Nat. Hazards Earth Syst. Sci., 11, 2075–2083, 2011
www.nat-hazards-earth-syst-sci.net/11/2075/2011/
doi:10.5194/nhess-11-2075-2011

Natural Hazards and Earth System Sciences

An entropy approach for evaluating the maximum information content achievable by an urban rainfall network

E. Ridolfi, V. Montesarchio, F. Russo, and F. Napolitano

Dipartimento di Ingegneria Civile, Edile e Ambientale, Sapienza Università di Roma, Rome, Italy

Received: 31 March 2011 – Revised: 16 June 2011 – Accepted: 22 June 2011 – Published: 28 July 2011

Abstract. Hydrological models are the basis of operational flood-forecasting systems. The accuracy of these models is strongly dependent on the quality and quantity of the input information represented by rainfall height. Finer space-time rainfall resolution results in more accurate hazard forecasting. In this framework, an optimum raingauge network is essential in predicting flood events.

This paper develops an entropy-based approach to evaluate the maximum information content achievable by a rainfall network for different sampling time intervals. The procedure is based on the determination of the coefficients of transferred and nontransferred information and on the relative isoinformation contours.

The nontransferred information value achieved by the whole network is strictly dependent on the sampling time intervals considered. An empirical curve is defined, to assess the objective of the research: the nontransferred information value is plotted versus the associated sampling time on a semi-log scale. The curve has a linear trend.

In this paper, the methodology is applied to the high-density raingauge network of the urban area of Rome.

1 Introduction

Rainfall height variability makes data collection a relevant task for hydrological purposes. Spatial rainfall information is usually provided by raingauge networks whose density is a key parameter for the proper observation of rainfall fields (Russo et al., 2005; Villarini et al., 2007). Several issues, such as hazard nowcasting (Lombardo et al., 2006, 2009), flood forecasting (Lopez et al., 2005; Russo et al., 2006; Montesarchio et al., 2009) and sewer-system monitoring (Giulianelli et al., 2006) are strictly dependent on the space-time rainfall resolution; the design and evaluation of rainfall networks are, therefore, of great importance. Several authors have dealt with the issues of assessment or design of water-quality monitoring networks (e.g., Ozkul et al., 2000; Mogheir and Singh, 2002; Mogheir et al., 2003, 2004) and raingauge networks (e.g., Bras and Rodriguez-Iturbe, 1985; Husain, 1989; Krstanovic and Singh, 1992a,b; Yoo et al., 2008).

Bras and Rodriguez-Iturbe (1985) illustrate the use of static linear estimation to evaluate the accuracy of possible raingauge configurations and both Krstanovic and Singh (1992a,b) and Yoo et al. (2008) use the concept of heuristic entropy to define the optimum number and the density of raingauges in the network, respectively. Krstanovic and Singh (1992a,b) assess the raingauge networks of Louisiana, USA, considering daily, two-day, weekly and monthly data-sampling intervals, whereas Yoo et al. (2008) evaluate the rainfall network of the Choongiu Dam Basin in Korea, using a mixed and a continuous distribution function applied to daily rainfall data. Informational entropy has been widely applied in hydrological and water resource fields for purposes such as the determination of parameters of a probability space subject to given constraints (Papoulis, 1991), the development of a univariate model for long-term stream flow forecasting (Krstanovic and Singh, 1991a,b), the derivation of the appropriate distribution of the studied variable (Papoulis, 1991; Koutsoyiannis, 2005) and the determination of a rainfall threshold value (Montesarchio et al., 2011). This study develops an entropy-based approach for evaluating the maximum content of information reached by the network at different sampling time intervals in an urban area, where the response to intense rainfall events is generally rapid. The

Correspondence to: E. Ridolfi
(elena.ridolfi@uniroma1.it)

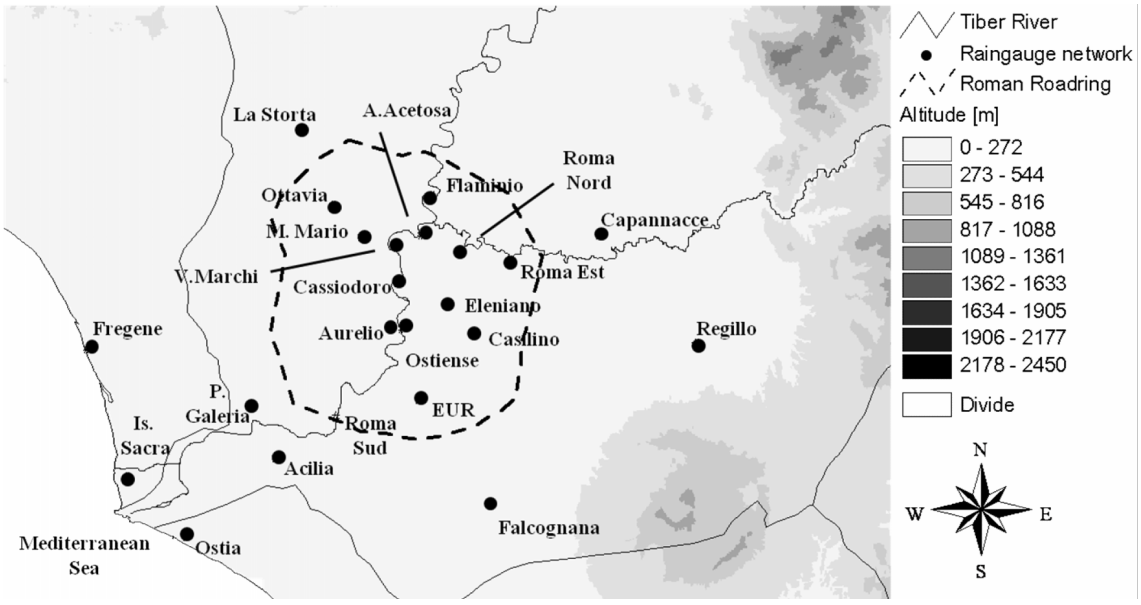

Fig. 1. Map of the rainfall network in Rome. Each raingauge is identified by name.

