U.S.- Italy Research Workshop on the
Hydrometeorology, Impacts, and Management of Extreme Floods
Perugia (Italy), November 1995

STATISTICAL ANALYSIS OF THE SPATIAL VARIABILITY OF VERY EXTREME RAINFALL IN THE MEDITERRANEAN AREA

P. Furcolo, P. Villani, and F. Rossi
Department of Civil Engineering
University of Salerno, Italy

Abstract. Natural hydrometeorological disasters in the Mediterranean region have occurred in the past essentially by outlying storm events characterized by considerable rainfall intensity and rare frequency. The characterization of this type of event is a crucial point in risk mitigation. Because of the rare occurrence of these extreme events they can be analyzed only on a regional frequency basis in order to reduce the uncertainty associated with parameter estimation at gauged sites and for risk assessment at ungauged sites. A statistical regional model includes: (i) a probabilistic model that can describe the extraordinary high floods and rainfall observed in the past (outliers) and (ii) a regionalization model that can take into account the observed spatial variability of the parameters of the probabilistic model. A regional model based on the TCEV distribution along with a geostatistical analysis of its parameters is presented here. The regional model assumes the observed variance stems from two sources: sampling variability due to uncertainties in at-site estimates, and spatial variability due to effective differences between sites. Traditional geostatistical techniques refer to the *exactitude property* in gauged sites whereby at-site estimates are affected by sampling uncertainties that can be predominant for high order parameters. An iterative geostatistical procedure is implemented, which makes it possible to obtain the spatial structure of the noiseless variate. Some initial results are shown with reference to a case study for an Italian region while the objective differentiation between areas with different risks constitutes one of the most important findings of the proposed regionalization procedure.

1. INTRODUCTION

The Mediterranean regions are characterized by the occurrence of short and very intense flood events (outliers) causing landslides and floods that have resulted in considerable loss of life in addition to severe damage to urban areas, artistic and architectural assets, and infrastructure. These events are extremely rare in terms of any one specific site, but their frequency is not negligible in a regional context and they can only be predicted in probabilistic terms through regional statistical analysis.

A regionalization procedure essentially requires the identification of:

(i) An at-site probabilistic model that can interpret the temporal variability of the annual maximum of the aggregate rainfall of different durations;

(ii) A regionalization model that can explain the spatial variability of the at-site probabilistic model's parameters.

The identification procedure always starts with the simplest model that reproduces the process' fundamental characteristics. As we shall see below, the simplest at-site probabilistic model of the annual maxima of rainfall of different durations, whose parameters must have a clear physical meaning, can be assumed to be the type one extreme value (EV1) distribution in which the variation coefficient does not vary with the duration.

The simplest regionalization model refers to the concept of a homogeneous region, according to which the relevant statistical parameters are constant within the region (*homogeneous climatic region*). For the EV1 at-site model, this procedure implies a variation coefficient that is constant throughout the homogeneous region, which makes it possible to obtain an efficient regional parameter estimate. In reality, in identifying a homogeneous region and deciding whether the regional estimate improves the at-site estimates, it is essential to allow for a certain level of heterogeneity and, therefore, take into account the fact that the model's parameters vary from station to station because of a spatial disturbance error. In this sense, regionalization serves to extend the lengths of historical sequences leading to a reduction in temporal sampling errors, though at the price of introducing spatial disturbance errors. In addition, the reduction of time sampling variance as the number of stations' increases is limited by the time sampling correlation between concurrent events.

If the parameter's spatial variability is completely neglected, we have a definition of a deterministic homogeneous region. Whereas, if we recognize the existence of this spatial variability but do not recognize any stochastic or deterministic structure in it, we have a statistically homogeneous region. In this approach, if the spatial variability constitutes a significant proportion of the overall observed variability, attempts are typically made to reduce this proportion by subdividing the examined region into smaller parts, each of which may then be said to be homogeneous. In this case, there is said to be a deterministic spatial variability structure.

A particular application of this statistical homogeneity model, consisting of a *hierarchical approach* for homogeneous regions, has been adopted for the regional analysis of rainfall and floods in Italy within the context of the VAPI project (Versace et al., 1989). The simple regionalization model for homogeneous regions has very few parameters and enables robust procedures for their estimation. This is a considerable advantage, especially when there are statistical parameters of a higher order whose estimators are characterized by a high sampling variability. Moreover, not recognizing a spatial variability structure may generate the risk of identifying homogeneous areas with a spatial extension greater than the actual one, resulting in a significant overestimate of the risk in some places and underestimation in others.

The aim of the present paper is to characterize the spatial variability of the rainfall maximum probability distribution parameters. An initial generalization of the model for homogeneous regions takes into account the presence of a sampling correlation between the stations (inter-station correlation) (Gabriele and Iiritano, 1994). A more general model for the spatial variability of the rainfall probability distribution parameters must also take into account the possibility of distinguishing between areas subject to a different hydrological risk within a single hydrographic region. In this case, the homogeneity of a region is no longer regarded as the homogeneity of the value of K_T, but in a wider sense as the homogeneity of the structure of the K_T distribution parameters. The probabilistic model adopted is first briefly mentioned along with the physical meaning of the model's parameters and a general model for the statistical analysis of the spatial structure of the extreme rainfall process is illustrated. Then it is shown that

different formulations of regional models proposed in the past can be referred as special cases of the model proposed herein. A case study is utilized to show how the present formulation of the model is suitable for regional analysis when a region is comprised of areas with a different hydrological risk.

2. THE AT-SITE PROBABILISTIC MODEL

2.1 THE POISSON – EXPONENTIAL COMPOUND PROBABILISTIC MODEL

As already mentioned, in choosing the at-site probabilistic model it is preferable to refer to models that reproduce the physical characteristics of the process in order to utilize the information available a priori for the subsequent regionalization procedure. In order to analyze the annual maximum of fixed duration rainfall, the class of models that are endowed with the requisites of simplicity and phenomenon descriptive ability refers to a rough model of the marked Poisson compound stochastic process, with identically distributed marks. In this class the simplest model assumes that the distribution of marks is exponential (Poisson - exponential compound process). The probabilistic model of the maxima thus obtained approximates a model commonly used in hydrological maxima frequency analysis, known as Gumbel's law or type 1 extreme value model (EV1), from which it differs in that only non-negative variables are taken into consideration.

The parameters of the referred model are closely linked to the base process and, in particular, are the Poisson occurrence rate $\Lambda(d)$, corresponding to the mean annual number of events with aggregation to the duration d, and the mean value $\theta(d)$ of the intensity of the peaks of duration d. The cumulative distribution function (CDF) of the annual maximum depth of rainfall of duration d is:

$$F_{X(d)}(x) = \exp\left[-\Lambda \cdot e^{-x/\theta(d)}\right] \qquad (x > 0) \qquad (1)$$

In order to complete the at-site probabilistic model, it is necessary to specify how its parameters vary with duration. Typically, for the mean $\theta(d)$ of the peaks of duration d, we find the self-similarity property, applies. It follows

$$\theta(d) = \theta(1) \cdot d^{n} \qquad (2a)$$

Furthermore, for durations that are not too long, we assume that

$$\Lambda(d) = \Lambda = \text{const} \qquad (2b)$$

Relations (2a) and (2b) taken together imply that the self-similarity property holds. For the mean of the annual maxima of rainfall of duration d it gives

$$\mu[X(d)] = \mu[X(1)] \cdot d^{n} \qquad (2c)$$

The set of relations (2) implies a process that is self-similar with the duration, in which the exponent $n \in [0,1]$ is characteristic of dissipating chaotic processes. The probabilistic growth factor varies with the return period as

$$K_T = \frac{X_T(d)}{\mu[X(d)]} \tag{3a}$$

where $X_T(d)$ is the T-year annual maximum rainfall depth with duration d and $\mu[X(d)]$ is the mean of X(d). For the referred distribution K_T varies with the return period T as

$$K_T = 1 + \frac{\sqrt{6}}{\pi} Cv \cdot \left(\log T - C\right) \tag{3b}$$

in which Cv = distribution variation coefficient and C = Eulero's constant = 0.5772….

Considering the stochastic interpretation of Gumbel's model, its variation coefficient (Cv) can be expressed as a function of the Poisson Λ occurrence rate as:

$$Cv = \frac{\pi}{\sqrt{6}} \left(\log \Lambda + C\right)^{-1} \tag{4}$$

A probabilistic model of this type thus presents the advantage of having an extremely small number of parameters with a clear physical meaning. However, in a series of investigations (Rossi et al., 1984), the EV1 model was seen to be unable to reproduce the sampling variability of the observed historical series, mainly because of the occurrence of extraordinary events in the series of annual maxima that had to be considered as outliers when compared to the EV1 distribution. It is well known that rainfall events in the Mediterranean area are generally related to baroclinic cyclones. Recent meteorological observations seem to show that in the Mediterranean at least some of these extraordinary events occur because of cyclonic events with a clearly tropical structure, which, as they evolve, tend to form hurricanes (Reale, 1996). Unfortunately, the historical data available on this sort of event is insufficient and thus an a priori separation of the meteorological events is not possible.

