Techniques for Supplementing Solar Radiation Network Data

A Report of Task 9: Solar Radiation and Pyranometry Studies

Volume 2

September 1992

INTERNATIONAL ENERGY AGENCY
Solar Heating & Cooling Programme

THE INTERNATIONAL ENERGY AGENCY
SOLAR HEATING AND COOLING PROGRAMME

International Energy Agency

The International Energy Agency, headquartered in Paris, was formed in November 1974 as an autonomous body within the framework of the Organination for Economic Cooperation and Development to establish cooperation in the area of energy policy. Twenty-one countries are presently members, with the Commission of tbe European Communities participating under a special arrangement.

Collaboration in the research, development and demonstration of new energy technologies to help reduce dependence on oil and to increase long-term energy security has been an important part of tbe Agency's programme. The IEA R&D activities are headed by the Committee on Research and Development (CRD) which is supported by a small Secretariat staff. In addition, four Working Parties (in Conservation, Fossil Fuels, Renewable Energy and Fusion) are charged with monitoring the various collaborative energy Agreements, identifying new areas for cooperation and advising the CRD on policy matters.

Solar Heating and Cooling Programme

One of the first collaborative R&D agreements was the IEA Solar Heating and Cooling Programme which was initiated in 1977 to conduct joint projects in active and passive solar technologies, primarily for building applications. The eighteen members of tbe Programme are:

Australia	Germany	Norway
Austria	Finland	Spain
Belgium	Italy	Sweden
Canada	Japan	Switzerland
Denmark	The Netherlands	United Kingdom
European Community	New Zealand	United States
France (observer)		

A total of eigbteen projects or "Tasks" have been undertaken since the beginning of the Programme. The overall programme is managed by an Executive Committee composed of one representative from each of the member countries, while the leadership and mangement of the individual Tasks is the responsibility of Operating Agents. These Tasks and their respective Operating Agents are:

*Task 1: Investigation of the Performance of Solar Heating and Cooling Systems - Denmark
*Task 2: Coordination of Research and Development on Solar Heating and Cooling - Japan
*Task 3: Performance Testing of Solar Collectors - United Kingdom
*Task 4: Development of an Insulation Handbook and Instrument Package - United States
*Task 5: Use of Existing Meteorological Information for Solar Energy Application - Sweden
*Task 6: Solar Heating, Cooling, and Hot Water Systems Using Evacuated Collectors - United States
*Task 7: Central Solar Heating Plants with Seasonal Storage - Sweden
*Task 8: Passive and Hybrid Solar Low Energy Buildings - United States
*Task 9: Solar Radiation and Pyranometry Studies - Federal Republic of Germany
*Task 10: Material Research and Testing - Japan
*Task 11: Passive and Hybrid Solar Commercial Buildings - Switzerland
Task 12: Building Energy Analysis and Design Tools for Solar Applications - United States
Task 13: Advanced Solar Low Energy Buildings - Norway
Task 14: Advanced Active Solar Systems - Canada
Task 15: Advanced Central Solar Heating Plants (In Planning Stage)
Task 16: Photovoltaics in Buildings - Germany
Task 17: Measuring and Modeling Spectral Radiation - Germany
Task 18: Advanced Glazing Materials - United Kingdom

*Completed Task

REPORT No. IEA-SHCP-9D-1

TECHNIQUES FOR SUPPLEMENTING SOLAR RADIATION NETWORK DATA

VOLUME 2: THEORY

Antoine Zelenka
Swiss Meteorological Institute, Krähbühlstrasse 58, CH - 8044 Zürich, Switzerland

Gerhard Czeplak
Deutscher Wetterdienst, Meteorologisches Observatorium Hamburg
Frahmredder 95, D - 22393 Hamburg 65, Germany

Vito D'Agostino
SASIAM, Tecnopolis CSATA Novus Ortus, P.O. Box 775, I - 70010 Valenzano / Bari, Italy

Weine Josefsson
Swedish Meteorological and Hydrological Institute, S - 60176 Norrköping, Sweden

Eugene Maxwell
National Renewable Energy Laboratory, 1617 Cole Boulevard, Gloden, CO 80401-3393, U.S.A.

Richard Perez
Atmospheric Siences Research Center, SUNY at Albany, 100 Fuller Road, Albany, NY 12205, U.S.A.

and (Chapter 11)

Maurizio Noia*, Corrado Ratto, and Roberto Festa
Università di Genova, Dipartimento di Física, Via Dodecaneso 33, I - 16146 Genova, Italy
*Prsent Affiliation: AGIP, S.p.a., Drilling Technologies, R&D Dept., Milano, Italy

September 1992

This report documents work performed within the IEA Solar Heating and Cooling Programme
Task 9: Solar Radiation and Pyranometry Sudies
Subtask D: Techniques for Supplementing Solar Radiation Network Data

Additional copies may be ordered from:

ENET
Swiss Federal Office of Energy
CH - 3003 Bern
Switzerland Cost: 60 CHF (3 Volumes)

Distribution category: Unrestricted

CONTENTS

VOLUME 1: REPORT

VOLUME 3: RESULTS

CHAPTER 10

10. Theory of spatial interpolation

Review prepared by: ***Vito D'Agostino*** *and Gerhard Czeplak (Sections 10.3.3 to 10.4.5)*
Tecnopolis CSATA Novus Ortus and DWD Meteorologisches Observatorium Hamburg respectively

Under sponsorship from: *Italian National Research Committee and Bundesministerium für Forschung und Technologie (BMFT) respectively*

The procedure of estimating the values of physical properties at unsampled sites within areas covered by existing observation networks is called interpolation, while outside the areas it is called extrapolation (note that in Chap. 3 "extrapolation" is used with a different meaning). The value of a variable between data points can be interpolated only by *fitting some plausible model of variation to the values at the data points* and then calculating the value at the requested location. The spatial distribution of a variable is often arranged in an array whose elements (or pixels, in analogy with satellite images) have the following characteristics:

- ground resolution in square meters, i.e., the pixel or element area;
- geographical position (e.g., latitude, longitude, Gauss-Boaga, UTM system, etc.)

In this case the estimation process is also called gridding.

The problem to be solved is one of choosing a model suitable for solar radiation data, taking into account the sensor's operational characteristics (ground- or satellite-based). In addition to the approaches, we discuss several aspects which influence the quality of the results:

- *spatial representativeness of the network stations.* The spatial distribution of network stations is usually arranged to meet specific needs, within limits imposed by economical constraints. For example, an agronomical solar radiation network will be organized according to the growth conditions required for crops and this cannot be optimal for solar energy resource assessment. Thus, the existing network and the information one tries to extract from it are not always compatible.

- *temporal representativeness of the time series for each station.* Several relationships between the fluctuations of solar radiation and the temporal resolution of the measurements should be considered. A high temporal resolution can more accurately determine variations produced by fast drifting clouds. Therefore, it is important to establish the maximum integration time that can be used without loss of information relevant to the application [10.1].

Principal component transformations (see Section 10.4.2) generate new variables which are pairwise uncorrelated. The principal components (variables) are linear combinations of the observed variables, chosen such that they are orthogonal. The first principal component is that linear combination which has the maximum variance among all linear functions derivable from the given variables. This condition leads to a unique set of coefficients. The second principal component is the linear combination having the maximum variance among all linear functions (of the given variables) that are orthogonal to the first principal component, and so on. If the measurements present a consistent variability, then we can identify the variable which is a significant contributor to variability between the individuals;

- *operative conditions of the stations.* Solar radiation measurements by pyranometers require homogeneity with respect to the operative conditions and with respect to the local characteristics of the stations. Standard quality control techniques (e.g., tracing of systematic behavior in time series) are described in [10.2].

Interpolation method performances are often evaluated with respect to implementation degree, user friendliness and execution time of the code. These are important factors, but there is much subjectivity in their weighting. Even when the weighting can be objective (e.g. execution time), then it refers more to the properties of a compiler and hardware rather than to the algorithm itself. Therefore, it is more advisable to select *accuracy, robustness* (insensitiveness to minor changes in parameters) and *smoothness* of the different methods as comparison criteria.

The finite areal extension of pixels forcibly introduces a spatial error due to the assimilation of the point to the pixel.

In order to test the accuracy of an interpolating surface, it is necessary to have a data sample on the real surface in order to use it for, e.g., the jacknife validation technique. Note at this stage that the report exclusively deals with interpolation, i.e., with surfaces which match the sustaining data (within measurement errors) at the measurement sites. So-called trend surfaces (low order polynomial surfaces which are fit by minimizing the sums of squared residuals) are not considered, because they can be of only marginal interest in the context of this report.

Most interpolation schemes require numerical input from the user as well as a sensitivity analysis. A reliable scheme must respond to small changes in the inputs with noticeable changes in the results.

The smoothness measure of a surface Z(x,y) can be defined according to the theory of splines as:

$$L_1(Z) = \int_A [Z'(x,y)]^2 dA \qquad \text{(first order)}$$

$$L_2(Z) = \int_A [Z''(x,y)]^2 dA \qquad \text{(second order)}$$

Essentially, the first order condition L_1 measures the flatness, and the second order condition L_2 the smoothness a surface. Numerical approximations of L_1 are calculated by estimating the slope of each surface element dA with subsequent averaging of the squared values over the surface A. The average of the squared differences of the elementary L_1 yields an approximation of L_2.

10.1 Probabilistic and deterministic approaches

Given the known set of multidimensional coordinates $x_i = (u_i, v_i, w_i, ...)$, $i=1,...,n$, where the first two elements u_i, v_i represent the geographical position, and the sampled value $z_i(x_i)$ of the solar radiation field Z(x), the aim of interpolation consists in computing any z_o value at any generic position x_o in the field.

It is possible to classify the interpolation methods as *local* and *global* with respect to the data set management, and as *deterministic* and *stochastic* with respect to the mathematical hypothesis. The first named classification separates the methods that utilize the whole data set in the estimation process from those which utilize only a data subset.

In the estimation process, the local approach is more flexible because it guarantees the management of a data set of limited extension. Methods relying on the local approach consider only the data points in a neighbourhood of x_o to estimate z_o. The neighbourhood may be composed of a pre-fixed number of points: three points at a time define a triangulation. Smoothness of the interpolating surface at the crossing from one sub-domain to another is enforced by imposed conditions on derivatives. These conditions are external to the estimation process in the sense that

they are given *a priori* without any relation to the behavior of the data. Surface regularity is enforced when large triangles result from an unfavorable distribution of the data.

Deterministic or stochastic classification of the approach depends upon the hypothesis applied to the data model. The deterministic approach utilizes given assumptions on the spatial variability of the data while the probabilistic approach allows the use of different hypotheses about this variability. Moreover, the latter hypotheses must be compatible with the distribution of the data.

Examples of deterministic methods of local type are the Shepard and the Akima interpolations. Shepard's method [10.3 and 10.4], also called gravity interpolation, is based on the weighted average of those z_i which fall in a specified neighbourhood of x_0. The weights are evaluated as a function of the inverse squared distance between x_0 and x_i (see Section 10.2.1). Triangulation is used in Akima's method [10.5], where quintic polynomial surfaces are fit through the data points on the triangles' edges. Constraints on derivatives on the triangles' sides are added to ensure surface continuity and derivability (see Section 10.2.2).

The probabilistic methods are all rooted in the domain of geostatistics. Geostatistics is developed by using the structural information contained in the spatial data: in this sense it is more flexible, powerful and general than classical interpolation techniques, which do not take advantage of the information about spatial variability embedded in the actual data. Geostatistics makes use of stochastic processes for obtaining the description of the data's structure. Particularly, the modelling step can be stated as in Table I.

Table I. Relationship between the qualitative spatial behaviour of a variable and the mathematical hypothesis.

Variable's behaviour	Mathematical hypothesis	Tools
Regular	Weak stationarity	Covariance
Less regular	Intrinsic	Variogram
Irregular	Intrinsic of order-k	Generalized covariance

For example, it is demonstrated that the general bivariate spline interpolation is a particular generalized covariance function model (Dubrule [10.6]). Also within the probabilistic approach, Gandin's optimal interpolation utilizes the weak stationarity hypothesis (Delfiner [10.7]).

The generality of geostatistics follows from the manifold of the available models. The cost of this generality is paid in terms of variogram or generalized covariance determination during the structural analysis step, which must precede the interpolation itself. Selection of the appropriate geostatistical interpolation technique depends upon the trend (or drift, i.e., the type of global variation) in the solar radiation field. Table II describes the main relationships. A further link is that Gandin's optimal interpolation is similar to ordinary kriging in the weak stationary hypothesis.

Table II. Relationship between the trend of the radiation field and the appropriate geostatistical interpolation technique.

Type of trend model	Kriging technique
Constant	Ordinary
Linear combination of basic functions	Universal

The complexity of the stochastic approach calls for adequate EDP technologies which are not necessarily required by the deterministic approach. Computer graphics is needed for the structural analysis step. A menu-driven software is strongly recommended. Parallel processors should be available for time consuming applications. Ultimately one should expect support by expert systems.

Most important is the fact that geostatistics provide an estimation variance which is most useful in assessing the uncertainty of the interpolated data. This variance of estimation is written as a function of the data values and of their position in the field. Formal introduction of additional positions therefore allows determination of the optimal location for new stations [10.13].

10.2 Description of the deterministic approaches

10.2.1 Weighted averages

This approach is classified as deterministic of local type. It consists in computing the weighted average

$$z_o = \sum_{i=1}^{n} w_i z_i \qquad (10.1)$$

of those measured values z_i that are located in the neighborhood of $x_o = (u_o, v_o)$. Shepard [10.3 and 10.4] suggests weighting functions as

$$w_i = [(1-D_i)^2/D_i^2] \Big/ \sum_{k=1}^{n} w_k \quad \text{for} \quad D_i < 1 \quad \text{and} \quad w_i = 0 \quad \text{for} \quad D_i > 1 \qquad (10.2)$$

with $D_i = d_i/R$, where d_i is the distance between x_o and x_i and R is the radius of the sampling circle. The choice of R depends upon the density of the data points and upon the characteristics of the topographic relief. The sampling circle should contain at least four to five sample points. Unfavorable distribution of the x_i can call for an increase of the neighborhood's size for obtaining a sufficient number of z_i in some locations x_o. This, in turn, may result in an overall increase of processing time as well as a loss of accuracy. Even when using equation (10.2) which guarantees a smooth decrease of the weights towards the periphery of the sampling circle, the resulting surface can present edges and steep slopes.

In regions without marked orography, d_i is set equal to the euclidean horizontal distance. Otherwise, Zelenka and Lazic [10.8] suggest the use a coefficient for weighting the squared elevation increments. The latter procedure is a surrogate to statistical interpolation methods, where the weights are based on (supposedly known) covariances between pairs of stations.

10.2.2 Akima interpolation

Akima's method consists in determining a bivariate fifth-degree polynomial in x and y over each of the triangular cells which can be built from the locations (x_i,y_i) of the data. Construction of a univocal triangulation in the reference plane is achieved by applying the "min-max" criterion of Lawson (e.g., [10.1]). Algorithms to produce the triangulation are presetented by Akima [10.5]: first connect the closest pair of points, then add a point at a time in ascending order of the distance from the midpoint of the closest pair. Each time a new point is added, a triangle is constructed and it is guaranteed that the new point lies outside the polygon constructed with the old points.

Interpolation in a triangle relies on the following assumptions and facts:

i) $z(x,y)$ is a quintic polynomial surface, viz.:

$$z(x,y) = \sum_{j=0}^{5} \sum_{k=0}^{5-j} q_{jk} x^j y^k \qquad (10.3)$$

where 21 coefficients are to be determined.

i i) The values of the function, as well as its first and second partial derivatives, can be evaluated at each vertex of the triangle. This gives three times six conditions.

i i i) The partial derivatives of the function differentiated in the direction normal to each side of the triangle are polynomials of degree three (at most) of the distance from the side. As, purely mathematically, these would be polynomials of degree four, this assumption adds one condition per side.

Both (i) and (ii) produce 21 equations for the q_{jk} and they also ensure the smoothness of the interpolated values. The smoothness of the resulting surface along the sides of the triangle can be proven by considering another cartesian coordinate system. The transformation between the x,y and the new system is linear, so that the partial derivatives are uniquely determined at each vertex as linear combinations of the former derivatives.

Besides the "fluttering" of this surface on the convex hull of the data points (see Sect. 3.2 in Volume 1), the main drawback of Akima's method is that it has no flexibility at all. The surface is unique for a given distribution of z. It appears advantageous only if applied in the local mode: then it is possible to modify the surface interactively while varying the size of the local data.

10.3 Description of the probabilistic approaches

10.3.1 Gandin's optimum interpolation

Gandin explained the method in terms of probabilistic hypotheses and, thus, it represents an excellent introduction to the geostatistical approach. It considers the (solar radiation) field as a random function $Z(x)$ sampled at $x_i = (u_i,v_i)$ locations where z_i is a value of the random variable $Z(x_i)$ for $i=1,...,n$.

Gandin developed the theory for the dispersion values $Y_i = z_i - E[Z_i]$ ($E[Z]$ is the expected value of Z; we recall that the E of a discrete random variable is its mean). Let z_o be the unknown field value at x_o:

$$z_o = E[Z_o] + Y_o \tag{10.4}$$

Estimation of the Y_o dispersion, Y_o^*, is performed by linear combination of the dispersion values Y_i (assuming $E[Z(x)]=0$):

$$Y_o^* = \sum_{i=1}^{n} \beta_i Y_i \tag{10.5}$$

The weights β_i are evaluated by minimizing the average error

$$\delta_o^2 = E[(Y_o - Y_o^*)^2 \tag{10.6}$$

or

$$\delta_o^2 = E[\, Y_o^2 - 2\sum_{i=1}^{n} \beta_i Y_o Y_i + \sum_{i=1}^{n}\sum_{j=1}^{n} \beta_i \beta_j Y_i Y_j \,] \tag{10.7}$$

which is equal to

$$\delta_0^2 = \sigma_{00} - 2\sum_{i=1}^{n} \beta_i \sigma_{i0} + \sum_{i=1}^{n}\sum_{j=1}^{n} \beta_i \beta_j \sigma_{ij} \tag{10.8}$$

where $\sigma_{ij} = E[Y_i Y_j]$ is the covariance between Z_i and Z_j.

Partial derivatives of (10.8) with respect to β_i are set to zero to give the minimum δ_o^2. This yields the following system of n equations

$$\sum_{j=1}^{n} \beta_j \sigma_{ij} = \sigma_{io} \qquad \text{for} \qquad i=1,...,n \tag{10.9}$$

which has a unique solution. The coefficients σ_{ij} of the system are evaluated experimentally from the data, but σ_{io} is unknown. Thus, it is necessary to fit a positive definite function of distance through the σ_{ij} and to use this function to get σ_{io}.

When the β_i verify (10.9), then δ_o^2 is minimum and its value becomes

$$\delta_o^2 = \sigma_{oo} - \sum_{i=1}^{n} \beta_i \sigma_{io} \tag{10.10}$$

When $E[Y_o - Y_o^*] = 0$ (unbiased condition), then the Y_o^* estimator (10.5) is exact and δ_o^2 represents the variance of the estimation error on Y_o.

When E[Z(x)] is different from zero, Gandin assumes that it is constant (E[Z(x)]=C) in the neighborhood of x_o. Then

$$Z_o^* = \sum_{i=1}^{n} \beta_i Z_i \tag{10.11}$$

and

$$\sum_{i=1}^{n} \beta_i Z_i = \sum_{i=1}^{n} \beta_i \{ E[Z_i] + Y_i \} = C \sum_{i=1}^{n} \beta_i + \sum_{i=1}^{n} \beta_i Y_i \tag{10.12}$$

from which the estimation error becomes

$$Z_o - Z_o^* = C \left(1 - \sum_{i=1}^{n} \beta_i \right) + \left(Y_o - \sum_{i=1}^{n} \beta_i Y_i \right) \tag{10.13}$$

This error doesn't depend upon C if

$$\sum_{i=1}^{n} \beta_i = 1 \tag{10.14}$$

is fulfilled.

For minimizing (10.13) it is necessary to introduce the classical Lagrange factor μ, which leads to the (n+1) times (n+1) system of linear equations

$$\sum_{j=1}^{n} \beta_j \sigma_{ij} = \sigma_{io} + \mu \quad \text{for} \quad i=1,...,n; \qquad \sum_{i=1}^{n} \beta_i = 1 \tag{10.15}$$

and, as above, when the β_j satisfy (10.15), the minimum variance of estimation becomes

$$\delta_o^2 = \sigma_{oo} - \sum_{i=1}^{n} \beta_i \sigma_{io} + \mu \tag{10.16}$$

10.3.2 Geostatistics

Geostatistics can be defined as a set of statistical tools applicable to the study of phenomena varying in space and time. The basic idea is to use both the spatial continuity of the data, which is a reflection of the large scale behavior of the underlying process, and the erratic variations at each point in the field. Modelling by means of the concept of random functions and by means of hypotheses on their probability distributions permits the description of both the structured and the erratic effects of the field, and it further allows the analyst to put physical interpretations into the model.

The probability distributions are inferred from the data by means of structural analysis which is, in turn, based on variogram modelling. Thus, the analyst is forced to make decisions that are consistent with the data. The behavior of a theoretical variogram is as follows: the closer two sample points are to each other, the less the field (in our case the insolation) should vary. In practice, many factors can modify this behavior. Sometimes it is possible to decompose the spatial structure along specific directions (identified by an angle with respect to the longitudinal axis, plus or minus a tolerance angle) along which the expected structure can be revealed. Then the increment of the variable is computed for each direction and for each pair of sample data. Use of a distance interval (LAG) allows computation of means of the squared increments within their respective LAGs. A statistically meaningful number of pairs is needed in each LAG, but especially for the first LAG, i.e., for the one closest to zero interdistance h. The graph in which the mean increments are plotted as a function of the LAGs represents the experimental variogram in that direction. A model consisting of positive definite forms must then be fit to the experimental variogram. Such a model is characterized by a SILL (maximum sample variance) and by a RANGE (maximum distance h within which pairs of points interact).

