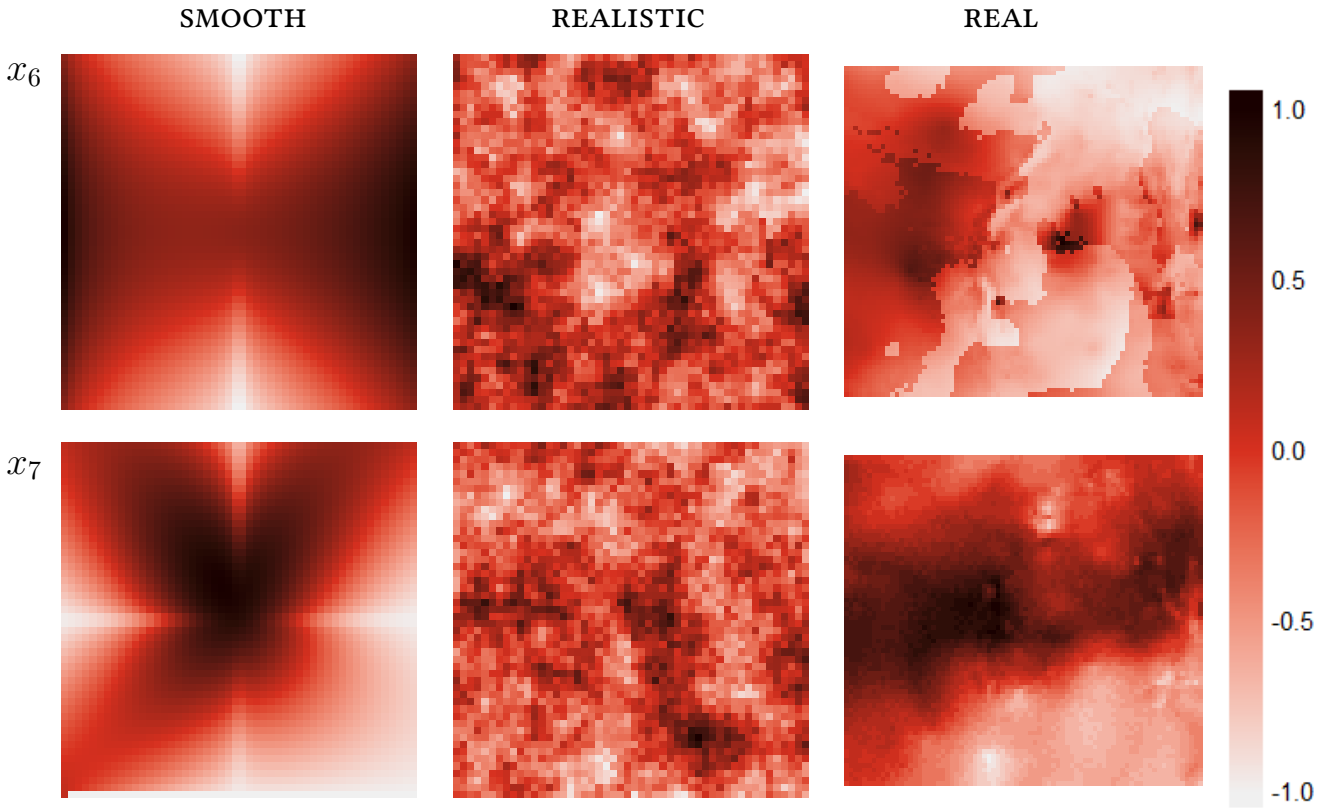


Figure 1: Maps of simulated smooth (1st) and realistic (2nd) and real world (3rd column) landscapes on (simulated) Longitude (lon) / Latitude (lat) grid. All values were rescaled to [-1,1]. Smooth landscapes have the functional forms of $x_1 = lon$, $x_2 = lat$, $x_3 = (lon - \overline{lon})^2$, $x_4 = (lat - \overline{lat})^2$, $x_5 = x_3^{x_4} x_4^{x_3}$, $x_6 = x_1^{x_1} x_3^{x_4}$, and $x_7 = x_2^{x_1} x_4^{x_3} log(x_5 + 1)$). Realistic landscapes are simulated unconditional Gaussian random fields from an exponential covariance model with variance = 0.1 and scale = 0.1. The real landscapes are bio-climatic variables downloaded from `http://www.worldclim.org` and cropped to the extent of 5N24E to 7S37E, with x_1 = annual mean temperature, x_2 = precipitation of coldest quarter, x_3 = mean diurnal range, x_4 = annual precipitation, x_5 = temperature seasonality, x_6 = precipitation of warmest quarter, and x_7 = isothermality. The grid sizes for the simulated landscapes are 50 × 50 cells, and for the real landscape 72 × 78 cells.