rainfall data are sampled in time intervals that allow the evaluation of the most relevant rainfall events for urban concentration times, i.e., every three, six and twelve hours. In addition, thirty-minute, hourly, daily, weekly, two-week and monthly sampling times are analysed. First, an overview of the methodology for rainfall entropy estimation is given. Then, using the entropy approach, transinformation and nontransferred information indices are evaluated, to measure the information content of rainfall data. Subsequently, isoinformation contours are plotted using the coefficient of nontransferred information. Finally, a comparative analysis is performed by grouping rainfall data according to seasonality. Both seasons are then combined for the annual data and compared.

2 Classical definition of information entropy

The concept of entropy was first introduced by Clausius in the context of thermodynamics; it can be interpreted as a measure of disorder in a system. To understand the heuristic aspect of entropy, consider a set of n events. Because it is not possible to know which of these n events will occur, the situation is uncertain. In information theory, entropy is a measure of the uncertainty associated with the occurrence of a certain event (Papoulis, 1991). Information entropy has been applied in different fields because of its important characteristics: versatility, strength and efficiency. An extensive review of applications of this theory in hydrologic and hydraulic fields can be found in Singh (1997).

The entropy $H(X)$ of a discrete-type random vector (RV) is defined as (Papoulis, 1991):

$$H(X) = E\left[\log_2(x)\right] = -\sum_i p_i \log_2 p_i \tag{1}$$

where $P(X = x_i) = p_i$ is the probability that the RV X takes the value x_i.

The above expression can be generalised for use with logarithms with bases other than 2. Natural logarithms are used in this work and the corresponding entropy values are measured in napiers.

For two discrete-type RVs, X and Y, assume the values x_i and y_j, respectively; their joint entropy is defined as (Papoulis, 1991; Krstanovic and Singh, 1992a):

$$H(X,Y) = -\sum_{i,j} p_{i,j} \log_2 p_{i,j} \tag{2}$$

where $p_{i,j} = p\left(X = x_i, Y = y_j\right)$ is the joint probability of a particular combination of the rainfall records of two raingauges (e.g., X and Y). $H(X,Y)$ represents the total amount of uncertainty associated with realisation x_i and y_j of the two raingauges.

The conditional entropy is (Papoulis, 1991; Krstanovic and Singh, 1992a):

$$H(X|Y) = H(X,Y) - H(Y) \tag{3}$$

The previous results can be generalised to an arbitrary number of RVs, as explained in Krstanovic and Singh (1992a). To optimize a rainfall network, it is important to evaluate how much information is repeated in two or more raingauges and how much has not yet been transferred. In this way, it

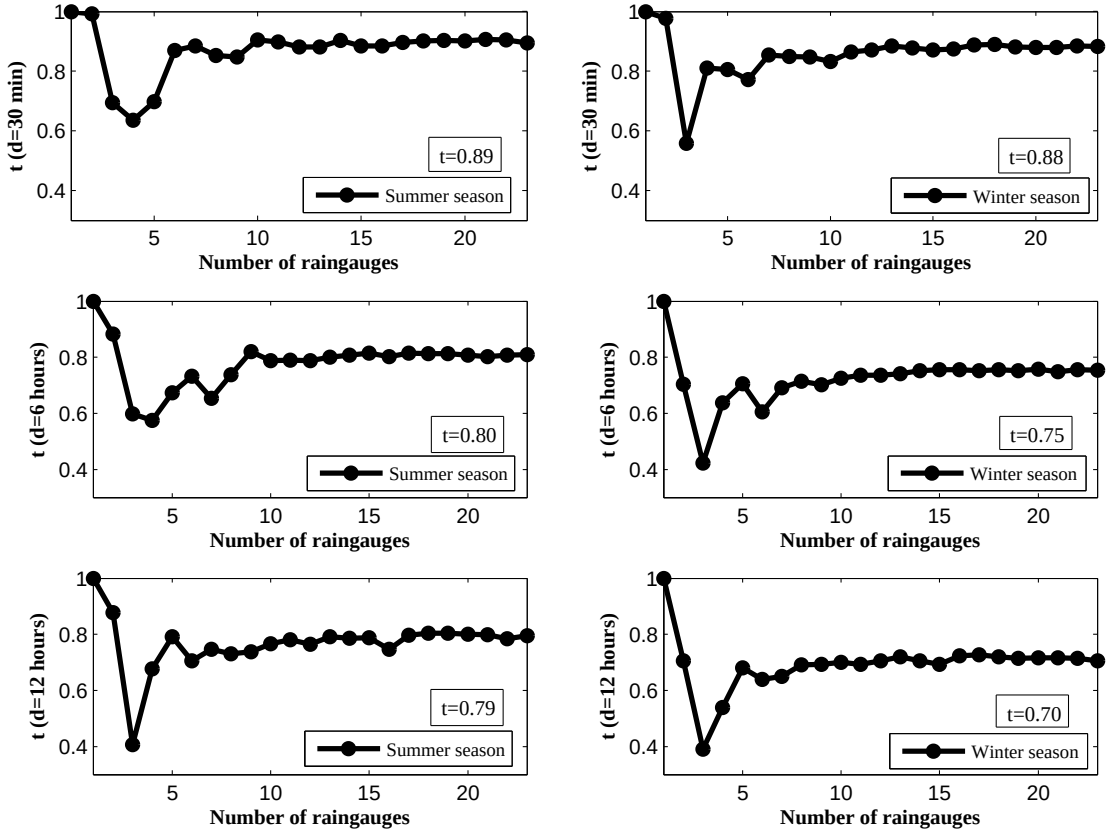

Fig. 2. Coefficient of nontransferred information for sampling time intervals of $d = 30$ min, 6 and 12 h for the summer (left) and winter (right) seasons in Rome. The maximum content of non-redundant information is quoted in the box for each sampling time. This result is the value assumed by t, corresponding to the last raingauge added to the network. The same values are presented on a semi-log scale in Fig. 5 for each season.