Anisotropic behavior of the variable is indicated when more than one variogram model has been identified for the different directions. The anisotropy should be removed when possible. In the most favorable situation, one model is indicated with different SILLs and (or) RANGEs. The structural analysis is fully employed in the subsequent, interpolation step.

However, prior to interpolation, cross validation with of all the selected variogram models ought be performed, so to select the one which is the most consistent with the available data. Two tests are used to make the selection: the first measures the absence of bias in the estimate (that is, neither under- nor over-estimation), the second makes a comparison of the drifts between pairs with known values and the drifts of the variances of estimation.

Interpolation based on the selected variogram model reconstructs the value of the given variable in each point of the field. The general technique can be subdivided into various kriging techniques which work as linear estimators of the random function. They are free of bias and have minimum variance.

10.3.2.1 Modelling

We consider the solar radiation field Z(x) varying in the area A: { Z(x), $x \in$ A}. The probabilistic approach consists of randomizing this set and of interpreting it as a realization of a set of dependent random variables Z(x) for $x \in$ A. A model able to describe such an interpretation is called a *random function* (RF) and the associated variables are called *regionalized variables*. The characterization of the random function Z(x) is determined by means of its distribution laws (i.e., by its moments), whose identification requires some regularity hypotheses in order to reduce the number of required parameters.

These hypotheses can be written in increasing order of generality for Z(x):

a) Stationarity of order two: which requires the existence of the first and second moments:

$$E[Z(x)] = m(x)$$

$$E[Z(x)Z(x+h)] = \sigma(x,x+h) + m(x)m(x+h)$$

with $\sigma(x,x+h)$ being the covariance between Z(x) and Z(x+h) and m(x) being the mean of Z(x). When the RF Z(x) exhibits a constant mean over the whole field and when the covariance between Z(x) and Z(x+h) depends only on h, then we can say that Z(x) has stationarity of order two:

$$E[Z(x)] = m$$

$$\sigma(x,x+h) = \sigma(h)$$

b) Intrinsic hypothesis: for phenomena without limited dispersion, thus without covariance, it is necessary to consider the increments of the RF rather than the RF itself,

$$E[Z(x) - Z(x+h)] = m(x) - m(x+h)$$

$$Var[Z(x) - Z(x+h)] = 2\,\gamma(x,x+h)$$

with Var used to denote the variance. The function $\gamma(x,x+h)$ is called a variogram and it plays a central role in geostatistics. The existence of covariance implies the existence of the variogram while the reverse is not true. When the variogram depends only on h we say that Z(x) is intrinsic:

$$(1/2)Var[Z(x) - Z(x+h)] = \gamma(h)$$

c) Intrinsic hypothesis of order-k: for phenomena with conspicuous irregular behavior we can use a generalization of the intrinsic hypothesis concerning the increments of Z(x). Matheron [10.9] describes the theory of intrinsic random functions of order-k (IRF-k) as it proceeds from the intrinsic hypothesis.

10.3.2.2 Structural analysis

The variogram γ of an intrinsic random function is by definition the mean square value of the differences Z(x+h)-Z(x) and it depends only on the h vector in a k-dimensional field. Thus, γ is a function of the modulus |h| of h and of the polar angle α between h and the longitudinal axis:

$$\gamma(h) = \gamma(\,|h|\,,\,\alpha\,)$$

For a fixed α the graph of $\gamma(h)$ reveals the directional features of the structure. The aim of the analyst is to implement a strategy for recognizing a structure in the sample that can be modelled with a theoretical variogram. The expected properties of a variogram - which directly proceed from its definition - are:

- if measurements are taken at arbitrarily close points x and x+h, the mean squared differences of Z(x+h)-Z(x) tend towards zero;
- the closer x+h is to x, the smaller is the squared increment of the variable;
- beyond a certain range of inter-distances the squared increments of the variable stabilize around a value comparable with the variance of the whole sample.

Variogram models have to satisfy certain conditions of mathematical consistency with the squared averages, without which one could possibly come up with negative variances. Thus the choice of the models is among the conditionally positive definite functions that ensure a positive variance. The variogram parameters are essentially three: The RANGE a, the SILL C, and the NUGGET EFFECT N. The current models are:

1) The **power model** (Fig.1)

$$\gamma(h) = N + \Omega\,|h|^{\beta} \qquad \text{with} \qquad 0 < \beta < 2$$

The linear model ($\beta = 1$) is the specific case that simply states that the variance of the increments is proportional to the distance between two locations in the field.

2) The **spherical model** (Fig.2)

$$\gamma(h) = N + C[(3/2)|h|/a - (1/2)|h|^3/a^3] \qquad |h| \le a$$

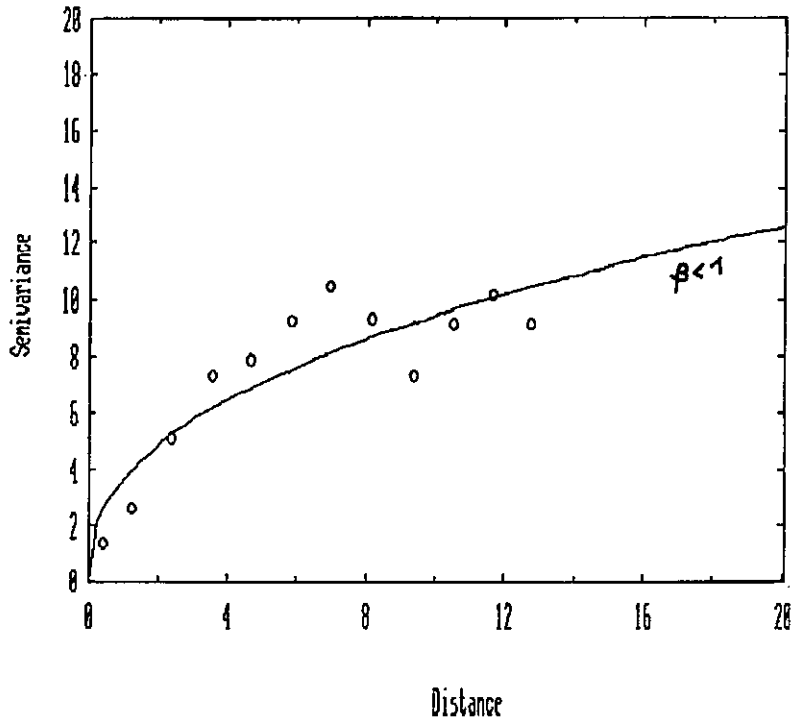

Figure 1: The power variogram model (full line). Scales are in arbitrary units.

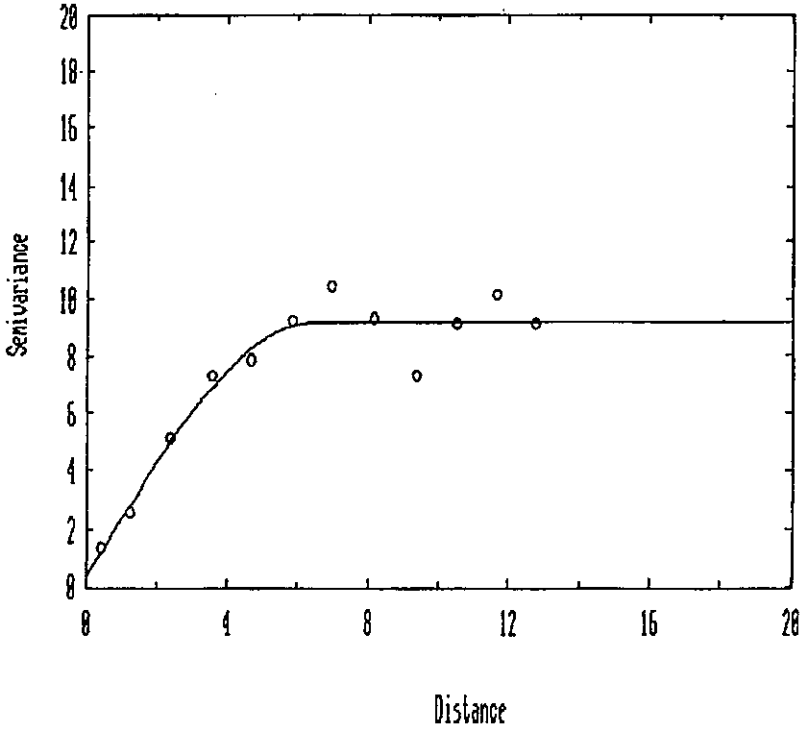

Figure 2: The spherical variogram model (full line). Scales are in arbitrary units.

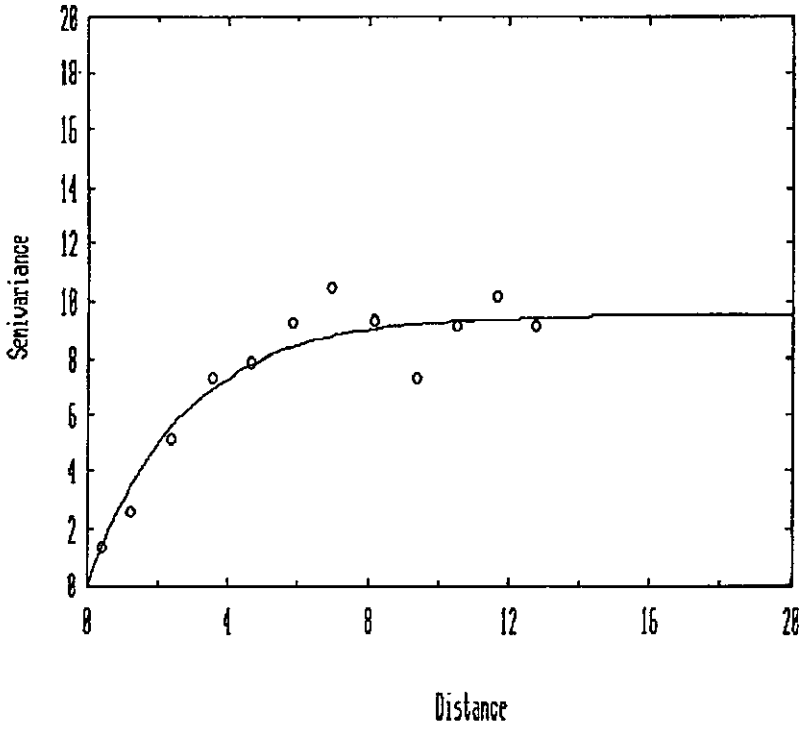

Fig. 3: The exponential variogram model (full line). Scales are in arbitrary units.

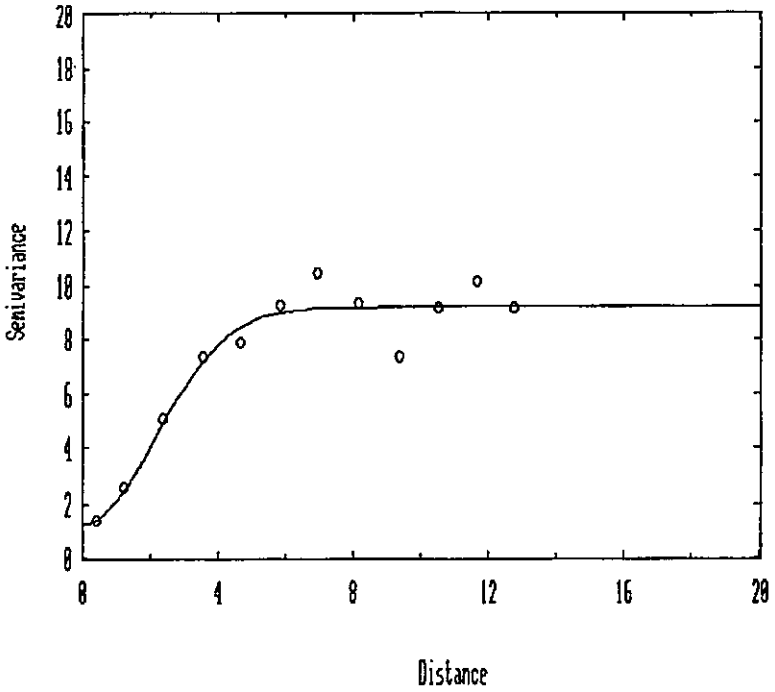

Fig. 4: The Gaussian variogram model (full line). Scales are in arbitrary units.

The tangent at the origin permits an assessment of the difference between the linear and spherical models. The spherical model gives more weight to distant points, i.e., it implies a more continuous behavior of the field.

3) The **exponential model** (Fig.3)

$$\gamma(h) = N + C\ [\ 1 - \exp(-|h|/a)] \qquad |h| \leq a$$

The abscissa where the tangent at the origin interects the SILL C is smaller than in the spherical model (1/3 of a versus 2/3 of a).

4) The **Gaussian model** (Fig.4)

$$\gamma(h) = N + C\ [\ 1 - \exp(-|h|^2/a^2\ ] \qquad |h| \leq a$$

This model is characteristic of an extremely continuous and smoothly differentiable phenomenon.

As presented here, the models have $\gamma(h) = C$ for $|h| > a$, and a nugget effect N equal to zero. The nugget effect is to be understood as a discontinuity of the variogram at the origin. Generally a nugget effect different from zero is due to one of the following causes:

i) to a component of the phenomenon whose range is smaller than the smallest interval for which the experimental variogram can be computed;

i i) to measurement uncertainties;

i i i) to location errors of sample points.

In many cases, physical ancillary information is more useful than short-distance sampling for assigning the right cause, because there is always a minimum inter-distance below which the variogram is simply not known. In fact, an analysis at shorter distances may reveal that the variogram continuously dips to zero and that an apparent discontinuity is merely due to the spacing of the sample points, which is too wide to detect structures at a smaller scale. Alternatively, the first observable variogram point may be aligned with the origin although the variogram has a bend and a nugget effect. Such a behavior is typical of location errors. Obviously there is a risk to be wrong in both cases. Therefore, the best hypothesis about the variogram may proceed from physical information about the variable, if such are available. Often, the variance of measurement errors is known from separate experiments and it may account for a part of the nugget effect. In other cases, we know that the variable is continuous and model the apparent nugget effect by a spherical variogram with a range smaller than the shortest observed lag. Practice, however, shows that computational robustness is gained when highly continuous models are used with a very small nugget effect.

In the proposed models 1 to 4, the different shapes also describe different decay rates of the influence between the field points as a function of their mutual spacing. When the variogram reaches the sill, it means that there is a distance (range) beyond which Z(x) and Z(x+h) are without correlation and the sill itself becomes the variance of the whole sample:

$$\gamma(h) = (1/2)\mathrm{Var}[Z(x+h)-Z(x)] = (1/2)[\mathrm{Var}(Z(x+h))+\mathrm{Var}(Z(x))] = (2\sigma^2)/2 = \sigma^2 \quad (10.17)$$

If the sample variance σ exists for h=0 (and if it is finite), then the following holds:

$$\gamma(h) = C(0) - C(h) \qquad (10.18)$$

where C(h) is the covariance function of Z(x).

The variogram model is fit to an experimental variogram (open circles in Figs. 1 to 4) which is computed from the data by taking

$$\gamma^*(h) = \frac{\sum_{k=1}^{n(h)} [\, Z(x_k + h) - Z(x_k)\,]^2}{2n(h)} \qquad (10.19)$$

where n(h) is the number of squared increments belonging to each distance class (LAG). Figure 5 illustrates the construction of the experimental variogram according to (10.19).

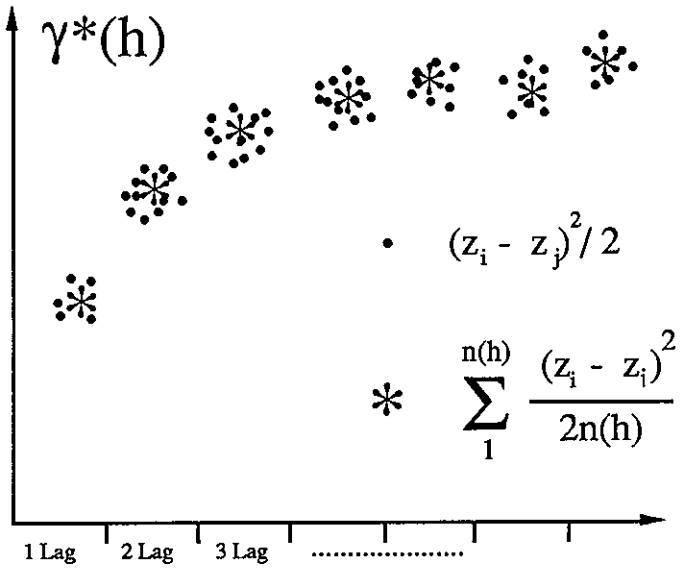

Fig. 5: Construction of the experimental variogram.

In other words, an experimental variogram is the ordered set of values (h, $\gamma^*(h)$) which can be presented as a graph of γ^* against h.

By definition of the average squared increments we have:

$$E[Z(x+h)-Z(x)]^2 = Var[Z(x+h)-Z(x)] + \{E[Z(x+h)-Z(x)]\}^2 \qquad (10.20)$$

and it follows that

$$\gamma^*(h) = (1/2)E[Z(x+h)-Z(x)]^2 = \gamma(h) + (1/2)[m(x+h)-m(x)]^2 \qquad (10.21)$$

Because in each x we have a random variable, it follows that m(x) is the mean of the random function Z(x): m(x) is also called trend or drift of the field. $\gamma^*(h)$ coincides with $\gamma(h)$ if, and only if the increments have zero mean or, equivalently, if m(x) is constant over the field. When a trend exists, $\gamma^*(h)$ is always an upward biased estimator of the variogram.

If there is a trend in the sample the experimental variogram doesn't reach a sill. Consider a linear trend in, for simplicity, a one dimensional field:

$$m(x+h)-m(x) = [a(x+h)+b]-[ax+b] = ah \qquad (10.22)$$

from which it follows that

$$\gamma^*(h) = \gamma(h) + a^2h^2/2 \qquad (10.23)$$

A parabolic term is added to the variogram. It doesn't produce effects for short distances, but for long ones it becomes the dominant term (Fig.6). The effect of a trend other than linear, as well as extension to more than one dimension, is more complex although the same conclusions are reached.

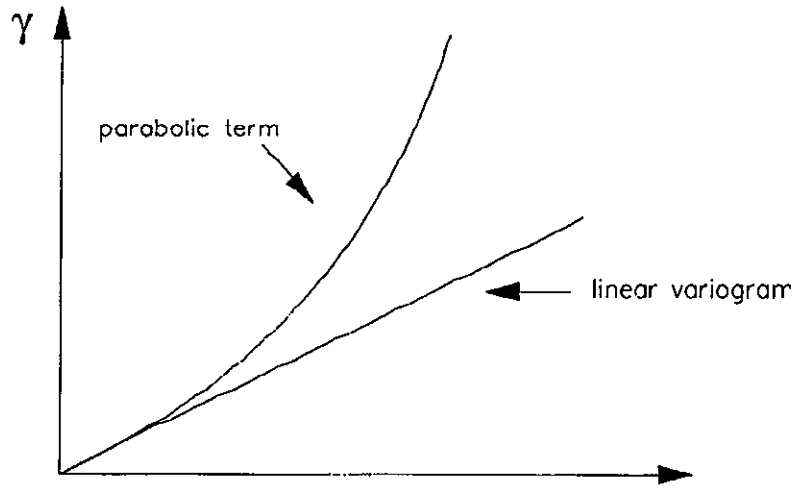

Figure 6: Linear variogram and linear variogram plus parabolic term.

10.3.2.3 Performing the structural analysis

In this section the reader should keep in mind that kriging is a method that needs support by an interactive software package. Several such packages are discussed in Section 10.5.

In a preliminary step, exploratory data analysis should visualize the spatial distribution of the sample and deliver information on:

- the minimum sample distance (which governs the definition of the LAG in the experimental variogram);
- the mean and the variance of the sample (which governs the choice of the SILL in the experimental variogram);
- the presence or absence of spatial trends (which ought to be revealed by the experimental variogram);
- the normality of the distribution of the population.

The goal of the structural analysis is the detection of the theoretical behavior to be expected from the variogram. The strategy is aimed at modelling the corresponding behavior and embedding it in the estimation process. Thus, the strategy tends to establish some rules which minimize the arbitrariness of the analysis, even though the subjectivity of the process cannot be avoided completely. The rules consist of the following steps:

s1: at first, an isotropic behavior of the phenomenon is assumed: all pairs of points within a tolerance angle $\pm 90^{o}$ with respect to the direction 0^{o} are considered;

s2: a LAG size is chosen slightly larger than the minimum sample distance;

s3: the experimental variogram is computed;

s4: the first three LAGS are analyzed according to the following criterion. The first LAG must contain a statistically meaningful number of points (pairs of stations) - see Fig. 5. If this is not the case, the size of the LAG must be increased and step s3 is repeated. When the first LAG behaves correctly, then the same analysis is perfomed for the second LAG and afterwards for the third one;

s5: the experimental variogram is examined to determine if its behavior can match one of the theoretical models. One expects a growth of the values in the first LAGs and a settlement of the values for the subsequent LAGs. The growth can occur with:

- a parabolic shape which reflects a highly continuous and differentiable field;
- a linear shape which reflects a continuous but less regular field;
- a discontinuity at zero which is termed nugget effect: the field is discontinuous and highly irregular. Here, continuity is to be understood in the mean square sense.

Selection of the slope at the origin and deciding about the inclusion of a nugget effect are usually performed while choosing the variogram model yielding the best fit. The RANGE is assessed visually and the SILL is placed where the variogram stabilizes. This SILL should be roughly equal to the overall variance of the data if the RANGE is smaller than the size of the region being studied. The fitting is done by trial and error, preferably using an interactive visual display. Very often, estimates of the variogram behave poorly and are difficult to interpret. In addition, the same points are used both to estimate the variogram and to run the kriging procedure (interpolation). To minimize the effects of correlation and bias, standard models are fit by eye to the experimental points; least squares fitting is not recommended.

