The process of integration of manual and automatic meteorological networks in the design of the Information System of Environmental Climatology of Andalusia (SICA). Problems associated with the selection of variables, validation of data entering the system and interpolation of gaps.

Comunicación presentada en III International Conference on Experiences with Automatic Weather Stations, Torremolinos, 19-21 Febrero 2003

Mª Fernanda PITA and Juan Mariano CAMARILLO
Department of Physical Geography and Regional Geographical Analysis
University of Seville
c/María de Padilla s/n, 41004. Seville (Spain)

1. Introduction

The Climatic System as a whole, as well as the values reached by the diverse variables which comprise it, are involved in a great deal of different physical-environmental processes which make up the natural system of a region. Losses on account of soil erosion are directly related to the intensity of the rainfall, and processes of atmospheric pollution can be intensified by situations of thermal inversion by stability. The appearance of certain forest pests is directly related to the thermal integral according to very well-known threshold values, and the evolution of forest fires can be modelled from meteorological data of temperature, wind direction and intensity. These are some examples which show the fundamental role played by climate in the multitude of processes which characterize the environmental functioning of a region.

Taking into account the multitude of processes which climate is subject to, it would seem necessary to create a cross tool for the handling of climatic and meteorological information, which could serve the various users involved in Environmental management. The setting up of this tool in Andalusia has been promoted by the Environment Commission (CMA) of the Local Government of Andalusia. To this effect, an agreement has been signed with the Ministry of Agriculture and Fisheries (CAP) of the Local Government of Andalusia and the National Institute of Meteorology (INM), which forms part of the Ministry of the Environment. The objective is to exchange and share existing climatic information in each one of these organisms. This shared information will be integrated in and managed by the Information System of Environmental Climatology of Andalusia (SICA).

The basic determining factors imposed on the System are as follow:

A. The need to integrate data from the main observation networks in Andalusia, even if they are heterogeneous in design and structure.

B. Transversality, in the sense that the variables selected and the construction processes for new variables (added variables), quality control of the data, as well as the final analytical products, should satisfy a multitude of users involved in management and environmental research in the Community. These users cover a wide spectrum, in which we can find bodies like the services of Planning and Organisation of Natural Resources, Forest Management, Agricultural Planning, Extinguishing of Forest Fires or Civil Protection.

C. Universality, in the sense that the aim of the System is to integrate as many climatic variables as possible. In addition, it has been designed with the whole group of potential users in mind, from those who belong to the area of environmental management, to Andalusian researchers, and, lastly, to the general public through the connection of the System to the propagation tools which the CMA has at its disposal.

D. Flexibility and openness of the System, since it has to guarantee the implementation of its own internal management tools to allow for the integration of new networks for observation, new stations, variables, formats, interpolating techniques, analysis or users. A great deal of the merit and usefulness of the System is to be found precisely in this principle. Through this, it can be adapted permanently to the diverse needs and technological innovations which may appear. Nevertheless, the integration of this principle implies a considerable effort since it requires a qualified System administrator who, accordingly, becomes the delegate between the management requirements and the mechanisms for acquisition, control and use of the meteorological and environmental information.

E. The quality of the data integrated in the System. Regardless of the control processes which may be implemented by each one of the organisms producing and supplying information, the design of the System introduces a whole range of methods and techniques to control outliers, aberrant and illogical data, which are univocal for each one of the variables in their different time scales.

All these determining factors or requisites of the System have shaped each one of the phases needed for its design and computer installation. In this paper we will only aim to describe some of the necessary tasks for the selection of the variables which make up the System, as well as the methods used to validate information and interpolate gaps.

