REPRESENTATION OF CONDITIONAL RANDOM DISTRIBUTIONS AS A PROBLEM OF "SPATIAL" INTERPOLATION

HANS VON STORCH

Institute of Hydrophysics

GKSS Forschungszentrum

D-21502 Geesthacht Germany

Abstract

The problem of specifying random distributions conditional upon external "independent" factors may be seen as a spatial interpolation problem of conditional moments in a generalized phase space. Different techniques solving this interpolation problem are presented, and the different requirements for applications for simulatic and forecast purposes are discussed. The design of universal empirical coordinates is outlined and the concept of data assimilation by means of forecast schemes is sketched.

1. The Interpolation Problem

It was Alexander von Humboldt who solved in the early 19th century the problem of presenting spatially distributed data – he invented the "isotherms" by drawing imaginary lines of constant temperature ir geographical map. These contour lines served chiefly the purpose of visualizing the quantitative data. Inside the 20 degree isotherm all stations report temperature larger than 20 degrees, whereas outside the temperature at all stations would be less than 20 degrees. The isotherm itself is imaginary; in principle there is such a line, but it can be determined only approximately; it is the art of spatial interpolation to describe this unknow unobservable line.

The idea of generalizing geographical maps was extended. While in Humboldt's case the coordinates geographical coordinates, in many modern situations "isolines" of various variables are drawn in a parame space with coordinates being physical or other variables. An example is given in Figure 1 showing th precipitation in Central England given as a function of vorticity and flow direction (from Osborn et al., 1998 This continuous " map " is derived from a finite number of precipitation values, which have been averaged ove all cases in small intervals of similar vorticity and flow direction. Clearly, the mathematical construction i limited to two dimensions but is useful for any number of dimensions – even though a graphical representation is limited to one, two and three dimensions.

The use of physical variables such as temperature, momentum, concentration of substances and the lik coordinates is relatively intuitive, but the use of generalized coordinates in state space representations need some familiarization. An example is given in Figure 2 taken from Biau et al. (1998) displaying precipitation a the rain gauge Orense in Northwest Spain as a function of the coefficients of two indices of regional air pressure distribution. Also in this cases the problem of structuring the data is a problem of multi-dimensio spatial interpolation.

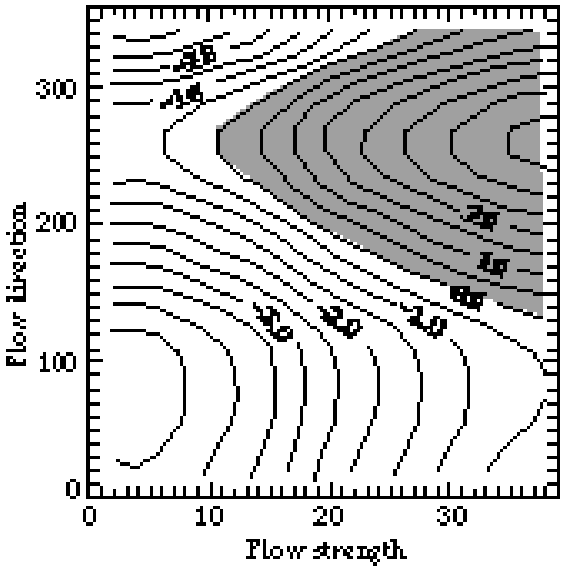

Figure 1: Mean precipitation anomalies (i.e. deviations from the long term mean; in mm/day) in Central England given as function of flow direction (degree) and flow strength (m/s). From Osborn et al., 1998

In general, we are asking for a "surface" S in a n-dimensional space with coordinates $\mathbf{x} = (x_1 \ldots x_n)$ *satisfying*

$$|| S(\mathbf{x}^k) - S^k || < \delta$$

at K data points with coordinates $\mathbf{x}^k$ *and an observed state* S^k. δ *represents a small acceptable deviation, which in many cases is simply zero. In case of kriging, when a nugget effect is taken into account,* δ *is nonzero.*