is possible to identify the raingauges that yield repetitive information. The transinformation index is used to compute common information provided by two or more variables and to evaluate their redundant information. For two stochastically dependent RVs X and Y, the difference between the sum of the marginal entropies and the joint entropy equals the transinformation index (Amorocho and Espildora, 1973; Harmancioglu and Yevjevich, 1985; Krstanovic and Singh, 1992a):

$$T(X,Y) = H(X) + H(Y) - H(X,Y) \tag{4}$$

This parameter defines the amount of information that is repeated in both RVs. If two RVs are independent, their transinformation coefficient equals zero because the variables considered have no information in common.

In the multivariate case, this index represents a measure of the repeated information that results when the i-th raingauge is added to the network and can be defined as (Krstanovic and Singh, 1992a):

$$T((X_1, X_2, \ldots, X_{i-1}), X_i) = \\ = H(X_1, X_2, \ldots, X_{i-1}) - H((X_1, X_2, \ldots, X_{i-1})|X_i), \tag{5}$$

The coefficient of nontransferred information permits the description of the amount of uncertainty remaining in the raingauge network when a new raingauge is added. In heuristic terms, it represents the nontransferred information through two or more variables. In the bivariate case, the nontransferred information index is defined as (Harmancioglu and Yevjevich, 1985; Krstanovic and Singh, 1992a):

$$t_2 = \frac{T_0 - T_i}{T_0}, 0 \le t_2 \le 0 \tag{6}$$

where T_0 is the upper limit of transferrable information between variables (in the bivariate case T_0 is equal to the marginal entropy of the RV with the maximum value of marginal entropy) and T_i is the common information between the considered variables.

3 Evaluation of the maximum content of non-redundant information achieved by the network

To evaluate the adequacy of a raingauge network, it is important to know how much information was actually transferred

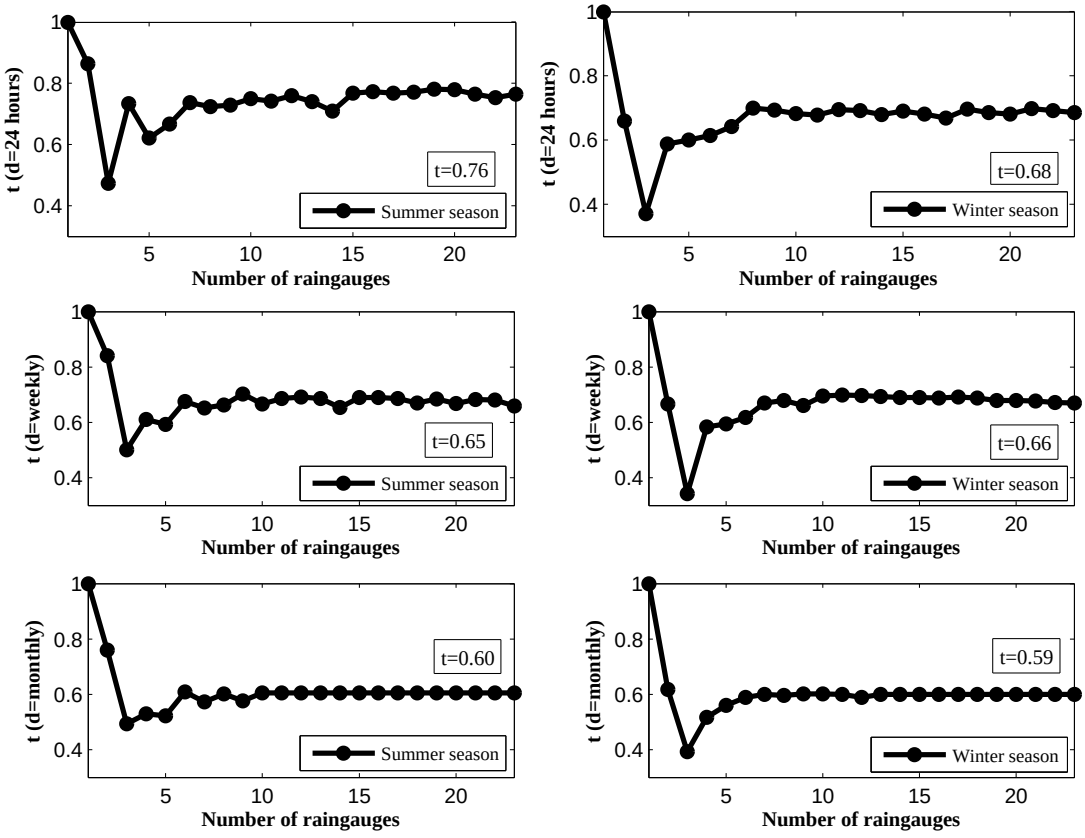

Fig. 3. Coefficient of nontransferred information for sampling time intervals of $d = 24$ h, 1 week and 1 month for summer (left) and winter (right) seasons in Rome. The maximum content of non-redundant information is quoted in the box for each sampling time. This result is the value assumed by t, corresponding to the last raingauge added to the network. The same values are presented on a semi-log scale in Fig. 5 for each season.

between variables and how much information remains to be transferred. In this study, the parameters of transferred and nontransferred information are used to answer these questions.

