

Mutual information analysis to approach nonlinearity in groundwater stochastic fields

*Original*

Mutual information analysis to approach nonlinearity in groundwater stochastic fields / Butera, Ilaria; Vallivero, Luca; Ridolfi, Luca. - In: STOCHASTIC ENVIRONMENTAL RESEARCH AND RISK ASSESSMENT. - ISSN 1436-3240. - STAMPA. - 32:10(2018), pp. 2933-2942. [10.1007/s00477-018-1591-4]

*Availability:*

This version is available at: 11583/2742271 since: 2019-07-16T11:33:06Z

*Publisher:*

Springer

*Published*

DOI:10.1007/s00477-018-1591-4

*Terms of use:*

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

*Publisher copyright*

Springer postprint/Author's Accepted Manuscript

This version of the article has been accepted for publication, after peer review (when applicable) and is subject to Springer Nature's AM terms of use, but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: <http://dx.doi.org/10.1007/s00477-018-1591-4>

(Article begins on next page)

# 1 ***Mutual information analysis to approach nonlinearity in groundwater*** 2 ***stochastic fields***

3 Ilaria Butera, Luca Vallivero, Luca Ridolfi

4 Department of Environment, Land and Infrastructure Engineering-Politecnico di Torino

5 Corresponding Author: Ilaria Butera phone +390110905673 email: [ilaria.butera@polito.it](mailto:ilaria.butera@polito.it)

## 6 ***Abstract***

7 In heterogeneous porous media, transmissivity can be regarded as a spatial stochastic variable.  
8 Transmissivity fluctuations induce stochasticity in the groundwater velocity field and transport features. In  
9 order to model subsurface phenomena, it is important to understand the relationships that exist between  
10 the variables that characterize flow and transport. Linear relationships are easier to deal with.  
11 Nevertheless, it is well known that flow and transport variables exhibit interdependences that become  
12 more and more nonlinear as the heterogeneity increases.

13 The aim of this work is to draw attention to the information contained in nonlinear linkages, and to show  
14 that it can be of great relevance with respect to the linear information content. Information theory tools  
15 are proposed to detect the presence of nonlinear components. By comparing the cross-covariance function  
16 and mutual information, the amount of linear linkage is compared with nonlinear linkage. In order to avoid  
17 analytical approximations, data from Monte Carlo simulations of heterogeneous transmissivity fields have  
18 been considered in the analysis. The obtained results show that the presence of nonlinear components can  
19 be relevant, even when the cross-covariance values are nil.

20

## 21 **Key-Words**

22 Nonlinearity, Mutual Information, Heterogeneous transmissivity fields, Groundwater stochastic fields

## 23 ***1. Introduction***

24 Groundwater is the most relevant source of high quality fresh water. However, groundwater is vulnerable:  
25 overexploitation and pollution constitute an increasing threat. In order to manage this precious resource,  
26 studies are necessary to obtain a better understanding of flow and transport phenomena. Over the past  
27 few decades, the difficulty of obtaining detailed knowledge about the spatial distribution of aquifer  
28 parameters, and hydraulic conductivity in particular, has led to the development of stochastic approaches  
29 in order to resolve groundwater issues by means of numerical and analytical studies (e.g., Dagan 1989;  
30 Dagan and Neuman 1997; Rubin 2003). Hydraulic conductivity is modelled as a spatial random function  
31 with given statistical proprieties, which are inferred from field data analysis and, as a consequence,  
32 hydraulic heads, velocity components and solute trajectories also become stochastic variables.

33 In the past, a great deal of attention was paid to linear stochastic theory (e.g., Dagan 1984, 1989; Rubin  
34 1991, 2003). In this case, the log-conductivity field is approximated by its first-order perturbation expansion  
35 and it is inserted into the continuity equation and Darcy's law. According to some hypotheses on the flow  
36 and the domain size, linear theory is able to provide analytical expressions for the first and second  
37 statistical moments of local variables (e.g. log-conductivity, head fluctuations, flow velocity components,  
38 whose values only depend on their location in the space) and transport variables (trajectory fluctuations,  
39 spatial moments of the plume) that depend on the entire transport process. The adoption of the first-order  
40 perturbation expression would limit the applicability of the linear theory to low levels of transmissivity  
41 heterogeneity, i.e., the log-conductivity variance should be less than one. Nevertheless, numerical and  
42 analytical studies (e.g., Bellin 1992; Hsu et al. 1996; Dagan et al. 2003) have shown that a linear theory can  
43 be applied to higher heterogeneity levels, because of the balance of higher-order terms. The increased  
44 range of applicability of the linear theory has drawn attention to the possibility of examining the properties  
45 of flow and transport, which had previously been investigated mainly through their statistical moments  
46 (e.g., Dagan 1984, 1989; Dagan and Neuman 1997; Rubin 2003).

47 Attention has also been paid to understanding the role of the higher-order terms that were omitted in  
48 linear theory approximations, and to the consequent nonlinear relationships between flow and transport  
49 variables (e.g., Dagan 1994; Hsu et al. 1996; Salandin and Fiorotto 1998). Hsu et al. (1996) developed  
50 second-order analytical expressions for fluid velocity covariance functions and for the covariance functions  
51 of trajectory fluctuations. They observed that the impact of second-order terms becomes appreciable in  
52 transport processes when the log-conductivity variance approaches two. Analyzing the frequency  
53 distributions of the velocity components and trajectory fluctuations and their values of the statistical  
54 moment, Salandin and Fiorotto (1998) clearly evidenced the effects of nonlinear terms in both flow and  
55 transport features.

56 As the effect of nonlinear terms increases with the heterogeneity level, several studies have been carried  
57 out on flow and transport in highly heterogeneous media (e.g., Dagan et al. 2003; Fiori et al. 2003; Jankovic  
58 et al. 2003; Gotovac et al. 2009; Meyer and Tchelepi 2010). These authors focused on the analysis of the  
59 probability density functions (pdfs) and the statistical moments of the characterizing quantities, such as the  
60 velocity components, trajectory fluctuations and travel time.

61 In this picture, the aim of the present work has been to draw attention to the nonlinear dependence that  
62 exists among some groundwater variables. Such a dependence can be relevant, even when the linear  
63 linkages are negligible, and it offers information that can be important in a number of problems, such as in  
64 conditioning techniques and inverse problems.

65 In order to shed light on the nonlinear linkages, cross-covariance functions have been analyzed and  
66 information theory tools (i.e., mutual information, Shannon 1948) have been applied. Mutual information  
67 tools capture nonlinear relationships, while cross-covariance functions only grasp the linear relationship

68 between variables. The analysis is performed by processing data obtained from Monte Carlo simulations. In  
 69 this way, analytical expressions are not used for the covariance and cross-covariance functions, and the  
 70 terms that are neglected in their derivation do not affect the analysis.

71 Information theory tools have been largely used in other fields, such as economics, biology, mathematics  
 72 and geophysics (e.g., Islam and Sivakumar, 2002; Pluim et al, 2003; Leydesdorff et al., 2006; Donges et al.,  
 73 2009; Kinney and Atwal, 2014). Information theory has already been used to deal with groundwater  
 74 problems: for example, Woodbury and Ulrych (1993, 1996, 2000) successfully applied the principle of  
 75 minimum relative entropy to forward probabilistic modelling and to recover the release history of a  
 76 groundwater contaminant, while Kitanidis (1994) proposed the dilution index which is an adaptation of the  
 77 entropy expression. Mishra et al. (2009) proposed the use of mutual information analysis as a global  
 78 sensitivity analysis technique, instead of stepwise regression analysis. Gotovac et al. (2010) have recently  
 79 applied the maximum entropy principle to obtain the complete characterization of the travel time pdf  
 80 (probability density function). Zeng and Wu (2012) also applied mutual information to detect the most  
 81 important uncertainty factors in groundwater levels for a specific case study. However, mutual information  
 82 has never been adopted to detect the role of nonlinear components in groundwater transport processes.