Alternatively, reference could be made to a statistical methodology in which the annual maximum X(d) is assumed to be the maximum of a Poissonian number of variables generated by a mixture of two exponentials [Rossi et al., 1984]. As we will see, this model is characterized by parameters that are strictly linked to the physical phenomenon being examined and possesses the necessary flexibility both for the interpretation of the outlier events present in the historical series and for the interpretation of data that does not respect the property of scale-invariance with duration.

2.2 THE TWO-COMPONENT PROBABILISTIC MODEL

As mentioned above, the need to take into account both the historical occurrence of extraordinary rainfall events and the non-applicability of meteorological procedures for the direct identification of such events resulted in the development of the two-component extreme value probabilistic model (TCEV) (Rossi et al., 1984). This distribution can be interpreted as the probability distribution of the annual maximum for a Poissonian process composed of a mixture

of two independent populations, each of which has an exponential distribution. One population is called the ordinary component and represents the rainfall annual maximum events that are most frequently recorded; the other is called the extraordinary component and is the population that generates the outliers. The form of the CDF in question is as

$$F_X(x) = \exp\left[-\Lambda_1 e^{-x/\theta_1} - \Lambda_2 e^{-x/\theta_2}\right] \tag{5}$$

in which Λ_i (i=1,2) represents the mean annual number of events belonging to the i-th component, while θ_i (i=1,2) represents the mean value of the single population. Thus, the parameters of this distribution have a clear physical meaning. This is an essential characteristic of the probabilistic model when operating in the context of climatically homogeneous regions. The distribution shape parameters are

$$\begin{aligned} \theta^* &= \theta_2 / \theta_2 \\ \Lambda^* &= \Lambda_2 / \Lambda_1^{1/\theta^*} \end{aligned} \tag{6}$$

in which Λ_1 represents the scale parameter.

Hosking (1990) demonstrated the utility, of estimators based on the L-moment ratios, in hydrological extreme analysis. Similarly to the definitions and the meaning of the ratios between ordinary moments, the coefficients of L-kurtosis, L-skewness and L-variation are defined as

$$\begin{aligned} \text{L}-\text{Kurtosis} &= \frac{5\cdot\left[2\cdot\left(2\beta_3 - 3\beta_2\right) + \beta_0\right]}{2\beta_1 - \beta_0} + 6 \\ \text{L}-\text{Cs} &= \frac{2\cdot\left(3\beta_2 - \beta_0\right)}{2\beta_1 - \beta_0} - 3 \\ \text{L}-\text{Cv} &= \frac{2\beta_1 - \beta_0}{\beta_0} \end{aligned} \tag{7}$$

in which β_r is the PWM of order r defined by

$$\beta_r = \int_0^\infty x \cdot F(x)^r \cdot f(x) \cdot dx \tag{8}$$

Starting from the TCEV shape parameters only, it is possible to evaluate the probability (p_2) that the annual maximum will come from the second component through the following relation (Beran et al., 1986):

$$p_2 = -\frac{\Lambda^*}{\theta^*} \cdot \sum_{j=0}^{\infty} \frac{(-1)^j}{j!} \cdot \Lambda_*^{\ j} \cdot \Gamma\left(\frac{j+1}{\theta^*}\right) \tag{9}$$

In particular, the TCEV shape parameters are in univocal correspondence with the skewness and kurtosis coefficients (or L-coefficients) while, with the shape parameters fixed, the

scale parameter depends solely on the variation coefficient (or L-coefficient). In conclusion, the probabilistic growth factor K_T depends only on the distribution scale and shape parameters. In the rest of the present paper, particular attention will be paid to the regional estimation of K_T. In the analyses by (Versace, 1994), the coefficients of kurtosis, skewness, and variation do not seem to vary with duration. Therefore, in an initial approximation, it is possible to consider as valid the hypothesis that, for the TCEV distribution, K_T is independent of duration.

3. DEFINITION OF THE PARAMETER'S SPATIAL VARIABILITY MODEL

3.1 THE SPATIAL VARIABILITY MODEL

Indicating with S(s) the parameter in question (L-Kur, L-Cs, L-Cv) and with Z(s) one of its estimates in the gauged site of coordinates s, we have:

$$Z(s) = S(s) + \varepsilon(s) \tag{10}$$

in which $\varepsilon(s)$ is an error term. Relation (10) expresses the fact that because of the limited length of the historical series, the estimated value of the parameter differs from the theoretical value by a quantity $\varepsilon(s)$, which represents the sampling error. Under the hypothesis of unbiased estimators, the field $\varepsilon(s)$ appears as a random field with zero mean and variance equal, in the site, to the variance of the estimator, depending on the probabilistic model and the length of the historical series at site. It follows

$$\begin{aligned} &E[\varepsilon(s)] = 0 \\ &Var[\varepsilon(s)] = \sigma_\varepsilon^2(s) \end{aligned} \tag{11}$$

In general terms, S(s) can be regarded as a random field in the region under investigation and consists of two components

$$S(s) = \mu(s) + W(s) \tag{12}$$

in which $\mu(s)$ = term of deterministic variability (large scale), and W(s) = random spatial field (intermediate scale) with mean E[W(s)] = 0, and variance Var[W(s)] = σ_w^2. Clearly, therefore, under the hypothesis of independence of the random at-site processes W(s) and $\varepsilon(s)$, the variance of the field Z(s) is given by the sum of the two variances σ_w^2 and $\sigma_\varepsilon^2(s)$:

$$Var[Z(s)] = \sigma_w^2 + \sigma_\varepsilon^2(s) \tag{13}$$

The complete characterization of the model also requires the identification of the spatial correlation structure of the single field components. To this end, reference is normally made to tools like the spatial covariogram or the spatial semi-variogram [Cressie, 1991]. The presence of any spatial correlation structures in the terms W(s) and $\varepsilon(s)$ gives, for the field Z(s), a variogram

formed of the sum of the variograms of W(s) and ε(s), assuming that the two above-mentioned terms are independent.

The combination of (10) and (12) gives a general regionalization model. A practical application of this model can be obtained, as stated from the outset, by starting from simplifying hypotheses, whose effects have to be assessed on each occasion, and may lead to a relaxation of the same hypotheses. A particular class of such models considers the constant deterministic term

$$\mu(s) = \mu_0 \tag{14}$$

which corresponds to a hypothesis of *first order stationarity*. Most of the regionalization models present in the literature refer to a field that is stationary or becomes stationary after a deterministic detrend operation. Moreover, the various models proposed differ according to the value assumed by the ratio $\sigma_w^2 / \sigma_\varepsilon^2$. When this ratio is close to zero we get the simple *homogeneous region* model. In this case we can see that sampling variance can explain the observed variance and, therefore, the uncertainty due to the estimation of the parameter completely masks any spatial variability present. A particular application of this model, consisting of a *hierarchical approach* for homogeneous regions has been adopted for regional flood analysis in Italy within the context of the VAPI project (Fiorentino et al., 1987). This model represents a term of comparison that is essential for the areas being studied and the following section will outline its main features and illustrate its relations with the general regionalization model proposed here.

3.2. APPROACH FOR HOMOGENOUS REGIONS

3.2.1 CHARACTERIZATION OF THE MODEL PARAMETERS

The regionalization model for homogeneous regions refers to an expression of the observed variable Z of the type:

$$Z(s) = \mu_0 + \varepsilon(s) \tag{15}$$

The model is characterized by the value assumed by the parameter μ_0 alone, but also requires the evaluation of the two parameters σ_w^2 and σ_ε^2, i.e. the spatial variance and the sampling variance. As mentioned above, the adopted model requires verification of the relation:

$$\sigma_w^2 << \sigma_\varepsilon^2 \tag{16}$$

Therefore the model requires the estimation of 3 regional parameters: the regional value μ_0 of S, the sampling variance σ_ε^2 and the spatial variance σ_w^2 (or the observed variance of the field Z).

3.2.2 ESTIMATION OF THE PARAMETERS OF THE REGIONAL MODEL

The parameter μ_0 can be estimated using two regional estimation techniques:

i) *spatial mean of the at-site estimates*: given the k values $Z(s_i)$ of at-site estimates of the parameter S, its regional estimate ($\hat{\mu}_0$) is given by:

$$\hat{\mu}_0 = \frac{1}{k} \cdot \sum_{i=1}^{k} Z(s_i) \qquad (17)$$

In this case the S parameter generally represents an ordinary or probability-weighted moment as opposed to a distribution parameter, and

ii) *maximization of the regional likelihood function*, as shown in Gabriele and Arnell (1991) for example.