First, the most important raingauge of the network is determined. According to the Principle Of Maximum Entropy (POME), the raingauge with the maximum value of marginal entropy, as defined in Eq. (1), is the central raingauge of the network.

Second, the conditional entropy of the central raingauge with respect to all the others is computed. The raingauge that gives the lowest redundant information is defined by:

$$\min[T(X,Y)] = \min[H(X) - H(X|Y)] \tag{7}$$

where X is the RV of the central raingauge and Y is the RV of the raingauge which has the least amount of information in common with the first. The latter is the second most important raingauge in the network. It is then necessary to calculate how much information will be provided to the network by adding additional raingauges to the two principals, one raingauge at a time. To find the i-th most important raingauge, it is necessary to retain the $(i-1)$ most important raingauges and compute their conditional entropies with respect to the i-th. The most important i-th raingauge is evaluated by minimising Eq. (5):

$$\begin{aligned} &\min\left[T((X_1,\ldots,X_{i-1}),X_i)\right] = \\ &= \min\left[H(X_1,\ldots,X_{i-1}) - H((X_1,\ldots,X_{i-1})|X_i)\right] \end{aligned} \tag{8}$$

By following this process step-by-step, it is possible to evaluate the order of importance of the raingauges in the network.

For every added raingauge in order of importance, the coefficient of nontransferred information is computed. In adding the i-th raingauge:

$$\begin{aligned} t_i &= \frac{H((X_1,\ldots,X_{i-1})|X_i)}{H(X_1,\ldots,X_{i-1})} \\ &= \frac{H(X_1,\ldots,X_i) - H(X_i)}{H(X_1,\ldots,X_{i-1})}. \end{aligned} \tag{9}$$

If, at step i, it is the case that $t_i \geq t_{i-1}$, then the new raingauge has repetitive information. In contrast, if $t_{i-1} > t_i$, the raingauge added at the i-th step has new information. The greater the difference between the values of the coefficient at any step, the greater the information gained by the network

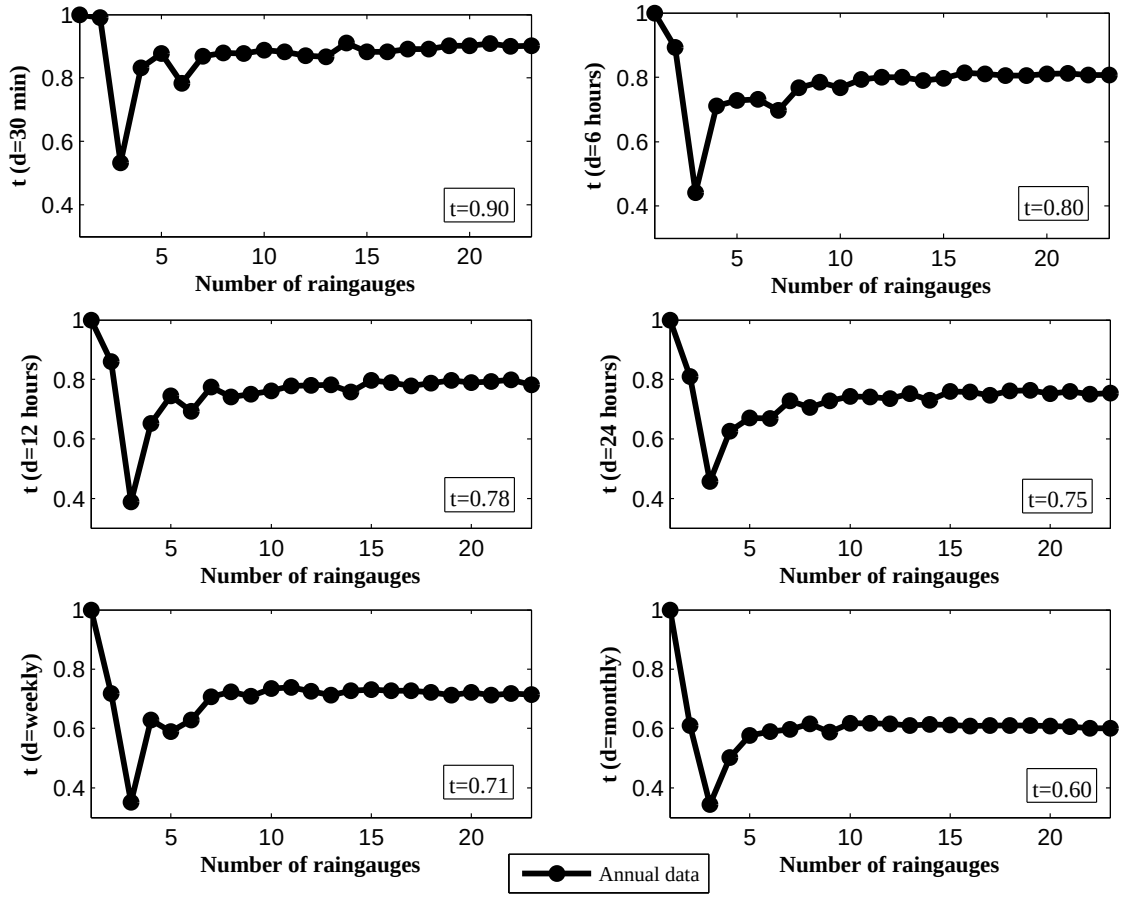

Fig. 4. Coefficient of nontransferred information for sampling time intervals of $d = 30$ min, 6, 12, 24 h, 1 week and 1 month for annual data in Rome. The maximum content of non-redundant information is quoted in the box for each sampling time. This result is the value assumed by t, corresponding to the last raingauge added to the network. The same values are presented on a semi-log scale in Fig. 5.

through the addition of a new raingauge at that step. If the addition of a new raingauge results in no new information, it can be assumed that the raingauge is contributing only redundant information. Its complementary relative measure, $(1-t_i)$, describes the amount of transferred information.