## 83 **2. Methods**

84 The Bravis-Pearson index,  $\rho$ , is known as the linear correlation coefficient or Pearson correlation  
 85 coefficient. It is a measure of the linear dependence of two random variables. Given two variables  $x$  and  $y$ ,  
 86 and assuming that  $N$  couples of  $(x_i, y_i)$ , data are available, the linear correlation coefficient is defined as

$$87 \quad \rho = \frac{Cov(x, y)}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2 \sum_{i=1}^N (y_i - \bar{y})^2}}, \quad (1)$$

88 where  $Cov(x, y)$  is the covariance between  $x$  and  $y$ ,  $\sigma_x$  and  $\sigma_y$  are the standard deviations of  $x$  and  $y$ ,  
 89 respectively, and  $\bar{x}$  and  $\bar{y}$  are the mean values.

90 Considering the Schwarz inequality  $|Cov(x, y)| \leq \sigma_x \sigma_y$ , it follows that  $|\rho| \leq 1$ : if  $|\rho| = 1$ , a perfect linear  
 91 relationship exists between  $x$  and  $y$  and the variables are fully correlated; instead, if  $|\rho| = 0$ , the variables are  
 92 not correlated. It is well known that a nil value of the linear correlation coefficient does not mean that the  
 93 variables are independent of each other: in fact, there can be a nonlinear relationship that has not been  
 94 captured by the linear correlation coefficient.

95 Entropy is a measure of the uncertainty of a system (Shannon, 1948). If  $x$  is a discrete random variable with  
 96 pdf  $p(x)$  and  $N$  data of  $x$  are available, its entropy is

$$97 \quad H(x) = -\sum_{i=1}^N p(x_i) \ln p(x_i) \quad (2)$$

98 and if two random variables are considered,  $x$  and  $y$  with joint pdf  $p(x, y)$ , their joint entropy is given by

99 
$$H(x, y) = - \sum_{j=1}^N \sum_{i=1}^N p(x_i, y_j) \ln p(x_i, y_j). \quad (3)$$

100 Considering that  $x$  and  $y$  can be dependent on each other, mutual entropy is defined. Mutual entropy  
 101 between  $x$  and  $y$  represents the reduction in uncertainty of  $y$  as a result of the information on  $x$  (and *vice*  
 102 *versa*), and it is expressed as follows:

103 
$$I(x, y) = H(x) + H(y) - H(x, y) = \sum_{i=1}^N \sum_{j=1}^N p(x_i, y_j) \ln \frac{p(x_i, y_j)}{p(x_i)p(y_j)}, \quad (4)$$

104 where it can be verified that  $I(x, y) \geq 0$ .

105 In mutual information analysis, two indicators are used to measure the dependence of two variables, the  $\mathcal{U}$   
 106 uncertainty coefficient (Theil, 1972) and the  $R$  coefficient (Granger and Lin, 1994), namely

107 
$$\mathcal{U}(x, y) = 2 \frac{I(x, y)}{H(x)H(y)} \quad (0 \leq \mathcal{U} \leq 1), \quad (5)$$

108 
$$R(x, y) = \{1 - \exp[-2I(x, y)]\}^{1/2} \quad (0 \leq R \leq 1). \quad (6)$$

109 The  $\mathcal{U}$  measure lies between 0 and 1: when the uncertainty coefficient is zero, it means that  $x$  and  $y$  are not  
 110 dependent on each other; if its value is unitary, the knowledge of  $x$  is able to completely predict  $y$ , and the  
 111 opposite is also true, i.e., a one-to one relationship exists between  $x$  and  $y$ . Similarly, if  $R$  is zero,  $x$  and  $y$  are  
 112 independent, while  $R$  is equal to one when there is an exact (linear or nonlinear) dependence relationship  
 113 between  $x$  and  $y$ .

114 It can also be verified (Cover and Thomas, 1991) that when the bivariate distribution  $p(x, y)$  of  $x$  and  $y$  is  
 115 Gaussian then

116 
$$R(x, y) = |\rho(x, y)|. \quad (7)$$

117 Property (7) makes use of the  $R$  indicator particularly interesting to investigate the linear/nonlinear  
 118 relationship between two variables when one variable has a Gaussian pdf and the other one is presumed to  
 119 be linearly related to the first one. In this case, if the  $R$  and  $\rho$  values are identical, the relationship is purely  
 120 linear, otherwise their displacement is a proxy of nonlinear terms in the relationship between the variables.  
 121 Variables that deviate from Gaussianity are also considered in the present manuscript. In order to compare  
 122  $R$  and  $|\rho|$ , we selected cases where either at least one variable is Gaussian or the correlation coefficient is  
 123 almost nil. In the first case, as only one of the two variables is Gaussian,  $R - |\rho|$  being different from zero  
 124 implies that the other variable is not a pure linear function of  $x$ . In the second case, an approximately zero  
 125 value of  $|\rho|$  entails that the whole dependence embedded in  $R$  can be ascribed to nonlinear dependencies,  
 126 regardless of the pdf of the variables.

127 It can also be observed that both  $R$  and  $|\rho|$  varies from 0 and 1, and typical values of  $R=0.6-0.7$  mark a  
 128 strong association between the variables (e.g., Mishra et al. 2009), while  $|\rho|=0.6-0.7$  means an important  
 129 correlation (i.e., linear dependence). In the same way,  $R=0.2-0.3$  marks a weak association between

130 variables, while  $|\rho|=0.2-0.3$  means a weak correlation (i.e., linear dependence). Therefore,  $R-|\rho|$  seems to be  
 131 significant when it is greater than 0.3-0.4, namely when  $R$  and  $|\rho|$  clearly describe a different degree of  
 132 association.

### 133 **3. Problem statement**

134 In this work, flow and transport phenomena have been considered through heterogeneous porous media.  
 135 A simple flow scheme is proposed, where the source of non-linearity is easily controlled by the log-  
 136 transmissivity heterogeneity level. Two-dimensional confined aquifers, without recharging, are considered:  
 137 the boundary conditions are constant in time and produce a flow that develops in the  $\{x_1, x_2\}$  plane (see  
 138 Fig. 1): the beds of the confined aquifer are plane and parallel (the thickness,  $B$ , of the aquifer is constant),  
 139 one of the principal anisotropy directions,  $x_3$ , is orthogonal to the confining beds and the hydraulic head  
 140 gradient in the  $\{x_1, x_2\}$  plane does not depend on  $x_3$  (e.g., de Marsily 1981). Heterogeneity is due to spatial  
 141 variations of transmissivity, while the effective porosity is considered to be constant.

142 The approach to the problem is stochastic: transmissivity is a spatial random function with statistical  
 143 features that are inferred from the data. The aquifer is considered a realization of an ensemble of  
 144 statistically equivalent aquifers.