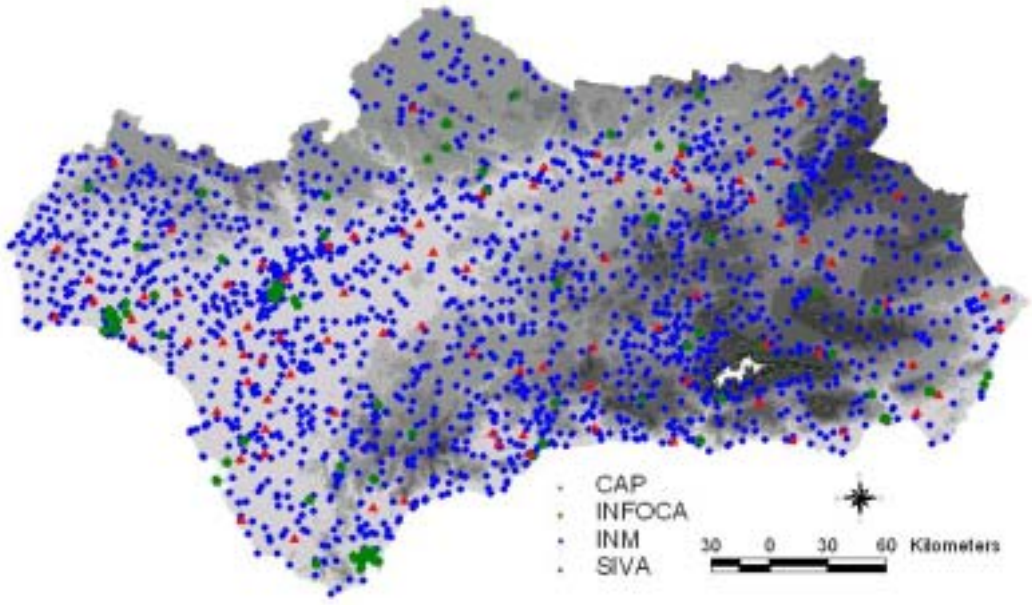
Before proceeding, we wish to emphasise the idea that the main difficulties in this task have arisen from two sources. On the one hand, they arise from the great diversity existent in the format of the source observations, especially the differences registered between data from automatic observation stations and conventional or manual stations. The measures taken have been based, firstly, on the adoption of a standard common format for each and every one of the selected variables. The second measure is the genesis of the System's own administration tool which allows for the inclusion of new formats, and their conversion into the common internal standard format. This solution simply requires knowledge of the original format of the incoming data, which is provided by the organisms producing the information.

Moreover, difficulties arise from the fact that the homogenising, validation and interpolation processes have to be developed in order for them to be applied routinely to a vast and complex system, with many different variables, in very different time scales and in a territory as diverse as the one existent in Andalusia. This limits the possibilities, as it rules out a great deal of precision, and leads to a certain degree of abstraction.

2. Information included in the SICA.

The System integrates meteorological and climatic information from three different observation networks in operation throughout the Autonomous Community. The majority of the information comes from the National Institute of Meteorology, with a regional spread of almost 2000 manual and 50 automatic stations, which are representative of regional climatic variability. In addition to these observation stations, there are those belonging to the Ministry of Agriculture and Fisheries (CAP) of the Local Government of Andalusia, with 170 automatic stations situated in areas devoted to agriculture, and especially those occupied by irrigation areas. The Ministry of the Environment (CMA) of the Local Government of Andalusia has 100 stations directed towards the prediction of forest fires, control of forest areas and monitoring of air quality (see figure 1).

FIGURE 1: NETWORKS OF WEATHER STATIONS IN ANDALUSIA



The evident predominance of manual stations existing within the National Institute of Meteorology and their almost total exclusivity where series of historical observation are concerned, has motivated us to use their formats as a source of inspiration for the design of our own system. But this also integrates the features to be found in the formats of the automatic stations, since these are destined to play a leading role in meteorological and climatic information in the near future. On the other hand, the System is open and flexible enough to make room at a later date for the new requisites for handling meteorological and climatic information which may arise in this field.

3. Selection of variables in the SICA.

3.1. Variables selected

Taking as our basis the variables observed in the different services which have supplied the SICA, we have used as our starting-point the inclusion of a total of 11 magnitudes, divided into 39 different variables. If we take into account the different time scales

which each one of the variables may adopt and we incorporate some parameters that may be derived from them, we obtain a total of 651 different observation series. These may be considered as variables, and are in fact the variables which comprise the system we are now going to present (see table 1).