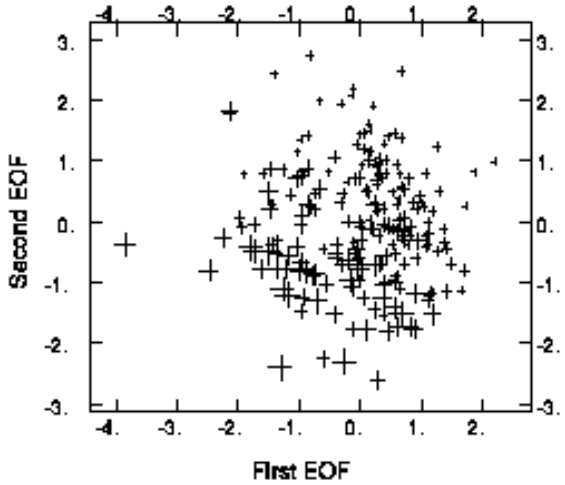

Figure 2: Monthly mean precipitation at Orense as a function of the coeffients of the first two EOFs of monthly mean air pressure (see Figure 4). Winter only.

Behind the concept discussed so far stands the view that there is a "true" surface S, with a unique values $S(\mathbf{x})$ *for any location* $\mathbf{x}$ *in our geographical or state space. This assumption is reasonable when we think of the height of a terrain and geographical coordinates (at least when we disregard geological changes). However when we think of the surface of the ocean, there is no longer a well defined height but a variable surface elevation may be described in probabilistic terms. In Osborn's case of the precipitation as a function of direction and strength (Figure 1), precipitation can not completely be understood as a function of the variables* $\mathbf{x}$*, but must in part be understood as being random. The character of randomness, though, may depend concurrent flow direction and strength. That is, precipitation may be modeled as a random variable R conditioned upon flow direction and strength, and Osborn's surface in Figure 1 is the conditional expectation of precipitation*

$$S(\mathbf{x}) = E(R|\mathbf{x}). \quad (2)$$

Similarly, one can plot surfaces of conditional percentiles or conditional moments instead of the cond expectation. In a sense, geographical maps displaying the spatial distributions of extreme events prov information about such conditional percentiles.

Estimation is based on temporal sampling. For example, in Osborn's case, the value S= -2.5 mm/day for strength of 20 m/s and a flow direction of 100° is the average of

rainfall amounts recorded on many days with a flow strength of about 20 m/s and a flow direction of about 100°.

Therefore, the results of the interpolation is an approximate or estimated surface S^E, which differs to some extend from the true surface. The purpose of the spatial interpolation is the determination of the continuou surface $S(\mathbf{x})$ and not the reproduction of the points S^k. Therefore, the success of S^E as an estimator of S may be determined by comparing the estimates $S^E(\mathbf{x})$ with the true $S(\mathbf{x})$ at a number of additional data points. "Additional" means that the data $S(\mathbf{x})$ have not used for determining S^E. Conventional measures of success are

- *the bias, in the mean or in the standard deviation, i.e., $B_M = < S^E(\mathbf{x}) > - < S(\mathbf{x}) >$ and $B_\sigma = \sigma^E - \sigma$ where the brackets <,> represent the averaging operation over all considered data points x. σ^E and σ are standard deviations of the $S^E(\mathbf{x})$ values and of the $S(\mathbf{x})$ values.*
- *the mean square root difference, i.e.,* $\mathrm{MSE} = \sqrt{<(S^E(x) - S(x))^2>}$
- *the squared correlation, i.e.,*
  $\rho = <S^E(\mathbf{x}) \cdot S(\mathbf{x})>^2 / <S^E(\mathbf{x})^2><S(\mathbf{x})^2>$
- *the represented variance*
  $\varepsilon = 1 - MSE / <S(\mathbf{x})^2>$

2. Solving the interpolation Problem

*Whenever spatial interpolation, be it in a conventional geographic set-up or in a state space set-up, is attempted, one has agree on some assumptions about the representativity of the data. These data usually represent the property S at some points **x**. In most cases the property is thought to be continuous so that the point observation is really representative for neighborhood of **x**. The size of the neighborhood is measured by a correlation length scale. In case of kriging sometimes a "nugget effect" is modeled which allows for a limited discontinuity of the data (Wackernagel, 1995).*

In the randomized set-up, it is assumed that the data points are drawn from random variables which are identical neighborhoods of the data points, so that several observations may be combined into estimates of local means, stand deviations, percentiles or extreme value statistics.

The interpolation itself can be done in various ways; they differ with respect to a-priori assumptions made about the structure of the surface.