145 The velocity field and the hydraulic head field are related to the transmissivity field through Darcy's law and  
 146 the continuity equation

$$147 \quad U_i(\mathbf{x}) = \frac{1}{nB} \sum_{j=1}^2 (T_{ij}(\mathbf{x}) J_j(\mathbf{x})) \quad i=1,2, \quad (8)$$

$$148 \quad \nabla \cdot (\mathbf{T} \nabla H) = 0, \quad (9)$$

149 where  $U_i$  is the seepage velocity component ( $i=1,2$ ),  $n$  is the effective porosity,  $\mathbf{J}(\mathbf{x})$  is the hydraulic head  
 150 gradient,  $T_{ij}$  is the point value of the transmissivity tensor  $\mathbf{T}$  and  $H$  is the hydraulic head.

151 Because of the stochasticity of transmissivity, the velocity components and hydraulic heads are also  
 152 stochastic. In each realization, the value of a variable at a given point is characterized by a fluctuation value:  
 153  $v_i(\mathbf{x}) = U_i(\mathbf{x}) - \langle U_i(\mathbf{x}) \rangle$ , where  $v_i(\mathbf{x})$  is the velocity component fluctuation in direction  $i$  ( $i=1,2$ ),  $U_i(\mathbf{x})$  is the velocity  
 154 value at location  $\mathbf{x}$  and the symbol  $\langle \rangle$  denotes the ensemble mean operator; similarly,  $h(\mathbf{x}) = H(\mathbf{x}) - \langle H(\mathbf{x}) \rangle$  is  
 155 the hydraulic head fluctuation.

156 Transport processes are affected by the stochasticity of the flow field. Considering the motion of a particle  
 157 released into an aquifer at time  $t=0$  in  $\mathbf{x}_0=(0,0)$ , its location at time  $t$  — given by  $\mathbf{X}(t) = \int_0^t \mathbf{U}(\mathbf{X}(t)) dt$  — is  
 158 stochastic: the ensemble mean location has coordinates  $\langle X_1(t) \rangle$  and  $\langle X_2(t) \rangle$ , along the  $x_1$  and  $x_2$  axes,  
 159 respectively, and the trajectory fluctuations are given in each realization by  $X'_1(t) = X_1(t) - \langle X_1(t) \rangle$  and  $X'_2(t) =$   
 160  $X_2(t) - \langle X_2(t) \rangle$ .

161 In this work, in order to take advantage of eq.(7) and to detect nonlinearity, the log-transmissivity field,  
 162  $Y(\mathbf{x})=\ln(T(\mathbf{x}))$ , is assumed to be a stationary second order field with a multivariate normal (MVN)  
 163 distribution, according to classic stochastic approaches (e.g., Delhomme 1979; Dagan 1984). As  $Y$  has an  
 164 MVN pdf, if the flow and transport variables are linearly related to  $Y$ , they also have an MVN pdf and the  
 165 bivariate distribution of these variables is normal: in this case, eq.(7) is verified and  $R=|\rho|$ . Instead, the non-  
 166 equality between  $R$  and  $\rho$  values points out the presence of nonlinear terms in the relationships between  
 167 the considered variables. The impact of nonlinear terms increases as the difference between  $R$  and  
 168  $\rho$  increases.

169 The numerical data processed in this work were obtained by means of Monte Carlo simulations of the  
 170 transmissivity field. The numerical approach is the same as the one that was used in previous works (e.g.,  
 171 Butera et al. 2009; Butera and Soffia 2017). A brief description of the Monte Carlo set up is presented  
 172 hereafter.

173 Two-dimensional heterogeneous transmissivity fields in the  $\{x_1, x_2\}$  plane, which model confined aquifers  
 174 with horizontal flow, were generated through the Fast Fourier Transform method (Gutjahr 1989). The log-  
 175 transmissivity field,  $Y(\mathbf{x})=\ln(T(\mathbf{x}))$ , is assumed (i) to be a stationary second order field, (ii) to have an MVN  
 176 pdf and (iii) to be characterized by an exponential covariance function  $C_Y(\mathbf{x}, \mathbf{x}') = \sigma_Y^2 \exp(-r / \ell_Y)$ , where  
 177  $r = |\mathbf{x} - \mathbf{x}'|$  and  $\ell_Y$  is the correlation length of log-transmissivity.

178 Since the impact of the nonlinear terms increases as the transmissivity field variance increases, two  
 179 heterogeneity levels were considered —  $\sigma_Y^2=0.16$  and  $\sigma_Y^2 =2$  — to reproduce weakly and mildly  
 180 heterogeneous aquifers. The generated transmissivity fields had a square shape with a size equal to  $42\ell_Y$   
 181 and were subdivided into  $252 \times 252$  blocks with a side size equal to  $\ell_Y/6$ ; a value of transmissivity was  
 182 assigned to each block. If the ensemble mean of the velocity values is computed, the flow is uniform, it  
 183 evolves along direction  $x_1$  and it is obtained by assigning an impervious boundary condition to the northern  
 184 and southern sides and fixed head values to the western and eastern sides (Fig.1).

185 Hydraulic heads,  $H(\mathbf{x})$ , were computed at the nodes of the transmissivity blocks, and eq. (9) was solved by  
 186 means of the Galerkin finite element method. In order to avoid boundary effects, the external frame (with  
 187  $6\ell_Y$  width) was no longer considered in the subsequent analysis of the local and transport variables. The  
 188 transport simulation considered an instantaneous release of the solute from a point source: the particle  
 189 trajectory was computed in the  $0\tau - 21\tau$  interval, using the particle tracking method, where  $\tau=t\langle U_1 \rangle / \ell_Y$  is  
 190 the dimensionless time and  $\langle U_1 \rangle$  is the ensemble mean velocity, which is uniform in space. In each  
 191 realization, a particle was released into  $\mathbf{x}_0=(0,0)$ , using the coordinate systems shown in Fig.1.

192 A total of 1500 realizations were performed for the smaller log-transmissivity variance and 3000 for the  
 193 higher . The number of simulations was chosen to ensure the convergence of the second moments of both  
 194 the velocity and trajectory components. The time step used in the particle tracking procedure is the

195 minimum between  $\Delta\tau_1$  and  $\Delta\tau_2$ , where  $\Delta\tau_1=0.2*\Delta x/v_{1max}$  ( $\Delta x$  is the grid side size, which is equal to  $l_x/6$ ,  
 196 and  $v_{1max}$  is the maximum velocity value along direction 1 in the simulation) and  $\Delta\tau_2=\tau-\tau_{rec}$  ( $\tau_{rec}$  is the  
 197 recording time).

198 Standard routines that implement eq. (1) were used to compute the linear correlation coefficients.  
 199 Different estimators were applied to evaluate the mutual entropy from a data series (e.g., see Papan and  
 200 Kugiumtzis 2008) and tested. The tests, which are not reported here, considered both noised linear series  
 201 and noised nonlinear series; the latter were obtained using Henon and Mackey-Glass models. Four different  
 202 estimators were implemented: the histogram method — i.e., eq. (4) was computed from the empirical  $p(x)$   
 203 and  $p(x,y)$ , and the results were affected to a great extent by the choice of the binning — and those based  
 204 on the  $k$  nearestneighbours, that is, the Kozachenko and Leonenko (1987) estimator and the Kraskov et al.  
 205 (2004) estimator. The tests considered time series with up to 4000 elements and with a normal bivariate  
 206 distribution: therefore,  $R$  was expected to be equal to  $|\rho|$ . The histogram method and the Kozachenko and  
 207 Leonenko method gave the worst performances, as they resulted to be affected to a great extent by the  
 208 series' number of the elements. The two algorithms proposed by Kraskov et al. (2004), which compute  
 209 mutual entropy without using eq.(4), were found to be equivalent and produced better results when the  
 210 free parameter was set equal to three. Accordingly, the second algorithm (i.e.  $l^2$ ) proposed by Kraskov was  
 211 used in the subsequent computations.