Table 1. - Integrating variables of the SICA

MAGNITUDES	NUMBER OF VARIABLES
Air temperature	48
Soil and sub-soil temperature	75
Rainfall	67
Atmospheric humidity	64
Cloudiness	22
Insolation	7
Evaporation	24
Evapotranspiration	12
Atmospheric pressure	20
Wind	204
Solar radiation	108
TOTAL	651

These figures clearly indicate that this is a very broad and detailed system including redundant information. In an attempt to save space, a good number of computing systems opt to eliminate redundant information and include the necessary algorithms to obtain data from the primary information included in the system where necessary. In our case, it has not been our main aim to save space in the system, and yet it has been our aim to provide the various users with a fast and easy access to consultations. Another key factor is to have fast and easy development of diverse and multiple applications. Logically, these facilities for consultation and application are counteracted by greater difficulties in the design of the system structure and its management. In the case of Andalusia, it has been estimated that our own server will be required to satisfy the physical demand for storage of the on-line data for a three-year period included in the control, pooling and analysis of data. Afterwards the data will then be stored on diskette.

Once this general principle was taken on board, the definitive selection of the variables to be considered in the system was carried out. The aim has been not to lose any of the information existent in the traditional observation networks, and to emphasise the variables which are seen to have greater promise for the future, even when they were not observed in depth in the traditional networks. This would be the case of the variables related to solar radiation or wind, for example.

3.2 Selected units of measurement

In general, the units selected for each variable have corresponded to those most frequently used in current observation networks, but the necessary algorithms to translate other possible units existing in other networks to these units have also been implemented. This process of transformation of units has been necessary on numerous occasions given that there are many formats existent in the different networks in this sense (see table 2).

Table 2. – Variables which incorporate mechanisms for conversion of units of measurements.

VARIABLES	UNIT SELECTED IN THE SICA
Variables associated with air temperature	ºC
Variables associated with total rainfall	mms
Variables associated with atmospheric pressure	HPa
Variables associated with vapour pressure	HPa
Variables associated with cloudiness	Octas
Variables associated with insolation	Minutes
Variables associated with wind direction	Degrees
Variables associated with wind speed	Kms/h
Variables associated with solar radiation	W/m^2 o KJ/m^2
Variables associated with hydrometeors	Logical field

3.3. Time scales

The most commonly used observation scales in traditional networks (daily, monthly and annual) have been maintained, and have been extended only by the inclusion of the total rainfall, which has also been developed to the scale of the hydrological year. The greatest problems have arisen in the "intradays" scale, in which very different frequencies of observations appear, such as ten minutes frequencies, which is the most common, but also 15 or 30 minutes, as well as the three or four hourly observations which may appear in conventional stations like the network of complete stations of the INM or the Synops network.

The SICA has opted for the ten-minute scale for "intradaily" data and for it to be pooled subsequently within the daily scale. As a consequence, all the variables observed with different frequencies have undergone processes of adaptation. In this sense, there are three types of different variables; firstly, the quantitative variables where pooling is achieved by using arithmetical average (temperature, humidity, etc); secondly, qualitative or dichotomic variables (cloud types, which are codified, or some meteors, for which it is only registered whether or not they were detected), and thirdly quantitative variables which are pooled by means of accumulation (total rainfall, evapo-

transpiration, etc). Each one of these requires different pooling methods, which are particularly difficult in the case of the latter.

It is useful to point out the processes required to adapt the daily variables of the INM itself, which ,in some cases, such as rainfall, have different systems for daily pooling (from 0 – 24 hours, or 7 o'clock one day to 7 o'clock the next) between which it was necessary to choose (in our case, we have opted for the conventional day, namely, from 0 to 24 hours of a day), as well as the necessary processes to adapt the three or four hourly variables of the Institute itself.

4. Methods for validation of data.

In the process of validation of data, the main aim of not to lose any potentially valid information has prevailed. To this end, the methods to be applied have been designed with sufficiently broad and open criteria so as to avoid the possibility of eliminating any accurate data, while accepting the risk that erroneous data could enter the system. On the other hand, under no circumstances are erroneous data eliminated from the system, but instead they are marked out as such, and they do not figure in subsequent treatments, but they are substituted by gaps or "no data". Furthermore, in spite of the automation of the process, this first phase of validation includes a second control process in the form of an *incident report*. This report, which should reflect all the errors and invalid data automatically detected, is only to be seen by the qualified system administrator, in such a way that he/she may evaluate the ultimate validity of the processes which have been carried out automatically. In the same way, this incident report should be stored directly and automatically by the System itself, thereby guaranteeing future access to it.