As already stated, the absence of a SILL in the experimental variogram indicates the presence of spatial trends in the sample. When the SILL exists, it should lie at about the value of the variance of the population (Equ. (10.17)). If the distribution of the experimental points obviously contradicts this expectation, it is recommended (as a first step) to increase the size of the LAG and to go back to s3. However, the LAG size later influences the resolution associated to the map of the field (whose establishement is the ultimate goal of the whole procedure!). The mapping scale must be larger than the LAG size of the structural analysis; otherwise difficulties arise with the correct reproduction of extreme values. Therefore, the step back to s3 is not always possible. Also, it may not lead to the expected result. So it is a matter of judgement when to abandon the isotropic assumption and to face one or several of the following conclusions:

- a more refined sampling is necessary (if possible);
- it is necessary to find and express a functional relationship between the given variable and some other variable dependent on it. The missing information can then be supplied by the latter variable;
- perform an anisotropic analysis (if possible);
- the field reflects pure randomness; this is the limiting case of total absence of structure and it is equivalent to the assumption of randomly distributed measurement errors in the deterministic approach.

Among these conclusions, the third one, i.e. *anisotropy*, can be addressed in a straightforward manner. Only squared deviations of pairs which have a definite orientation (within a reasonable tolerance angle), e.g. North-South or East-West, are considered and the study is repeated from step s2 for each direction.

Departures from isotropy are classified into two categories: elliptic or geometrical anisotropy, which may be corrected by an affine coordinates transformation, and stratified anisotropy, which is more complex to deal with.

The RANGE now, of course, depends on the direction and there can be several RANGEs reflecting different scales of structures. In Fig. 7 the variograms have the same SILL in two directions, but different RANGEs a_1 and a_2. In Fig. 8 the variograms are linear, but they have different slopes.

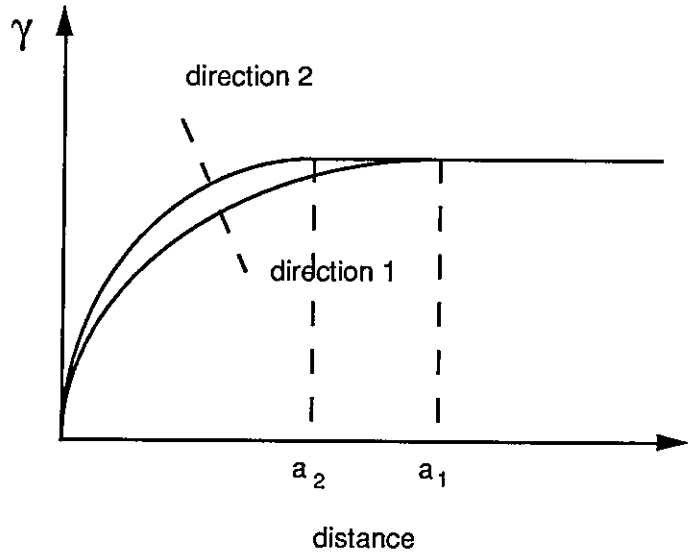

Figure 7: Variograms with same sill but different ranges in two different directions.

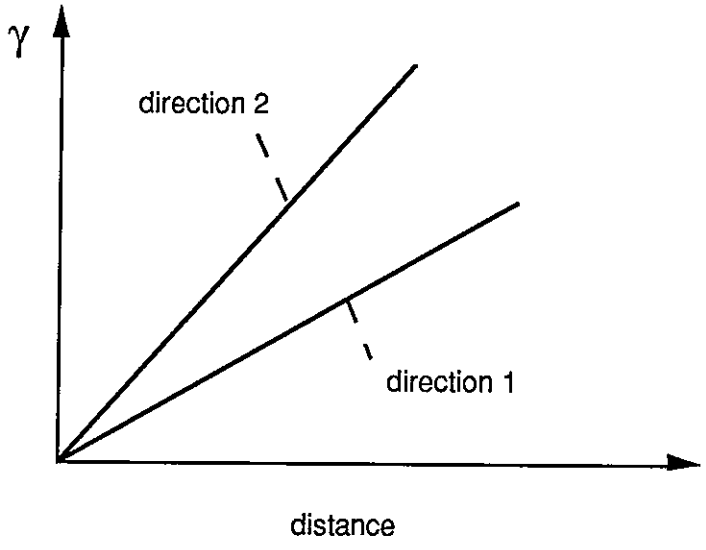

Figure 8: Linear variograms with different slopes in two different directions.

When the anisotropy can be traced to only two main directions that are perpendicular one to the other and when, moreover, these directions are coincident with the 0° and 90° axes, then the ratio

$$k = a_1 / a_2$$

permits the restoration of isotropy. When the main directions do not coincide with the coordinate axes, the formulae get a little more complicated since they involve the angle Φ between the x-axis and the main anisotropy axis Ox' (Fig. 9). The anisotropy corrected variogram $\gamma(h)$ is of the form

$$\gamma(h) = \gamma_1 (\sqrt{} (h'Ah))$$

where the transformation

$$h'Ah = [(x_1 - x_2)\cos\Phi+(y_1 - y_2)\sin\Phi]^2 + k^2 [(y_1 - y_2)\cos\Phi - (x_1 - x_2)\sin\Phi]^2$$

allows the description of the variability between pairs (x_1,y_1) and (x_2,y_2) in the x,y system with a unique formula. This is the elliptic anisotropy.

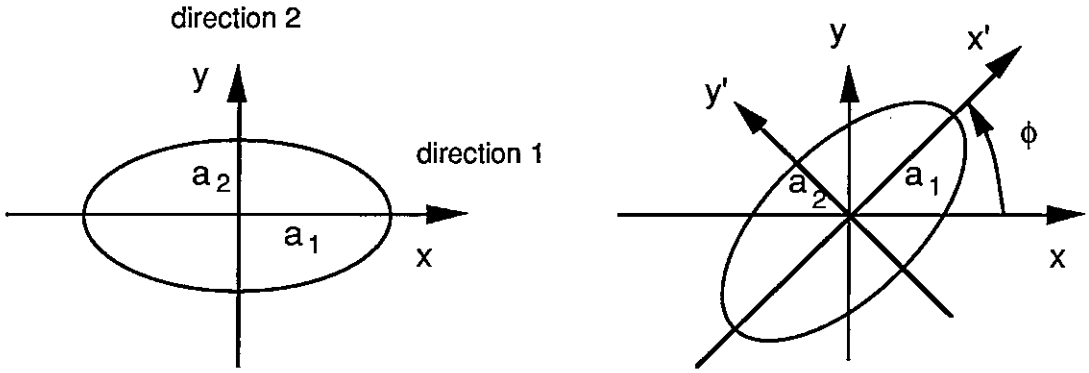

Figure 9: Main structure directions in the elliptic anisotropy.

Other anisotropies may be considered as a linear combination of variograms, one for each direction of anisotropy

$$\gamma(h) = \Sigma \gamma_i(h)$$

For example, in 3-D space the vertical direction often plays a select role: there are more variations across the strata than within them. The variogram is then split into two terms, the one being possibly isotropic, the other depending only on depth.

In general a reasonably good fit can be obtained with the sum of two or, at most, three basic variogram models. Anisotropies are the greatest source of difficulty in structural analysis and it may be difficult to distinguish between drift and anisotropy.

10.3.2.4 Kriging techniques

Current kriging techniques rely on linear estimators of the random function Z(x)

$$Z^*(x) = \Sigma \beta_i Z_i \qquad (10.24)$$

Kriging techniques determine the weights β_i in a stochastic way. (In the deterministic approach - for the formal similarities refer to Section 10.2.1 - the weights are pre-determined in terms of a, not necessarily euclidean, distance measure).

Kriging is the name given to a group of minimum-error-variance algorithms (Best LInear Unbiased Estimators (BLUE)) which produce unique solutions since the matrix of the coefficients depends on positive definite variogram models. The properties of the techniques can be listed as follows:

1) Existence and uniqueness of the solution is ensured;
2) The variance attached to the estimator is smaller than any other;
3) The kriging estimator is unbiased;
4) The kriging estimator is an exact interpolator (in the absence of a nugget effect due to measurement uncertainties, it honors the data points);
5) Kriging takes into account:
   - the structural distance of the information (the closer the sample point to the location being estimated, the greater the weight it should receive),
   - the configuration of the data (clustered, redundant data receive less weight),
   - structural features (continuity) captured with the selected variogram model $\gamma(h)$.

As already stated above (Sections 10.1 and 10.4.1) the understanding of a pattern of spatial dependence between n sample points requires a model. These models are reviewed below.

A) Ordinary kriging

In the absence of a trend, E[Z(x)] is constant and a formalism develops which is similar to Gandin's. However increments are considered instead of residuals, that is, the intrinsic hypothesis is accepted. Then $Var[Z(x+h)-Z(x)] = 2\gamma(h)$, resp. $Var[Z(x_i)-Z(x_j)] = 2\gamma(h_{ij})$.

Again, let Z_o be the unknown field value at x_o. Lack of bias requires

$$E[Z_o - Z_o^*] = 0 \qquad (10.25)$$

from which it follows that

$$\sum_{i=1}^{n} \beta_i = 1 \qquad (10.26)$$

Further, we require the estimator to be "best" in the sense that the estimation variance is minimum. In other words, we wish to determine the weights of Z_o^* such that

$$\sigma^2 = E\{[Z_o - Z_o^*]^2\} \Rightarrow \text{minimum} \qquad (10.27)$$

under the constraint (10.26). It is readily shown that the estimation variance σ^2 is

$$\sigma^2 = 2\sum_{i=1}^{n} \beta_i \gamma_{oi} - \sum_{i=1}^{n}\sum_{j=1}^{n} \beta_i \beta_j \gamma_{ij} \qquad (10.28)$$

so, by Lagrange formalism, it is possible to write the following system

$$Q(\beta_i,\mu) = \sigma^2 + 2\mu(\sum_{i=1}^{n} \beta_i - 1) \Rightarrow \min \qquad \text{and} \qquad \sum_{i=1}^{n} \beta_i = 1$$

Setting the partial derivatives of Q to zero we obtain the OK (Ordinary Kriging) system of n+1 equations for the n+1 unknowns β_i plus μ:

$$\sum_{j=1}^{n} \beta_j \gamma_{ij} + \mu = \gamma_{oi} \quad \text{for } i=1,...,n \qquad \text{and} \qquad \sum_{j=1}^{n} \beta_j = 1 \tag{10.29}$$

While the γ_{ij} can be computed from the data, the γ_{oi} need the variogram model to be evaluated. The weights β_i which satisfy (10.29) yield the variance of estimation

$$\sigma^2_{min} = \sum_{i=1}^{n} \beta_i \gamma_{oi} - \mu \tag{10.30}$$

B) Universal kriging

For modelling a non-constant E[Z(x)] Matheron [10.10] suggests linear combinations of basic functions $f_k(x)$ (usually $k \leq 3$)

$$E[Z(x)] = m(x) = \sum_{k=0}^{l} a_k f_k \tag{10.31}$$

However, this drift cannot be validated within the data range because of the insensitivity of the coefficients of the basic functions. The non-stationary model (drift plus residuals)

$$Z(x) = m(x) + Y(x) \tag{10.32}$$

has the variogram

$$2\gamma(h) = [m(x+h) + m(x)]^2 + 2\gamma_R(h) \tag{10.33}$$

and it is obvious that the whole estimation procedure relies on the knowledge of the variogram of the residuals $\gamma_R(h)$. If one is fortunate enough, the sample will contain some area where, for some direction (and limited values of h), the trend might be negligible. Then $\gamma_R(h) = \gamma(h)$ will follow from the structure analysis of this region. Otherwise a tedious, and perhaps unstable, iteration process must be undertaken: various estimates of (10.31) will be substracted from (10.32) until the next $\gamma(h)$ will no longer markedly differ from the previous one. This last variogram will be assumed as being $\gamma_R(h)$.

The BLU conditions then give the following UK (Universal Kriging) system of linear equations:

$$\sum_{j=1}^{n} \beta_j \gamma_{R,ij} + \sum_{k=0}^{l} \mu_k f_k^j = \gamma_{R,io} \qquad \text{for } i=1,\ldots,n$$

$$\sum_{i=1}^{n} f_k^i = f_k^o \qquad \text{for } k=0,\ldots,l \tag{10.34}$$

where f_k^i represent the values of the f_k at the locations x_i. The order of the UK system is (n+l+1) times (n+l+1) and the μ_k are Lagrange multipliers associated to the m(x) condition (the last l+1 equations of the system). The UK system provides a unique solution when the basic functions f_k are linearly independent of the set of data, that is, when

$$\sum_{k=0}^{l} c_k f_k = 0 \tag{10.35}$$

which implies that $c_k = 0$ for each k=0 to l. The variance of estimation then becomes

$$\sigma^2 = \gamma_{R,oo} - \sum_{i=1}^{n} \beta_i \gamma_{R,io} + \sum_{k=0}^{l} \mu_k f_k^o \tag{10.36}$$

Note that if l=0, then $m(x) = a_o$ while $f_o(x) = 1$ by convention, and the UK system reduces to the OK system (stationary case).

Finally it ought be pointed out that UK, despite its name, is the least "universal" of all the kriging algorithms, because it is the most constrained.

C) IRF-k kriging and Cokriging

Additional kriging techniques are: **IRF-k kriging, cokriging and disjunctive kriging.**

Disjunctive kriging introduces non-linear estimators which are beyond the scope of this report. It is mentioned here because the interested reader will find it referenced in the literature. For the same reason, we note an entire field of geostatistics, namely **simulation**: knowledge about its distribution functions permits the simulation of additional realizations of a RF Z(x) [10.12].

The other techniques metioned above deserve more description as they have been applied in the course of our validation work.

IRF-k (Intrinsic Random Function of order k) kriging has the advantage of automatic recognition of the spatial structure (see Section 3.4 in Volume 1). It is applicable when the intrinsic hypothesis of order k holds and it filters out any underlying polynomial trend [10.11]. However, it requires samples of sufficient size for the structure recognition to be operative. The structural tools are termed *generalized covariance* functions K(h), and among these a great deal of attention has been devoted to polynomial forms. It is demonstrated that the spline interpolation is equivalent to IRF-k kriging with the generalized covariance

$$K(h) = |h|^2 \ln(|h|)$$

a fact which confirms the generality of the geostatistical approach.

Cokriging relies on the concept of coregionalization, which simultaneously uses two (or even more) correlated random functions in such a way, that the spatial information contained in each of them is of particular utility to the estimation process.

Journel [10.12] notes the convenience of this approach in the case when one of the random functions of primary interest has been undersampled with respect to an auxiliary RF. Theoretically this is justified, because the cokriging variance is shown to be less than the mere kriging variance, and this means that the estimation process becomes more accurate. Further, the spatial information of the better sampled, auxiliary variable is transferred to the primary one through the *cross-variogram* function. However, this implies a significant number of locations where both variables are sampled. Matheron [10.11] discusses the theoretical fundamentals of the coregionalization. We develop here only some aspects of the equations.

We assume that Z_1 and Z_2 are two variables modelled through the RF concepts. Stationarity of first and second order is postulated for the increments of the functions (intrinsic hypothesis), which allows simplification of the probability distributions of Z_1 and Z_2. In this case, we call variogram (γ_1, resp, γ_2) the expectation of the squared increments of Z_1, resp. Z_2, and cross-variogram γ_{12} (which is equal to γ_{21}) the expectation of the increment products between Z_1 and Z_2.

If Z_1 and Z_2 are physically correlated and if the sample values $\{Z(x_{11}),...Z(x_{1n})\}$ and $\{Z(x_{21}),...,Z(x_{2m})\}$, with $m < n$ exist, then, it is possible to define an estimator of Z_1 as a linear combination of the available sample values provided that $x_{1k} = x_{2k}$ for $k < m$:

$$Z_1^*(x_o) = \sum_{i=1}^{n} \beta_i Z_{1i} + \sum_{j=1}^{m} \alpha_j Z_{2j} \qquad (10.37)$$

where $Z_{1i} = Z_1(x_{1i})$ and $Z_{2j} = Z_2(x_{2j})$.

The weights β_i and α_j are again determined by requiring that Z_1^* be the best unbiased (BU) estimator of Z_1. This technique differs from mere kriging because the estimator also takes the Z_2 values into account. As a consequence, the cokriging (coK)-system is complicated by the indexes which increase according to the number of variables involved in the estimation.

As derived from the "Best" condition (minimum variance), the coK-system contains I rows (for I=1,..,n)

$$\sum_{i=1}^{n} \beta_i \gamma_1(x_{1i} - x_I) + \sum_{j=1}^{m} \alpha_j \gamma_{12}(x_{2j} - x_I) + \mu_1 = \gamma_1(x_0 - x_1)$$

and J rows (for J=1,...,m)

$$\sum_{i=1}^{n} \beta_i \gamma_{21}(x_{1i} - x_J) + \sum_{j=1}^{m} \alpha_j \gamma_2(x_{2j} - x_J) + \mu_2 = \gamma_{21}(x_0 - x_j)$$

(where μ_1 and μ_2 are the usual Lagrange factors), while the "zero bias" condition further adds the equations

$$\sum_{i=1}^{n} \beta_i = 1 \qquad \text{and} \qquad \sum_{j=1}^{m} \alpha_j = 0$$

By solving the coK-system we obtain the estimate (10.37) and the cokriging variance σ_{coK} in the x_o location

$$\sigma^2_{coK} = \mu_1 + \sum_{i=1}^{n} \beta_i \gamma_1(x_{1i} - x_o) + \sum_{j=1}^{m} \alpha_j \gamma_{12}(x_{2j} - x_o) \qquad (10.38)$$

If Z_1 is correlated to Z_2 then σ_{coK} is generally less than the ordinary kriging variance: the additional term involving γ_{12} allows more flexibility in minimizing (10.38).

For an interval around a given distance h, and, if appropriate, for each direction class, an experimental cross-variogram is estimated from the sample data using the equation:

$$\gamma^*_{12}(h) = \frac{\sum_{k=1}^{n(h)} [Z_1(x_k+h) - Z_1(x_k)][Z_2(x_k+h) - Z_2(x_k)]}{2n(h)} \qquad (10.39)$$

with n(h) being the number of pairs having a distance within the designated lag. If $Z_1 = Z_2$ for any x, then $\gamma_{12}{}^*$ becomes the experimental (auto-)variogram $\gamma_1{}^*$ (or $\gamma_2{}^*$) (see Eq. (10.19)).

The experimental cross-variogram $\gamma_{12}{}^*(h)$ follows from those locations where values of both Z_1 and Z_2 are sampled. An authorized model (e.g., linear, spherical, exponential, gaussian, etc.) must be fit to it either visually, or by some other good procedure. This model can be validated by the jacknife technique.

For the (auto-)variograms γ_1 and γ_2, the allowed theoretical models ensure conditions of mathematical consistency without which one would possibly come up with negative variances. Similarly, for a cross-variogram, the Cauchy-Schwarz relation

$$|\gamma_{12}(h)| \leq \sqrt{[\gamma_1(h)\gamma_2(h)]}$$

must be fulfilled for all $h \geq 0$ to ensure a positive cokriging variance in all circumstances.

References

[10.1] C.R. Rao: 1964. "The use and interpretation of principal components analysis in applied research", Indian Jour. of Statistics 2, 17-210.

[10.2] P.F. Abbott: 1986. "Guidelines on the quality control of surface climatological data". World Climate Programme, WCP-85. WMO/TD-No. 111, WMO, Geneva.

[10.3] D.S. Shepard: 1968. "A two-dimensional interpolation function for irregularly spaced data", Proc. 23rd Nat. Conf. Assoc. Computing Machinery, pp. 517-524.

[10.4] D.S. Shepard: 1984. Computer mapping: the SYMAP interpolation algorithm. In G.L. Gaile and C.J. Willmott Eds.: 1984, Spatial Statistics and Models. D. Reidel.

[10.5] H. Akima: 1978. "A method of bivariate interpolation and smooth surface fitting for irregularly distributed data points", ACM Transactions on Mathematical Software, Vol. 14/2, pp. 148-1510.

[10.6] O. Dubrule: 1984. "Comparing splines and kriging", Computers and Geosciences, Vol. 10, No. 2-3, pp. 327-338.

[10.7] P. Delfiner: 1973. "Analyse objective du géopotentiel et du vent géostropique par krigeage universel", Notes Internes de l'Établissement d'Études et de Recherches Météorologiques, Paris.

[10.8] A. Zelenka and D. Lazic:1987, "Supplementing network global irradiance data". Advances in Solar Energy Technology, Proc. 1987 Biennial Congres ISES, Hamburg. Vol. 4, pp. 3861-3865, Pergamon 1988.

[10.9] G. Matheron: 1978. "The intrinsic random function and its application", Adv. Applied Stat., 4.

[10.10] G. Matheron: 1982. "The theory of regionalised variables and its applications", Les Cahiers du Centre de Morphologie Methématique de Fontainebleau, No. 5, 211p., École des Mines de Paris.

[10.11] P. Delfiner *et al.*: 1987. "Basic Introduction to Geostatistics, Part I", Course Manual of Applied Geostatistics for the Petroleum Industry, Centre de Géostatistique EMP, Fontainebleau.

[10.12] A. Journel and C. Huijbregts: 1978. "Mining Geostatistics", Academic Press, London.

[10.13] F. Szidarovsky: 1983. "Optimal observation network in geostatistics and underground hydrology". *Appl. Math. Modelling*, 7, pp. 25-32.

10.3.3 Principal Components Analysis

10.3.3.1 Introduction

The principal components analysis, often called the method of empirical orthogonal functions (EOF), is a pure statistical method to evaluate scalar quantities. It is free from any deterministic assumption and independent from the application in which it might be involved. The EOF method is mainly confined to the evaluation of eigenvalues and eigenvectors of symmetric matrices. Because of the extensive computations required, it cannot be used without computers.