212 The  $R$  parameter was found to be very sensitive to  $l$  fluctuations (numerical error) close to zero. In order to  
 213 smooth the spurious fluctuations of the  $R$  parameter in Figs. 2-4, a moving average was applied in some  
 214 cases, paying attention not to affect the trend of the data. The size of the moving average window is  
 215 specified in the figure captions.

#### 216 **4. Results and Discussion**

217 The results obtained after processing the data of the Monte Carlo experiments are reported hereafter. The  
 218 analysis considers both local variables (log-transmissivity, velocity and hydraulic head), whose fluctuations  
 219 constitute spatial random functions, and non-local variables (i.e., trajectories), whose fluctuations at a  
 220 given time are the result of a path through the heterogeneous field.

221 The behaviour of the absolute value of the linear correlation coefficient and parameter  $R$  is compared in  
 222 Fig.2, which refers to the following local variables:  $Y(\mathbf{x})$  (log-transmissivity),  $v_1(\mathbf{x})$  (fluctuation of the  
 223 longitudinal velocity component),  $v_2(\mathbf{x})$  (fluctuation of the transversal velocity component) and  $h(\mathbf{x})$   
 224 (hydraulic head fluctuation). The  $Y$  values in all the frames in Fig. 2 were measured in  $\mathbf{x}_0=(0,0)$ , while the  
 225 other variables were sampled at  $\mathbf{x}_i=(x_i,0)$  locations, along the longitudinal axis and passing through the  
 226 origin of the reference system (see Fig. 1).

227 Figs 2a and 2b refer to the relationship between log-transmissivity and hydraulic head for two  
 228 heterogeneity levels. A good agreement between  $R$  and  $|\rho|$  can be noted for both of the heterogeneity

229 levels. This agreement denotes the absence of significant nonlinear linkages between these variables, that  
 230 is, up to  $\sigma^2_Y=2.0$ .  $R$  is below the correlation coefficient at some points; this is a numerical artefact that  
 231 occurs for small  $R$  values, due to its high sensitivity to small errors in the computation of  $I$ . Figs 2c and 2d  
 232 show the relationship between log-transmissivity and the longitudinal velocity component. A good  
 233 agreement between  $R$  and  $|\rho|$  can be noted for the smaller heterogeneity variance (Fig. 2c), which excludes  
 234 the presence of important nonlinear relationships between those variables, while the agreement is not so  
 235 good in Fig. 2d when  $Y$  and  $v_1$  are at the same location, thus indicating that nonlinearity occurs.

236 Figs 2e and 2f are more interesting. It can be seen that while the correlation function is close to zero for  
 237 almost every point, the  $R$  curve has a peak at zero, that is, when  $Y(\mathbf{x})$  and  $v_2(\mathbf{x})$  are measured at the same  
 238 location, or at a short distance aligned along the flow direction, the dependence between the variables is  
 239 fully nonlinear. This fact shows that, although the correlation function is zero, transversal velocity  
 240 components depend on the  $Y$  fluctuations, which modify the flow field around it: the dependence is weak  
 241 for  $\sigma^2_Y=0.16$  ( $R=0.28$ ), but is quite important for  $\sigma^2_Y=2.0$  ( $R=0.59$ ). It is worth noting that the nonlinear  
 242 relationship denoted by  $R$  cannot be captured by higher-order analytical covariance functions, which could  
 243 resemble the numerical covariance values.

244 The difference between  $R$  and  $|\rho|$  shown in Figs. 2 denotes that even when the log-transmissivity field has  
 245 been generated with a Gaussian pdf, the velocity components and hydraulic head appear to deviate from a  
 246 Gaussian distribution. This fact is in agreement with the results of numerical analyses (e.g., Bellin et al.  
 247 1992, Salandin and Fiorotto 1998). The behaviour of the  $R$  parameter and the behaviour of the linear  
 248 correlation coefficient are shown in Figs 3 and 4, considering the trajectory fluctuations (the non-local  
 249 variable) and a local variable. Fig. 3 shows the relationship between the log-transmissivity fluctuations at  
 250  $\mathbf{x}_0=(0,0)$  (i.e. the solute injection point) and the trajectory fluctuations at a given time. The results for the  
 251 largest heterogeneity level are shown only up to  $\tau=8.4$ , because some particles exit from the numerical  
 252 domain for larger times, and the trajectory statistics cannot be computed, as only the slower particles  
 253 would be considered. The use of mutual-information-based tools for a higher heterogeneity level allows us  
 254 to capture the presence and the importance of nonlinear relationships compared to linear ones. In fact,  
 255 considering both  $X_1(t)$  and  $X_2(t)$ , there is a rough agreement between  $R$  and  $|\rho|$  for the lower heterogeneity  
 256 level (Figs. 3a and 3c), while the  $R$  parameter is clearly above  $|\rho|$  for the higher heterogeneity level (Figs. 3b  
 257 and 3d). This is much more evident in Fig. 3d, which considers  $Y(0,0)$  and  $X_2(\tau)$ : although the correlation  
 258 value is almost nil,  $R$  varies from 0.46 to 0.18, thus showing a moderate association between the variables  
 259 for early travel times.

260 The presence of nonlinear terms is also evident in Fig 4, where selected cases are shown in order to depict  
 261 the role of nonlinear dependences between velocity and trajectory fluctuation. Figs. 4a and 4b illustrate the  
 262 behaviour of  $R$  and  $|\rho|$  for  $v_2(0,0)$  and  $X_1(\tau)$ , while Figs. 4c and 4d show the behavior of  $R$  for  $v_1(0,0)$  and  
 263  $X_2(\tau)$ . In these cases, the correlation coefficient is almost nil, thus denoting the absence of any significant

264 linear relationships between the variables, while  $R$  is not nil, that is, all the relationships between the  
265 variables are nonlinear. The  $R$  value decreases with time, as expected, and it shows that there is a good  
266 association between the variables at an early time for the higher heterogeneity level, which is not captured  
267 by the correlation coefficient. The results shown in Fig. 4 suggest that nonlinearity plays a key role in the  
268 relationship between velocity and trajectory components, when different directions ( $v_1-X_2$ ;  $v_2-X_1$ ) are  
269 considered. These nonlinear dependences contain important information that can be useful to both  
270 understand the phenomena and to solve, for instance, conditioning and inverse problems.

## 271 **5. Conclusions**

272 In this work, information theory tools have been applied to draw attention to the presence of non-  
273 negligible nonlinear terms in the relationships between the flow and transport variables that take place in  
274 heterogeneous porous formations. The analysis was aimed at pointing out the nonlinear interdependence  
275 that exists between variables, and its weight with respect to the linear interdependence. Multi-Gaussian  
276 transmissivity fields were considered to take advantage of the relationship that exists between the mutual  
277 information  $R$  parameter and the correlation coefficient  $|\rho|$  for normal bivariate distributions: in this case,  
278 linear and nonlinear contributions can clearly be identified. In order to protect the analysis from analytical  
279 approximation effects, numerical data from Monte Carlo experiments were used.