Another fundamental determining factor in the design of these methods was the fact that they are meant to be applied routinely and in real time (as the data enters the system) on a large number of very diverse variables, which, in their turn, develop in a very diverse territory. This reduces the number of variables on which they will be applied to only those which guarantee a reasonable success rate due to their specific characteristics. Moreover, it re-enforces the need to endow the methods with a certain flexibility, and determines whether these are absolute and not relative tests in each case, namely, proofs that may not use information from neighbouring observation stations, but are limited to using the information existent in their own station. On the other hand, this also reduces to a degree the potentiality of the total process, whose sole aim is to eliminate the worst mistakes and the most striking outliers, but in no way aims to guarantee a data bank which is absolutely free from errors, or exempt researchers or system managers from applying more precise and specific validation processes when they may be necessary.

The validation methods applied may be grouped into three different types:

A. *Ranges.* These assume that each variable is assigned with a range of possibility, which has been inspired by the knowledge of the natural variability of each one of them in the region. The historical minimums and maximums of each one of the variables have been the basic instrument to assign them with generous, but realistic filters. However, we are aware that in some of the particularly changeable variables throughout the region (for example, rainfall), diversity imposes very open ranges, thereby reducing the potentiality of the control. In the future, nothing will prevent each one of the

observation stations from being assigned with specific ranges based on individual behaviour. So far, this task has been impossible and, on the other hand, we have deliberately avoided establishing ranges inspired by standard deviations in the series (values exceeding three or four standard deviations are commonly used as a limit) given that many of them did not comply to normal curves, which reduced guarantees and the method's own potentiality.

The ranges established for variables associated with the insolation magnitude are worth a special mention. This is because the potential maximum hours of sunlight varies throughout the year in their case, and is determined by the latitude of the observation station. Moreover, in the "intradaily" variables, it is necessary to take into account the interval of time on which the observation is based, which, on occasion, turns out to be the real limit – in a ten minute series, the highest value which the insolation can reach is ten minutes.

Also worthy of mention are the variables which implicitly contain a time dimension (such as the intensity of the rain, which is defined as the total rainfall per unit of time), for which it was necessary to establish flexible ranges so that they could adapt to any time interval of measurement. (In the case of rainfall, for instance, we established a generic maximum limit of 400 mms/h, which would have to be adapted and would vary, logically, depending on the time interval under consideration. ten-minute rainfall, hourly, daily, etc.)

B. *Logical filters*: These receive their name from the fact that they take their inspiration from basic principles of logic and, more precisely, from the principle of non-contradiction, and their exact function is to prevent any of the data entering the system from defying this principle. Applying this, and being aware of the functioning of the variables and the processes of receiving data, it is easy to detect any aberrant data, such as the existence of values for negative rainfall, or days when the maximum temperature is lower than the minimum, or rainy days with a total absence of cloudiness, etc. Logical filters have not been applied to all the variables because not all of them were adapted to this type of method, although a great number of them were (see table 3).

C. *Increases:* Finally, in the case of some variables, filters have been created by establishing a limit in the increase experienced by these variables using previous observations with regard to the observation being considered. The variables which best adapt to this type of method are the continuous ones which show a certain inertia, such as the intradaily temperature. In this case, a top value can be established in the increase experienced from one period of observation to the next in such a way that the increases which exceed this threshold may be considered symptomatic of measurement errors. In the case of variables in which inertia is even more noticeable, such as the case of the temperature of the subsoil, these limits in increase can be fixed, even with regard to values registered the day before. (See table 3)

Table 3. – Variables with validation methods in addition to the method of range assignment

METHODS OF VALIDATION	VARIABLES

Logical filters	Average daily temp. Minimum daily temp. Maximum daily temp. Total rainfall for 10 mins. of obs. Precipitation in the form of rain Precipitation in the form of snow Precipitation in the form of hail Temp. of the damp therm. Temp. of the dewfall
Maximum increase from the previous observation	Temp. of the dry therm.
Maximum increase from the previous day	Temp. at 0.05 m below ground Temp. at 0,12 m below ground Temp. at 0,15 m below ground Temp. at 0,20 m below ground Temp. at 0,15 m above ground
Variable range according to the day of the year	Intradaily insolation Daily insolation Monthly insolation

If we bear in mind that, in addition to these filters, all the variables are limited by specifics ranges, it could lead us to think that many of the possible errors which may be registered in the data would be detected before entering the system.