An estimate of the sampling variance can be obtained by generating synthetic series constrained to the estimated regional value $\hat{\mu}_0$, whose value is equal to the mean number of series observed and the computing the variance of the estimated parameter for each series. From Eq. (13), if the observed value of VAR(Z) is known, by subtraction we obtain an estimate of σ_w^2. A region is assumed to be homogeneous with respect to the S parameter when the observed variability of Z can be globally attributed to the sampling variance, i.e. considering the whole series and checking that the extreme values are randomly distributed throughout the region. Otherwise, the region in question is subdivided into a number of climatic subregions that can be considered to be homogeneous as defined above. This division into homogeneous climatic subregions is achieved using a heuristic procedure that aims to preserve the meaning of a homogeneous climatic region and uses a hypothesis validation criterion based on a significance test (Beale test) of the reduction of the mean quadratic error due to a unit increase in the number of subareas considered (Rossi and Villani, 1994).

In order to identify homogeneous regions it should be noted that, if it is true that (according to the above section 3.1) the observed variability of a parameter in a region depends on spatial variability and sampling variability, it is also true that, the first term is seen to assume a relative importance that increases with the order of the parameter itself. This empirical observation is justified by the fact that the sampling variance increases with the order of the parameters and also because the parameters of a higher order are mostly linked to climatic parameters characterizing larger regions. In view of this consideration, the VAPI project has adopted an estimation procedure over three hierarchical levels (Fiorentino et al., 1987). In practice, a region that is homogeneous in respect of a certain statistical parameter is assumed to be homogeneous with regard to all the parameters of a higher order.

At the first two levels of regionalization, the procedure used for drawing up reports on Flood Evaluation in Italy refers to a statistical homogeneity model that acknowledges the presence of a spatial variability (σ_w^2) inside the homogeneous areas, but this variability is assumed to have no structure. Under this hypothesis, Gabriele and Iiritano (1994) developed an estimation technique based on maximum regional likelihood by mutually binding the estimation of the shape parameters to that of the scale parameter, with subdivision into sub-areas fixed at the second level. This also provides the iterative estimation algorithm with greater numerical stability. The same authors also consider the presence of a sampling correlation between the stations.

3.3. EFFECTS OF THE SAMPLING CORRELATION

The validation procedure adopted for the statistically homogeneous regions model illustrated above assumes that the random field $\varepsilon(s)$ is white noise, i.e. there is no sampling correlation between the historic series recorded at different sites. This is verified when the sampling correlation vanishes within a spatial scale smaller than the minimum distance between the stations and corresponds, in terms of the observed variogram, to a pure nugget whose value is the same as the value of the observed variance. Otherwise it is necessary to take into account the inter-station correlation, i.e. a structure of spatial autocovariance of the field $\varepsilon(s)$. In this case, the theoretical variation of the estimator, for instance obtained through the generation of synthetic series, is greater than the observed value of the variance of $\varepsilon(s)$.

Under the hypothesis of a statistically homogeneous region and in the presence of inter-station correlation, the variance of the Z parameter can be obtained through the following relation (Gabriele and Iiritano, 1994):

$$\mathrm{Var}(Z) = \sigma_\varepsilon^2 \cdot \left(1 - \bar{\rho}_X^r\right) \tag{18}$$

in which $\bar{\rho}_X$ is the mean sampling correlation coefficient between the historic series, while *r* is the order of the estimated statistical parameter. Relation (18) implies that, in order to validate the homogeneous region hypothesis, a small sampling variance has to be taken into account.

In terms of the variogram, we can observe a correlation structure due to inter-site dependence. Therefore, as the observed variance will be equal to the mean value of the experimental variogram, it is lower than the sill value which is, instead, equal to the theoretical sampling variance. The presence of a spatial structure on the disturbance error term eliminates the independence hypotheses underlying the regional maximum likelihood estimator. On the other hand, these hypotheses do not affect the structure of the regional estimator based on the spatial mean.

The application of the statistical homogeneity model, whether the interstation correlation is taken into account or not, has made it possible in Italy to identify large homogeneous regions measuring about 10,000 km^2, at least at the first level of regionalization (Ferrari et al., 1994). Occasionally, however, this procedure can hide the effective spatial variability structure of the parameter in question, as will be shown below in the illustration of the application of the regionalization model to a case study. A more general regionalization procedure that takes into account the possible presence of a random spatial structure in the S parameter is illustrated below.

3.4. MODEL OF A RANDOM SPATIAL FIELD WITH AN AUTOCORRELATION STRUCTURE

3.4.1 CHARACTERIZATION OF THE MODEL PARAMETERS

In a more general model, it is necessary to take into account the presence of a random intermediate scale term and its covariance structure. In this case, maintaining the hypothesis of first order stationarity, the regional model becomes:

$$Z(s) = \mu_0 + W(s) + \varepsilon(s) \tag{19}$$

The presence of a spatial correlation in both the sampling disturbance term and the spatial disturbance term entails the following expression of the observed variance in the field:

$$\mathrm{Var}(Z) = \sigma_W^2 \cdot (1 - \overline{\rho}_W) + \overline{\sigma}_\varepsilon^2 \cdot (1 - \overline{\rho}_\varepsilon) \tag{20}$$

in which σ_W^2 = variance of the field W(s), constant over the whole region (2nd order stationarity hypothesis); $\overline{\rho}_W$ = mean regional value of the spatial correlation of the field W(s); $\overline{\sigma}_\varepsilon^2$ = mean regional value assumed by the sampling variance; and $\overline{\rho}_\varepsilon$ = mean regional value of the spatial correlation of the field $\varepsilon(s)$.

At this point, a number of hypotheses can be made on the variability of σ_ε^2. Under the hypothesis of a statistically homogeneous region, we have $\sigma_\varepsilon^2(s) = \overline{\sigma}_\varepsilon^2 = \text{constant}$. In the more general case, the sampling variance of the at-site estimator depends on the value assumed by the S parameter in the same point (*heteroschedastic model*), that is to say a relationship of the following type is hypothesized:

$$\sigma_\varepsilon^2(s) = g(S(s)) \tag{21}$$

The complete definition of the spatial variability structure of the variable Z(s) corresponds to the definition of a semi-variogram model that, in the more general case of a heteroschedastic model, is defined by the following expression:

$$\gamma_Z(s_i, s_j) = \sigma_W^2 \cdot \left(1 - \rho_W\left(\left|s_i - s_j\right|\right)\right) + \frac{\sigma_{\varepsilon,i}^2 + \sigma_{\varepsilon,j}^2}{2} - \rho_\varepsilon\left(\left|s_i - s_j\right|\right) \cdot \sqrt{\sigma_{\varepsilon,i}^2 \cdot \sigma_{\varepsilon,j}^2} \tag{22}$$

The presence of heteroschedasticity means that the hypothesis of intrinsic stationarity at the second moment no longer holds for the field Z(s). It is therefore necessary to operate in terms of the variogram of the field of the variable $S(s) = \mu_0 + W(s)$, which is assumed to be intrinsically stationary. Since $S(s) = Z(s) - \varepsilon(s)$, several authors refer to S(s) as the *noiseless* variable (Cressie, 1991; Bourgault, 1994).

3.4.2 ESTIMATION OF THE PARAMETERS OF THE REGIONAL MODEL

Once the spatial variability model has been defined, it is typically used to make regional estimates by means of a BLUE spatial estimator. This is known as *kriging* (Matheron, 1963) and makes use of the estimate of the so-called experimental variogram. In the case in question, the model represented by (22) requires the characterization of the variance and shape of the spatial correlation of the two error terms, one of which is non stationary at the second order. The problem can be solved by means of the following procedure:

- identification of the link between the at-site sampling variance and the value assumed by the field S(s). This operation can be carried out a priori once the at-site probabilistic model has been identified;

- identification of the form of the sampling error's autocorrelation structure, for instance through direct analysis of the spatial structure of the maximum annual rainfall events;
- iterative procedure for inference on the field S(s) starting from the observations on the field Z(s).

The latter procedure starts from the estimate of $\overline{\sigma}_{\varepsilon}^2$ as the nugget of the empirical variogram of Z when $\varepsilon(s)$ is white noise, or as $\overline{\sigma}_{\varepsilon}^2 = g(\hat{\mu}_0)$. In the first step, it is hypothesized that the sampling variance is constant in space and thus, the variogram of S can be easily derived from that of Z (de Marsily, 1986). After this, the following iterative scheme is followed:

1. estimate of the field S(s) using *noiseless kriging* starting from the observations $Z(s_i)$;
2. calculation of the values of $\sigma_{\varepsilon,i}^2$ starting from the estimated values of $S(s_i)$;
3. estimate of the empirical variogram of S starting from that of Z, after assigning the values of $\sigma_{\varepsilon,i}^2$;
4. return to point 1 until solution convergence is obtained.