- *The strongest assumption specify the global structure. A frequent case refers to multiple regression, which is based on the assumption that the surface is a (linear) plane. Approaches using Canonical Correlation Analysis and Redundancy Analysis belong into this category (cf. von Storch and Zwiers, 1999).*
- *The fitting process is considerably more flexible if the surface is assumed only locally linear, or cubic etc., with certain continuity assumptions. Straight forward linear interpolation is an example; Brandsma and Buishand (1997) have used cubic splines for specifying precipitation as a function of temperature. Also kriging belongs into this category.*
- *Nonlinear fitting algorithm such as fuzzy logic (Faucher et al., 1999) as well as neural networks are also in use. In these approaches, it is somewhat unclear what the underlying assumptions are.*
- *The weakest assumption are required for the analog technique (Zorita and von Storch, 1999), which represents the surface as piecewise constant plateaus around the data. The technique is also known as Voronoi nets (cf. Stoyan et al.,*

1997).

We have applied some of these techniques for the problem of mapping monthly mean precipitation at Orense in wint (December, January and February) as a function of the monthly mean air pressure distribution in the European / North Atlantic sector. The detailed description of the air pressure field is provided by 285 grid points every 5^o latitude and longitude on an area extending from 70^{o}W to 20^{o}E and 25^{o}N to 70^{o}N. For each winter month in the years 1899 to 1989 one such vector of 285 components. The dynamical concept is that the amount of rainfall may be related to variations in the large-scale pressure distribution. Of course, not all variations in rainfall can be traced back to pressure pattern variations, but it reasonable to view local precipitation as a random variable which is conditioned upon the prevailing air pressure. In oth words, we argue the statistics of rainfall, in particular the mean and the standard deviation at some location, vary with weather type as given by the mean air pressure pattern.

Of course, one can not map rainfall as a function of 285 coordinates. Therefore two indices are calculated for each month which characterize the large-scale features of the air pressure distribution. These indices are the coefficients of the firs Empirical Orthogonal Functions (EOFs; for a detailed introduction, refer to von Storch and Zwiers, 1999; Preisendorfer, 1988) or principal components of the air pressure field. EOFs are a system of orthogonal vectors which are adapted considered vector field, and are most powerful in representing variance. In that sense they represent something like uni efficient coordinates. The concept behind EOFs is sketched in the Appendix; there the first two EOFs used here are displayed.

The rainfall in Orense as a function of these first two EOF coefficients is displayed in Figure 2; for each month one cross plotted, with the size of the cross proportional to the amount of rainfall. Obviously the rainfall constitutes not a smooth surface; this is meaningful when we consider rainfall as a conditional random variable, and each observed rainfall amount one realization of this family of conditional random variables. Even if two month are close in terms of the EOF coefficients, the rainfall may differ substantially.