More detailed and more mathematical descriptions have been presented elsewhere by Lorenz [10.3.1], Obukhov [10.3.2], Erdmann & Fechner [10.3.3], Fechner [10.3.4], and Fechner [10.3.5]. In the present case, we have put our emphasis on practical applications of this method by giving a concrete example.

10.3.3.2 Motivation and Initial Data

In our case, the initial data consist of a matrix made up of values of a single variable (global irradiation) taken at different locations over a specified period of time. Because measured data like these are correlated with each other, we had a surplus of information. Uncorrelated or independent data are orthogonal in the mathematical sense. Therefore, a transformation must be performed to convert the measured data with its redundant information into uncorrelated orthogonal variables. This greately reduces the amount of data and leads to independent variables which can explain the measured data.

To say it in other words: a certain physical variable is examined which is influenced by other variables which are not known. The initial matrix of measured data, which are correlated with each other, contains those influencing variables implicitly, so that the redundance of information makes it impossible to evaluate these influencing variables independently. The transformation of the measured data matrix into a so-called correlation matrix (which is also the covariance matrix because of the normalization) reduces the redundant data to those variables which are independent and orthogonal.

The method presented has its origin in works by Stidd [10.3.6] and Granger [10.3.7].

10.3.3.3 The Covariance Matrix

The measured data matrix A_{ij} consists of i rows (stations) and j columns (dates). The covariance matrix K_{ij} is calculated (see [10.3.1] and [10.3.3]) as follows

$$K_{ij} = \frac{1}{N} \sum_{n=1}^{N} f_{ni} f_{nj} \qquad (1)$$

with N = number of stations where the columns (dates) are multiplied and then the rows are summed. The rectangular measured data matrix A_{ij} becomes a square covariance matrix K_{ij} of j x j elements where the diagonal elements K_{jj} are proportional to the spatial variance.

10.3.3.4 Calculating the EOFs

From this covariance matrix the eigenvalues and eigenvectors (orthogonal functions) are calculated, by either using Jacobi's method [10.3.8] or by using the Fortran subroutine FO2ABF from the computer subroutine library NAG-Lib [10.3.9] in one's own computer program.

The output of the eigenvalues and the respective eigenvectors are arranged in declining order of magnitude. The eigenvectors make up an eigenvector matrix E.

The eigenvalues are summed. The contribution of one eigenvalue on the total sum expressed as percentage is identical to the contibution of the respective eigenvector on the total variance. In the present case the first four eigenvectors explain 99 % of the total variance. This shows that arranging eigenvalues and eigenvectors in a declining order of magnitude places them them in their order of importance.

Lorenz [10.3.1] proved that empirical orthogonal functions are the best among all linear functions that can be derived from measured data, because they explain the largest part of the total variance.

This is the reason one can limit the following calculations to the most important eigenvalues without losing essential information.

10.3.3.5 The Multipliers

Achieving a country-wide mapping of a meteorological quantity requires a matrix M of multipliers. The idea of multipliers stems from Stidd [10.3.6] and is defined as

$$M = E' \; A \tag{2}$$

were E' is the transpose of E. As the first 4 columns of E explain 99 % of the total variance, in equation (2) only the first 4 rows of E' are used making M a (4 x N)-matrix. For each of the N stations it contains the first 4 (most important) multipliers.

These multipliers are only exact at the station locations where the data were taken. To get multipliers for locations between the stations, each of these 4 multipliers has to be interpolated separately over a digital terrain map. At each grid-point of the map, a surch radius R of a certain length is described. The values of all multipliers of stations with a smaller distance to the grid point than R are weighted by the inverse square distance, and an average value is calculated at that grid point. This procedure yields one single map for each multiplier in M (4 maps in the present case).

10.3.3.6 Mapping

To calculate a quantity (global irradiation in our case) from a matrix M, equation (2) is multiplied by E on the left-hand side

$$E \; E' \; A = E \; M \tag{3}$$

From the orthogonality condition, the transpose of an orthogonal matrix is identical to its inverse ($E' = E^{-1}$), and equation (3) becomes

$$\begin{aligned} E \; E^{-1} \; A &= E \; M \\ A &= E \; M \end{aligned} \tag{4}$$

According to (4) the "measured data matrix" A is made up of two parts. One part is only time-dependent, the rows of the eigenvector matrix E represent the months in a year, and the other part, the multiplier matrix M, is only space-dependent.

Matrix A in equation (4) is not identical to the initial (measured data) Matrix A in equation (2), because when calculating A in (4), only the first 4 (most important) columns of E and the first 4 rows of M were used.

To obtain values for the variable of interest for the entire digital terrain map, one has to multiply the 4 eigenvectors in E with the value of the multiplier at the same grid point of the 4 corresponding multiplier maps. The sum of these 4 products is the final value of the variable at one point. Repeating this procedure yields the entire map.

References

[10.3.1] Lorenz, E.H., Empirical Orthogonal Function and Statistical Weather Prediction. Sci. Rep. No. 1, Dept. of Meteor., M.I.T., Cambridge, MA, 1956.

[10.3.2] Obukov, A.M., The Statistically Orthogonal Expansion of Empirical Functions. Izvest. Geophys. Ser. 1960, engl. translation, 288-291.

[10.3.3] Erdmann, H., Fechner, H., Die vertikalen nat rlichen Orthogonalfunktionen einer 19-jhrigen Reihe von halbtgigen Radiosondendaten der Station Erlangen/Stuttgart. Meteor. Rdsch. 28 (1975), 110-121.

[10.3.4] Fechner, H., Darstellung meteorologischer Felder mit endlichem Definitionsgebiet durch Reihen orthogonaler Funktionen. Ber. Inst. f. Meereskd. Nr. 44, Kiel 1978.

[10.3.5] Fechner, H., Empirische Orthogonalfunktionen. Promet 13 (1983), no. 1/2, 41-47.

[10.3.6] Stidd, C.K., The Use of eigenvectors for Climatic Estimates. J. Appl. Meteor. 6 (1967) 255-264.

[10.3.7] Granger, O.E., Climatology of Global Solar Radiation in California and an Interpolation Technique Based on Orthogonal Functions. Sol. energy 24 (1980),153-168.

[10.3.8] Jacobi, C.G.J., Über ein leichtes Verfahren, die in der Theorie der Säculastörungen vorkommenden Gleichungen numerisch aufzulösen. Jour. Reine Angew. Math. 30 (1846),51-95.

[10.3.9] Numerical Algorithms Group, The NAG Fortran Manual - Mark 12. Vol. 4, 1st Ed., Oxford 1987, Sect. F02ABF.

10.4 Cluster analysis

10.4.1 Introduction

Cluster analysis is a method of classification generally used in the fields of biology, medicine, climatology etc.. It arranges n items, each of which is distinguished by different characteristics (variables), in classes or clusters according to a given procedure. This procedure aims at forming classes such that items of a given class are as similar as possible, while they are as different as possible from items of other classes. In other words, the total set of n items is divided in sections (clusters) which are internally as homogeneous as possible, but externally as heterogeneous as possible. Detailed mathematical foundations are given by, e.g., Ward [10.4.1], Wishart [10.4.2], Vogel [10.4.3] and Anderberg [10.4.4].

There are several different cluster analysis methods. The one presented here is particularly well suited for climatology (Wenzel [10.4.5], Reich [10.4.6]).

10.4.2 Procedure

Carrying out a cluster analysis consists of 5 steps:

(1) Selection of items to be classified (solar radiation network stations);

(2) Selection of those variables like temperature, wind, sunshine duration etc., which characterize each station and are suited for classification;

(3) Definition of a criterion of similarity among stations and among clusters;

(4) Definition of a suitable method for finding the optimum number of clusters;

(5) Qualification of the classification results.

10.4.3 Mathematical Background

10.4.3.1 Standardization

Standardization avoids dissimilar internal weighting caused by different variability and it allows comparisons of characteristics which are measured in different units. Standardization is usually achieved by means of a Z-transformation

$$Z_i = (X_i - X_m)/S$$

where X_i is the variable or item (of characteristic feature), X_m is the arithmetic mean and S the standard deviation of the X_i. By this transformation each variable gets the same measuring unit and varies within the same order of magnitude.

One has to decide whether the cluster analysis is to be accomplished with non-standardized or with standardized data, i.e., whether to allow for dissimilar internal weighting ar to avoid this weighting. The former is not necessarily a drawback, as in our case, monthly or daily sums of irradiance with a wide range of variability may offer a better chance for distinguishable clusters.

10.4.3.2 Correlation characteristics

The basis of all cluster analysis methods is a data matrix of n elements (rows) and of m characteristics (columns). This matrix can be presented in an m-dimensional space if its columns are conceived as orthogonal axes in a Euclidian space. Complete othogonality of these axes is achieved if the pertinent characteristics, or variables, are not correlated with each other, that is, if they vary independently. If this is not the case, the Euclidian distances in the cluster analysis will be distorted. This distortion leads to an internal weighting of the characteristics corresponding to their degree of correlation.

To find out whether the characteristics are correlated, the correlation matrix has to be evaluated and subjected to the test of significance. If redundant characteristics are implied, the matrix has to be reduced in such a way that only independent characteristics remain. This may be achieved by a factor analysis which combines redundant characteristics in a new, fictitious characteristic. In the case considered in Chapter 5, factor analysis was not employed.

10.4.3.3 Similarity

Objects or clusters can be merged if their characteristics are similar. To define a similarity scale, a so-called distance function has to be chosen that - from a spatial point of view - represents the distance between two objects within the m-dimensional space of variables.

If there is only one variable per object, the distance function equals the difference of the values. If an object has several variables, all distances have to be combined to a distance G defined as

$$G_{ij}^2 = \frac{1}{m} \sum_{l=1}^{m} (X_{il} - X_{jl})^2$$

where m is the total number of characteristics, X_{il} is the measured value of the i-th object, and G_{ij}^2 is the mean square Euclidian distance between object i and object j.

Before evaluating the distance between an object and a cluster, or between two clusters, the average mean of the characteristics of all objects per cluster, the so-called centroid, has to be calculated. The latter can be interpreted as the center of a cloud of spots, the spots being at the spatial positions of the objects. Calculation of the distances between all these n objects or clusters, yields a symmetrical n-dimensional distance matrix, where it is assumed that the similarity of 2 objects, or clusters, is greatest if the distance between them is minimal.

10.4.4 Ward's method

Ward's cluster analysis method [10.4.1] is one of the so-called hierarchical agglomerative methods. It is also known as the method of the error sum of squares and it is widely used. Starting from the data matrix, the first fusion step consists of merging in one new cluster the 2 nearest objects. The next fusion step is to choose the cluster with the least error sum of squares. Merging cluster P with cluster Q is defined as

$$EQS = \frac{pq}{p+q} \sum_{j=1}^{m} (\bar{P}_j - \bar{Q}_j)^2$$

where p and q are the number of objects in cluster P and cluster Q respectively, and the P and Q with overbars are their respective centroids. At the end of this procedure, all objects are combined in one single cluster.

10.4.5 Defining the optimum number of clusters

Obviously, it has to be decided at which fusion step this heirachical classification procedure should be stopped. It can be seen from the increase of the error sum of squares how similar or dissimilar two clusters are at the moment when they are merged into one new cluster. A sudden increase of EQS at a given fusion step indicates the formation of a heterogeneous cluster by fusion of two formerly homogeneous clusters. Thus, plotting the error sum of squares as function of the number of clusters permits identification, by means of the so-called "elbow-criterion" (where EQS shows sudden increases), those points where the described clustering procedure might be stopped.

References

[10.4.1] Ward, J.W.: Hierarchical Grouping to Optimize an Objective Function. *Jour. Amer. Statist. Ass.* **58**, 236-244, 1963.

[10.4.2] Wishart, D.: Clustan User's Manual. Clustan Ltd., Edimbourgh, 1984.

[10.4.3] Vogel, F.: Probleme und Verfahren der numerischen Klassifikation. Vandenhoeck u. Ruprecht Verlag, Göttingen, 1975.

[10.4.4] Anderberg, M.R.: Cluster Analysis for Applications in Probability and Mathematical Statistics 19, New York, 1973.

[10.4.5] Wenzel, J.: Klimaklassifikation mit Hilfe von cluster- und diskriminanzanalytischen Verfahren dargestellt am Beispiel von China. Diplom-Arbeit, Geogr. Inst. Univ. Mainz, 1988.

[10.4.6] Reich, T.: Die räumliche Struktur des Niederschlagfelds. Zwei Hierarchische Gruppierungsverfahren und Dendogramme. *Zeitschr. Meteorol.* **36**, 38-53, 1986.

10.5 Technological supports

In what follows, we review some of the software products useful for an adequate interpolation process. We examine three classes of products with respect to spatial interpolation performance: *ARC-INFO* as a Geographic Information System (GIS), *UNIRAS* as a development shell of scientific applications and *BLUEPACK-3D* as a geostatistical system.

GISs are able to store, manage, combine and display different geographical data types (spatial data, raster images, cartographic data). Geographical data describe objects from the real world in terms of:
-their position in a given coordinate system,
-their attributes (e.g., solar radiation, elevation),
-their spatial interrelation with each other (topological relations), which describe how they are linked together or how one can "travel" between them.
Generally, GISs consist of the following modules: data input and verification, data storage and database management, and mapping. Data input covers all aspects of transforming data captured in the form of existing maps, field observations, and sensors (including aerial photography, satellites, and recording instruments) into a compatible digital form. Data storage and data management determines how position data, linkage (topology), and attributes of geographical elements (points, lines, areas representing objects on the earth's surface) are structured and organized, both with respect to the way they must be handled in the computer and how they are perceived by the users of the system. Data are displayed and the results of the analysis are reported to the user in the form of maps.

ARC-INFO

ARC-INFO is a GIS distributed by Environmental Systems Research Institute (ESRI) which contains a Surface Modelling and Display Module called TIN. The latter acronym stands for Triangulated Irregular Network; a set of adjacent, non-overlapping triangles between points having x,y coordinates and z values. The TIN data structure includes topological relationships between points and their closest neighbors (i.e., which points define a triangle and which triangles are adjacent to each other). Therefore, the TIN model allows for an efficient generation of various space models. The data which represent the surface are organized into two forms: a series of x,y,z points and a series of triangles which connect these points to their neighbors. For example, each triangle can be conceived as a facet of a surface. The information which defines each of these triangles as a set of nodes and the triangle's adjacency to other triangles forms the topological structure of TIN. For each triangle, TIN records a triangle number, a list of the 3 nodes defining the triangle, the x,y coordinates of each node, the surface ordinate (z value) of each node, and a list of each triangle's three adjacent triangles. *Akima*'s interpolation method is applied in *ARC-INFO* and the triangulation method is guided by *Lawson*'s criterion.

A three-dimensional model of the surface is generated from a TIN or from a lattice file. Both models are based on point data which associate a z value item with an x,y location. Smooth surfaces generated in this way become the models upon which to perform surface analysis.

UNIRAS

UNIRAS Inc.'s product is based on rastering techniques and comes in two forms. One is interactive, supporting non-specialist computer users, the second has an extensive subroutines library which enables professional programmers to integrate high quality graphics in new or existing applications.

Interpolations implemented in the Application Graphic Library (AGL) perform gridding operations based on a variety of methods, most of which have been mentioned in the foregoing sections. AGL-Interpolations are used in connection with AGL-GRIDS, AGL-BLOCKS, AGL-CONTOURS and AGL-CONTOURS PLUS for gridding and smoothing geographical data. AGL-INTERPOLATIONS comprises a collection of FORTRAN 77 subroutines from which the user selects the required combination. The following techniques are supported:

- a combination of double linear or weighted average, slope calculation, quadratic interpolation and smoothing. Faults and regions are optionally supported;
- bivariate interpolation with smooth surface fitting based on *Akima*'s procedure. Use of local functions in the search results in lower processing costs;
- fit of polynomial surfaces of any degree to a small region of the grid area. The user has control over the gridding procedure and may choose the parameters to fit the density of the actual points;
- gridding with a method similar to the first method. Search, however, is performed using fast grid operations. The larger the original data set, the larger the gain in computing time. Faults and regions are, again, optionally supported;
- gridding based on minimum curvature. This technique is based on a surface which has minimum deformation energy while passing through all the data points;
- AGL-Kriging is a library of routines for estimating semi-variograms for 2-D and 3-D data points and plotting semi-variograms in 2-D, for Universal Kriging and for 2-D and 3-D kriging interpolation.

BLUEPACK-3D

A general purpose program for optimal interpolation of a variable distributed in a one, two or three dimensional space and sampled at points which may be arbitrarily arranged. In addition to the estimates *BLUEPACK-3D* calulates the estimation variance which can be used to specify confidence limits. *BLUEPACK-3D* can automatically identify the variable's structure based on the theory of intrinsic random functions. It performs not only kriging but also stochastic simulation in 3-D by the "turning bands" method. The automatic structure identification permits good use of *BUEPACK-3D* in mapping applications.

The products reviewed above are implemented on a variety of computers: from personal computing to mainframe and graphic workstations. The class of GIS develops the deterministic approach in terms of Akima's interpolation or in terms of moving averages. Thus, *ARC-INFO* has the ability to conveniently manage cartographic information, but it doesn't include the geostatistical approach. *UNIRAS* is able to manage geostatistical tools and deterministic methods in a sophisticated graphical environment. It doesn't support specialized input/output devices but provides shells to develop such applications. *BLUEPACK-3D* includes complex geostatistical analysis and it contains some of the GIS functions (from data acquisition to mapping). It is designed specifically for the end-user.

CHAPTER 11

11. THEORY OF SATELLITE-BASED METHODS

Review prepared by: *Maurizio Noia, Corrado F. Ratto and Roberto Festa*
Department of Physics, University of Genoa

Under sponsorship from: *National Research Committee and MURST*

Satellite-based estimates of the irradiation available at the Earth's surface were introduced in Section 2.3 in the course of the presentation of the data. In order to allow discussion of the validation results, Section 4.1 included first principles of the conversion of the counts registered by the satellite's radiometers into irradiation at ground level. The present Chapter elaborates the theory of the conversion methods, describes basic examples and systematically points out the differences among them. In a certain sense, this Chapter describes the physical content and the numerical background of Fig. 4.1 in Chapter 4.

11.1 Introduction

Some of the first attempts to utilize satellite data in estimating incident solar radiation at the earth's surface are found in the works of Hanson [1], Vonder Haar and Ellis [2,3,4]. All these methods were based on the images provided by polar satellites. A major drawback of these early insolation studies was the limited temporal coverage provided by the polar orbiting meteorological satellites (only one or two daylight images per day). This meant that cloud cover variations could not be taken into account within the framework of these early studies. The problem of cloud cover variability was solved through use of data from geostationary meteorological satellites. These data have high temporal resolution, as these satellites are capable of producing an image every 30 minutes with acceptable spatial resolution (1-5 kilometers at the subsatellite point).

To estimate the solar irradiance incident on the ground by satellite images, two different approaches were developed. The first approach is represented by the methods based on a statistical regression between the digital counts measured by the satellite and the corresponding solar radiation value measured at the earth's surface. These are commonly termed ***statistical methods.*** The second approach is represented by the methods based on radiative transfer models that explicitly describe the physical processes (i.e. scattering and absorption) operating in the earth-atmosphere system. These are usually referred to as ***physical methods.***

In this chapter we describe the concepts and approximations on which both statistical and physical methods are based and we review some of them, namely those developed by Hay and Hanson [5], Tarpley [6], Justus, Paris and Tarpley [7] and Cano [8], for the statistical methods, and those developed by Gautier, Diak and Masse [9], Möser and Raschke [10], Dedieu, Deschamps and Kerr [11] and Marullo, Dalu and Viola [12], for the physical methods. We also report the precision of all these methods as assessed in several applications by the authors themselves and/or by other scientists (e.g. Raphael [13] and Raphael and Hay [14]).

11.2 Physical basis of all methods

The solar radiation incident on the top of the atmosphere interacts with the atmospheric components (i.e. molecular gases and aerosol particles) and with the clouds. As a consequence, a part of this radiation is backscattered toward space, a part is absorbed, while the rest reaches the

ground. The radiation reaching the ground interacts with the earth's surface: a part of this radiation is absorbed by the ground, while the remainder is reflected back toward space. Therefore the solar radiation emerging from the atmosphere is composed of the solar radiation backscattered by the atmosphere before reaching the ground, and by that reflected by the earth's surface and transmitted again through the atmosphere, as illustrated in figure 1.

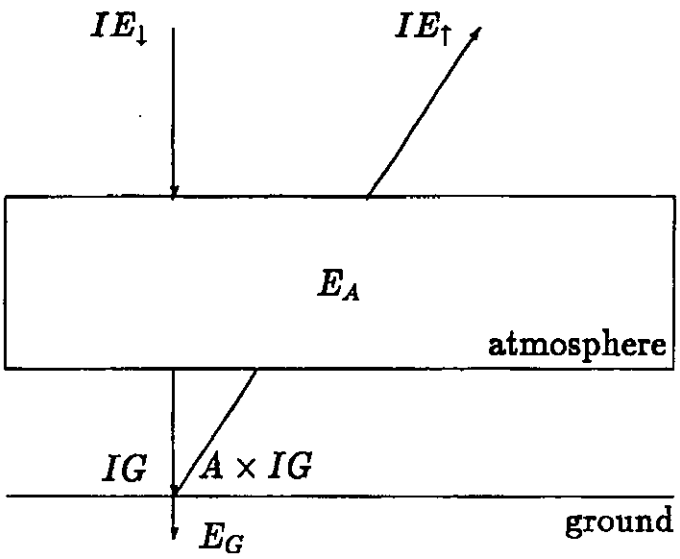

Figure 1: Solar radiation fluxes in the atmosphere - earth system

Considering the flux density conservation in an earth-atmosphere column, we may write (see figure 1):

$$IE_{\downarrow} = IE\uparrow + E_A + E_G \qquad (1)$$

where:

$IE_{\downarrow}$ represents the flux density of the downward solar radiation incident on the top of the atmosphere;
$IE\uparrow$ represents the flux density of the upward solar radiation emerging from the atmosphere, which is proportional to the brightness measured by the satellite;
E_A is the solar radiation flux per unit surface absorbed by the atmosphere;
E_G is the solar radiation flux per unit surface absorbed by the ground.