280 The obtained results show that nonlinear components can be relevant for mildly heterogeneous aquifers  
281 ( $\sigma^2_\gamma=2$ ) and that the use of covariance/cross-covariance functions can be somewhat limiting to investigate  
282 the relationships that exist between groundwater variables and to manage field data. Nonlinear  
283 relationships are less important in weakly heterogeneous aquifers ( $\sigma^2_\gamma=0.16$ ), but they show that, in some  
284 cases, variables with nil correlation coefficients are not independent.

285 The unsuitability of the covariance/cross-covariance functions to address nonlinearity can be extended to  
286 non-Gaussian transmissivity fields (Gomez-Hernandez and Wen 1998, Riva et al. 2017); however, in this  
287 case, a direct comparison of the mutual information  $R$  parameter and the correlation coefficient  $|\rho|$  cannot  
288 be made.

289 According to the Authors, mutual information-based tools could be applied extensively in groundwater  
290 analyses in order to shed light on the nonlinear relationships that exist among groundwater variables. Such  
291 tools could improve the understanding of the subsurface flow and of transport phenomena and their  
292 forecasting, and could thus be used to support other statistical approaches, e.g. geostatistical methods that  
293 are based on covariance functions and which can only be applied in the case of linear relationships  
294 between variables.

295 Future developments of the research could include the impact of small deviations from Gaussianity on the  
296  $R-|\rho|$  metric, the effect of molecular diffusion, the role of the covariance structure of nonlinear dependence,

297 other nonlinearity sources (e.g., more complex flow fields), and methods to incorporate the information  
 298 contained in nonlinear relationships in model developments and parameterizations.

299

### 300 *Acknowledgements*

301 The Authors would like to thank Tomas Aquino and an anonymous reviewer, whose comments have  
 302 helped to improve the manuscript.

303

304

### 305 *References*

306 Bellin A, Salandin P, Rinaldo A (1992) Simulation of dispersion in heterogeneous porous formations:  
 307 statistics, first-order theories, Convergence of Computation. *Water Resources Research*. 28(9): 2211-2227

308 Butera I, Soffia C (2017) Cokriging transmissivity, head and trajectory data for transmissivity and solute path  
 309 estimation. *Groundwater*. 55(3): 362-374; doi: 10.1111/gwat.12483

310 Butera I, Cotto I, Ostorero V (2009) A geostatistical approach to the estimation of a solute trajectory  
 311 through porous formations. *Journal of Hydrology*. 375(3-4): 354-355; 10.1016/j.jhydrol.2009.06.029

312 Cover TM, Thomas JA (1991) *Elements of Information Theory*. New York, John Wiley & Sons.

313 Dagan G (1984) Solute transport in heterogeneous porous formations. *Journal of Fluid Mechanics*. 145:  
 314 151-177.

315 Dagan G (1989) *Flow and Transport in Porous Formations*. Springer Verlag.

316 Dagan G (1994) An exact nonlinear correction to transverse macrodispersivity for transport in  
 317 heterogeneous formations. *Water Resources Research*. 30(10): 2699-2705.

318 Dagan G, Neuman SP (1997) *Subsurface flow and transport : a stochastic approach*. Cambridge University  
 319 Press.

320 Dagan G, Fiori A, Jankovic I (2003) Flow and transport in highly heterogeneous formations: 1. Conceptual  
 321 framework and validity of first-order approximations. *Water Resources Research*. 39(9): 14-1, 14-11. doi:  
 322 10.1029/2002WR001717

323 Delhomme JP (1979) Spatial variability and uncertainty in groundwater flow parameters: a geostatistical  
 324 approach. *Water Resources Research*. 15: 269-280

325 De Marsily G (1986) *Quantitative Hydrogeology*. Academic Press.

326 Donges JF, Zou Y, Marwan N, Kurths K (2009). The backbone of climate. *EPL*, 87(4), 48007. doi:  
 327 10.1209/0295-5075/87/48007

- 328 Fiori A, Jankovic I, Dagan G (2003) Flow and transport in highly heterogeneous formations: 2. Semianalytical  
329 results for isotropic media. *Water Resources Research*. 39(9): 15-1, 15-9. doi: 10.1029/2002WR001719
- 330 Gomez-Hernandez JJ, Wen X-H (1998) To be or not to be multi-Gaussian? A reflection on stochastic  
331 hydrogeology. *Advances in Water Resources*. 21(1): 47-61.
- 332 Gotovac H, Cvetkovic V, Andrievic R (2009) Flow and transport statistics in highly heterogeneous porous  
333 media. *Water Resources Research*. 45, W07402. doi: 10.1029/2008WR007168.
- 334 Gotovac H, Cvetkovic V, Andrievic R (2010) Significance of higher moments for complete characterization of  
335 the travel time probability density function in heterogeneous porous media using the maximum entropy  
336 principle. *Water Resources Research*. 46, W05502. doi: 10.1029/2009WR008220.
- 337 Granger CWJ, Lin J (1994) Using the mutual information coefficient to identify lags in nonlinear models.  
338 *Journal of Time Series Analysis*. 15: 371-384.
- 339 Gutjahr AL (1989) Fast Fourier transforms for random field generation: project report for Los Alamos Grant  
340 to New Mexico Tech.
- 341 Hsu KC, Zhang D, Neuman SP (1996) Higher-order effects on flow and transport in randomly heterogeneous  
342 porous media. *Water Resources Research*. 32(3): 571-582.
- 343 Islam MN, Sivakumar B (2002) Characterization and prediction of runoff dynamics: a nonlinear dynamical  
344 view. *Advances in Water Resources*, 25(2): 176-190. doi: 10.1016/S0309-1708(01)00053-7
- 345 Jankovic I, Fiori A, Dagan G (2003) Flow and transport in highly heterogeneous formations: 3. Numerical  
346 simulations and comparison with theoretical results. *Water Resources Research*. 39(9): 16-1, 16-13. doi:  
347 10.1029/2002WR001721
- 348 Kitanidis PK (1994) The concept of the dilution index. *Water Resources Research*. 30(7):2011-2026.
- 349 Kinney JB, Atwal GW (2014) Equitability, mutual information, and the maximal information coefficient.  
350 *PNAS*, 111(9): 3354-3359. doi: 10.1073/pnas.1309933111
- 351 Kozachenko LF, Leonenko NN (1987) Sample estimate of the entropy of a random vector. *Problems of*  
352 *information transmission*. 23(1): 95-101.
- 353 Kraskov A, Stogbauer H, Grassberger P (2004) Estimating mutual information. *Physical Review*, E  
354 69(066138). doi: 10.1103/PhysRevE.69.066138
- 355 Leydesdorff L, Dolfsma W, Van der Panne G (2006) Measuring the knowledge base of an economy in terms  
356 of triple-helix relations among “technology, organization, and territory”, *Research Policy*, 35(2): 181-199.