Table 1: 12 datasets (= 2 landscapes * 2 distributions * 3 SAC) causes as minimal set of simulated data.

dataset	landscape	distribution	SAC cause
110	smooth	Gaussian	reference
111			SAC onto response variable
112			omitted predictor
120		Bernoulli	reference
121			SAC onto response variable
122			omitted predictor
310	real	Gaussian	reference
311			SAC onto response variable
312			omitted predictor
320		Bernoulli	reference
321			SAC onto response variable
322			omitted predictor

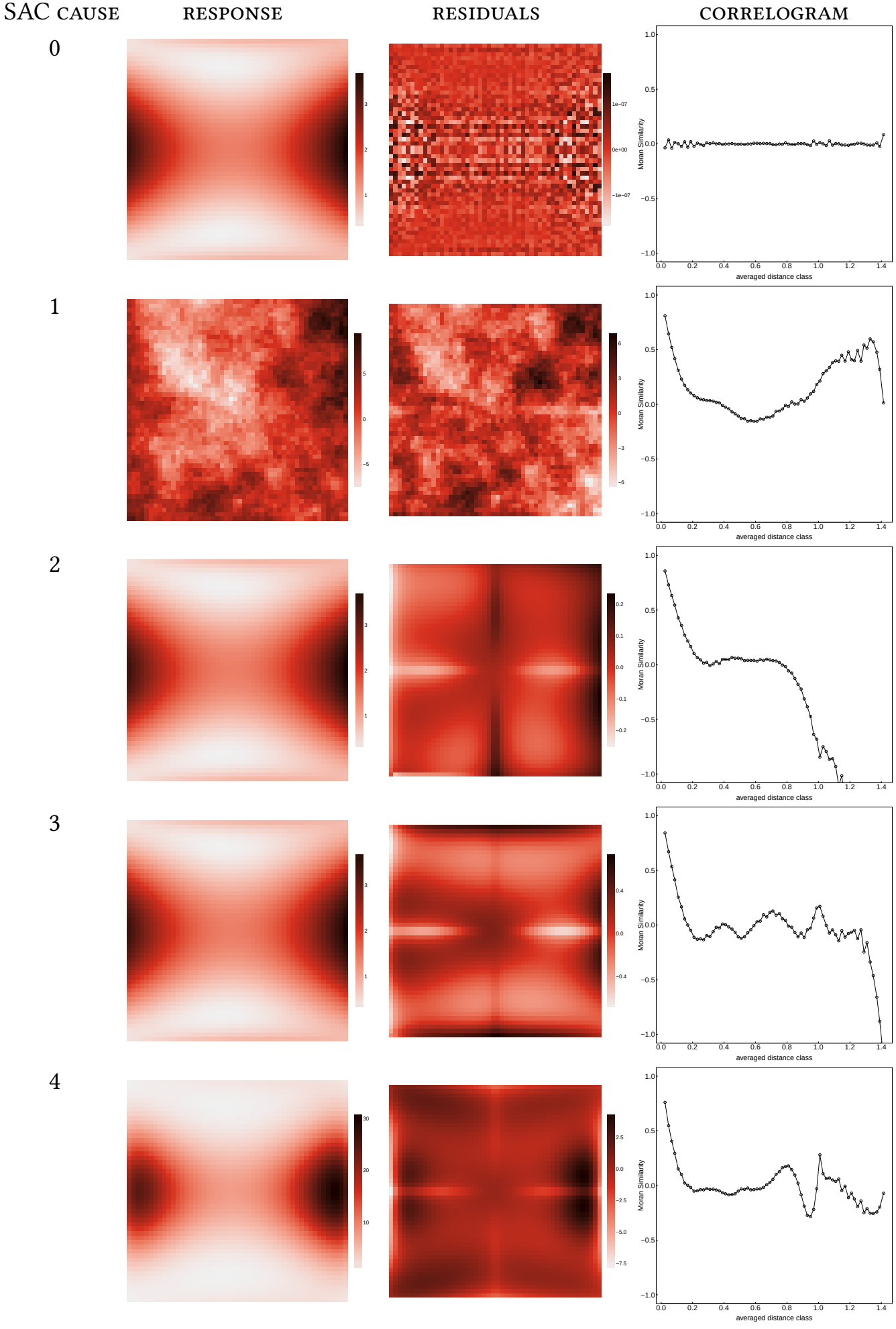


Figure 2: Response and residuals maps and correlograms for smooth landscapes, Gaussian distribution and five different spatial autocorrelation causes.