The solar radiation absorbed by the ground can be expressed in terms of the *surface albedo* A and the solar irradiance at the surface IG:

$$E_G = IG(1-A) \qquad (2)$$

Hence, substituting equation (2) in (1) and solving for IG, we obtain:

$$IG = (IE_{\downarrow} - IE\uparrow - EA)/(1 - A)\ . \qquad (3)$$

$IE_{\downarrow}$ depends on the solar zenith angle and on the sun-earth distance and can be computed by means of the following equation:

$$IE_{\downarrow} = F_{cs} (r_o/r)^2 \cos\theta \qquad (4)$$

where:

F_{cs} = 1367 W/m^2 is the *solar constant*;
r_o and r are, respectively, the mean and the actual sun-earth distance;
θ is the solar zenith angle.

Notice that, since the radiometers on board the satellites are sensitive to visible radiation in a window (e.g. 0.40 - 1.10μm for METEOSAT) narrower than the spectral range of pyranometers, some authors (e.g. Marullo *et al.* [12]) use a correspondingly reduced value for the solar constant.

If we were at anytime able to estimate the energy E_A absorbed by the atmosphere and to know *a priori* the surface albedo A, then equation (3) could be used to estimate IG from the values of $IE\uparrow$ measured by the satellite radiometer. Unfortunately, the problem is more complex since it is not possible to know the surface albedo A for every point of the region studied and since the value of E_A is very variable depending on the atmospheric conditions.

One of the most important factors affecting both the incoming solar radiation at the earth's surface and the solar radiation emerging from the atmosphere, is the change in the solar zenith angle corresponding to a change in the air mass through which the solar radiation propagates. The second largest cause results from the presence of clouds. The presence of water and ice droplets in clouds strongly increases both absorption and scattering of solar radiation. Changes in atmospheric water vapor or aerosol content can be considered as second order effects.

Since changes in solar zenith angle can be exactly computed, if we can outline cloud regions in the satellite images and evaluate with reasonable accuracy their effects on solar radiation, then we might be able to estimate insolation in both clear and cloudy conditions. Indeed the flux density $IE\uparrow$ measured by the satellite is very sensitive to clouds, since clouds can be intense reflectors of visible solar radiation. So the upward flux density $IE\uparrow$ associated with a cloudy region is, in general, much larger than that associated with a clear sky area (except for snow or ice covered regions) and this is used to detect the presence or absence of clouds. In the following we shall see that some models also use the values of the flux of infrared radiation reaching the satellite in order to improve the distinction between cloud cover and some kinds of land.

11.3 General description of the methods

The methods to estimate solar irradiance at ground, as we have already pointed out, may be divided into statistical and physical types. Statistical methods are based on one or more relationships, treated as statistical regressions, between the solar radiation measured by means of a pyranometer in a meteorological station and the simultaneous digital count value provided by the satellite for the pyranometer site. This relationship is used to estimate the surface solar irradiance for the entire region. The independent variables used in the regression equations include:
cosine of the solar zenith angle, $\cos\theta$, that gives, among others, information about flux incident on the top of the atmosphere;
cloud cover index, n, to describe the amount of the cloud cover over the studied region;
atmospheric transmittance, $T(\theta) = IG/IE_{\downarrow}$, to account for the attenuation produced by the atmospheric components;
brightness, B, measured by the satellite and assumed proportional to $IE\uparrow$, that gives information about the upward solar radiation emerging from the atmosphere;

clear-sky brightness, B_0, that gives information about the ground albedo;
maximum brightness, B_{max}, that gives information about the cloud albedo.

The physical methods are based exclusively on physical considerations that explicitly represent the radiant energy exchanges taking place within the earth-atmosphere system. The main quantities considered in the physical methods are:
scattering and *absorption coefficients* of the clear atmosphere components;
cloud albedo and *absorption* coefficients;
surface albedo.

The main advantage (see Table 1) of the statistical methods is their simplicity and, consequently, their operational efficiency. Indeed they use the digital count values provided by the satellites to estimate surface irradiance and do not need to convert such values into flux density of the upward solar radiation emerging from the atmosphere. Furthermore, these methods usually do not need complementary meteorological data (i.e. temperature, humidity and precipitable water).

Tab1e 1: Advantages and disadvantages of both statistical and physical methods.

	Statistical	Physical
ADVANTAGES	• operational efficiency • no need for meteorological data • no need for calibration	• generality • no need for ground solar radiation data
DISADVANTAGES	• need ground solar radiation data • lack of generality	• need meteorological data • need calibration

The main limits of the statistical methods are the need for ground solar radiation data and the lack of generality. Remember that the regression equation coefficients, determined for the places corresponding to the ground solar radiation data, are used to estimate the solar radiation reaching

the ground throughout the region studied. Furthermore, there is no guarantee that these coefficients would have the same values in other areas. The main advantage offered by the physical methods is their generality, since they may be applied everywhere. Besides, they do not need solar radiation data measured at the surface. However, some of the physical methods need complementary meteorological data to estimate the various quantities relative to the interactions of the solar radiation with the atmosphere. Another limit of the physical methods is that they convert the satellite digital counts into the corresponding flux density of the upward solar radiation emerging from the atmosphere. This conversion requires an accurate and updated calibration of the instrument.

A difficulty common both to statistical and to physical methods (but more significant for statistical) is that the space and time scales of satellite images are different from those of ground-based measurements. In particular, two problems arise when comparing the satellite data to ground solar radiation measurements. The first problem results from errors in locating the pyranometers sites on the images. The second problem is that satellite data are instantaneous measurements over a small solid viewing angle, while the ground measurements are integrated over time (usually one hour) over a solid angle of 2π. The solution found to both problems is the use of target areas encompassing numerous pixels. This not only reduces the spatial problems but also those connected to the time discrepancy. As Tarpley claims in the paper cited below, an instantaneous measurement averaged over a 50 km target is likely to be equivalent to a time averaged point measurement made by a pyranometer, at least to the extent that cloud distribution is uniform within the targets and that changes in cloud type and thickness are small over an hour.

Tarpley [6] subdivides the GOES satellite images relative to the considered region into target areas made up of 7 x 6 8-km resolution pixels (≈ 50 x 50 km), centered at the intersections of 0.5° latitude and longitude lines. Justus, Paris and Tarpley [7] considered target areas of 5 x 5 8-km resolution pixels (≈ 40 x 40 km). with targets centered on a 1° x 1° latitude-longitude grid spacing.
Gautier et al. [9] averaged the estimates deduced from the satellite images over 8 x 8 2.5-km resolution pixels (≈ 20 x 20 km) centered on a given ground station.
Raphael [13,14] calculated the radiation on a 5 x 5 pixel array (≈ 4 km East-West x 8 km South-North) and assumed that three adequately weighted (with respect to time) images should represent the radiation conditions for one hour.

11.4 Statistical methods

The simplest statistical method is that developed by Hay and Hanson [5], here referred to as the HH method. The main feature of this method, which uses only visible satellite data, is its treatment of the atmosphere as a "black box". However, despite what we wrote about the general features of the statistical methods, the HH method requires calibration of the satellite data.

The method developed by Tarpley [6] (referred to as the T method) also uses only visible satellite data. This method is based on a discrete cloud index that distinguishes three sky conditions: clear, partly cloudy, and cloudy. Hence, for every condition, Tarpley developed a different regression equation. These regression equations have some physical basis so that the T model is considered as the more physical among statistical methods. A simplification of the T method was developed by Justus, Paris and Tarpley [7]. This method, here referred to as the JPT method, does not require the estimation of a cloud index and is based on one regression equation only, valid for all cloud conditions.

The last statistical method we will review is that developed by Cano et al. [8] (referred to as the C method). The C method describes the cloud amount with a cloud index that, in this case, is a continuous quantity. This method appears to be very simple with sound physical basis.

11.4.1 HH model

The HH model [5] was part of a study undertaken to map the distribution of shortwave radiation incident at the sea surface, for the Global Atmospheric Research Program Atlantic Tropical Experiment, using satellite observations in the visible region of the spectrum (0.55 - 0.75μm). The HH method is based on a simple linear relationship between the *visible radiance*, $SR=IE\uparrow/IE_{\downarrow}$, i.e. the radiation measured by the satellite sensor and normalized to the extraterrestrial irradiance, and the atmospheric shortwave trasmittance T, defined as the ratio between the irradiance at the surface, IG, and the extraterrestrial irradiance, $IE_{\downarrow}$. This expression is:

$$T = a - b\,SR \tag{5}$$

where a and b are the regression coefficients. Notice that the digital counts provided by the satellite have to be transformed into the visible radiance via calibration procedures.

Equation (5), remembering equation (4), can also be written as:

$$IG = F_{cs}(r_o/r)^2 \cos\theta\,[a - b\,SR]\,. \tag{6}$$

The values of the regression coefficients a and b determined by Hay and Hanson are:

$$a = 0.79, \qquad b = 0.71$$

Raphael and Hay [14] (see sect. 11.6) find a better agreement with their data base using the following values for a and b:

$$a = 0.788, \qquad b = 1.078$$

Note that the atmospheric transmittance changes as the content of the atmosphere changes. An increase in the concentration of aerosols, water vapor and cloud cover, produces increased scattering in the atmosphere: this results in a decrease in the atmospheric transmittance and ground measured insolation and, inversely, to an increase in the satellite measured radiance. On the other hand, an increased absorption will reduce the radiation measured at the surface and, at the same time, the radiance measured at the satellite. Thus an inverse relationship between atmospheric transmittance and satellite measured radiance as in equation (5) should easily capture the variations in scattering which influence the irradiance measured at the surface, but will be less reliable when absorption is significant.

The relationship breaks down under conditions of high surface albedo (e.g. a snow or ice covered surface). The high albedo will increase both the brightness measured by the satellite (except under heavily overcast conditions) and the radiation measured at the surface through the effect of multiple reflections. This would lead to a significant underestimation of the radiation at the surface, particularly under a clear sky.

11.4.2 T model

The method developed by Tarpley [6] represents a part of the outcome of an experiment conducted in the summer of 1977 over the U.S. Great Plains. In this study GOES images were used. The modelling procedure requires the development of a clear sky (i.e. minimum) brightness parametrization, the determination of a cloud index and other image related quantities (mean target brightness and mean cloud brightness) and the determination of the atmospheric trasmittance. All these quantities are used in the final regression equations.

11.4.2.1 Clear-sky brightness parametrization

To determine cloud characteristics, it is necessary to know when each pixel is cloud free. Tarpley calculated an experimental clear-brightness value by regression against functions of solar zenith

angle and the azimuth angle between the sun and the satellite. The clear-sky brightness parameterization was provided by the following equation:

$$B_0 = a + b\cos\theta + c\sin\theta\cos\phi + d\sin\theta\cos^2\phi \quad (7)$$

where:
B_0 = minimum brightness;
θ = local solar zenith angle;
ϕ = azimuth angle between sun and satellite, illustrated in figure 2.

The second term on the right hand side of equation (7) accounts for the geometry of the incident flux; the following two terms are introduced to account for changes in target brightness due to shadowing at the surface and anisotropic scattering.

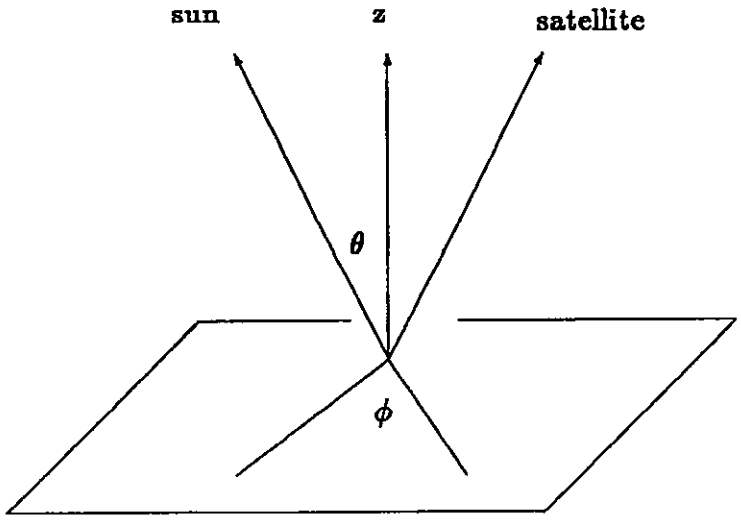

Figure 2: Solar and satellite angles used in the clear sky brightness equation.

To compute the clear-sky brightness regression coefficients a, b, c and d more than 100 observation targets were collected by Tarpley during 27 days just prior to the test period. For each day and for each target, mean standard deviation and the solar azimuth and zenith angles were computed and stored. Tarpley adopted the following automatic cloud detection and elimination procedure.

- A pass was made through the data discarding cases with solar zenith angles > 85°, since observations at large zenith angles were likely to be anomalous.
- Cases with excessive target standard deviations were discarded: this criterion was used as an indication of clouds in the target. The selection with respect to standard deviation produced cloud-free data for determining a reasonable clear-sky brightness prediction.
- The "residuals", defined as the difference between the measured and the predicted brightness, were calculated for each target. Residuals from the "first guess" clear-sky brightness equation were used to identify the remaining cloud-contaminated targets.
- Thresholds were fixed to eliminate targets contaminated by small or partially transparent clouds. Residuals were compared with thresholds: residuals exceeding these thresholds were discarded from the data base.
- After three cycles, this procedure yielded a cloud-free data set and reliable regression coefficients.

Notice that targets containing features with markedly contrasting brightness, such as steep mountains, coastal areas and great lakes, could not be accurately fit to the regression equation.

11.4.2.2 Determination of the cloud index and other image related quantities

Cloud amount in each target is determined via a two-threshold method, originally developed by Shenk and Salomonson [15]. This method classifies the pixels into three classes, clear, partly cloudy and cloudy, separated by two threshold values (in [15] these classes are defined with respect to count values on a 6-bit scale (GOES) ranging from 0 to 63):
a *clear/partly cloudy* threshold T_1 is defined as the clear brightness, B_0, plus three counts: a pixel is defined as clear if its count value is $\leq T_1$;
a *partly cloudy/cloudy* threshold T_2 is defined as the clear brightness, B_0, plus five counts. Any pixel with value $> T_1$ but $\leq T_2$ is considered to be partly cloudy. Pixels with values $> T_2$ are classified as *cloudy*.

The cloud index n is determined via the following expression:

$$n = (0.5N_2 + N_3)/(N_1+N_2+N_3) = (N_2 + 2N_3)/(2N) \quad (8)$$

where N is the total number of pixels in a target area, while N_1, N_2 and N_3 are the numbers of pixels in clear, partly cloudy and cloudy categories, respectively.

The mean target brightness B_m, another image related quantity, is simply defined by Tarpley as the mean brightness of a 7 x 6 pixel array. The last image related quantity is the mean cloud brightness B_{cld}, used as an indicator of cloud thickness: mean cloud brightness is computed by averaging the brightness values of all pixels in the 7 x 6 array that are brighter than T_2.

11.4.2.3 Atmospheric transmittance determination

In order to obtain the final regression equation between the solar radiation received at the earth surface and the brightness seen by the satellite, Tarpley introduced an atmospheric transmittance T, obtained by means of the following equation:

$$T = T_{ws} \times T_{wa} \times T_r \quad (9)$$

where T_{ws}, T_{wa} and T_r denote the visible transmission due to water vapor scattering, water vapor absorption and Rayleigh scattering, respectively. The scattering transmission functions (Davies et al. [16]) are:

$$T_{ws} = 1 - 0.00225\ w\ m \quad (10)$$

and

$$T_r = 0.0972 - 0.0826\ m + 0.00933\ m^2 \quad (11)$$

where w is the equivalent column of precipitable water (cm) and m is the optical air mass. The water vapor transmission function (Mc. Donald [17]) is:

$$T_{wa} = 1 - 0.077(w\ m)^{0.3}\ . \quad (12)$$

The expression for m used by Tarpley is:

$$m = \exp(-z/8243)/\cos\theta \quad (13)$$

where z is the station's elevation expressed in meters, and θ is the solar zenith angle. Expression (13) is not valid at large zenith angles. Raphael and Hay [14], therefore, prefer the formula developed by Kasten [18]:

$$m = \exp(-z/8243)/[\cos\theta + 0.15/(93.8805^{o} - \theta^{o})^{1.253}] \quad (14)$$

11.4.2.4 Regression equations

To properly account for the different physical processes depleting the incoming solar radiation under clear, partly cloudy and cloudy conditions, the T model uses three different equations. According to the cloud amount in the target, hourly solar irradiation is estimated as follows:

$$HG = a + b\cos\theta + c\,T + d\,n + e\,(B_m/B_0)^2 \quad (n < 0.4)$$

$$HG = a + b\cos\theta + c\,n\,(B_{cld}/B_n)^2 \quad (0.4 \le n < 1) \quad (15)$$

$$HG = a + b\cos\theta + c\,(B_{cld}/B_n)^2 \quad (n = 1)$$

where:

HG is the hourly surface irradiation;
$B_n = B_0(\theta = 45^{o}, \phi = 105^{o})$ is the normalized clear-sky brightness;
while n, B_m, B_{cld}, B_0, T and θ were already defined in the two preceding sections.

Table 2: Regression coefficients entering equations (15) as determined by Tarpley (1979) and - in parenthesis - by Raphael and Hay (1984). Units: $kJm^{-2}h^{-1}$.

Coefficients	clear $n < 0,4$	partly cloudy $0,4 \le n < 1,0$	cloudy $n = 1,0$
a	- 809.54 (-195.67)	-400.79 (-199.30)	-274.73 (-49.80)
b	3646.91 (3722.93)	3959.34 (4047.97)	3672.04 (2187.16)
c	1155.10 (85.98)	-319.13 (-329.30)	-314.10 (-168.80)
d	-438.90 (151.10)		
e	-266.78 (-90.86)		

Satellite brightness measurements were constrained by Tarpley to enter the regressions as ratios to compensate for time variations in reflected radiance, geographic variations in albedo, and to minimize effects of instability in the visible sensor. Indeed, the observed brightness of a clear

target will vary on a completely clear day due to changing illumination angles, while the ratio between the measured target brightness and the predicted brightness is relatively constant for different targets and for different times of the day. In the case of cloudy targets, the measured cloud brightness was divided by a "normalized" surface brightness to reduce calibration errors and, perhaps, partially compensate for albedo differences from target to target.

The values of the regressions coefficients a, b, c, d and e determined by Tarpley are reported in Table 2, in kJ m^{-2} h^{-1}, together with the values obtained by Raphael and Hay [13,14] in their experiment (see section 11.6).

From Table 2 it is also evident that identification of both substantially clear scenes and cloudy conditions is possible only if the dependence on the solar zenith angle is accounted for within the images.

11.4.3 JPT model

The model developed by Justus, Paris and Tarpley [7] was part of the Agriculture and Resources Inventory Surveys through Aerospace Remote Sensing (AgRISTARS) program and was applied to produce insolation maps of the United States, Mexico and South America based on GOES images. The JPT model is a simplification of the T model, presenting only one regression equation valid for all cloud conditions and without need for supplementary meteorological data such as the precipitable water content of the atmosphere.

11.4.3.1 Clear-sky brightness values estimate

To account for variations in the directional reflectance of the earth-atmosphere system due to variations with earth-location position of viewing angle and solar illumination angle, Justus and coworkers determine values of the clear-sky brightness for each satellite viewing time and for each of the target areas by the following procedure which they term "minimum brightness technique".

The authors assume to already know, for a particular time of day and for each target-area, a brightness under clear-sky conditions, B_0', and the corresponding observed mean target brightness B_m (see section 11.4.2.2). Two thresholds, B_{min} and B_{max}, are pre-determined, which indicate the likely presence of clouds at the time of observation ($B_m > B_{max}$) or insufficient scene illumination for insolation estimates ($B_m < B_{min}$). Notice that in this model, brightness is in 6-bit digital counts but the numerical values to attribute to B_{min} and B_{max} are not given in the cited paper.

For each new image a new clear brightness value B_0 is determined in the following way:

$B_0 = B_0'$ if $B_m \geq B_{max}$

$B_0 = w_1 B_0' + (1-w_1) B_m$ if $B_0' < B_m < B_{max}$

$B_0 = B_m$ if $B_0' - 2 < B_m \leq B_0'$

$B_0 = w_2 B_0' + (1-w_2) B_m$ if $B_{min} \leq B_m \leq B_0' - 2$

$B_0 = B_0'$ if $B_m < B_{min}$

where the weights w_1 and w_2 can assume values between 0 and 1. The values empirically determined by Justus, Paris and Tarpley, are respectively 0.99 and 0.90.

The first equation (likely presence of clouds at the time of observation) leaves the clear brightness unchanged if the observed counts exceed the predetermined threshold B_{max}. The second

equation, with w_1 only slightly less than 1, allows a small increase in clear-sky brightness to occur at intermediate values of B_m, in order to allow the clear-sky brightness to increase with seasonal variations in surface reflectance or solar illumination angle. The third equation replaces the current clear-sky brightness value with any lower count (i. e. clearer sky) values encountered. The fourth equation, with w_2 greater than 0, does not allow the clear-sky brightness value to decrease by a large amount from one observation to another: this feature is an attempt to suppress spurious effects on B_0 values caused by satellite navigation errors, image line dropouts or other problems, which occasionally occur in the satellite imagery. Finally, the fifth equation (insufficient scene illumination for solar radiation estimates if B_m drops below the predetermined minimum threshold B_{min}.) leaves the clear-sky brightness unchanged.