In this case, the kriging function effectively assumes the role of spatial filter rather than that of a spatial estimate (Bourgault, 1994).

A particular case of the illustrated procedure is known in literature as *kriging with uncertain data* (de Marsily, 1986) if the sampling error spatial variability structure is neglected. In this case, the problem of the inference on S starting from Z is easier to solve as the effect of the field $\varepsilon(s)$ on the experimental variogram of Z(s) consists of adding a term of pure nugget. As said above, it is an apparent nugget as, in actual fact, the observed structure has such a small spatial scale that it cannot be highlighted in the resolution with which the data is possessed. In this case, the variogram of S(s) is actually obtained by subtracting the mean variance of the field $\varepsilon(s)$ from the variogram of Z(s).

In the iterative estimation procedure presented here, the calculation of $\sigma_{\varepsilon,i}^2$ (carried out in point 2) actually requires a procedure for estimating the parameters of the CDF starting from the estimated values of the statistical parameters $S(s_i)$. If the L-moment ratios are used as S parameters, this procedure has been highly error-prone in the past as the observed combinations of L-kur and L-Cs often fall outside the TCEV's theoretical feasibility space. Furthermore, in some areas of the domain of this space, the numerical estimate algorithm presents a marked instability which makes the estimates obtained somewhat unreliable (Gabriele and Arnell, 1991).

In this paper, some of the limits of the traditional estimate procedure based on PWMs have been overcome. Indeed, it has been observed that the instability arises for sites in which Λ^* assumes very low values, at the limit tending towards 0, or when θ^* tends to the value 1. Both cases correspond to the degeneration of the TCEV distribution into the well-known EV1 distribution. For this reason, a procedure has been implemented whereby, when the system solution tends towards the above limit values, it is assumed that the solution itself corresponds to an EV1 probability distribution.

4. A CASE STUDY: THE PIEDMONT AND LIGURIA REGIONS OF NORTHERN ITALY

4.1. DESCRIPTION OF THE REGION

The regionalization procedures so far illustrated have been applied for the regional study of the annual maxima of daily rainfall in a hydrographic region comprising the basins of the river Po and its tributaries as far as the Tanaro and the basins of Liguria. This hydrographic region includes areas that are highly dishomogeneous orographically and stretch from the southern coastal plain to the Ligurian Apennines and from the western Alps and to the eastern side of the Po river plain, including an area in the north with numerous lakes.

The sample considered is formed of a series of annual maxima of daily rainfall recorded at 277 rainfall measurement stations, whose sites are indicated in Fig.1. The minimum length recorded is 45 years for Piedmont and 30 years for Liguria, giving the series length has an average of 48 years.

Following Gabriele and Iiritano (1994), an estimate of the mean regional value of the spatial correlation of the field $\varepsilon(s)$ was obtained by analyzing just the stations' corresponding annual maxima events. The correlation was found to be extremely low: $\overline{\rho}_{\varepsilon}$ =0.06. As mentioned above, since we are interested in the values of $\overline{\rho}_{\varepsilon}^{\,r}$ (r=2,3,4) for the effects of the inter-station correlation on the moments of order higher than the first, these effects can be considered to be utterly negligible in the region in question.

4.2 DESCRIPTION OF PREVIOUS REGIONALIZATION STUDIES

This region is part of a larger region of Po river basins examined by Brath and Rosso (1994) in the context of the VAPI project conducted by G.N.D.C.I. (Gruppo Nazionale Difesa dalle Catastrofi Idrogeologiche). In a preliminary phase, the above authors refer to the standard procedure adopted in the VAPI project and to a data sample, including the sample considered in the present paper, and found it is impossible to reject the hypothesis that the whole region is homogeneous at the first regionalization level, according to the shape parameters of the TCEV distribution estimated through maximization of the regional probability function, i.e.: $\theta^* = 1.983$, $\Lambda^* = 0.249$. At the second regionalization level they separate the basins of Liguria from those of the river Po to give the shape and scale parameters obtained by Brath and Rosso (1994) for the two regions considered: $\theta^* = 1.907$, $\Lambda^* = 0.218$, $\Lambda_1 = 31.6$ for the Po basins, and $\theta^* = 2.231$, $\Lambda^* = 0.262$, $\Lambda_1 = 35.6$ for the Ligurian basins.

This preliminary analysis, based solely on a comparison between the observed variance and the sampling variance of the parameters and without checking whether the empirical spatial distribution of the outliers was truly homogeneous, was subsequently expanded upon by the same authors.

Subsequent analyses on the same area, conducted exclusively on the annual maxima of the flood flows, identify homogeneous regions under the hypothesis that, in each region, the annual maxima of the flood flows have a self-similar dependence with the area of the basin and that the exponent of this dependence identifies a dissipation mechanism typical for the region [De Michele and Rosso, this issue]. These analyses still have to separate the part of Liguria that faces onto the Tyrrhenian Sea from the rest of the region and consider the whole area between

the Dora Baltea and the Tanaro as a homogeneous region, distinguishing it from the other Po basins.

4.3. THE REGIONALIZATON STUDY FOR STATISTICALLY HOMOGENEOUS REGIONS

In the region in question, the values of the first three L-moment ratios were examined for each station according to the at-site estimate procedure proposed above. On the basis of these estimates, the mean regional value was assessed for each of these under the hypothesis of a homogeneous region, as reported in the first column of Table 1. In particular, with the regional values of the L-skewness and L-kurtosis parameters fixed at the first level, the regional values are estimated for the shape parameters in the TCEV distribution, and are equal to $\theta^* = 1.806$, $\Lambda^* = 0.244$. At the second regionalization level, under the hypothesis that the whole region is homogeneous and after fixing the regional value of L-Cv, the regional value of the TCEV distribution's scale parameter is estimated and is equal to $\Lambda_1 = 23.8$.

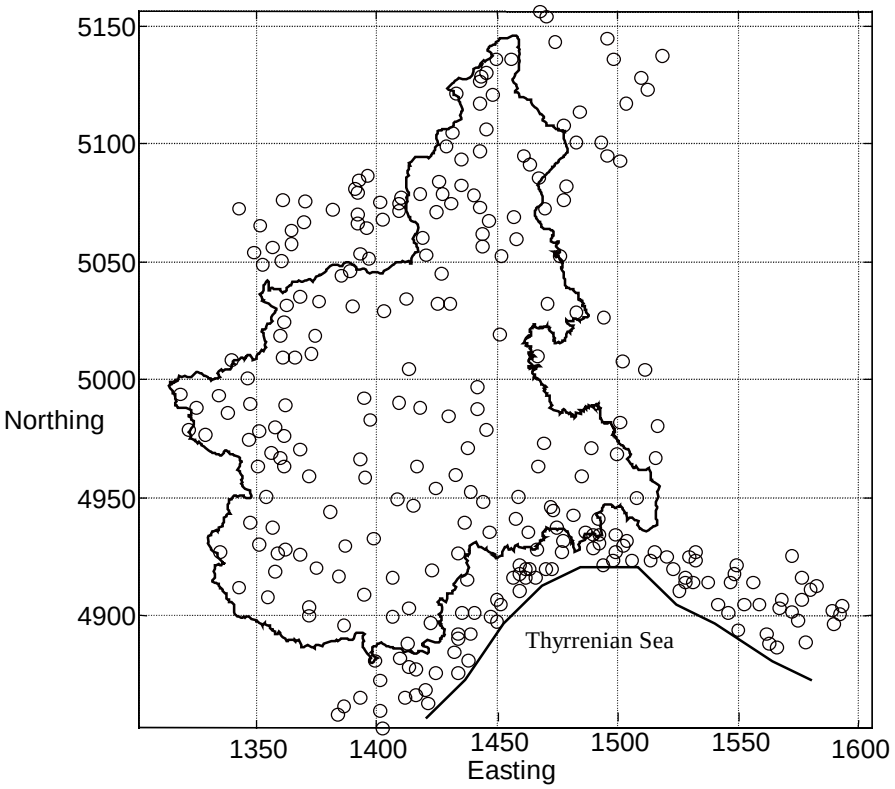


Figure 1 Locations of the rain gauges in the study region.

Table 1 Observed regional values for some statistical parameters and simulated theoretical values under the hypothesis of regional homogeneity; $<\mu_0>$ = observed spatial average, V(Par) = observed spatial variance for the parameter Par, $\mu_{0,gen}$ = synthetically generated mean value, μ_V = mean of synthetically generated variance of Par, σ_V = variance of synthetically generated variance of Par, H = heterogeneity statistics.