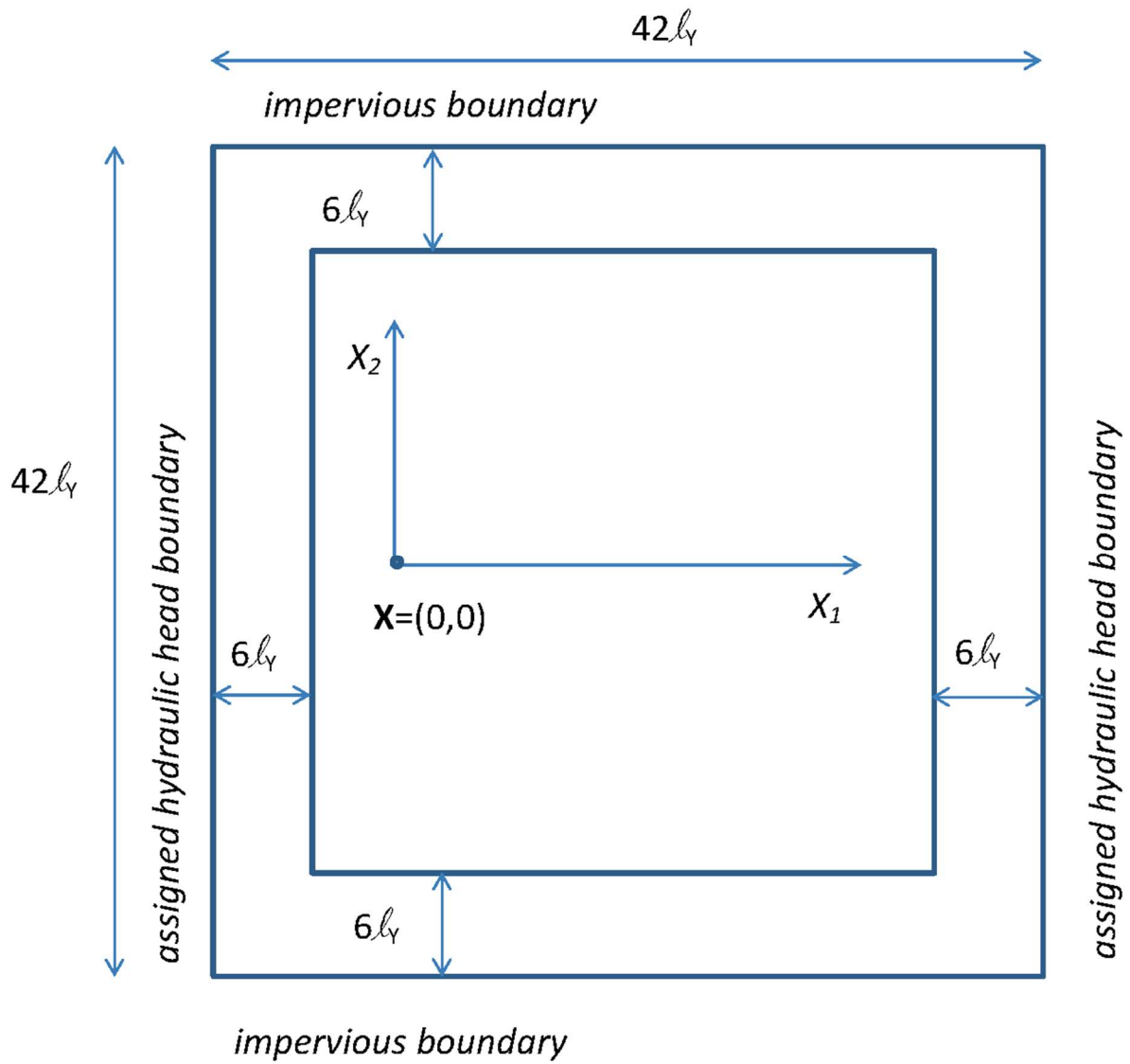
- 357 Meyer DW, Tchelepi HA (2010). Particle-based transport model with Markovian velocity processes for  
358 tracer dispersion in highly heterogeneous porous media. *Water Resources Research*. 46, W11552. doi:  
359 10.1029/2009WR008925
- 360 Mishra S, Deeds N, Ruskauff G (2009) Global sensitivity analysis techniques for probabilistic groundwater  
361 modelling. *Ground Water*, 47: 730-747. doi: 10.1111/j.1745-6584.2009.00604.x
- 362 Papana A, Kugiumtzis D (2008) Evaluation of mutual information estimators on nonlinear dynamic systems.  
363 *Nonlinear Phenomena in Complex System*. 11(2): 225-232.
- 364 Pluim JPW, Maintz JBA, Viergever MA (2003) Mutual-information-based registration of medical images: A  
365 survey. *IEEE Trans. Medical Imaging*, 22(8): 986-1004. doi: 10.1109/TMI.2003.815867
- 366 Riva M, Guadagnini A, Neuman SP (2017) Theoretical analysis of non-Gaussian effects on subsurface flow  
367 and transport. *Water Resources Research*. 53: 2998-3012. doi: 10.1002/2016WR019353
- 368 Rubin Y (1991) Prediction of tracer plume migration in disordered porous media by the method of  
369 conditional probabilities. *Water Resources Research*. 27(6): 1291-1308.
- 370 Rubin, Y., 2003. *Applied Stochastic Hydrogeology*. Oxford University Press.
- 371 Salandin P, Fiorotto V (1998) Solute transport in highly heterogeneous aquifers. *Water Resources Research*.  
372 34(5): 949 -961.
- 373 Shannon CE (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*. 27, 379-  
374 423.
- 375 Theil H (1972) *Statistical Decomposition Analysis*. Amsterdam, North-Holland Publishing Co.
- 376 Zeng XK, Wan D, Wu JC (2012) Sensitivity analysis of the probability distribution of groundwater level series  
377 based on information entropy. *Stochastic Environmental Research and Risk Assessment*. 26: 345-356. doi:  
378 <https://doi.org/10.1007/s00477-012-0556-2>
- 379 Woodbury AD, Ulrych TJ (1993) Minimum relative entropy: Forward probabilistic modeling . *Water*  
380 *Resources Research*. 29(8): 2847-2860.
- 381 Woodbury AD, Ulrych TJ (1996) Minimum relative entropy inversion: Theory and application to recovering  
382 the release history of a groundwater contaminant. *Water Resources Research*. 32 (9): 2671-2681.
- 383 Woodbury AD, Ulrych TJ (2000) A full-Bayesian approach to the groundwater inverse problem for steady  
384 state flow. *Water Resources Research*. 36 (8): 2081-2093.
- 385
- 386
- 387