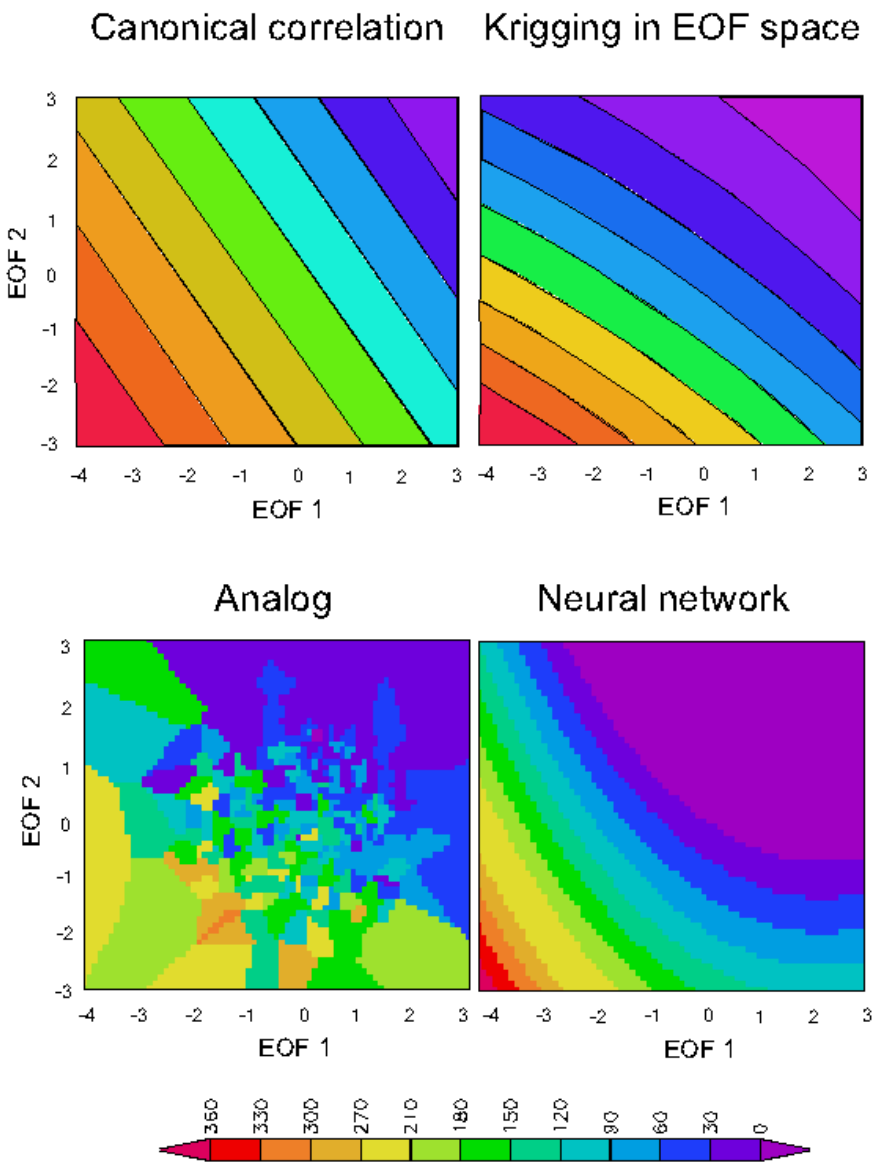

Figure 3: Interpolation of the data points shown in Figure 2 with different techniques: Canonical Correlation (multiple regression), kriging, analog and neural network.

We have applied different interpolation techniques to the data displayed in Figure 2; the results are shown as "isolines" Figure 3. The three techniques Canonical Correlation (or multiple regression), kriging (geostatistics) and neural nets re smooth surfaces - estimates of the conditional means. Maximum rainfall is described by all three approaches for both coefficients being negative, and minimum rainfall when both are positive. Otherwise there are some differences, and it is hard to tell by simply looking at the maps which of the techniques returns the best fit.

TABLE 1. Skill of different techniques for interpolation rainfall in Orense (Kriging) Canonical Correlation (CCA), neural nets (NN)

	Kriging	CCA	NN	analog
ρ	0.81	0.70	0.68	0.69
MSE	6.4mm	7.0mm	7.2mm	7.4mm
ε	65%	51%	44%	38%
B_σ	5mm	21mm	13mm	–2mm

The interpolation was done with data from 1899 to 1968; thus the remaining data from the winters 1969-1989 may se additional, independent data to measure the skill of the schemes. The result is listed in Table 1. The best results are obtained by kriging, with a correlation of 0.81 between the locally observed monthly mean rainfall and the estimated rainfall derived by determining the air pressure distribution indices in the diagrams in Figure 3, and 65% of the month-to-month variability of the local data represented. Also the mean square root error is minimum with 6.4 mm. The bias of the mean is in all c negligible; the standard deviation of 70 mm, however, is underestimated by the three smooth interpolations.

The analog technique scores least in terms of the correlation between the rainfall and its estimate, the mean square root error and the represented variance; however it is superior in terms of the standard deviation of the time series of estimated rainfall.

3. Application of the interpolated "surface"

The purpose of interpolation is "to guide people in unknown terrain", i.e., to guess the state ot the system in "locations" n visited no far. Such guesses can be of very different format, depending on the user's needs.

- *In many cases, and in particular in the case of forecasts, the randomness is considered an unavoidable nuisance, since it is intrinsically unpredictable. Therefore, not the actual value is specified but the conditional expectation. Ideally, the specification is: When the coordinates are close to* $\boldsymbol{x}$*, on average, our interpolated variable will have a value of* $S^E(\boldsymbol{x})$*. The actual value will with some probability be within the interval* $S^E(\boldsymbol{x})$ $\pm\Delta$*, with some level of uncertainty* Δ*. Thus, for forecasting problem, techniques returning smooth surfaces are superior and the analog technique should not be used.*

- *On the other hand, often not "best predictions" are needed but "weather generators", i.e., methods which generate time series with statistics as observed. In that case the purpose is "simulation", and the capability of the system to generate the right level of variability (and other aspects such as the autocorrelation function; length of dry and wet spells) becomes essential. The analog method fulfills this request automatically, if designed properly (cf. Zorita et al., 1995).*

  The aspect of simulating conditional random variables is implicit in almost all parameterisation schemes that are used in dynamical models to account for the net effect of unresolved processes (for instance clouds or gravity wave atmospheric models) on the resolved scales (for a discussion, see von Storch, 1997). In practice, howe parameterisations are formulated as conditional expectations.