Parameter	OBSERVED		TCEV			
	$<\mu_0>$	V(Par)	$\mu_{0,gen}$	μ_V	σ_V	H
L-kurtosis	0.1715	0.004395	0.1691	0.004089	0.0002851	1.372
L-skewness	0.2087	0.006960	0.2022	0.006173	0.0005196	1.527
Skewness	1.1845	0.4565	1.1550	0.4002	0.0534	1.0545
L-variation	0.1959	0.001181	0.1954	0.0005246	0.0000426	15.427

The regionalization model adopted in the VAPI project hypothesizes that the theoretical sampling variance represents the origin of all the observed variability. On the basis of this hypothesis, at each regionalization level, the VAPI procedure verifies that the value $\overline{\sigma}_\varepsilon^2$, obtained through the generation of synthetic series, represents a significant percentage of the value of the observed variance Var(Z).

In order to verify the performance of this regionalization model, a series of simulations were carried out on 300 regions of 277 stations with the same number of characteristics as the observed one. The second column in Table 1 reports the observed value of the variance in the whole region in question for the various parameters. The subsequent columns report the results of the simulation experiment.

Comparing the mean generated value ($\mu_{0,gen.}$) with the mean value observed in the region ($\hat{\mu}_0$), we can see how the estimators of the L-moments have a much smaller distortion than those of the ordinary moments. Regional homogeneity is verified by comparing the value of Var(Z) with the mean value μ_V of the regional variances generated. Note how when S coincides with Cs, L-Cs or L-kurtosis, μ_V always represents at least 88% of the observed variance, which makes the regional homogeneity hypothesis seem acceptable at the first level. Figure 2 shows the comparison between the observed distribution and the theoretical generated distribution of the values of L-Cs and it seems to point out that the regional homogeneity hypothesis is acceptable. This is unlike the case of the parameter L-Cv, in which μ_V explains less than 45% of the observed variance.

An objective test of regional homogeneity was proposed by Hosking and Wallis (1993) who used the heterogeneity statistic H defined as follows:

$$H = \frac{\text{(observed variance V) - (theoretical mean } \mu_V)}{\text{(theoretical standard deviation } \sigma_{Vthe})} \qquad (23)$$

Lower values of H correspond to an enhanced adaptation of the probabilistic model to the spatial homogeneity hypothesis. The authors recommend that for H<1 the homogeneity hypothesis cannot be rejected and for H>2 the homogeneity hypothesis cannot be accepted.

Table 1 reports the values of μ_V, σ_V and H that were calculated with reference to the various statistical parameters used. Note how in all cases H>1 with regards to the first level parameters, while for L-Cv H>>2, thus confirming that the homogeneity hypothesis at the second level should be rejected.

The procedure recommended in the VAPI project consists of subdividing the region into homogeneous subregions. As mentioned above, Brath and Rosso (1994) achieve this subdivision heuristically; while De Michele and Rosso (this issue) utilize a different technique applied to the hydrometric data alone to achieve a result that is substantially analogous in terms of the identification of homogeneous regions.

For the second level regional analysis, the present paper refers to a cluster analysis procedure based on the k-means criterion (Everitt, 1981) based on the statistics of the annual maxima of daily rainfall, considering the statistical parameters L-kurtosis, L-Cs, and L-Cv as variables. The result of this procedure led to a subdivision in which the groups are completely mixed (see Fig.3a). The results improve when only the L-Cv parameter is considered together with the spatial coordinates of the stations. In this case, however, the introduction of the spatial coordinates means that every group is attributed with a regional value of L-Cv that, in actual fact, differs only by 5%, passing from 0.19 to 0.20. This means that the subregions are identified not as a function of the statistical parameter, but only according to their location (see Fig.3b). However, in this case, practically all the Po basins belong to a single homogeneous region. The result of the application of this procedure is that the probabilistic growth factor, and with this also the hydrometeorological risk, is unique throughout the Po region examined.

4.4. THE REGIONALIZATION PROCEDURE WITH SPATIAL VARIABILITY OF THE FIELD

Alternatively, we can analyze the hypothesis in which the high sampling variance in actual fact masks a spatial variability of the probabilistic growth factor, with the result that different areas in the same region are effectively subjected to a different hydrometeorological risk. To this end, the spatial field model with an autocorrelation structure is adopted, using the inference procedure described above. The first step of this procedure is to estimate the experimental variograms for the statistical parameters on which the probabilistic growth factor depends (i.e. a fourth order moment). This presents a constant variogram typical of a field comprising solely of white noise (Fig.4a). The regional estimate of L-kurtosis was thus obtained by referring to a straightforward moving average window procedure. Although the L-Cs parameter is characterized by high sampling noise, it has an evident spatial correlation structure (Fig.4b). Furthermore, the variogram of L-Cv displays a well-defined spatial autocovariance structure, partly because of the smaller sampling variance (Fig.4c).

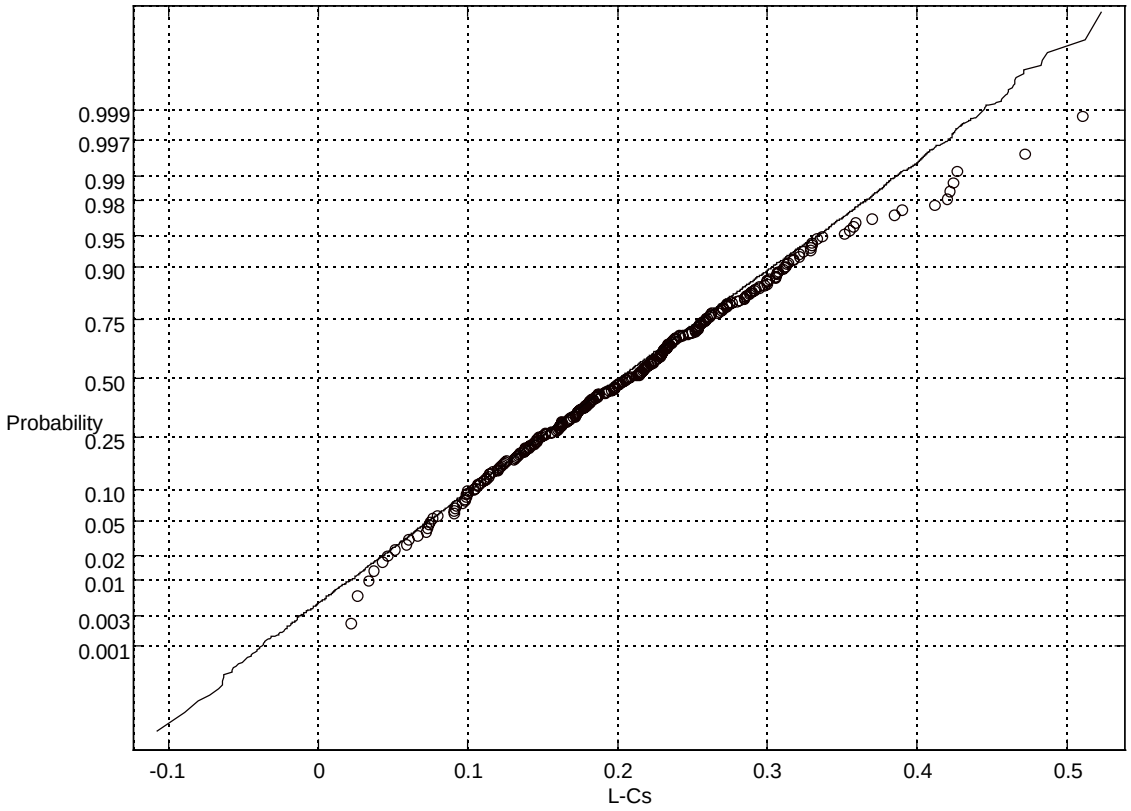


Figure 2. Normal probability plot of observed (circle) and theoretical (line) distributions for L-Cs.

For the latter L-moments, the iterative noiseless kriging procedure described above was applied in order to get an estimate of the real theoretical value of the desired parameter in any part of the region starting from the sampling values observed on the measurement sites. The results of this procedure have been reported as contour plots for both the value of the shape parameter θ_* of the TCEV distribution (Fig.5) and for the spatial distribution of the probabilistic growth factor for an assigned return period (Fig.6).

These results show the presence of two areas with a higher pluviometric risk: one in the south-eastern part of the examined region, astride the Bormida basin; the other in western Piedmont, including the upper Orco basin. These areas present K_T values that are above 3.6 for T=1000 years (fig.6).

Whereas for much of the central plain the value of θ_* is almost equal to 1, which means that the probability distribution of the annual maximum of daily rainfall is of the type EV1. Therefore, in the absence of the extraordinary component, this area is subject to a lesser hydrometeorological risk. For instance, for T=1000 years, K_T is less than 2.8.