4 Case study: the urban area of Rome

The methodology is applied to a case study of the metropolitan area of Rome. The target area contains twenty-four stations in its raingauge network. The dataset covers a period of eighteen years from 1992 to 2009 and has a ten-minute time resolution.

The historic rainfall sequences are divided into a summer season (1 April to 30 September) and a winter season (1 October to 31 March). Both seasons were combined to produce annual data. For analysis, records were sampled in time intervals of 30 min, 1 h, 3 h, 6 h, 12 h, 1 day, 1 week, 2 weeks and 1 month. The raingauge named Castello Vici was eliminated from the analysis because great amounts of data were missing. Therefore, twenty-three raingauges were included in the analysis. A map of the rainfall network examined is shown in Fig. 1.

4.1 Data analysis

The entropy approach, as explained in the previous sections, involves univariate and multivariate discrete probabilities. To evaluate the marginal probability of each RV, divide the values of the considered RV into v categories (class intervals). The number of these class intervals has been defined in Mogheir et al. (2003):

$$v = 1 + 1.33\log(n) \tag{10}$$

where n is the length of the time series of the considered variables. Once computed, the frequency associated with each class (i.e., how many values of the time series fall in each class interval), is divided by the number of elements of the time series itself (i.e., n); the marginal probability is thereby determined.

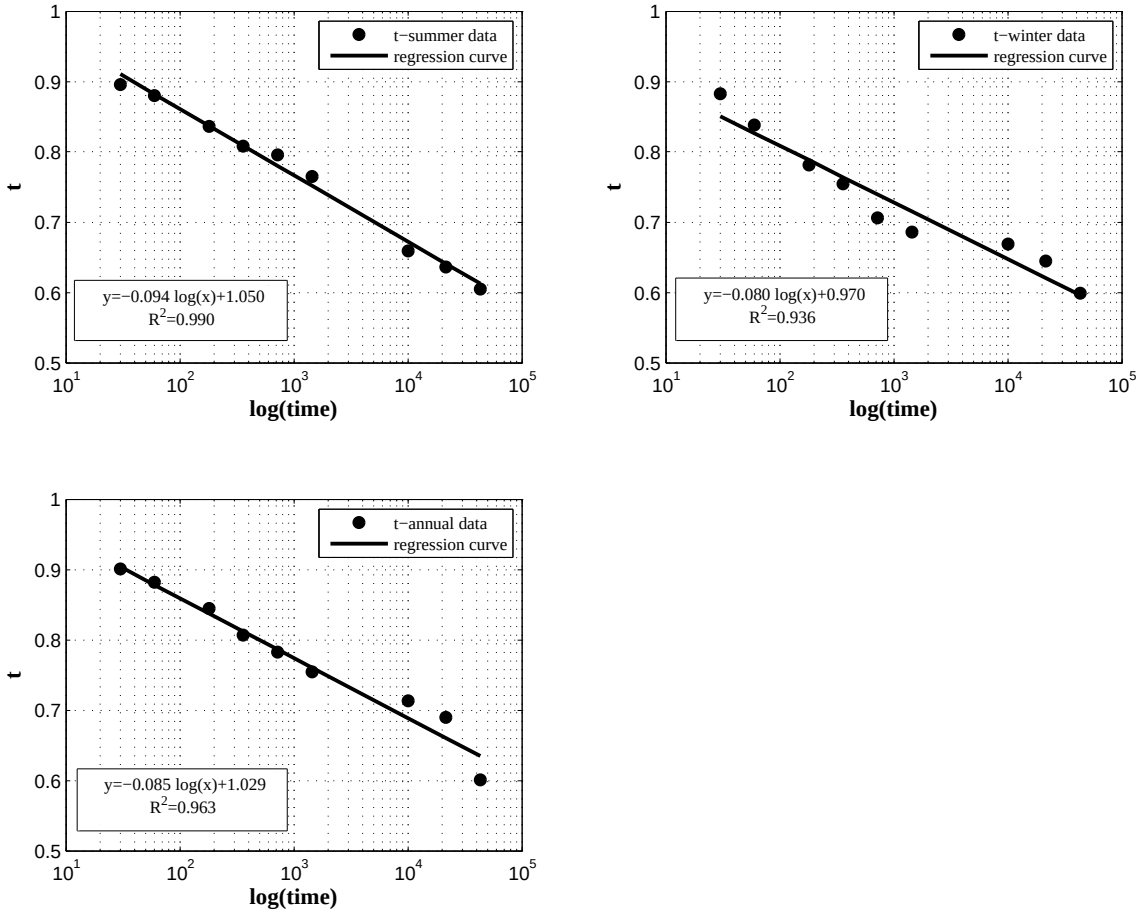

Fig. 5. Simple linear regression curve for maximum non-redundant information curve for summer and winter seasons (above) and annual data (below) on a semi-log scale. The equation and the coefficient of determination (R^2) of each regression curve are given in the box below each plot.

To evaluate the bivariate or multivariate probabilities, it is necessary to construct either a two- or n-dimensional contingency table, respectively. The bivariate case will be presented here (Mogheir et al., 2003). Divide the values of the RV X into v categories. The random variable Y is assumed to have u categories. The marginal frequencies are denoted by f_i and f_j, and the joint frequency is f_{ij}. The joint frequency corresponds to the number of elements of the X RV that fall in the i-th class when elements of the Y RV fall in the j-th class. An extensive description and an example of this methodology can be found in Mogheir et al. (2003).