4. The "analysis" problem

In environmental sciences usually only a few point observations in space and time are available from which an analysis of state of the system everywhere in space and time is to be derived. The best known example is the routine weather map, whic was originally prepared manually by skillful and experienced meteorologists. Later, this often difficult task was done ir objective manner by insightfully merging the limited observational evidence with dynamical understanding, i.e., realisti models. This "data assimilation" represents a hot topic in meteorology an oceanography undergoing fast development Robinson et al., 1998).

There are several different approaches; here we only sketch a prototypical approach, namely the Kalman filter (e.g., Jone 1988; Honerkamp, 1994).

*It is assumed that the system is characterized by a state variable $\underline{y}_t$ which is a three dimensional continuous distribution. Its dynamics are described by the dynamical model **F**. This model is acknowledged to be imperfect so that forecasts prepared with **F** exhibit errors $\underline{a}$, of which a covariance matrix is known. Thus*

$$\underline{y}_t = \mathbf{F}(\underline{y}_{t-1}) + \underline{a}_t$$

*On the other side, observations are available at time t. We write these observations as one vector $\underline{d}_t$, whose dimension is much smaller than the dimension of the state variable $\underline{y}_t$. The observations are thought to be related to the state variable through an "observation model" **O** Also this model is imperfect and allows for errors $\underline{b}$ with a known covariance matrix:*

$$\underline{d}_t = \mathbf{O}(\underline{y}_t) + \underline{b}_t$$

*The task is to estimate the present state $\underline{y}_t$ consistently from the previous state $\underline{y}_{t-1}$ and the present observations $\underline{d}_t$. This is also in some general sense an interpolation task. It is solved by first calculating the "first guess" $\underline{y}_t^*$ from the previous state $\underline{y}_{t-1}$ with the help of the dynamical model **F**:*

$$\underline{y}_t^* = \mathbf{F}(\underline{y}_{t-1})$$

Then, the observations are forecasted.

$$\underline{d}_t^* = \mathbf{O}(\underline{y}_t^*)$$

and compared with the actual observations $\underline{d}_t$. The final estimate is then a linear combination of the first guess and a term reflecting the mismatch of the predicted observations and the real observations:

$$\underline{y}_t = \underline{y}_t^* + \mathbf{K}\,(\underline{d}_t - \underline{d}_t^*)$$

*The challenging part of the exercise is the determination of the operator **K**. In case of the Kalman filter it is determined algebraically from **O** and the covariance matrices of $\underline{a}$ and $\underline{b}$. A simpler solution is the "nudging" technique, which can be applied when the observations are local values of the state variable. Then **K** has a small positive value at those locations where observation are available and is zero otherwise.*

Acknowledgement

Eduardo Zorita supplied the Orense example; Ilona Liesner and Beate Gardeike prepared the manuscript.

Appendix: Empirical Orthogonal Functions

Empirical Orthogonal Functions (EOFs) or principal components are powerful tools for representing a time-variabl

distribution $\mathbf{Y}$*. Mathematically it represents a new orthonormal coordinate system, i.e., the original field is expanded* $\mathbf{Y_t} = \mathcal{R}_i a_{i;t} \underline{e_i}$*. The* $\underline{e_i}$ *are fixed vectors with the same dimension as* $\mathbf{Y}$*. They are the EOFs. Because of their orthogonality, the time coefficients are simply given as the scalar product of the original field* $\mathbf{Y}$ *and the EOFs:* $a_{i;t} = \mathbf{Y_t}^T \underline{e_i}$*. The first EOF is determined such that the error* $E_1 = \mathcal{R}_t (\mathbf{Y_t} - a_{1;t} \underline{e_1})^T (\mathbf{Y_t} - a_{1;t} \underline{e_1})$ *is minimum - that is the case when* $\underline{e_i}$ *is the eigenvector of the covariance matrix* $\mathcal{R}_t \mathbf{Y_t} \mathbf{Y_t}^T$ *with the largest eigenvalue. The second EOF is the eigenvector with the second largest eigenvalue and so forth. The coefficients are pairwise uncorrelated, and their variance equals the eigenvalues associated with the EOFs.*