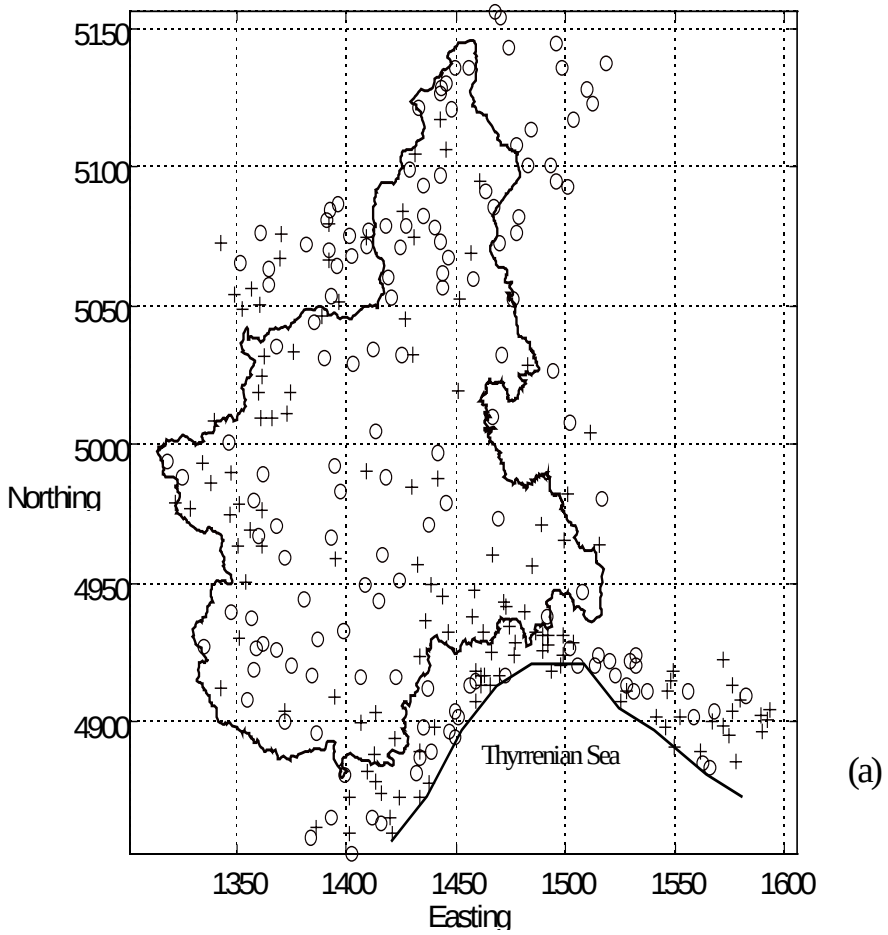
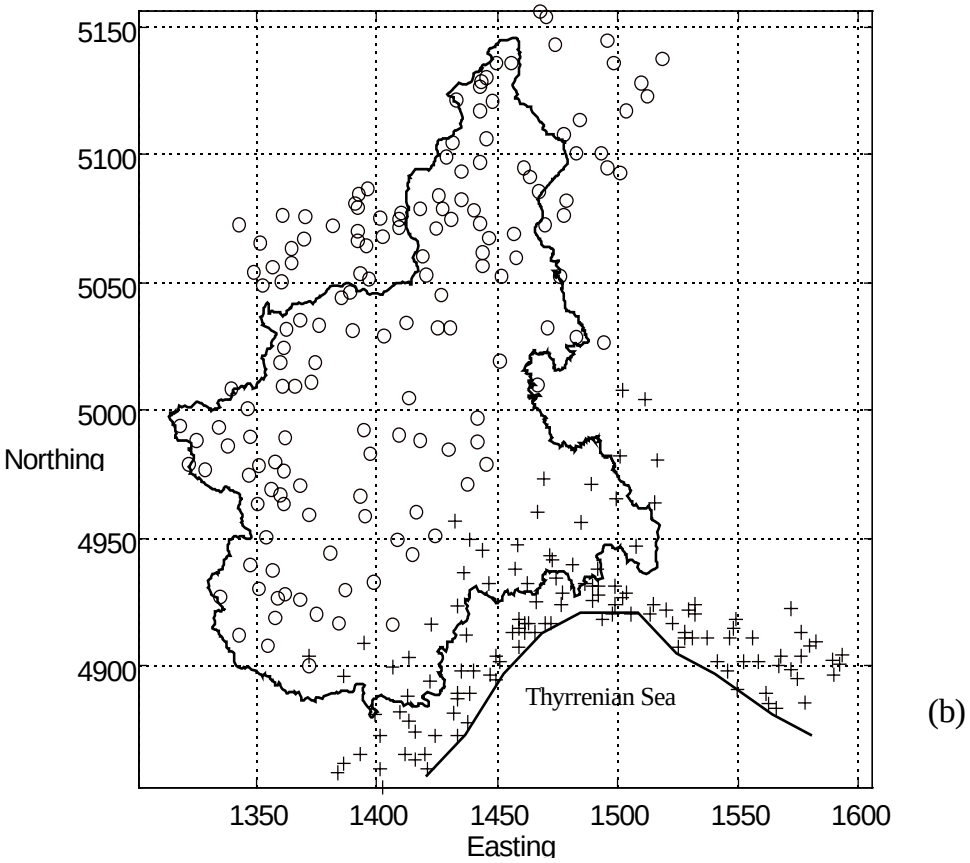


Figure 3. Identification of homogeneous regions using cluster analysis: k=2: a) Clustering by L-kurtosi, L-Cs and L-Cv ; and b) Clustering by L-Cv and spatial coordinates

In this case, there is an obvious advantage in using a technique for analyzing the spatial variability of the statistical parameters of the probability distribution of the annual maxima of daily rainfall, which makes it possible to achieve an objective discrimination of the hydrometeorological risk in areas with different climatic behaviours in the same region. Neglecting this variability would have meant attributing the same risk to every site in the region. For instance, for T=1000 years, the spatial homogeneity hypothesis would give a value K_T = 3.21 for the whole region with an underestimate ranging from 11% to 20% in the regions at greater risk, and an overestimate ranging from 15% to 50% in the regions at lesser risk.

5. FINAL REMARKS

Natural hydrometeorological disasters in the Mediterranean region are essentially affected by two factors: (i) rivers generally with shorter hydrologic response times than social



(b)

Figure 3 (Continued)

response time (**flash floods** with peak times of less than 6 hours); and (ii) disastrous floods caused in the past by outlying storm events characterized by considerable rainfall intensity and rare occurrences (**hydrometeorological risk**). The characterization of this type of event is a crucial point in risk mitigation, both for planning intervention and for the identification of the main flooding areas. Because of their rare occurrence, these extreme events can be processed only on a regional frequency basis in order to reduce parameter uncertainty estimation in gauged sites and for risk evaluation in ungauged areas. A statistical regional model includes: (i) a probabilistic model, which can explain the extraordinarily high floods and rainfall observed in the past (outliers); (ii) a regionalization model, which can take into account the observed spatial variability of the probabilistic model's statistical parameter.

Here, a regionalization model based on the TCEV distribution probabilistic model has been presented along with a geostatistical analysis of its parameters. The two-component probabilistic model makes it possible to carry out a statistical evaluation of the likelihood that the annual maximum for floods or rainfall will come from the extraordinary component (e.g. in Italy, this has been estimated at around 26% (Fiorentino et al. 1987). This likelihood is essentially a climatic characteristic of the region in question. For instance, in Great Britain, where rainfall is more regular than in the Mediterranean, it is virtually negligible (about 3%) (Rossi et al. 1986).

In South Africa, it has been observed (Pegram and Adamson 1988) that the likelihood of the annual maximum for rainfall over 3 days at any one site belonging to the extraordinary component varies from about 35% along the coastline to less than 15% inland. This variability has been interpreted in terms of synoptic scale climatology.