4.2 Interpretation of nontransferred information results

The objective of this study is to determine the maximum non-redundant information content that the network collects at different sampling time intervals. First, the central raingauge was defined. Then, at each step of the raingauge selection process, another raingauge with the minimum value of transferred information was added to the network, and the coefficient of nontransferred information was evaluated as in Eq. (9).

The coefficient of nontransferred information was plotted for every sampling time interval.

In both seasons (Figs. 2 and 3) and for an annual aggregation time (Fig. 4), the last value of the nontransferred information index is greater for the 30-min sampling time than for the monthly sampling time. The amount of non-redundant information provided by the whole network decreases with the increase in the rainfall sampling time. In fact, for greater sampling of time intervals, the information that raingauges provide is more redundant than that obtained using smaller sampling times. The reason for this redundancy is that for greater sampling times, the rainfall field is more uniform over the area and raingauge measures are nearly similar.

In the winter season, the index reaches a smaller value than in summer (Figs. 2 and 3). In urban Rome, winter raingauges provide more valuable information because of the regional climate: because winter is more rainy than summer,

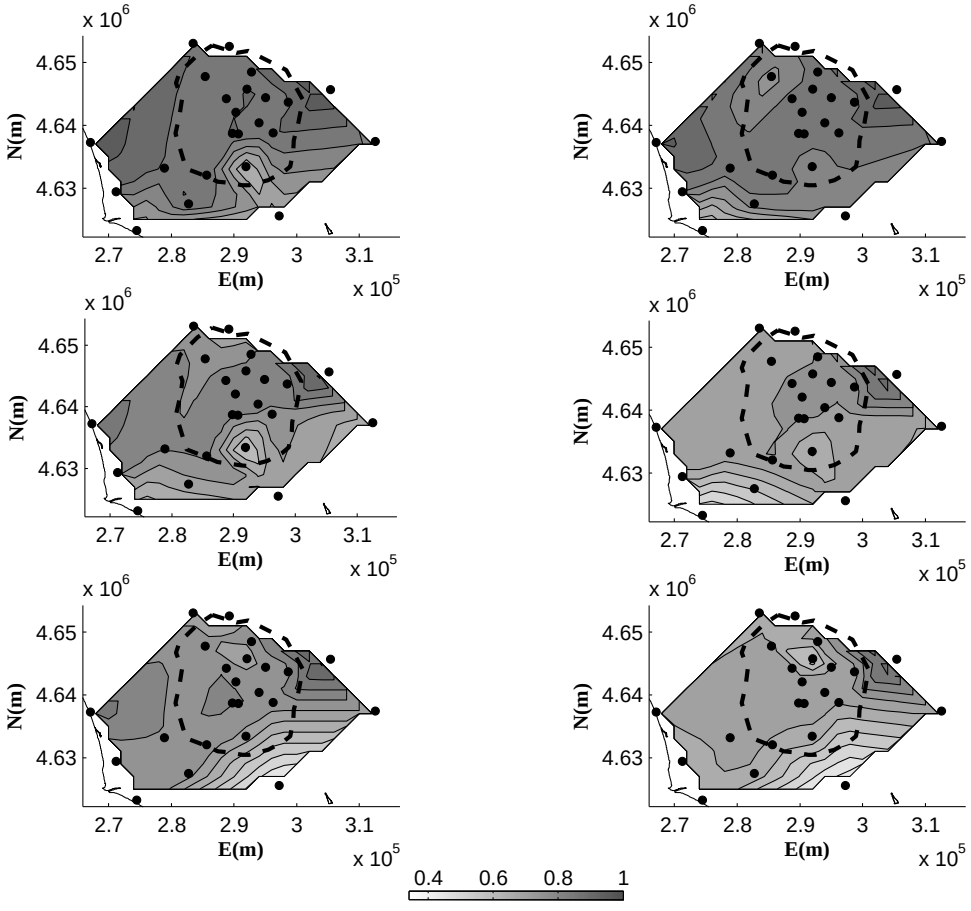

Fig. 6. Rome area isoinformation contours for sampling time intervals of $d = 30$ min, 6 and 12 h for summer (left) and winter (right) seasons. The colour bar varies from 0.34 to 1 and shows the value achieved by the nontransferred information index. The dashed line represents the Roman roadring.

the network provides a greater quantity of data during that season, and these data are relatively more redundant.

In the winter season in the weekly sampling time interval and in both seasons in the monthly sampling time interval (Fig. 3), the nontransferred information coefficient, after a little variation, remains constant.

For the monthly sampling time interval, in the summer season (Fig. 3) the 11th raingauge does not provide an information gain to the network. Its nontransferred information index has the same value as that of the previous raingauge. The same behaviour can be noticed for all the following raingauges. In the winter season (Fig. 3), the nontransferred information index reaches a constant value corresponding to the 14th raingauge.

For the weekly sampling time interval (Fig. 3) in the winter season, the nontransferred information index remains constant and corresponds to the last two raingauges.

For the annual data from the monthly sampling time interval (Fig. 4), the constant value of the nontransferred information index is reached for the last raingauge added.

It can, therefore, be inferred that the network converges on a constant value of nontransferred information that represents the maximum value of information that the network can achieve. This behaviour can be observed for each sampling time: the t_i oscillates around a constant value in correspondence to the last raingauges added to the network. The lower the magnitude of the difference between a t_i and a t_{i-1} value, the less the information that is added by the i-st raingauge.