Usually, only the low-indexed EOFs are used, as they represent most of the variability of the original field $\mathbf{Y}$ *but with considerably less coordinates. In practice the EOFs are calculated form a limited sample of fields* $\mathbf{Y}$*. Some care is needed when dealing with sampling uncertainty. For details see von Storch and Zwiers (1999).*

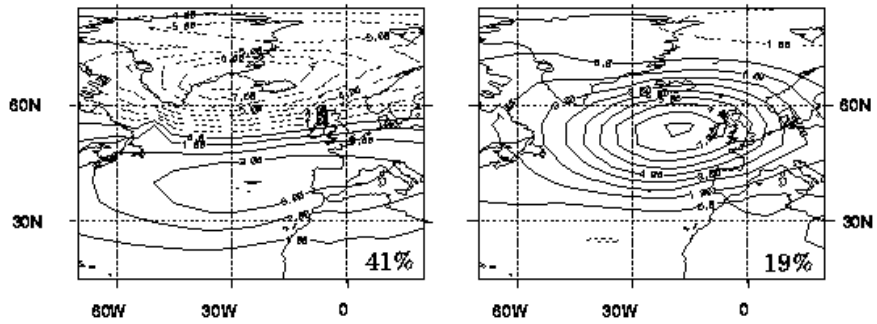

Figure 4: First two EOFs of monthly mean air pressure in winter. Units: hPa

For monthly mean air pressure in winter, the first two EOFs for the North Atlantic / European sector are displayed in Figure 4. Prior to calculating the EOFs, the time mean air pressure distributions have been subtracted so that the analysis deals with "anomalies", i.e., deviations form the long term mean. Their time coefficients have been used as coordinates in Figures 2 and 3. The first EOF represents 41% of the pressure variance, whereas the second accounts for 19%. When both EOFs have negative coefficient then the first EOF indicates negative pressure over Orense in NW Spain and the second EOF a cyclon flow. Both monthly mean features are favorable for the formation of rainy weather.

5. References

Biau, G., E. Zorita, H. von Storch and H. Wackernagel, 1999: Estimation of precipitation by kriging in EOF space. - J. Climate (in press)

Faucher, M., W. Burrows and L. Pandolfo, 1998: Empirical-Statistical reconstruction of surface marine winds along the western coast of Canada. Clim Res. (in review)

Honerkamp, J., 1994: Stochastic dynamical systems: concepts, numerical methods, data. VCH Publishers, ISBN 3-527-89563-9, 535 pp

Jones, R.H., 1985: Time series analysis – time domain. In: Murphy and R. W. Katz (Eds.) : Probability, Statistics, and Decision. Making in the Atmospheric Sciences. Westview Press, Boulder and London; ISBN 0-86531-152-8, 223-260

Preisendorfer, R.W., 1988: Principal Component Analysis in Meteorology and Oceanography. Elsevier, Amsterdam, 426 pp.

Robinson, A.R., P.F.J. Lermusiaux and N. Q. Sloan III, 1998: Data assimilation. In. K.H. Brink, A.R. Robinson (eds): The Global Coastal Ocean. Processes and Methods. The Sea Vol. 10. John Wiley & Sons, 541-593

Stoyan, D., H. Stoyan and U. Jansen, 1997: Umweltstatistik: Statistische Verarbeitung und Analyse von Umweltdaten. Teubner Stuttgart, Leipzig, ISBN 3-8154-3526-9, 348 pp.

von Storch, H., 1997: Conditional statistical models: A discourse about the local scale in climate modelling. In P. Müller and D. Henderson (Eds): "Monte Carlo Simulations in Oceanography" Proceedings 'Aha Huliko'a Hawaiian Winter Wor University of Hawaii at Manoa, January 14-17, 1997, 59-58

von Storch, H., and F.W. Zwiers, 1999: Statistical Analysis in Climate Research, Cambridge University Press ISBN 0 521 45071 3 hardback, 528 pp (in press)

Wackernagel, H., 1995:Multivariate Geostatistics, Springer Verlag, 270 pp ISBN 3-540-60127-9

Zorita, E., J. P. Hughes, D. P. Lettenmaier, and H. von Storch, 1995: Stochastic characterization of regional circulation patterns for climate model diagnosis and estimation of local precipitation. - J. Climate 8,1023-1042

Zorita, E. and H. von Storch, 1999: The analog method - a simple statistical downscaling technique: comparison with more complicated methods. - J. Climate (in press)