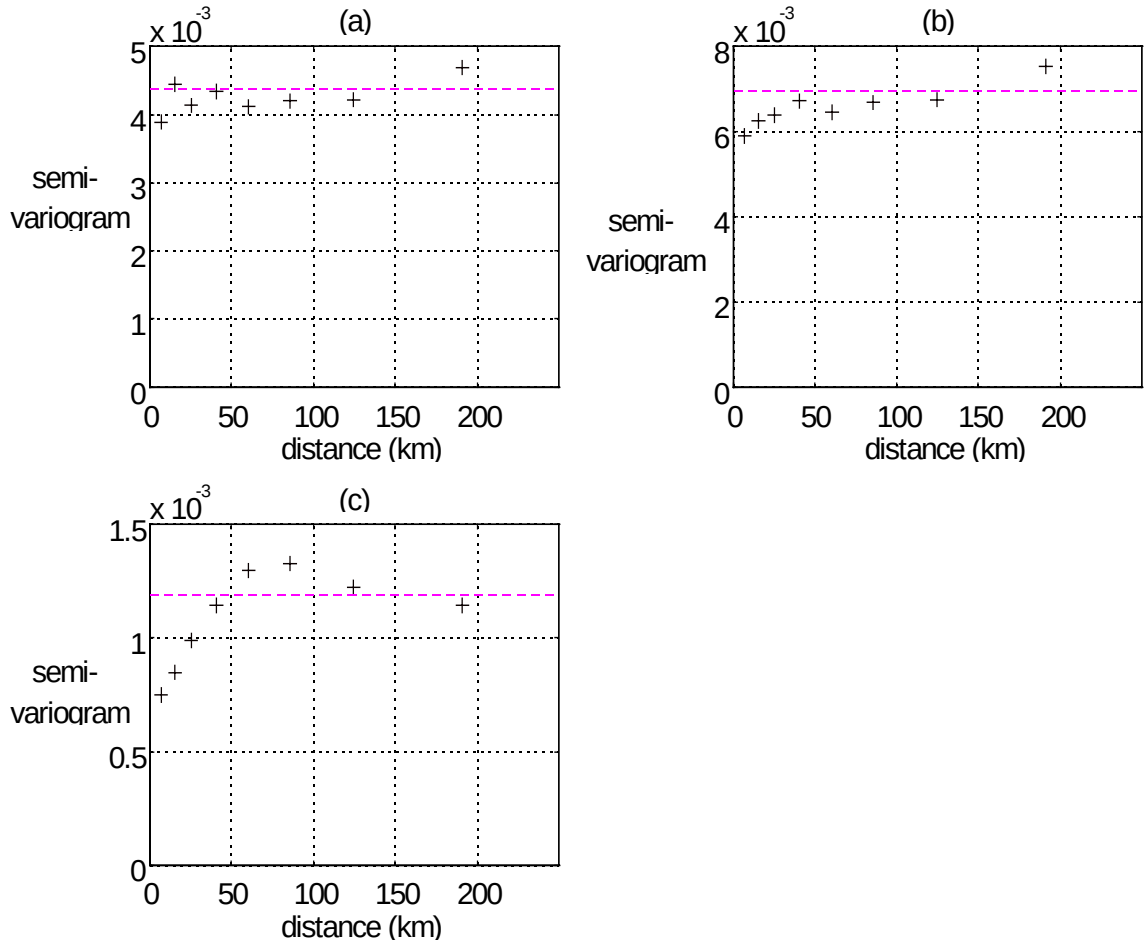


Figure 4. Experimental (observed) semi-variogram (+) and observed variance (dashed line) for: a) L-kurtosi ; b) L-Cs ; and c) L-Cv.

Attention has thus been focused on the spatial variability model for the statistical parameters of the TCEV probability distribution. A very general spatial variability model has been introduced that takes into account the presence of deterministic and aleatory components on a different spatial scale. In particular, it has been shown that the regionalization models usually adopted in flood frequency analysis stem from a particular case of the proposed model adopted in flood frequency analysis stem from a particular case of the proposed model called the *deterministic spatial homogeneity model*, in which the primary source of uncertainty regards the parameter's sampling estimate. Therefore, it is thus necessary to estimate a single regional value

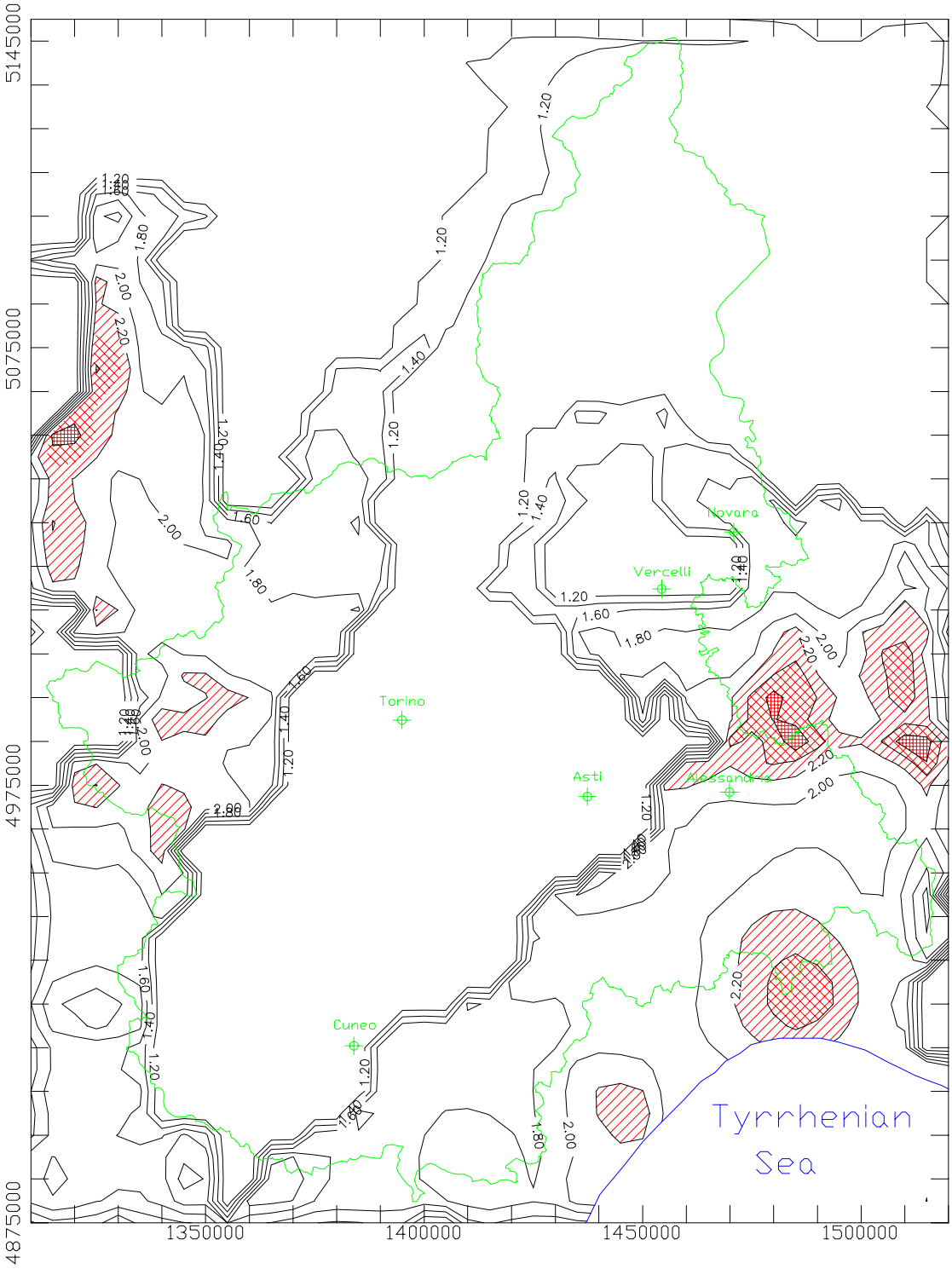


Figure 5. Contour plot of θ_*. Shaded areas with $\theta_* > 2.20$.