11.4.3.2 Regression equation

For each target area, the observed mean target B_m and clear-sky brightness B_0, determined as described above, are used to estimate hourly irradiation. The algorithm, which replaces equation (15) of Tarpley, is represented by the following equation:

$$HG = F_o (r_o/r)^2 \cos\theta \, [a + b \cos\theta + c \cos^2\theta] + d \, (B_m^2 - B_0^2) \qquad (16)$$

where:

F_o = 4921kJm^{-2}h^{-1} is the hourly total value for the solar constant, while the values of the regression coefficients a, b, c and d given by the author are reported in Table 3.

Table 3: Regression coefficients entering in equation (16) as proposed by Justus et al.(1986): a, b, and c are dimensionless, d is in kJm^{-2}h^{-1}.

a	b	c	d
0.4147	0.7165	-0.3909	-1.630

The last term in the right hand side of equation (16) is the cloud correction term. It was set by the authors proportional to the difference between the squares of the observed radiance and of the clear-sky radiance, since they used satellite data from GOES whose nominal calibration has radiances proportional to the square of brightness counts.

11.4.4 C model

The method developed by Cano [19] and Cano *et al.* [8], subsequently slighty modified by Moussu *et al.* [20], represents the basis for the HELIOSAT project of École Nationale Supérieure des Mines de Paris, Sophia-Antipolis (France). This approach was developed to treat METEOSAT images in the visible channel.

The basic concept of the C method is that the amount of cloud cover over a given area, statistically determines the global radiation for that area. Thus the processing is divided into the following steps:
- a map of the reference ground albedo is constructed and constantly updated, giving information about the clear-sky planetary albedo for every pixel;
- a cloud cover index map is derived from the comparison of the current satellite image and the reference albedo map;
- the atmospheric transmittance factors are computed using ground pyranometric data and a statistical linear regression between these factors and the cloud cover index at the same location;
- finally, the atmospheric transmittance factors are used to construct the global radiation map.

11.4.4.1 Construction of the reference albedo map

The first step of the C approach consists in the determination of a *reference* albedo, ρ, defined as the ratio between the visible digital counts from the satellite, $IE\uparrow$, and the flux IG incident on the ground. The term "reference" is used to underline that this quantity is not a physical albedo since the visible counts are not converted, through calibration, to a true physical flux.

According to the Bourges model [21], Cano *et al.* describe the global radiation IG_o incident on the ground at instant t under clear-sky conditions in the following way:

$$IG_o = 0.7\ IE_{\downarrow}\ (t)\ (\sin\theta)^{1+a} \tag{17}$$

where θ is the solar zenith angle and *a* is a parameters that characterizes the total atmospheric transmittance. Its value ranges from 0.14 to 0.17 and the value retained in the C method is 0.15. Then the solar irradiance received by the satellite in clear-sky conditions can be written as:

$$IE\uparrow = \rho_o\ IG_o = 0.7\ \rho_o\ IE_{\downarrow}\ (\sin\theta)^{1.15} \tag{18}$$

where ρ_o is the reference albedo for clear-sky conditions.

Note that equation (18) does not take into account the solar radiation attenuation after its reflection on the ground. It was subsequently corrected by Moussu et al. [21] and became:

$$IE\uparrow = 0.7\ \rho_o\ IE_{\downarrow}\ (\sin\theta\ \sin\theta_s)^{1.15} \tag{19}$$

where θ_s represents the satellite's zenith angle.

The reference albedo ρ_o is evaluated at each pixel in a recursive procedure by normalizing the the measured radiances $IE\uparrow$ by the radiances resulting from the clear-sky model, the cloudy cases being eliminated at each step. This is made possible by the fact that the cloud albedo is greater than the ground albedo, except over snow and some desert soils. If the cloud albedo is greater than the ground albedo, the presence of a cloud in the field of view of the sensor will result in increased measured radiance. Assuming that, given a satellite image time-series, the clouds are non-stationary, the minimum value observed at each pixel will provide the clear sky situation (reference albedo map).

The iterative procedure usually converges to the mean reference albedo value: important exceptions are snow, some desert surfaces and stationary clouds in the examined time-series. Then, cloud detection can be performed by the difference between the cloud induced response and the corresponding view of the ground under a clear-sky.

11.4.4.2 Evaluation of the cloud cover index

The inverse of the algorithm used to derive the reference map of the ground albedo is applied to the construction of a completely cloudy image while retaining only the cloudy cases. The histogram of this purely cloudy image provides an estimation of the average albedo of the top of the opaque clouds, ρ_c.

The cloud cover index $n^t(i,j)$ at pixel (i,j) for a given image at time t is defined as a function of the reference albedo, $\rho_o(i,j)$, of the apparent albedo at the same point in the image, $\rho^t(i,j)$, and of the average albedo of the cloud tops, ρ_c as

$$n^t(i,j) = [\rho^t(i,j) - \rho_o(i,j)]/[\rho_c - \rho_o(i,j)] \qquad (20)$$

This cloud cover index ranges from 0 to 1 and it can be interpreted as the percentage of the cloud cover per pixel.

Expression (20) for the cloud cover index is not valid when the ground albedo is of the same order of the cloud albedo. This is the case for snow or ice covered areas and for some desert areas. In this case Cano [19,8] proposed a different definition of the cloud cover index using the radiance $R^t(i,j)$ measured by the satellite in the thermal infrared spectral band. The new expression for the cloud cover index is then

$$n^{t\prime}(i,j) = [R^t(i,j) - R_s]/[R_c - R_s] \qquad (21)$$

where R_c is the radiance of the opaque cloud tops and R_s is the radiance of the surface covered by snow or ice. Both these quantities are determined using histograms of the thermal infrared images; however no further study of this alternative approach was performed, as far as we know .

11.4.4.3 Statistical relationship between cloud cover index and ground measured transmission

The C method uses a linear regression between the satellite determined cloud cover index and the ground measured transmission factor $T^t(i,j)$, also known as the clearness index, and defined as the ratio of global radiation reaching the ground, IG, to the extraterrestrial irradiance $IE_{\downarrow}$,i.e.,

$$T^t(i,j) = IG(i,j)/IE_{\downarrow}(i,j) \qquad (22)$$

Considering the cloud cover index $n^t(i,j)$ as the percentage of a pixel covered by clouds, the global irradiance at ground may be written as a linear combination:

$$IG^t(i,j) = n^t(i,j)\ IG_c(i,j) + [1 - n^t(i,j)]IG_o(i,j) \qquad (23)$$

where IG_c and IG_o are the global irradiance for overcast and clear sky conditions, respectively.

To each of these extreme conditions one can associate a transmission factor, $T_c(i,j)$ respectively $T_o(i,j)$, which is supposed to be constant for a given hour . This realistic hypothesis leads to the following linear relationship:

$$T^t(i,j) = n^t(i,j)T_c(i,j) + [1 - n^t(i,j)]T_o(i,j)$$

$$= n^t(i,j)[T_c(i,j) - T_o(i,j)] + T_o(i,j)$$

$$= a(i,j)\ n^t(i,j) + b(i,j) \qquad (24)$$

where a(i,j) and b(i,j) are the regression coefficients .

These are determined by comparing the transmission factors deduced from the solar radiation data measured at ground with the corresponding cloud cover index derived from satellite images . Once these coefficients are known at ground stations, methods of interpolation, e.g. kriging, may be applied in order to define the complete field of coefficients for the area studied . However practice has shown that a and b vary only little in space but more with the viewing geometry. Therefore, the operational versions of the C method treat them as constants which change only between the morning, noon and afternoon images.

11.4.4.4 Determination of hourly global radiation at ground

Estimation of the hourly global radiation follows from the definition of the transmission factor

$$HG(i,j) = T^t(i,j)HG_{\downarrow} \qquad (25)$$

where $HG_{\downarrow}$ stands for the extraterrestrial radiation incident on the top of the atmosphere $IE_{\downarrow}$ integrated over an hour.

11.5 Physical methods

One of the best known physical methods is that developed by Gautier, Diak and Masse [9], hereafter referred to as the GDM method . The main feature of this method is the formulation of two differents models for *clear-sky* and *cloudy conditions* . This method was recently (1987) considered by Marullo, Dalu and Viola [12] for the development of a new physical method (MDV) The MDV method, in analogy with the GDM method, presents two different models: a *standard atmosphere model* and a *real atmosphere model*. The *standard atmosphere model* is similar to the GDM clear-sky model, while the real atmosphere model differs from the GDM cloudy-sky model in the description of aerosol and cloud effects .

The third method to be described was developed by Möser and Raschke [10] (MR), about the same time as the GDM model. The MR method is based on radiative transfer model calculations more elaborate than the simple ones used in the other methods described here. The radiative transfer model presents, as input parameters, several climatological quantities related to the atmospheric components and to the clouds. In particular, the inclusion of physical parameters related to the clouds (e .g., cloud top heights and optical depth of clouds) in the radiative transfer model requires the use of the thermal infrared data provided by geostationary satellites .

The fourth physical method described in this review is that developed by Dedieu, Deschamps and Kerr [11] (DDK). The main feature of this method is the formulation of a unique model valid both for clear and cloudy conditions .

11.5.1 GDM method

The GDM method is based on the following three assumptions:
- water vapor is the principal atmospheric component responsible for the depletion of solar radiation by absorption;

- aerosol effects, while important, are not considered because they are complex, variable and not well known;
- scattering from the ground is assumed to be isotropic.

The GDM method uses different models for cloud-free and cloudy atmospheres.

11.5.1.1 Clear-sky model

The solar radiation emerging from the atmosphere and measured by the satellite sensor, as we already noted, is composed of the radiation emerging from the atmosphere before reaching the ground and of solar radiation scattered by the earth's surface and propagating back through the atmosphere . Thus, in the absence of clouds, the solar radiation reaching the satellite is expressed in the following way (see figure 3):

$$IE\uparrow = IE_{\downarrow}\,\alpha + IE_{\downarrow}\,(1-\alpha)[1-a(u_1)][1-a(u_2)](1-\alpha_1)A \qquad (26)$$

where:
α = scattering coefficient for beam radiation (dimensionless);
α_1 = scattering coefficient for diffuse radiation (dimensionless);
$a(u_1)$, $a(u_2)$ = absorption coefficients for slant water vapor paths, u_1 for solar zenith angle and u_2 for satellite zenith angle (dimensionless);
A = surface albedo .

The scattering coefficients α, as a function of the solar zenith angle θ, and α_1, as a function of the satellite zenith angle θ_s, have been taken from Coulson [22] and are reported in Table 4 . The absorption coefficients $a(u_1)$ and $a(u_2)$ were expressed by the analytical expression derived by Paltridge [23] as a function of atmospheric precipitable water u:

$$a(u_1) = 0{,}14(u_1/\cos\theta)^{0.44} \quad \text{for } u_1 < 0.05 \text{ cm}$$
$$a(u_1) = 0{,}099(u_1/\cos\theta)^{0.34} \quad \text{for } u_1 > 0.05 \text{ cm}$$
$$a(u_2) = 0{,}14(u_2/\cos\theta_s)^{0.44} \quad \text{for } u_2 < 0.05 \text{ cm}$$
$$a(u_2) = 0{,}099(u_2/\cos\theta_s)^{0.34} \quad \text{for } u_2 > 0.05 \text{ cm}$$

Table 4: Scattering coefficients α and α_1 for different zenith angles.

Zenith angle (deg)	α	α_1
66,4	0,095	0,076
53,1	0,074	0,076
36,9	0,060	0,076
23,1	0,055	0,076
0,0	0,053	0,076

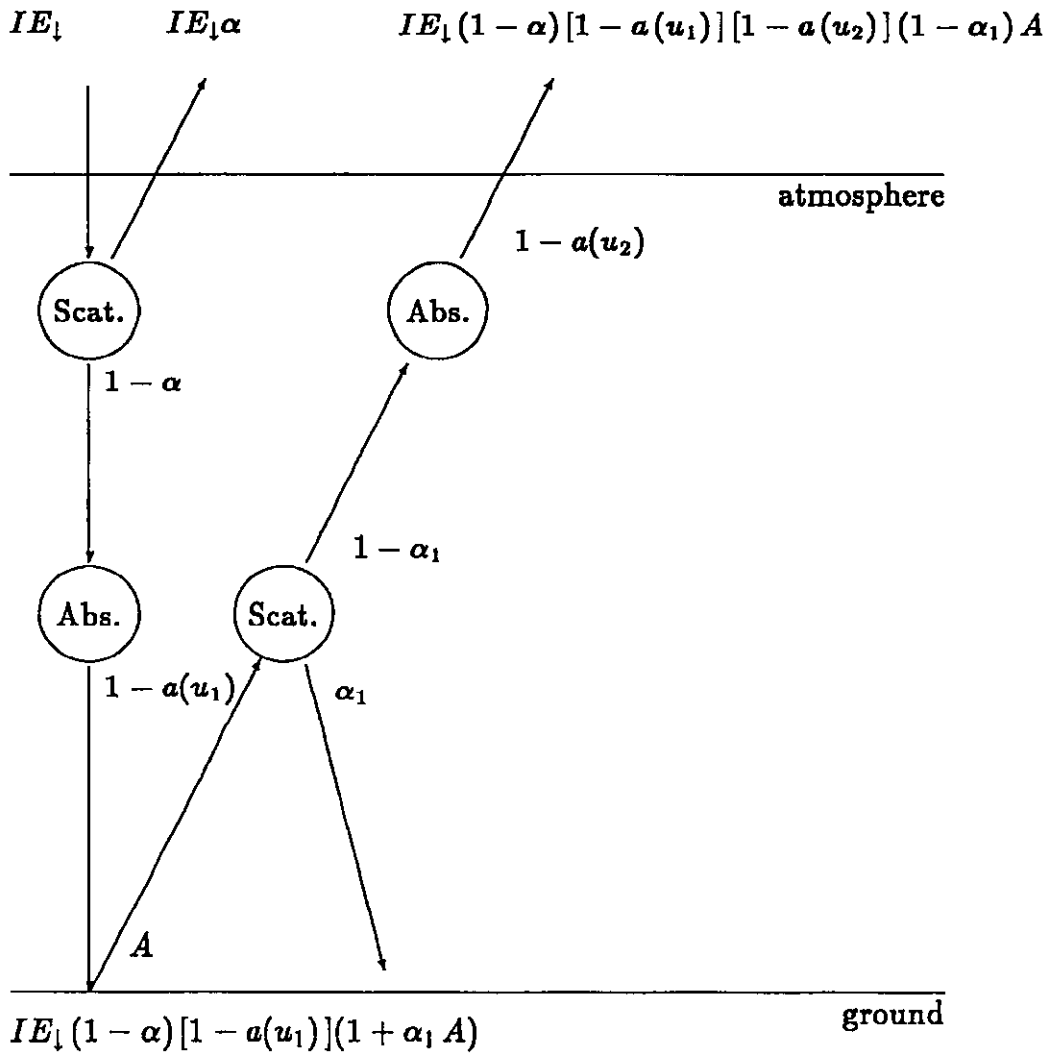

Figure 3: Scheme of the clear-sky model by Gautier, Diak and Masse (1980)

The atmospheric precipitable water u is parameterized via an empirical function of the surface dew-point temperature t_d, developed by Smith [24]:

$$\ln(u) = [0{,}1133-\ln\lambda] + 0{,}0393 t_d$$

where λ is a parameter depending on the season and the latitude.

The value of the solar radiation $IE_{\uparrow}$ can be derived, by means of a calibration procedure, from the visible satellite measurements, so that equation (26) can be solved with respect to the surface albedo A:

$$A = \{IE_{\uparrow} - IE_{\downarrow}\,\alpha\}/\{IE_{\downarrow}\,(1-\alpha)[1-a(u_1)][1-a(u_2)](1-\alpha_1)\} \qquad (27)$$

With this surface albedo, the incident solar radiation at the ground becomes:

$$IG = IE_{\downarrow}(1-\alpha)[1-a(u_1)](1+A\alpha_1) \qquad (28)$$

This original model was subsequently revised by Gautier and Frouin [25], to include ozone absorption, aerosols and multiple reflections. The aerosol effects, resulting essentially from scattering processes, have been parameterized according to Tanre et al. [26]. The formula for the solar radiation reaching the ground becomes:

$$IG = IE_{\downarrow}\{C_1 \exp(-C_2/\cos\theta)(1 - Oz_1)(1 - Oz_3)[1 - a(u_1)]\}/(1 - C_3 A) \qquad (29)$$

where

Oz_1 and Oz_3 are ozone absorption coefficients;
C_1, C_2 and C_3 are empirical constants derived from the surface visibility: the values used by Gautier and Frouin, who considered only surface visibility of 23 km and 5 km, are reported in table 5.

Table 5: Coefficients for aerosol effects proposed by Gautier and Frouin (1985)

visibility	C_1	C_2	C_3	C_4
05 km	0,91	0,1165	0,129	0,206
23 km	0,99	0,0644	0,093	0,148

11.5.1.2 Cloudy atmosphere model

The calculation of incident radiation in a cloudy atmosphere is more difficult than for clear-sky conditions. The scattering and absorption processes outlined for the clear atmosphere can take place above, between and below cloud levels. Because clouds scatter, absorb and reflect radiation in ways that even complex models describe with some difficulty, Gautier et al. [9] chose a simple treatment, valid for stratiform low and middle clouds: these clouds are the ones which attenuate solar radiation the most over space and time.

To describe the solar radiation received by the satellite under cloudy conditions, the authors consider separately the absorption above and below the clouds for both the downwelling and the upwelling paths, as depicted in fig. 4. The solar radiation received by the satellite is then approximated in the following way:

$$\begin{aligned} IE_{\uparrow} &= IE_{\downarrow}\alpha + \\ &+ IE_{\downarrow}(1-\alpha)[1-a(u_1)_t](1-\alpha_1)A_c[1-a(u_2)_t] + \\ &+ IE_{\downarrow}(1-\alpha)[1-a(u_1)_t](1-A_c)^2[1-a(u_1)_b] \times \\ &\times A(1-\alpha_1)[1-a(u_2)_t](1-abs)^2[1-a(u_2)_b] \end{aligned} \qquad (30)$$

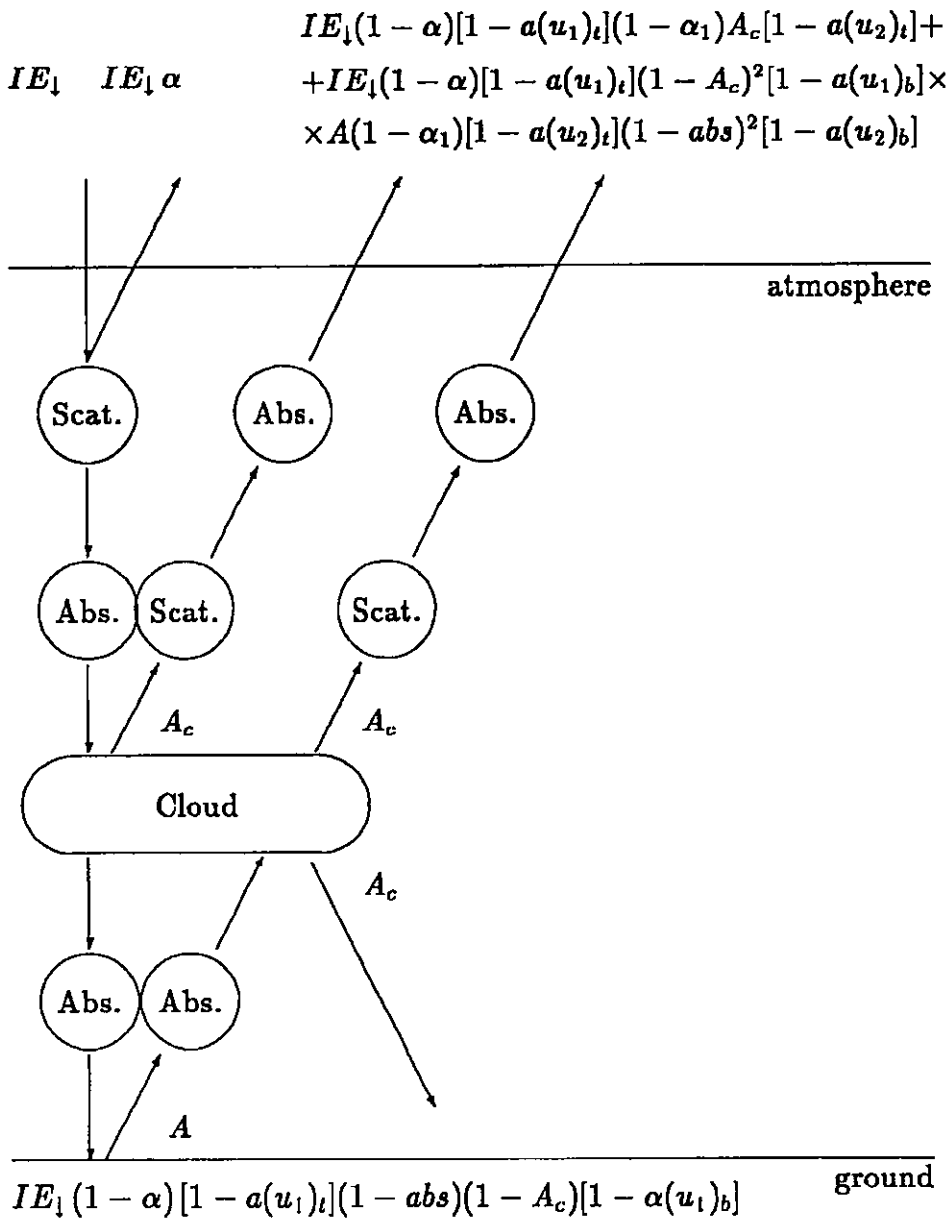

Figure 4: Scheme of the cloudy atmosphere model by Gautier, Diak and Masse (1980)

The new (dimensionless) variables introduced in this formulation are:
$a(u_1)_t$ and $a(u_2)_t$ = absorption coefficients for slant water vapor paths above clouds;
$a(u_1)_b$ and $a(u_2)_b$ = absorption corfficients for slant water vapor paths below clouds;
abs = cloud absorption coefficient;
A_c = cloud albedo.