For each sampling time interval, the constant value is achieved by considering a different number of raingauges.

The maximum value of information reached at each sampling time can be described with an empirical curve.

For both seasons and for the annual aggregate data, the latter value of the nontransferred information index is plotted against the corresponding sampling time interval on a semi-log scale (Fig. 5). The plot shows scale-invariance. The relation between the two variables is linear on the semi-log scale. It can, thus, be inferred that once this curve is known for a given network, the maximum value of nontransferred information can be obtained for the other sampling times.

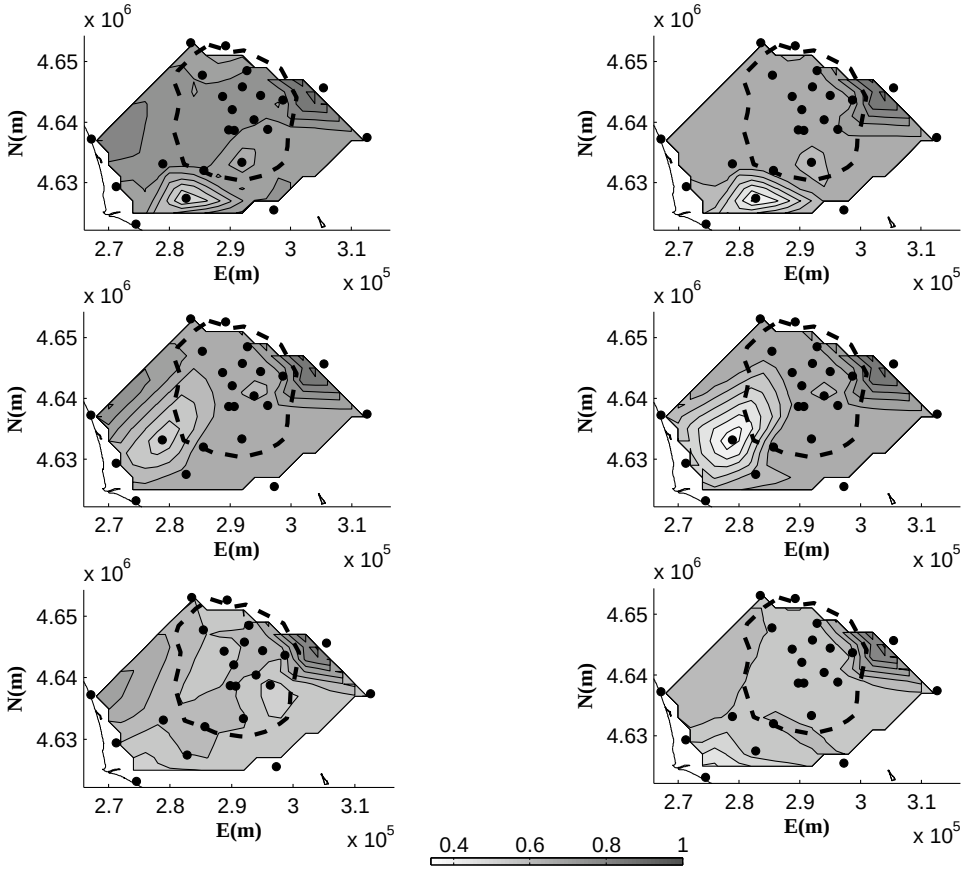

Fig. 7. Rome area isoinformation contours for sampling time intervals of $d = 1$ day, 1 week and 1 month for summer (left) and winter (right) seasons. The colour bar varies from 0.34 to 1 and shows the value achieved by the nontransferred information index. The dashed line represents the Roman roadring.

4.3 Isoinformation contours

In evaluating the coefficient of nontransferred information between the central and the other raingauges, it is possible to construct isoinformation contours, the lines of equal common information between the i raingauges considered. The value of t_i is evaluated with Eq. (9). Contours of isoinformation plotted on a map of the area analysed provide a complementary view of the results represented in Figs. 6 and 7.

These contours encompass the central raingauge, where t_i has a value of 1 and no information has been transferred. Summer and winter evaluations are compared for each sampling time interval. As explained in the previous section, the network reaches a lower value of nontransferred information with smaller sampling time intervals in winter. It can be observed that in the summer season, the central raingauge is Capannacce (Figs. 6 and 7), located in the northeast near the Sabatini Mountains. In summer, significant amounts of information are accrued in the rainiest area (located near the mountains), whereas in the urban and coastal areas, less information is collected. For sampling times of 3, 12 and 24 h, 1 day, 1 week and 1 month, the raingauge that gives the lowest nontransferred information index value is located on the coast, on the other side of the network. In winter, the principal raingauge is not the same for each sampling time interval. Therefore, the rainfall information is distributed over the region, not concentrated in a single area.

5 Conclusions

This work presents an evaluation of the rainfall network of the metropolitan area of Rome using entropy and defines an empirical method to assess the maximum non-redundant information achievable by a rainfall network at different sampling time intervals.

The rainfall records are divided according to seasonality and into different sampling intervals. For each season and each sampling time, the raingauge that contains the greatest rainfall information is identified.

Data from the summer and winter seasons are merged to obtain an annual rainfall record. The results for the summer

and winter seasons and for the yearly aggregation are then compared. For longer sampling time intervals, an equal number of raingauges accumulates less information in summer than in winter. Rainfall is more frequent in winter than in summer and yields more information. The entropy coefficients, a measure of information, illustrate this difference. The comparison of the nontransferred information indices for each sampling time interval indicates that if the sampling time interval is greater, then the redundant information reached by the network is also greater.

For smaller sampling times, the index of nontransferred information is greater than that of the other sampling times because the rainfall measures are less similar.