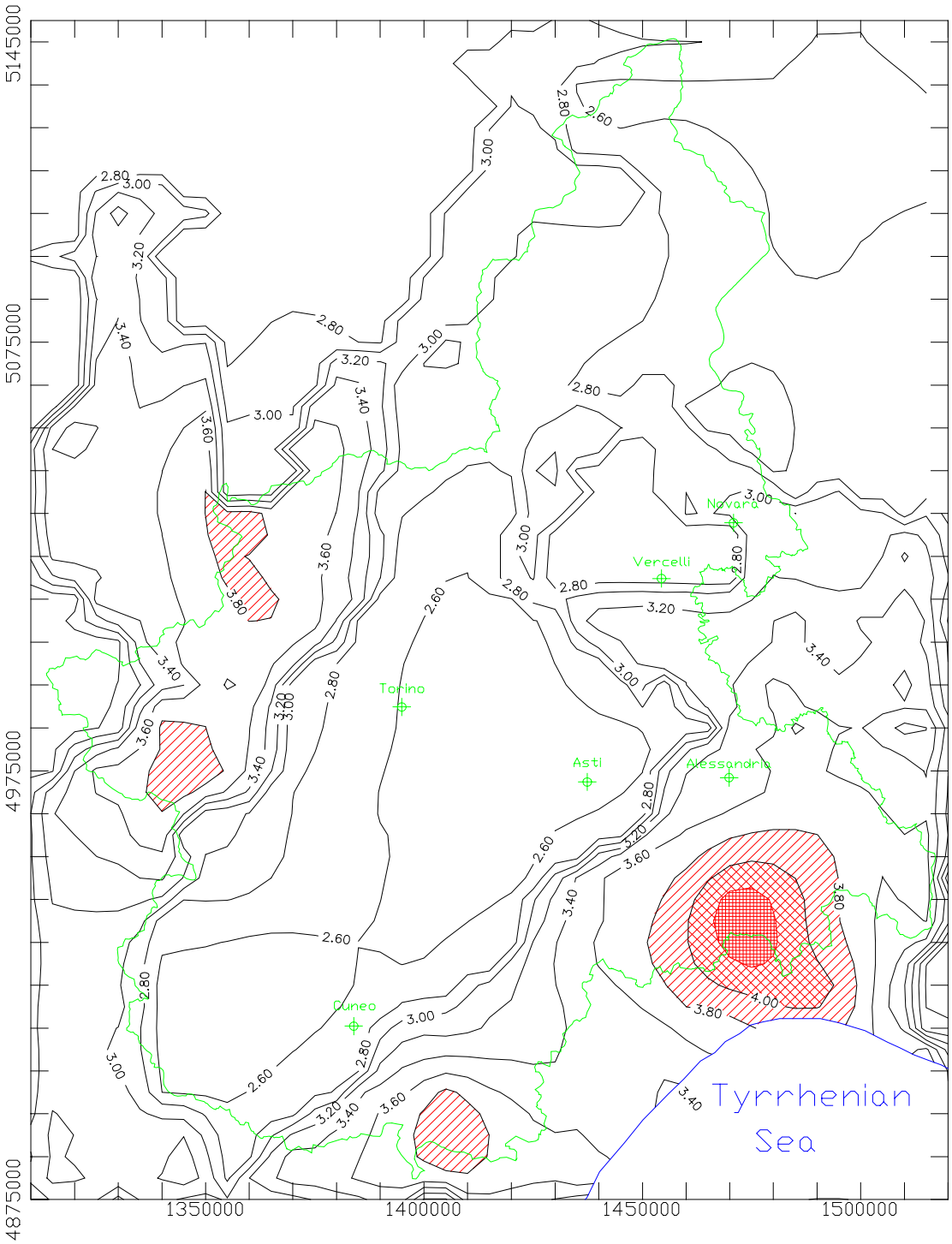


Figure 6. Contour plot of probabilistic growth factor K_T, for T = 1000 years. Shaded areas with $K_T > 3.8$.

for the statistical parameter being studied. This approximation improves as the estimator's sampling variance, which can be obtained through simulation of synthetic series of the region in question, approaches the variance value observed for the region. This also indicates that the

greater the parameter's statistical order, the more its estimation variability will hide any variability in the region, as already proposed in the hierarchical scheme of Fiorentino et al. (1987). In order to assess the goodness of this approximation, reference can be made to a *homogeneity statistic* introduced for this very purpose by Hosking and Wallis (1993).

Should this approximation turn out to be unsatisfactory for the objectives in question, the whole region can be subdivided into homogeneous groups, inside each of which the deterministic homogeneity hypothesis is adopted. Cluster analysis techniques can be used in order to carry out the above subdivision. This procedure has a number of disadvantages: the lack of a physical meaning for the subdivision, the fact that groups are spatially distributed in a leopard skin pattern, and uncertainties in the regional estimates at the edges of the clusters.

The proposed spatial variability application makes it possible to overcome these limitations by explicitly taking into account the effects of the spatial covariance structure of the local disturbance factors (intermediate scale spatial *variance*) and the presence of *interstation correlation* effects that ensure that the sampling uncertainties in different sites are correlated to one another. To this end, it has been necessary to define a procedure for estimating the value of the relevant statistical parameter at a certain site in the region starting from observations of this parameter taken in gauged sites and affected by sampling uncertainty. This procedure is iterative and starts from a relation defined a priori between the parameter's sampling variance and the value of the parameter itself, depending on the probabilistic model used. The usual geostatistical techniques, such as *noiseless kriging* (Cressie, 1991; de Marsily, 1986), can then be used to infer the regional values of the parameter in question and their spatial covariance structure.

This procedure has been applied with reference to the distribution of the annual maximum for daily rainfall in a region of northern Italy, including part of the Po basin and the basins of Liguria. It has been shown how, even when the deterministic homogeneity hypothesis can be considered satisfactory according to the usual acceptance criteria, the spatial variability analysis of the TCEV distribution scale and shape parameters makes it possible to say that there are areas with a different hydrometeorological risk and enables their objective identification. In the present case study, neglecting this spatial variability may lead to overestimation and underestimation errors and therefore, the deterministic homogeneity hypothesis remains conservative. The reasons for preferring the proposed procedure, in addition to the obvious reasons of an economic nature regarding the mis-sizing of engineering works resulting from such a procedure, include the observation that there are regions in which the hypothesis of a homogeneous region leads to substantial underestimations of the hydrometeorological risk (by up to 20%), as in the case of the basins recently hit by the flood events of 1994.

The presence of a spatial variability in the higher order statistical parameters of probability distributions for annual maxima of rainfall or floods raises a series of problems essentially linked to the possibility of characterizing this variability in climatic terms. In this connection, it might be particularly useful to have: (i) physically-based analyses regarding different meteorological events that explain the presence of different hydrological events; (ii) statistically-based analyses regarding the space-time structure of such events, classified according to type; (iii) assessment of the effects of such event characteristics on the analysis of the maxima at site and equalized for the area. Thus it is clearly necessary that the results presented here in terms of a statistical model of the spatial variability of statistical parameters of the hydrological maxima distribution are interpreted on a physical and conceptual basis and, likewise, that these results can represent the basis for constructing stochastic event models capable of correctly representing the phenomenon.

Acknowledgements: This work was partly funded by CNR 95.00267.PF42 grant and under the agreement between Regione Piemonte and CUGRI, 16/06/95 no. 131-47100.

6. REFERENCES

Beran, M.A., Hosking, J.R.M. and Arnell, N.W. (1986) *Comment on "TCEV distribution for flood frequency analysis"*, in *Water Resour. Res.*, 22 (2), 263-266.

Bourgault, G. (1994) *Robustness of noise filtering by kriging analysis*, in *Mathematical Geology*, 26 (6), 733-752.

Brath, A., and Rosso, R. (1994) *Valutazione delle piene nel bacino padano e nella Liguria tirrenica*, in *La valutazione delle piene in Italia (Bozza preliminare)*, B, GNDCI-CNR, Roma.

Cressie, N.A.C. (1991) *Statistics for spatial data*, John Wiley & Sons, Inc., USA.

de Marsily, G. (1986) *Quantitative Hydrogeology*, Academic Press, USA.

De Michele,C and R. Rosso (1996) *Self-similarity as a physically basis for regionalization of flood probabilities*, this issue.

Everitt, B. (1981) *Cluster Analysis*, Social Science Research Council, Halsted Press, U.K..

Ferrari, E., Gabriele, S. and Versace, P. (1994) (edited by) *La valutazione delle Piene in Italia*, GNDCI-CNR, Roma.

Fiorentino, M., Gabriele, S., Rossi, F. and Versace, P. (1987) *Hierarchical approach for regional flood frequency analysis*, in *Regional Flood Frequency Analysis*, (V.P. Singh ed), 35-49, D. Reidel, Dordrecht, Holland.

Gabriele, S. and Arnell, N.W. (1991) *A hierarchical approach to regional fllod frequency analysis*, in *Water Resour. Res.*, 27 (6), 1281-1289.

Gabriele, S. and Iiritano, G. (1994) *Alcuni aspetti teorici ed applicativi nella regionalizzazione delle piogge con il modello TCEV*, GNDCI-CNR, Roma.

Hosking, J.R.M. (1990) *L-moments: Analysis and estimation of distribution using linear combinations of order statistics*, in *J. R. Statist. Soc.*, 81, 158-171.

Hosking, J.R.M. and J.R. Wallis (1993) *Some useful statistics in regional flood frequency analysis*, *Water Resour. Res.*, 1993.

Matheron, G. (1963) *Principles of geostatistics*, in *Economic Geology*, n.58, 1246-1266.

Pegram, G. and Adamson, P. (1988) *Revised risk analysis for extreme storms and floods in Natal/KwaZulu*, in *Die Siviele Ingenieur in Suid-Afrika*, 1, 15-88.

Reale, O. (1996) *Tempeste autunnali sul Mediterraneo: inquadramento sinottico-dinamico ed aspetti previsionali in un General Circulation Model (GCM)*, in *Comunicazione in Tempeste Mediterranee*, Savona.

Rossi, F. and Villani, P. (1994) *A project for regional analysis of flood in Italy*, in *Coping with floods*, (G. Rossi, N. Harmancioglu and V. Yevjevich eds), 227-251, Pre-proc. of NATO - ASI, Kluwer Academic, Dordrecht, Holland.

Rossi, F., Fiorentino, M., Versace, P. (1984) *Two-Component Extreme Value distribution for flood frequency analysis*, in *Water Resour. Res.*, 20 (2), 847-856.

Rossi, F., Fiorentino, M., Versace, P. (1986) *Reply to the Comment on "Two-Component Extreme Value distribution for flood frequency analysis"*, *Water Resour.Res.*, 22(2), 267-269.

Versace, P., Ferrari, E., Gabriele, S. and Rossi, F. (1989) *Valutazione delle piene in Calabria*, IRPI-CNR, Geodata, Cosenza.