The three terms of equation (30) represent, respectively: the shortwave radiation scattered from the atmosphere to the satellite; the shortwave radiation reflected from the cloud to the satellite; and the shortwave radiation propagating through the cloud, backscattered by the ground, through the cloud, and to the satellite (see figure 4).

For low and middle clouds the authors estimated an average of 30% of the atmospheric water vapor is above the cloud level (i.e. $a(u_1)_t \approx 0.3\ a(u1)$), thus assuming that most of the water vapor is located under the cloud base.

Cloud absorption was estimated on the basis of the visible brightness measured by the satellite that indicates the presence (high values) or the absence (low values) of clouds. The authors chose a simple linear relationship between cloud absorption and visible brightness ranging from zero for no cloud to a maximum of 0.2 for very deep clouds.

With the radiant flux $IE\uparrow$ derived from the counts measured by the satellite, and accepting that the surface albedo A is known from equation (27), equation (30) can be solved for the cloud albedo A_c. The incident shortwave at the surface is then:

$$IG = IE_{\downarrow}(1-\alpha)[1-a(u_1)_t](1-A_c)(1-abs)[1-a(u_1)_b] \quad (31)$$

After adding aerosol, ozone and multiple reflection effects (see the preceding section), Gautier and Frouin [25] expressed the shortwave radiation incident at the surface in cloudy conditions, in the following way:

$$IG = \{IE_{\downarrow}C_1 \exp(-C_2/\cos\theta)(1-A_c-abs)(1-O_{z1})(1-O_{z3})[1-a(u_1)]\}/[(1-C_3A)(1-C_4A_c)] \quad (32)$$

where O_{z1} and O_{z3} are the ozone absorption coefficients and C_1 to C_4 are the empirical constants derived from the surface visibility already introduced in the context of equation (29).

11.5.1.3 Clear air and cloudy atmosphere models

The formulation of two different models for clear and cloudy conditions requires a procedure to decide when an image pixel is clear or cloudy, in order to know what model must be applied. This is performed in the GDM model by means of a brightness threshold value determined via a minimum technique (adding a confidence margin - whose numerical value is not given in the original paper - to the minimum digital count), for every pixel and every hour, using images relative to a series of days. Indeed the effect of a cloud, within the field of view of a visible radiometer, is to increase the measured brightness. When the pixel brightness is lower than the threshold, the pixel is declared clear and the calculations are made using the clear air model. Otherwise the cloud model is used.

Raphael and Hay [13,14] claimed that the procedure of taking minimum brightness for several images at the same time of day does not take into consideration variations in albedo with changing zenith angle and season. Thus, in their application, this part of the GDM approach was replaced by the Tarpley minimum brightness determination.

11.5.2 MDV method

The MDV method also relies on two different models: a standard atmosphere model and a real atmosphere model. The first one is similar to the clear-sky model of the GDM method, from which it differs only because it describes also the effects due to a standard aerosol loading in the atmosphere. Then, deviations from this ideal case are evaluated in a simple way , by assuming the presence of a reflecting non-absorbing layer, which simulates the effects of non-standard aerosol loading or the presence of clouds.

11.5.2.1 The model for a standard atmosphere

Marullo *et al.*, in agreement with Gautier *et al.* [9], Preuss and Geleyn [27] and Chen and Ohring [28], expressed the planetary albedo for a standard atmosphere, $PA_S = IE\uparrow/IE\downarrow$ as a linear function of the ground albedo A:

$$PA_S = a + b\,A \qquad (33)$$

To estimate the planetary albedo PA_S for a standard atmosphere at several solar zenith angles, Marullo *et al.* used a series of clear sky data. Since only one clear sky day was available to the authors, the time averaged values of PA_S were obtained from a spatial ensemble average, taken over the whole Italian peninsula, assuming that the surface albedo does not dramatically change from place to place. Then, about 500 values of PA_S were combined in groups having similar solar zenith angles, and this lead to the parametrization:

$$PA_S = \alpha(90-\theta)^{\gamma} + \beta \qquad (34)$$

with: $\alpha = 0.957$, $\beta = 0.111$ and $\gamma = -0.66$

From the definition of PA_S the radiation density flux at the satellite is:

$$IE_{S,\uparrow} = PA_S\, IE_{S,\downarrow}$$

where $IE_{S,\downarrow} = 888.7\,(r_o/r)^2 \cos\theta$ [W/m^2] is the fraction of the solar irradiance at the top of the atmosphere *in the METEOSAT visible range* (0.40 - 1.10μm).

In MDV's approach, equations (34) and (33) are used to evaluate the ground albedo A directly from satellite measurements. This means that the raw data (counts) have to be converted into radiances (W m^{-2} sr^{-1}) with a known calibration function, and then into energy fluxes (W m^{-2}) by the usual assumption that upwelling radiances at the top of the atmosphere are isotropic (which means that flux = π x radiance).

Assuming A is known, the solar radiation reaching the ground in standard atmospheric conditions, IG_S, is written, following Gautier *et al.* (equation (28)), as a function of the atmospheric transmissivity $T(\theta)$:

$$IG_S = IE\downarrow T(\theta)[1+\alpha_1(\theta)A] \qquad (35)$$

with $IE\downarrow$ following equation (4) and with

$$T(\theta) = (1-\alpha)[1-a(u_1)]$$

In order to study the functional relationships $T(\theta)$ and $\alpha_1(\theta)$, Marullo *et al.* used a radiative transfer model developed by Schmetz [29]. They adopted a temperature profile given for a summer mid-latitude standard atmosphere. The water vapor content was fixed at 25 kg/m^2 with a scale height of 2 km, while the aerosol content was separated in three distinct layers with

different physical characteristics as shown in table 6. This situation is considered to occur often over the Italian peninsula under clear sky conditions.

Table 6: Three-layer atmospheric aerosol profile (surface visibility 23 km) in the MDV standard clear-sky model.

Aerosol profile
$0 \div 2\ km$ maritime
$2 \div 10\ km$ continental
$10 \div 50\ km$ stratospheric

The radiative transfer model was used to calculate $T(\theta)$ and $\alpha_1(\theta)$ for different values of the solar zenith angle θ and the results were approximated by the following empirical relationships:

$$T(\theta) = a_1\ \theta^{\gamma 1} + b_1$$
$$\alpha_1(\theta) = a_2\ \theta^{\gamma 2} + b_2 \qquad (36)$$

with the following values for the coefficients:

$$\begin{array}{llll} a_1 = 0,194\ 10^{-6} & & a_2 = 0,9\ 10^{-8} & \\ b_1 = 0,6678 & & b_2 = 0,460 & \\ \gamma_1 = 3,29 & & \gamma_2 = 4,21 & \end{array} \qquad (37)$$

Whenever the atmospheric conditions are similar to those of the standard model, the solar radiation at the surface is estimated with equations (35) and (36).

11.5.2.2 The model for a real atmosphere

In most cases real atmospheric conditions differ from the fixed standard conditions used in the clear-sky model, and these differences have to be taken into account for a correct evaluation of the real solar irradiance. Marullo, Dalu and Viola assume that any significant deviation of the flux measured at the satellite from the flux evaluated for standard conditions must mainly depend on variations in the atmospheric particle loading. The presence of these particles is simulated by a non-absorbing thin reflecting layer which is assumed to be located above the absorption and scattering mechanisms acting in the standard atmosphere (see figure 5). In fact, as we shall see, the "reflection" of this layer can even be negative.

Denoting whith $IE_{\uparrow,r}$ the radiation reflected by the artificial aerosol layer, the solar irradiance at ground becomes:

$$IG = (IE_{\downarrow} - IE_{\uparrow,r})\ T(\theta)\ [1 + \alpha_1(\theta)\ A] \qquad (38)$$

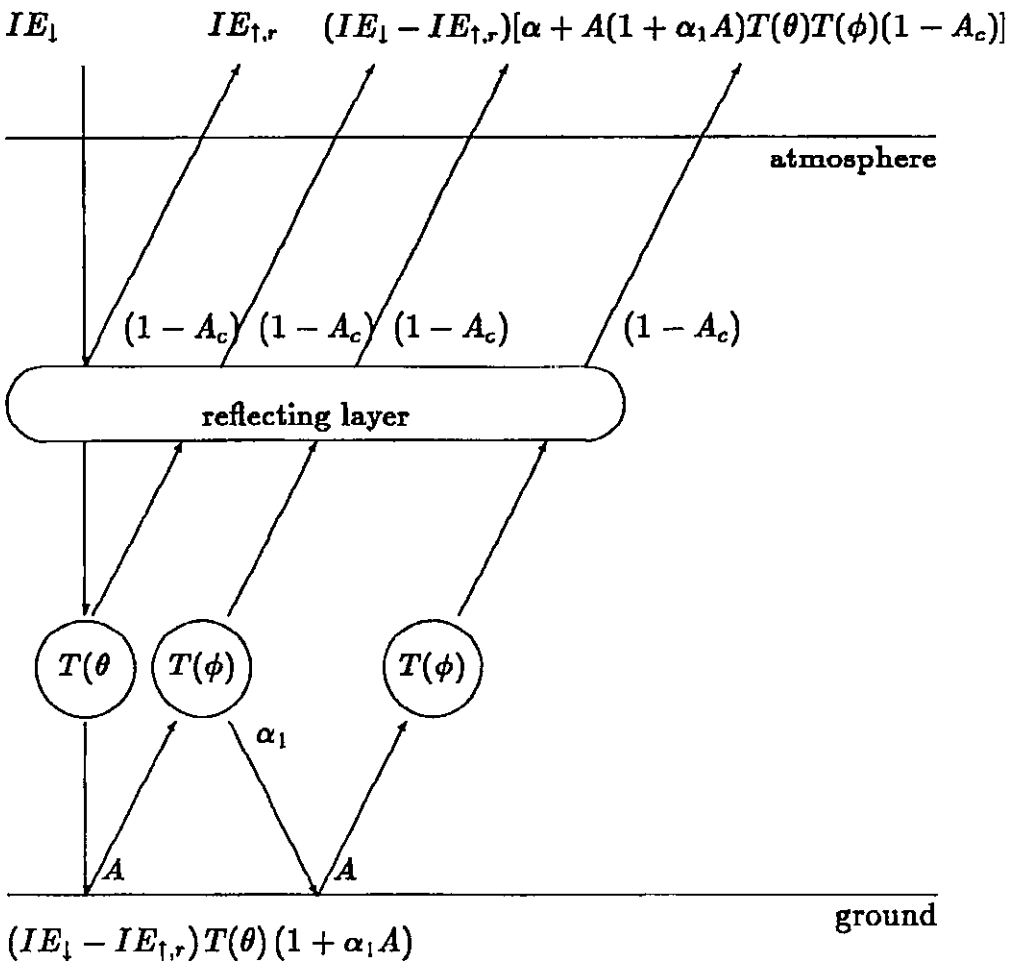

Figure 5: Scheme of the real atmosphere model by Marullo, Dalu and Viola (1987)

where $T(\theta)$ and $\alpha_1(\theta)$ are still given by equations (35) and (36).

The reflecting layer leads to a variation of the radiance measured by the satellite which reads (see figure 5):

$$IE\uparrow = IE\uparrow,r + (IE\downarrow - IE\uparrow,r)[\alpha + A(1 + \alpha_1 A)T(\theta)T(\theta_S)(1 - A_C)] \qquad (39)$$

where $T(\theta_S)$ is the total transmissivity of the atmosphere at the satellite view angle, and A_C is the albedo of the reflecting layer.

The first term in equation (39) is, by definition, the radiation reflected by the layer. The second term is the fraction of the remaining direct beam scattered back by the atmosphere. The third term is the solar radiation reflected by the ground and reaching the satellite after depletion by the fraction A_C on the ground-side of the reflecting layer.

By definition, $A_C = IE\uparrow,r/IE\downarrow$ and this value can be retrieved from equation (39) as soon as the satellite counts are converted into the flux $IE\uparrow$.

Since the METEOSAT spectral range (0.40 - 1.10 μm) does not coincide with the pyranometer's spectral range, $T(\theta)$, $T(\theta_s)$ and $\alpha_1(\theta)$ are still given by equation (36) but with:

$$\begin{array}{llll} a_1 = -0{,}30610^{-6} & a_2 = 0{,}410^{-8} \\ b_1 = 0{,}758 & b_2 = 0{,}463 \\ \gamma_1 = 3{,}21 & \gamma_2 = 4{,}44 \end{array} \qquad (40)$$

Also A_c, i.e. $IE\uparrow,_r$, is at first retrieved in the restricted spectral range. For the estimation of IG by equation (38) to hold, a conversion to the full range is necessary. Based on the radiative transfer code of Schmetz [29], Marullo *et al.* show that the "measured" A_c ought to be multiplied by about 0.9 to get the full range value.

The modulating effect of the thin reflecting layer is best understood while comparing the possible measured $IE\uparrow$ with the one valid for the standard cloudless atmosphere, $IE\uparrow,_s$:

a) if $IE\uparrow,_s = IE\uparrow$, then $IE\uparrow,_r = 0$: the atmospheric conditions are standard;

b) if $IE\uparrow,_s > IE\uparrow$, then $IE\uparrow,_r < 0$: the atmosphere is extremely clean and dry;

c) if $IE\uparrow,_s < IE\uparrow$, then $IE\uparrow,_r > 0$: either clouds or large turbidity, or both, fill the atmosphere.

Inversely, these relations govern the selection during the estimation of IG:

If the planetary albedo PA deduced from satellite images is within a 95% confidence interval around the standard atmosphere value PA_s given by equation (34), the difference is assumed to be due to small variations of the surface albedo, and the procedure for the standard atmosphere is retained. If PA is beyond the 95% confidence limit, the deviation is considered as significant, and IG is obtained from equation (38) with $IE\uparrow,_r$ from equation (39).

11.5.3 MR model

The method developed by Möser and Raschke [30] was part of the Project F of the European Community Solar Energy Research and Development Programme. This method, applied to METEOSAT images, is based on more realistic and, thus, more complex radiative transfer calculations than the other physically based methods. For instance it needs the inclusion of more parameters to describe the state of the atmosphere.

Furthermore the MR method - at least in its original formulation - also requires the use of the thermal infrared data, necessary to estimate the cloud top height, which is an input variable of the radiative transfer model. Fortunately, the MR method does not require the calibration of the satellite data.

11.5.3.1 Radiative transfer model utilisation

Möser and Raschke used the radiative transfer model developed by Kerschgens *et al.* [31], based on a two-stream approximation. The main variables used in this model are the solar zenith angle θ, the cloud top height h_c, the optical depth of clouds τ_c and the ground albedo A. Additional information required by the model includes the boundary layer structure (containing the aerosol effects), the climatological profiles of temperature, pressure, humidity and ozone and the droplet size distribution in the clouds.

Calculations performed with this model showed that solar irradiance at ground depends to a much lesser extent on the absorption by aerosols, water vapor and ozone than on cloud cover. So Möser and Raschke expressed the solar irradiance at ground in the following way:

$$IG = IG_o \times IG_n + (1 - IG_n)IG_c \qquad (41)$$

where:
IG_o = solar irradiance at ground in clear-sky conditions, when the irradiance reaches its maximum value;
IG_c = solar irradiance at ground in overcast conditions, when the irradiance reaches its minimum value IG_{min};
IG_n is the normalized irradiance at ground defined by the following equation:

$$IG_n = (IG - IG_{min})/(IG_o - IG_{min}) \qquad (42)$$

Note that IG_n can assume all the values between 1, in clear-sky conditions, and 0, in overcast conditions. Therefore, it is also called "fractional cloud cover".

As already previously stated, IG_n is mainly a function of the solar zenith angle and of the optical depth of clouds, viz.

$$IG_n = IG_n(\theta,\tau_c)$$

By analogy, it is possible to write the flux density of the solar radiation emerging from the atmosphere in the following way:

$$IE\uparrow = IE\uparrow,_{max} \times IE\uparrow,_n + (1 - IE\uparrow,_n)IE\uparrow,_{min} \qquad (43)$$

where:
$IE\uparrow,_{max}$ is the flux density of solar radiation emerging from the atmosphere in overcast conditions, when it reaches its maximum;
$E\uparrow,_{min}$ is the flux density of solar radiation emerging from the atmosphere in clear-sky conditions, when it reaches its minimum;
$IE\uparrow,_n$ is the normalized flux density of solar radiation emerging from the atmosphere.

Obviously, $IE\uparrow,_n$ is defined by the following equation:

$$IE\uparrow,_n = (IE\uparrow - IE\uparrow,_{min})/(IE\uparrow,_{max} - IE\uparrow,_{min}) \qquad (44)$$

Note again that $IE\uparrow,_n$ is equal to 0 in clear-sky conditions and to 1 in overcast conditions. It is therefore evident that the normalized irradiance IG_n, the normalized flux density $IE\uparrow,_n$ and the cloud optical depth τ_c are correlated. In particular, when τ_c increases then $IE\uparrow,_n$ also increases, while IG_n decreases and *vice-versa* when τ_c decreases. For these reasons Möser and Raschke described IG_n by means of $IE\uparrow,_n$ instead of τ_c:

$$IG_n(\theta, \tau_c) = IG_n(\theta,IE\uparrow,_n)$$

The main advantage of this formulation is that the value of $IE\uparrow,_n$ can be estimated with the visible satellite data as will be described below.

Another practical approximation introduced by Möser and Raschke is to assume that the solar irradiance at ground in overcast conditions is equal to zero, that is:

$$IG_{min} = 0 \quad (45)$$

With this assumption equation (41) may be written as:

$$IG = IG_o \times IG_n(\theta, IE\uparrow,n) \quad (46)$$

Here IG_o is completely determined by the radiative transfer model that determines also the expression of IG_n as a function of $IE\uparrow,n$.

5.3.2 Utilization of satellite data

The goal of Möser and Raschke was to correlate $IE\uparrow,n$ to the visible data provided by the satellite, while the thermal infrared data were used to estimate the cloud top height as we already noted. The visible data are used to estimate a normalized radiance L_n defined by:

$$L_n = (cB - cB_{min})/(cB_{max} - cB_{min}) \quad (47)$$

where:
c is the calibration factor (which must not be explicitly known);
B is the current digital counts provided by the satellite;
B_{min} and B_{max} are respectively the minimum (corresponding to clear-sky conditions) and the maximum (corresponding to overcast conditions) digital counts. These limits are determined by sequences of images covering 15-30 days in a way very similar to the C method.

Assuming isotropy of the radiation emerging from the atmosphere and multiplying equation (47) with π, one obtains:

$$L_n = (\pi B - \pi B_{min})/(\pi B_{max} - \pi B_{min}) = IE\uparrow,n \quad (48)$$

The solar irradiance at ground is so determined by substituting into equation (46) the value of $IE\uparrow,n$ determined by satellite data.

11.5.4 DDK method

The main feature of the method developed by Dedieu, Deschamps and Kerr [11] is the possibility of describing the solar irradiance at ground by means of a unique expression valid both for clear and cloudy conditions. The formulation was obtained by considering, initially, a model for clear-sky conditions and a model in which only the effect of clouds on solar radiation is considered, and by assuming that the attenuation of solar radiation due to the atmospheric components, except clouds, is the same both in clear and cloudy conditions.

11.5.4.1 Clear-sky model

The solar irradiance at the earth's surface in clear-sky conditions, IG_o, is modelled as follows:

$$IG_o = IE\downarrow T_o(\theta) \quad (49)$$

where $T_o(\theta)$ is the clear-sky transmission factor, accounting for gaseous absorption and for Rayleigh and Mie scattering.

Dedieu *et al.* computed $T_o(\theta)$ using the formulae of Lacis and Hansen [32] for atmospheric components together with the Tanre *et al.* [26] radiative transfer model for aerosol effects. The standard midlatitude mean values were used. Thus, water vapor content varies from 1.0 g cm^{-2} in spring to 2.5 g cm^{-2} in summer, while ozone content is assumed constant and equal to 0.35 atm cm; aerosol optical depth has been set to 0.132 at a wavelength of 1 μm.

11.5.4.2 Cloudy model

Then, Dedieu *et al.* consider a cloudy scattering atmosphere with no absorption and molecular scattering, assuming isotropy of the radiance reflected by the cloud layer and the surface. Additionally, the authors consider multiple reflections occurring between the clouds base and the ground (see figure 6). Under these conditions the planetary albedo $PA=|E\uparrow/|E|$ becomes:

$$PA = A_c + [(1-A_c)^2 A]/(1-A_c A) \qquad (50)$$

where A_c and A are the cloud and surface albedo, respectively, while the shortwave irradiance is then given by:

$$IG_c = |E\downarrow(1 - PA)/(1 - A) \qquad (51)$$

Note that equations (50) and (51) are extended over the whole solar spectrum assuming that there are no spectral changes in the scattering/reflecting properties of the clouds and of the surface.

Equation (51) clearly demonstrates that the effects of cloud variability on IG_c can be simply derived from the continuous monitoring of the planetary albedo PA as recorded by the satellites. Nevertheless, the surface albedo A must be known. It was determined by the authors (in a standard way) as the minimum of the planetary albedo observed by the satellite in a time series of images, long enough to have a good probability of cloud-free conditions.

Obviously, equation (51) must be handled with care over highly reflective surfaces, such as snow covered areas, where A may be close to 1. This is a consequence of the fact that clouds are difficult to distinguish from snow in the visible images.