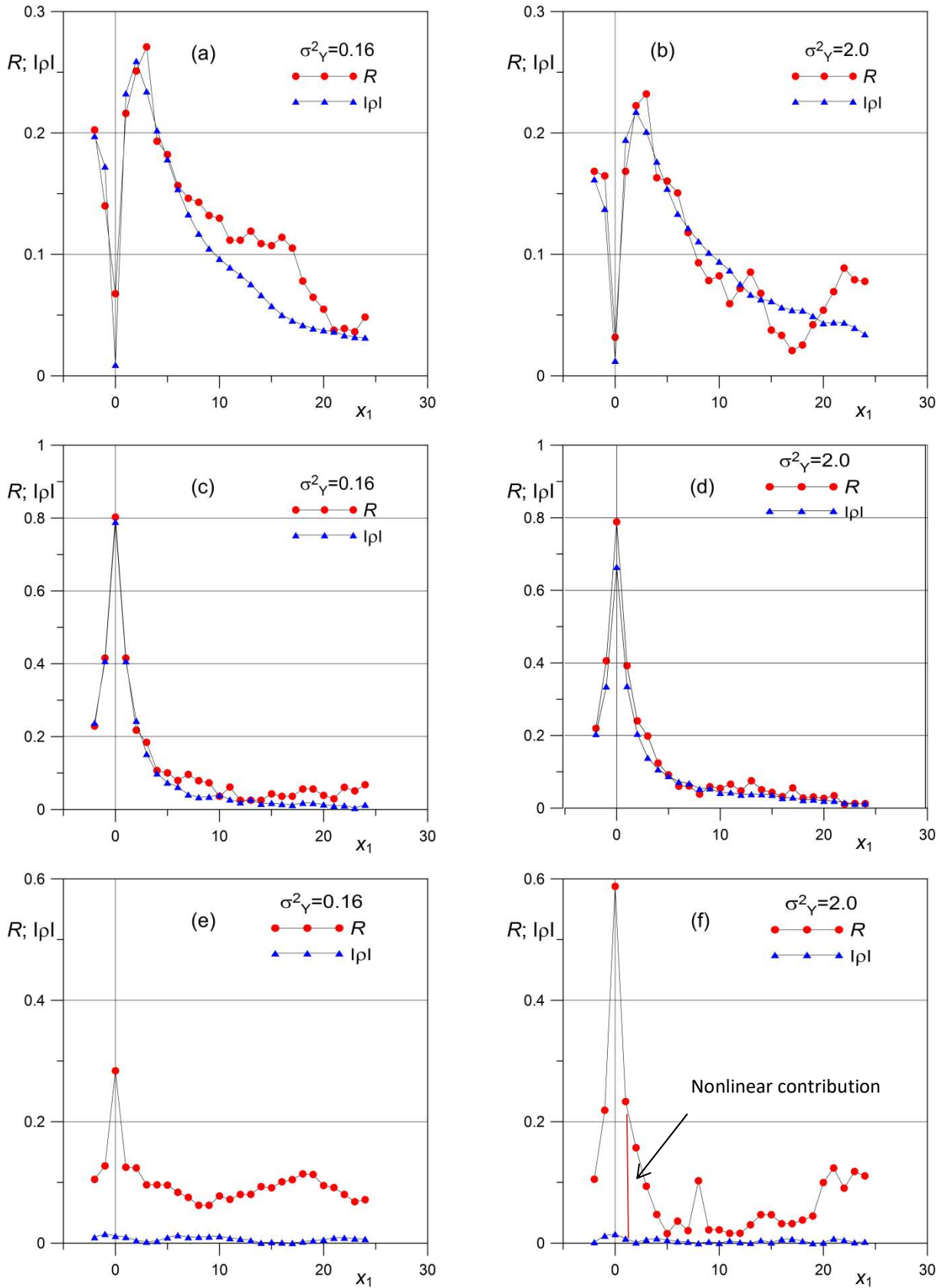
388 **Figures**

389

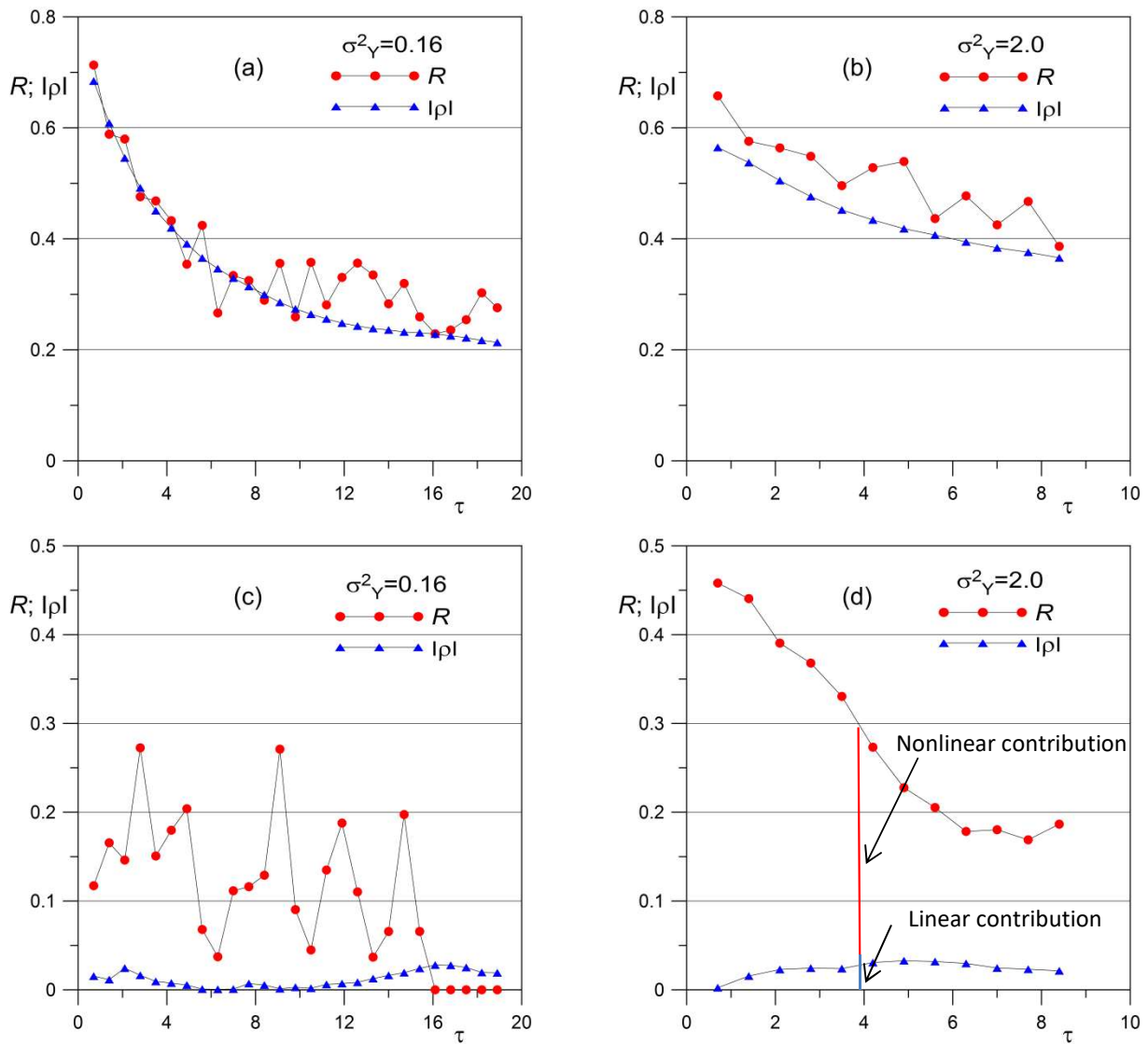
390

391

**Fig.1.** Sketch of the numerical domain.



**Fig. 2.**  $R$  (circle) and  $|\rho|$  (triangles) versus  $x_1$ , for different couples of variables. a)  $Y(0,0)-H(x_1)$ ,  $\sigma^2_\gamma=0.16$ ; b)  $Y(0,0)-H_1(t)$ ,  $\sigma^2_\gamma=2.0$ ; c)  $Y(0,0)-v_1(x_1)$ ,  $\sigma^2_\gamma=0.16$ ; d)  $Y(0,0)-v_1(x_1)$ ,  $\sigma^2_\gamma=2.0$ ; e)  $Y(0,0)-v_2(x_1)$ ,  $\sigma^2_\gamma=0.16$ ; f)  $Y(0,0)-v_2(x_1)$ ,  $\sigma^2_\gamma=2.0$ . The moving average window in the (a-b) panels is over five points for  $R < 0.18$  and  $x_1 > 0$ , while no moving average is applied to the (c-f) panels.