5.– **Methods for interpolation of gaps**.

In spite of the difficulty and risk implicit in the design of gap interpolation methods for routine application to very heterogenous data, it has seemed appropriate to include some in the system, given that there are several occasions when continuity is required in the observation series for later treatments, or to obtain the derived parameters. Nevertheless, they have only been used when there was a high guarantee of reliability, and the application benefits were very clear. On the other hand, the existence of any interpolated data is always known about in advance.

The recommended methods of interpolation depend on the time scale under consideration, and the guidelines for spatial/temporal behaviour of the variable in question. In general, for monthly and annual data, in which spatial variability is by now somewhat reduced, and as long as the variables lend themselves to it, it is recommendable to interpolate from the values adopted by the better correlated variables with the variable to be interpolated. The specific procedure for interpolation consists of the substitution of the missing data by the weighted average of the values inferred from the three better correlated series with the one which is to be interpolated, as long as there is a minimum of 10 pairs of values and the Pearson correlation coefficients reach levels higher than 0,75.

For the majority of the intradaily data, the polinomic functions are adjusted to the daily cycle of the variable in question. In the case of daily data (and in some cases intradaily data), interpolation is carried out by means of the arithmetical average of the values obtained by the same variable in the days preceding and following the missing data (see table 4).

Anyway, very few variables are subject to interpolation of gaps as we are aware of the wide margins of error in which the majority of these interpolations are to be found. Hence they are only recommendable in cases where the risk/benefit analysis is very favourable to the latter.

Table 4. – Methods of interpolation of gaps assigned to the SICA variables.

METHODS OF INTERPOLATION	VARIABLES
Arithmetical average of the two previous intradaily values and the two subsequent to the value to be interpolated.	Temp. 0,05 below ground level Temp. 0,10 below ground level Temp. 0,15 below ground level Temp. 0,20 below ground level Temp. 0,15 above ground level Dry therm. temp.(2º method.)
Arithmetical average of the two previous daily values and the two subsequent to the value to be interpolated	Average daily temp. 0,05m below ground level Average daily temp. 0,10 m below ground level Average daily temp. 0,15 m below ground level Average daily temp. 0,20 m below ground level Average daily temp. 0,15 m above ground level Average daily temp Minimum daily temp Maximum daily temp. Daily evaporation in Piché Daily evaporation in tank Daily evapotranspiration in lysimeter
Order 4 polinomic function for intradaily data which follow daily cycles.	Temp. of dry thermometer Evaporation with Piché Evaporation in tank Evapotranspiration in lysimeter
Weighted average of inferred values from the three best correlated series.	Average monthly and annual temp. Average temp. of the monthly and annual min. Average temp. of the monthly and annual max. Average monthly and annual daily temp. increase Total monthly and annual rainfall Number days rainfall per month and year Monthly and annual Piché evaporation Monthly and annual tank evaporation Monthly and annual lysimeter evapotranspiration

6. Conclusions

Climatic information by its very nature, and, especially, on account of its extraordinary abundance, presents enormous management difficulties. Such difficulties are increased when the management systems try to introduce quality control processes for data, or interpolation of gaps Above all, this is the case if these processes have to be developed routinely and have to be linked to other processes to obtain derived data or the carrying out of diverse applications. What is more, when the formats of the data entering the system are different, not only are the difficulties increased, but also the possibility of committing errors rises considerably. On account of all this, we are extremely grateful for the attempts being made at the moment to create basic common norms for measurement and treatment of the meteorological and climatological information. These norms, as well as subscribing to the conditions in which observations are carried out, should also pay special attention to the time scales for observation, units of measurement, and treatment given to erroneous or missing data.