11.5.4.3 Solar irradiance estimate

Assuming that molecular transmission under cloudy conditions is about the same as under clear skies, both processes, i.e., cloud effects in equation (51) and atmospheric transmission in equation (49), are combined by Dedieu *et al.* in a simplified way to derive the final equation to be used:

$$IG = |E\downarrow T(\theta)(1 - PA)/(1 - A) \qquad (52)$$

where $T(\theta)$ is set equal to $T_o(\theta)$ in practice. This assumption is justified as follows:
- most of the ozone absorption occurs above the clouds;
- saturation of water vapor bands makes water vapor absorption only weakly sensitive to the increase of water vapor in the presence of clouds (Davies et al. [33]).

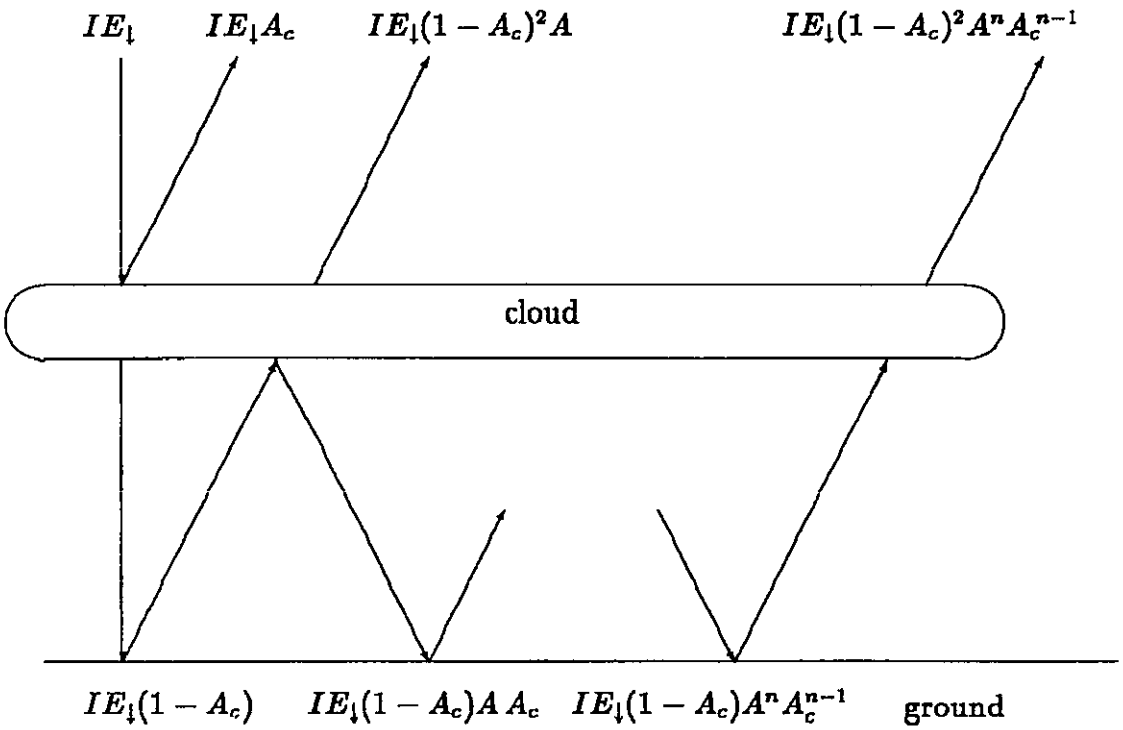

Figure 6: Scheme of the cloudy atmosphere model by Dedieu, Deschamps and Kerr (1987)

The albedos PA and A are defined by DDK as *window albedos* (i.e. albedoes observable in an atmospheric window) with no gaseous absorption, while the actual albedos are smaller because of molecular absorption.

Note that a significant aerosol effect is implicitly included in equation (52). Indeed the main aerosol effect is to increase solar radiation backscattering towards the satellite and, as a consequence, the planetary albedo PA. Therefore a strong concentration of aerosol is treated - by this model - as a cloud.

The clear sky transmission factor $T_o(\theta)$ is estimated knowing the values of the climatological parameters used in the radiative transfer model, while PA and A are determined from satellite data. PA is computed, after calibration, as the ratio between the solar radiation received by the satellite and the extraterrestrial radiation. A is determined by a minimum technique similar to that used by Gautier *et al.* [9].

11.6 Comparison between models

All the models were tested by their authors and by other scientists with experimental data . Few studies made a comparison of different models using a single homogeneous data set. A detailed assesment of three models (HH, T and GDM, in our notation) has been performed by Raphael [13] and Raphael and Hay [14].

The ground radiation data for these studies comes from a 12-station pyranometric network spanning a 45 x 70 km section of British Columbia, Canada. Twenty-one days were selected to include variable cloud cover conditions in all seasons. Only the visible data, in the 0.55 - 0.75μm region measured by GOES-2 satellite were used: these were in the form of counts on an 8-bit scale ranging from zero to 255 counts, with a resolution of four counts. The resolution of the satellite was of the order of 1.4 km x 1.4 km, but the accuracy of the earth location of the satellite imagery (the accurate alignement of points on the image to the same points on the earth's surface) was within ±2 pixels.

Czeplak, Noia and Ratto [34] compared the T and the MR models. Ground radiation data for this study comes from the 29-station pyranometric German network and encompasses daily totals for November 1986.

11.6.1 HH model

The HH model [5] was originally developed and tested using data from the tropical Atlantic: the estimates were within ±22% of the measured radiation, on an hourly basis, improving to ±8% on a daily basis.

Raphael and Hay [13,14] pointed out the inability of the regression coefficients developed for the tropical Atlantic to describe the conditions at midlatitude locations, mainly under the partly cloudy and overcast conditions, due to the generally different cloud regimes. The HH model with coefficients revised by Raphael underestimated the measured radiation on the partly cloudy days but was without significant bias for the clear and the overcast days.

11.6.2 T model

Estimates of daily totals with the T model were reported within 10% of the mean measured radiation. In the application done by Raphael and Hay [13,14], this model was shown to systematically underestimate the measured radiation for the partly cloudy and clear days. On the other hand, overestimation under overcast conditions is attributed to the inadequate handling of cloud absorption. Increases in the rms error from clear through overcast conditions were pointed out.

After Raphael and Hay's revision (regression coefficients entering in equations (15)) the bias for clear sky conditions was reduced to zero, with a concomitant decrease in the rms error to a value near ±5%. This result is - in this experiment - better than those obtained with the HH and GDM models.

Little improvement was attained for partly cloudy and overcast conditions. The value selected as a threshold for separating partly cloudy from overcast conditions is considered as very critical.

11.6.3 JPT model

In their paper [7], the authors report that their estimates are within ±16.2% of the average measured radiation on an hourly basis (7200 hourly totals from the eastern and central U.S.A. for August to December 1980) and within ±9.5% on a daily basis (about 1000 daily totals).

11.6.4 C model

The results obtained by Cano et *al.* [8] were affected by a r.m.s error of about 83 Wm^{-2} for global hourly values within sequences of about 7 days in May 1979 (29 French sites) and April 1982 (80 European sites). No average measured value is reported so that this rmse cannot be expressed in percent.

11.6.5 GDM model

In the original application of their model (184 site-days), Gautier et *al.* [9] found that clear day estimates were within 5% of the mean measured radiation, while the estimates for cloudy and completely overcast days were within 14% and 15% of the mean measured radiation, respectively. For all days, combined estimates were within 8% of the mean measured radiation.

In their application, Raphael and Hay [13,14] found that on the average this model overestimated the measured radiation under clear skies, and underestimated it under partly cloudy conditions. The model also overestimated measured radiation under overcast conditions, except in summer. The discrepancy for partly cloudy conditions is attributed (as for T model) to the problem in defining the correct cloud threshold. The overestimation in overcast conditions is again attributed to the inadequate handling of cloud absorption.

On the other hand the GDM model, applied by Raphael and Hay to a sequence of nine days demonstrated its superior performance with respect to models HH and T in simulating hourly radiation under partly cloudy and overcast conditions.

11.6.6 MDV model

Performance assessment of this method [12] was evaluated with measured daily totals for 2 days at 14 Italian sites. Estimates with the model resulted in a rmse of 8.5% of the measured average (10.4 $MJm^{-2}day^{-1}$) for a wide variety of sky conditions (14% clear sky, 43% partially cloudy, 43% overcast).

11.6.7 DDK model

The authors [11] obtained estimates within an error of ±19.5% of the measured average on an hourly basis. This error decreased by averaging the results over a period of one month. In this case the standard error was within ±6.7%, corresponding to about ±109 Wm^{-2}. The absolute errors were minimum for clear sky conditions (±50 W m^{-2}) and maximum for overcast sky (±150 Wm^{-2}) with an intermediate value (±100 Wm^{-2}) for partly cloudy skies.

11.6.8 MR model

The results obtained by Möser and Raschke [30] for hourly totals in June 1979 were affected by a standard error of about 72 Wm^{-2} (19% of the measured average) considering all sky conditions, decreasing to 60 Wm^{-2} (16%) for overcast conditions while reaching 82 Wm^{-2} (22%) for broken cloudiness. Daily totals for 81 European sites in April 1982 were reproduced with an rmse of 11.7% of the measured average of 16.48 MJm^{-2} .

Acknowledgements

This review is partly based on M. Noia's thesis ("Valutazione della radiazione solare al suolo a partire da immagini da satelliti geostazionari", Università di Genova 1990). Comments and suggestions by Dr. A. Zelenka, Swiss Meteorological Institute, Zurich, were greatly appreciated.

Nomenclature

A surface albedo of earth

A_c cloud albedo

abs cloud absorption

$a(u_1)_t$ and $a(u_2)_t$ absorption coefficients of short wave radiation above clouds

$a(u_1)_b$ and $a(u_2)_b$ absorption coefficients of short wave radiation below clouds

B brightness, i.e. digital counts measured by the satellite, proportional to the upwelling solar radiation (emerging from the atmosphere)

B_m mean target brightness, i.e. mean brightness of a pixel array [T and JPT]

B_{max} maximum brightness, giving information about the cloud albedo [T]

B_{max} threshold brightness [JPT]

B_{min} threshold brightness [JPT]

B_n normalized clear sky brightness, i.e. B_o ($\theta = 45^o$, $\theta_s = 105^o$) [T]

B_o clear sky brightness (minimum brightness)

B_o, observed clear sky brightness [JPT]

E_A solar radiation flux absorbed by the atmosphere

E_G solar radiation flux absorbed by the ground

F_{cs} solar constant

F_o hourly value for the solar constant [JPT]

HG hourly solar radiation incident on the earth's surface

$HG_{\downarrow}$ hourly solar radiation incident on the top of the atmosphere [C]

$IE_{s,\uparrow}$ flux density of the upwelling solar radiation emerging from atmosphere and received by the satellite in a standard atmosphere [MDV]

$IE_{s,\downarrow}$ flux density of the fraction of the solar radiation incident on the top of the atmosphere in the visible range (0.40 - 1.10μm) [MDV]

$IE\uparrow$ flux density of the solar radiation emerging from atmosphere and received by the satellite

$IE\uparrow,_{max}$ flux density of solar radiation emerging from the atmosphere in overcast conditions [MR]

$IE\uparrow,_{min}$ flux density of solar radiation emerging from the atmosphere in clear-sky conditions [MR]

$IE\uparrow,_n$ normalized flux density of solar radiation emerging from the atmosphere (dimensionless, $0 \leq IE\uparrow,_n \leq 1$) [MR]

$IE\uparrow,_r$ bacscattered radiation by the aerosol layer [MDV]

$IE_{\downarrow}$ flux density of the solar radiation incident on the atmosphere (Wm^{-2})

IG flux density of the solar radiation incident on the earth's surface (Wm^{-2})

IG_c flux density of the solar radiation incident on the earth's surface in overcast atmospheric conditions [C, MR and DDK]

IG_n normalized flux density of the solar radiation incident on the earth's surface (dimensionless, $0 \leq IG_n \leq 1$) [MR]

IG_s flux density of the solar radiation incident on the earth's surface in standard atmospheric conditions [MDV]

IG_o flux density of the solar radiation incident on the earth's surface in clear atmospheric conditions [C and MR]

m optical air mass

n cloud cover index ($0 \leq n \leq 1$)

PA planetary albedo, i.e., $IE\uparrow/IE\downarrow$

PA_s planetary albedo in standard atmospheric conditions, i.e., $IE_{s,\uparrow}/IE_{s,\downarrow}$ [MDV]

r actual sun-earth distance

r_o mean sun-earth distance

R^t radiance measured by the satellite in the thermal infrared spectral band at time t [C]

R_c radiance of the cloud tops measured by the satellite in the thermal infrared spectral band [C]

R_s radiance of surfaces covered by snow or ice, measured by the satellite in the thermal infrared spectral band [C]

SR visible radiance, another name for the planetary albedo [HH]

T_c atmospheric transmittance for downwelling solar radiation for overcast sky conditions [C]

T_r visible transmission due to Rayleigh scattering [T]

T^t atmospheric transmittance for downwelling solar radiation at time t [C]

T_{wa} visible transmission after water vapor absorption [T]

T_{ws} visible transmission due to water vapor scattering [T]

$T(\theta_s)$ transmittance for upwelling solar radiation, i. e., $IE\uparrow/(AxIG)$ [MDV, GDM]

$T(\theta)$ atmospheric transmittance for downwelling solar radiation, i. e., $IG/IE\downarrow$ [MDV,GDM]

T_o atmospheric transmittance for downwelling solar radiation for clear-sky conditions [C]

T_1 clear sky/partly cloudy sky threshold [T]

T_2 partly cloudy sky/cloudy sky threshold [T]

u_1 slant water vapor paths for solar zenith angle [GDM]

u_2 slant water vapor paths for satellite zenith angle [GDM]

α atmospheric scattering coefficient for beam radiation [GDM]

α_1 atmospheric scattering coefficient for diffuse radiation [GDM]

ϕ azimuth angle between sun and satellite

θ solar zenith angle

θ_s satellite zenith angle

ρ^t reference albedo, i.e., $IE\uparrow/IG_o$ at time t [C]

ρ_c average reference albedo of the cloud tops [C]

ρ_o reference albedo for clear sky conditions [C]

References

[1] U.S. Department of Commerce. *Studies of Cloud and Satellite Parametrization of Solar Radiation at the Earth's Surface*, 1971. Proc. Miami Conference on Remote Sensing.

[2] T.H. Vonder Haar. Solar Insolation Microclimate Determined Using Satellite Data. *In Solar Energy Data Workshop NOAA,* 1973.

[3] T.H. Vonder Haar and J.S. Ellis. Solar Energy Microclimate as determined from Satellite Observations. *Optics in Solar Energy Utilization*, **68**:18 - 28, 1975.

[4] J.S. Ellis and Vonder Haar. "Solar Radiation Reaching the Ground Determined from Meteorological Satellite Data". *Technical Report NTIS N79-27770/3GA, NASA,* 1979.

[5] J.E. Hay and K.J. Hanson. "A Satellite-Based Methodology for Determining Solar Irradiance at the Ocean Surface During Gate". *Bullettin of the American Meteorological Society*, **59**:1549, 1978.

[6] J .D. Tarpley. "Estimating Incident Solar Radiation at the Surface, from Geostationary Satellite Data". *Journal of Applied Meteorology,* **11**:72-1181, 1979.

[7] C. Justus, M.V. Paris, and J.D. Tarpley. "Satellite-Mesured Insolation in the United States, Mexico, and South America". *Remote Sensing of Enviroment*, **20**:57-83, 1986.

[8] D. Cano, J.M. Monget, M. Albuisson, H. Guillard, N. Regas, and L. Wald. "A Method for the Determination of the Global Solar Radiation from Meteorological Satellite Data". *Solar Energy* **37**:31-39,1986.

[9] C. Gautier, G. Diak, and S. Masse. "A Simple Physical Model to Estimate Incident Solar Radiation at the Surface from GOES Satellite Data". *Journal of Applied Meteorology,* **19**:1005-1012, 1980.

[10] W. Möser and E. Raschke. "Mapping of Global Radiation and of Cloudiness from Meteosat Image Data. Meteorol. Rdsch." **36**, April1983.

[11] G. Dedieu, P.Y. Deschamps, and Y.H. Kerr. "Satellite Estimation of Solar Irradiance at the Surface of the Earth and of Surface Albedo Using a Physical Model Applied to Meteosat Data". *Journal of Climate and Applied Meteorology* **26**:79-87, 1987.

[12] S. Marullo, G. Dalu, and A. Viola. "Incident Short-Wave Radiation at the Surface from Meteosat Data". *Il Nuovo Cimento,* **10C**:77-90, 1987.

[13] C. Raphael. "Models for Estimating Solar Irradiance at the Earth's Surface from Satellite Data: an Initial Assessment". Technical Report, Atmospheric Environment Service, Downsview Ontario, 1983. (Unpublished manuscript).

[14] C. Raphael and J.E. Hay. "An Assessment of Models which Use Satellite Data to Estimate Solar Irradiance at the Earth's Surface". *Journal of Climate and Applied Meteorology,* **23**:832-844, 1984.

[15] W.E. Shenk and V.V. Salomonson. "A Simulation Study Exploring the Effects of Sensor Spatial Resolution on Estimates of Cloud Cover from Satellite". *Journal of Applied Meteorology,* **11**:214 220, 1972.

[16] J.A. Davies, W. Schertzer, and M. Nunes. "Estimating Global Solar Radiation". *Boundary Layer Meteorology*, **9**:33-52, 1975.

[17] J .E. Mc Donald. "Direct Absorption of Solar Radiation by Atmospheric Water Vapor". *Journal of Meteorology*, **17**:319-328, 1960.

[18] F. Kasten. "A New Table and Approximation Formula for the Relative Optical Air Mass". *Arch. Meteor. Geophys. Bioklim.*, **B14**:206-223, 1966.

[19] D. Cano. "Étude de l'Ennuagement par Analyse de Séquences d'Images de Satellite. Application à l'Évaluation du Rayonnement Solaire Global au Sol". PhD thesis, École Nationale Supérieure des Télécommunications, 1982.

[20] G. Moussu, L. Diabaté, D. Obrecht and L. Wald. "A Method for the Mapping of the Apparent Ground Brightness Using Visible Images from Geostationary Satellites". *Int. J. Remote Sensing*, **10**:1207-1225, 1989.

[21] G. Bourges. "Courbes de Fréquence Cumulées de l'Irradiation Solaire Globale Horaire Recue par une Surface Plane". Technical Report, Centre d'Énérgetique de l'École Nationale Supérieure des Mines de Paris, July 1979.

[22] K.L. Coulson. "Characteristics of the Radiation Emerging from the Top of a Rayleigh Atmosphere, 1 and 2". *Planet. Space Sci.*, **1**:256-284, 1959.

[23] G .W. Paltridge. "Direct Measurements of Water Vapor Absorption of Solar Radiation in the free Atmosphere". *Journal of Applied Meteorology*, **30**:156-160, 1973.

[24] W.L. Smith. "Note of the Relationship between Total Precipitable Water and Surface Dew Point". *Journal of Applied Meteorology*, **5**:726-727,1966.

[25] C. Gautier and R. Frouin. "Downward Longwave Irradiance at the Ocean Surface Using Satellite Data." In *Proc. of the 3rd International Collection on Spectral Signature of Objects in Remote Sensing,* Les Arcs, France, 1985.

[26] D. Tanré, M. Herman, P.Y. Deschamps, and A. De Leffe. "Atmospheric Modelling for Space Measurements of Ground Reflectances, Including Bidirectional Properties". *Applied Optics*, **18**:3587-3594, 1979.

[27] H.J. Preuss and J.F. Geleyn. "Surface Albedos Derived from Satellite Data and their Impact on Forecast Models". *Arch. Meteorol. Geophys. Bioklimatol.*, **A 29**:345-356, 1980.

[28] T.S. Chen and G. Ohring. *Journal of Atmospheric Science*, **41**:158, 1984.

[29] J. Schmetz. "On the Parametrization of the Radiative Properties of Broken Clouds". *Tellus A.*, **36**:417, 1984.

[30] W. Möser and E. Rasche. "Incident solar radiation over Europe estimated from meteosat data". *Journal of Applied Meteorology*, **23**:166-170, 1984.

[31] M. Kerschegens, U. Pils, and E. Raschke. "A Modified Two Stream Approximation for Computations of The Solar Radiation Budget in a Cloudy Atmosphere". *Tellus*, **30**, 1978.

[32] A. Lacis and J.E. Hansen. "A Parametrization for the Absorption of Solar Radiation in the Earth's Atmosphere". *Journal of Atmospheric Science*, **31**:118-133, 1974.

[33] R. Davies, W.L. Ridgway, and K. Kyungeak. "Spectral Absorption of Solar Radiation in Cloudy Atmosphere: a 20 cm^{-1} Model". *Journal of Applied Meteorology*, **41**:2126-2137, 1984.

[34] G. Czeplak, M. Noia, and C.F. Ratto. "An Assessment of a Statistical Method to estimate Solar Irradiance at the Earth's Surface from Geostationary Satellite Data". *Renewable Energy*, 1:737-743, 1992.

Appendix

List of Participating Experts

Gerhard Czeplak
Deutscher Wetterdienst, Meteorologisches Observatorium Hamburg
Frahmredder 95
D - 22393 Hamburg 65
Germany

Vito D'Agostino
SASIAM, Tecnopolis CSATA Novus Ortus
P.O. Box 775
I - 70010 Valenzano / Bari
Italy

Weine Josefsson
Swedish Meteorological and Hydrological Institute
S - 60176 Norrköping
Sweden

Eugene Maxwell
National Renewable Energy Laboratory
1617 Cole Boulevard
Gloden, CO 80401-3393
U.S.A.

Richard Perez
Atmospheric Siences Research Center
State University of New York at Albany
100 Fuller Road
Albany, NY 12205
U.S.A.

Antoine Zelenka (Subtask Leader)
Swiss Meteorological Institute
Krähbülstrasse 58
CH - 8044 Zürich
Switzerland

Temporary Expert:

Corrado Ratto
Università di Genova
Dipartimento di Física
Via Dodecaneso 33
I - 16146 Genova
Italy