**Fig. 3.**  $R$  (circle) and  $|\rho|$  (triangles) behaviour versus  $\tau$  for different couples of variables. a)  $Y(0,0)-X_1(t)$ ,  $\sigma^2_\gamma=0.16$ ; b)  $Y(0,0)-X_1(t)$ ,  $\sigma^2_\gamma=2.0$ ; c)  $Y(0,0)-X_2(t)$ ,  $\sigma^2_\gamma=0.16$ ; d)  $Y(0,0)-X_2(t)$ ,  $\sigma^2_\gamma=2.0$ . The moving average window is over three points for  $R < 0.25$  in the (a,c) panels, while no moving average is applied to the (b,d) panels.

392

393

394

395

396

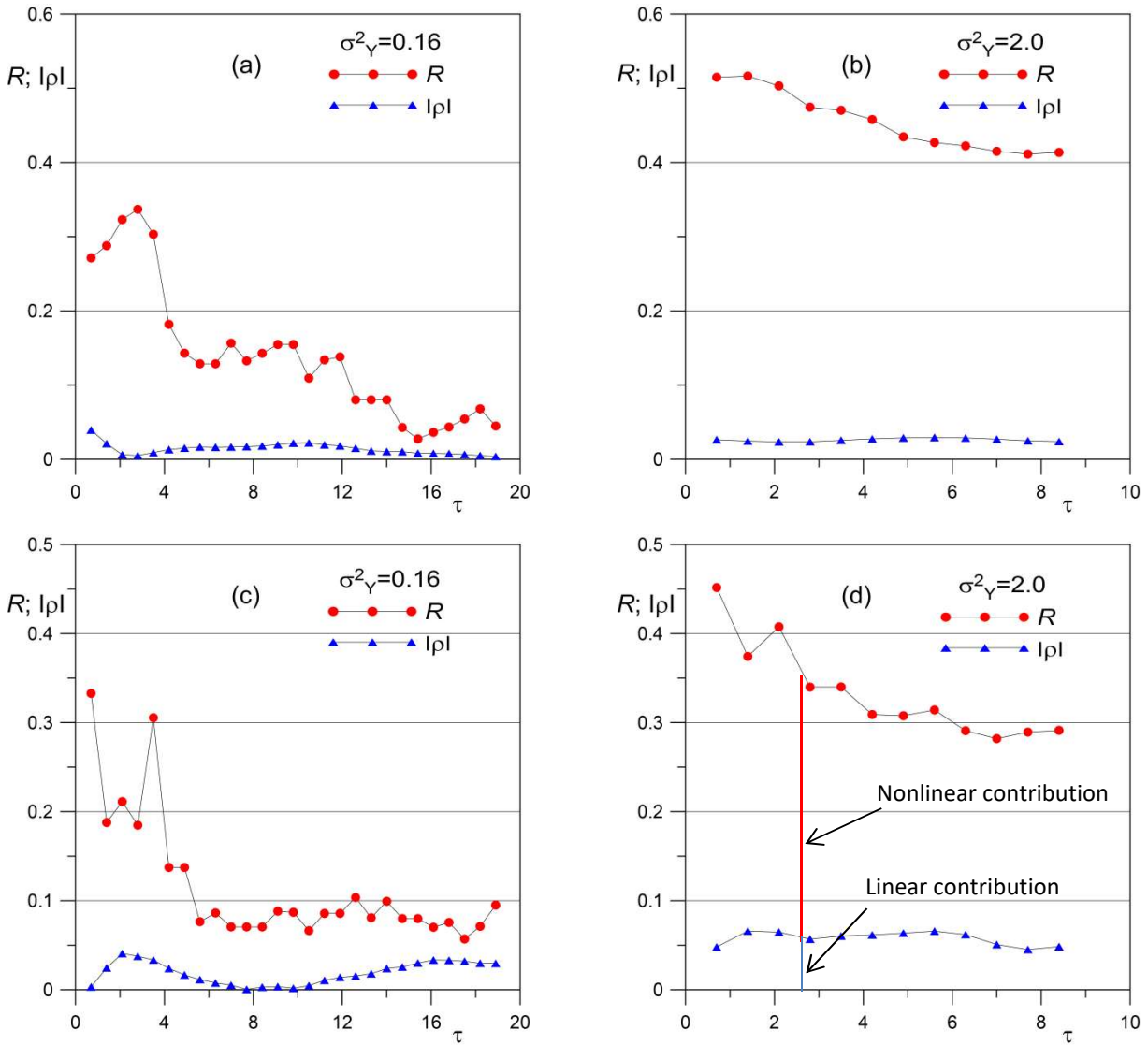
397

398

399

400

401



**Fig. 4.**  $R$  (circle) and  $|\rho|$  (triangles) behaviour versus  $\tau$ , for different couples of variables. a)  $v_2(0,0)-X_1(t)$ ,  $\sigma^2_\gamma=0.16$ ; b)  $v_2(0,0)-X_1(t)$ ,  $\sigma^2_\gamma=2.0$ ; c)  $v_1(0,0)-X_2(t)$ ,  $\sigma^2_\gamma=0.16$ ; d)  $v_1(0,0)-X_2(t)$ ,  $\sigma^2_\gamma=2.0$ . The moving average window in the(a,c) (b,d) panels is over five (three) points for  $R < 0.3$ .

402

403

404

405

406

407

408

409

410

411 **Figure captions**

412 **Fig.1.** Sketch of the numerical domain.

413 **Fig. 2.**  $R$  (circle) and  $|\rho|$  (triangles) versus  $x_1$ , for different couples of variables. a)  $Y(0,0)-H(x_1)$ ,  $\sigma^2_Y=0.16$ ; b)  $Y(0,0)-H_1(t)$ ,  
 414  $\sigma^2_Y=2.0$ ; c)  $Y(0,0)-v_1(x_1)$ ,  $\sigma^2_Y=0.16$ ; d)  $Y(0,0)-v_1(x_1)$ ,  $\sigma^2_Y=2.0$ ; e)  $Y(0,0)-v_2(x_1)$ ,  $\sigma^2_Y=0.16$ ; f)  $Y(0,0)-v_2(x_1)$ ,  $\sigma^2_Y=2.0$ . The  
 415 moving average window is over five points for  $R<0.18$  and  $x_1>0$  in the (a-b) panels, while no moving average is applied  
 416 to the (c-f) panels.

417 **Fig. 3.**  $R$  (circle) and  $|\rho|$  (triangles) behaviour versus  $\tau$ , for different couples of variables. a)  $Y(0,0)-X_1(t)$ ,  $\sigma^2_Y=0.16$ ;  
 418 b)  $Y(0,0)-X_1(t)$ ,  $\sigma^2_Y=2.0$ ; c)  $Y(0,0)-X_2(t)$ ,  $\sigma^2_Y=0.16$ ; d)  $Y(0,0)-X_2(t)$ ,  $\sigma^2_Y=2.0$ . The moving average window is over three  
 419 points for  $R<0.25$  in the (a,c) panels, while no moving average is applied to the (b,d) panels.

420 **Fig. 4.**  $R$  (circle) and  $|\rho|$  (triangles) behaviour versus  $\tau$ , for different couples of variables. a)  $v_2(0,0)-X_1(t)$ ,  $\sigma^2_Y=0.16$ ;  
 421 b)  $v_2(0,0)-X_1(t)$ ,  $\sigma^2_Y=2.0$ ; c)  $v_1(0,0)-X_2(t)$ ,  $\sigma^2_Y=0.16$ ; d)  $v_1(0,0)-X_2(t)$ ,  $\sigma^2_Y=2.0$ . The moving average window in the (a,c)  
 422 (b,d) panels is over five (three) points for  $R<0.3$ .

423