The maximum non-redundant information values and the corresponding sampling times are linearly related on a semi-log scale. Thus, once the equation of this curve is known for a given network the non-redundant information content is uniquely defined.

This important behaviour can be observed for both seasonal and for yearly sampling. Additional research into the physical meaning of this behaviour is underway.

Because this paper represents a preliminary study, these conclusions should be tested using different climatic conditions.

Acknowledgements. The authors would like to thank Ufficio Idrografico e Mareografico of Lazio Region for providing pluviometric data. In addition, authors would like to thank the reviewers for their useful comments and suggestions, which helped to improve the manuscript.

Edited by: A. Bartzokas
Reviewed by: two anonymous referees

References

Amorocho, J. and Espildora, B.: Entropy in the assessment of uncertainty in hydrologic systems and models, Water Resour. Res., 9(6), 1511–1522, 1973.

Bras, R. L. and Rodriguez-Iturbe, I.: Random functions and hydrology, Addison Wesley, Reading, Mass., 1985.

Giulianelli, M., Miserocchi, F., Napolitano, F., and Russo, F.: Influence of space-time rainfall variability on urban runoff, Proceeding of the 17th IASTED International Conference Modelling and Simulation, Montreal, QC, Canada, 546–551, 2006.

Harmancioglu, N. and Yevjevich, V.: Transfer of hydrologic information along rivers partially fed by karstified limestones, IAHS-AISH Publication, 161, 115–131, 1985.

Husain, T.: Hydrologic uncertainty measure and network design, Water Resour. Bull., 25(3), 527–534, 1989.

Koutsoyiannis, D.: Uncertainty, entropy, scaling and hydrological stochastics, 1: marginal distributional properties of hydrological processes and state scaling, Hydrol. Sci. J., 50(3), 381–404, 2005.

Krstanovic, P. F. and Singh, V. P.: A univariate model for long-term streamflow forecasting 1. Development, Stoch. Hydrol. Hydraul., 5(3), 173–188, 1991a.

Krstanovic, P. F. and Singh, V. P.: A univariate model for long-term streamflow forecasting 2, Application, Stoch. Hydrol. Hydraul., 5(3),189–205, 1991b.

Krstanovic, P. F. and Singh, V. P.: Evaluation of rainfall networks using entropy, I: Theoretical development, Water Resour. Manag., 6, 279–293, 1992a.

Krstanovic, P. F. and Singh, V. P.: Evaluation of rainfall networks using entropy II: Application, Water Resour. Manag., 6, 295–313, 1992b.

Lombardo, F., Napolitano, F., and Russo, F.: On the use of radar reflectivity for estimation of the areal reduction factor, Nat. Hazards Earth Syst. Sci., 6, 377–386, doi:10.5194/nhess-6-377-2006, 2006.

Lombardo, F., Montesarchio, V., Napolitano, F., Russo, F., and Volpi, E.: Operational applications of radar rainfall data in urban hydrology, IAHS-AISH Publication, 327, 258–266, 2009.

Lopez, V., Napolitano, F., and Russo, F.: Calibration of rainfall-runoff model using radar and raingauge data, Adv. Geosci., 2, 41–46, 2005, http://www.adv-geosci.net/2/41/2005/.

Mogheir, Y. and Singh, V. P.: Application of information theory to groundwater quality monitoring networks, Water Resour. Manag., 16, 37–49, 2002.

Mogheir, Y., de Lima, J. L. M. P., and Singh, V. P.: Assessment of spatial structure of groundwater quality variables based on the entropy theory, Hydrol. Earth Syst. Sci., 7, 707–721, doi:10.5194/hess-7-707-2003, 2003.

Mogheir, Y., de Lima, J. L. M. P., and Singh, V. P:. Characterizing the spatial variability of groundwater quality using the entropy theory, I. Synthetic data, Hydrol. Process., 18, 2165–2179, 2004.

Montesarchio, V., Lombardo, F., and Napolitano, F.: Rainfall thresholds and flood warning: an operative case study, Nat. Hazards Earth Syst. Sci., 9, 135–144, doi:10.5194/nhess-9-135-2009, 2009.

Montesarchio, V., Ridolfi, E., Russo, F., and Napolitano, F.: Rainfall threshold definition using an entropy decision approach and radar data, Nat. Hazards Earth Syst. Sci., 11, 2061–2074, doi:10.5194/nhess-11-2061-2011, 2011.

Ozkul, S., Harmancioglu, N., and Singh, V.P.: Entropy-based assessment of water quality monitoring networks, J. Hydrol. Eng., 5, 90–100, 2000.

Papoulis, A.: Probability, Random Variables and Stochastic Processes, 3rd Edn., McGraw-Hill, New York, 1991.

Russo, F., Napolitano, F., and Gorgucci, E.: Rainfall monitoring systems over an urban area: the city of Rome, Hydrol. Process., 19(5), 1007–1019, 2005.

Russo, F., Lombardo, F., Napolitano, F., and Gorgucci, E.: Rainfall stochastic modelling for runoff forecasting, Phys. Chem. Earth, 31(18), 1251–1261, 2006.

Singh, V. P.: The use of entropy in hydrology and water resources, Hydrol. Process., 11, 587–626, 1997.

Villarini, G., Lang, J. B., Lombardo, F., Napolitano, F., Russo, F., and Krajewski, W. F.: Impact of different regression frameworks on the estimation of the scaling properties of radar rainfall, Atmos. Res., 86(3–4), 340–349, 2007.

Yoo, C., Jung, K., and Lee, J.: Evaluation of rain gauge network using entropy theory: comparison of mixed and continuous distribution function applications, J. Hydrol. Eng., 13, 226–235, 2008.