Research, Digitisation, and Homogenization of Long-Term Data Series

Jean-Marc MOISSELIN[1], Olivier MESTRE[2]
[1]*Météo-France, Direction de la Climatologie*
42 av. G. Coriolis
31057 Toulouse Cédex
France
jean-marc.moisselin@meteo.fr
[2]*Météo-France, Ecole Nationale de la Météorologie*
42 av. G. Coriolis
31057 Toulouse Cédex
France
olivier.mestre@meteo.fr

Résumé - Le programme de **recherche en données anciennes** entamé en 1994 à Météo-France a permis l'enrichissement du patrimoine climatologique français, plus spécialement en moyennes mensuelles de températures minimales et maximales et en cumuls mensuels de précipitations. L'effort de saisie a essentiellement porté sur la période 1880-1950 jusque là pauvre en données. L'étude des changements climatiques à partir des séries brutes est hasardeuse en raison des nombreuses ruptures dues aux déplacements de postes, changements de capteurs ou modifications de l'environnement de mesure. L'**homogénéisation** des longues séries apparaît dés-lors comme une étape indispensable avant d'établir les tendances à long terme. Les outils d'homogénéisation, s'appuyant sur les méthodes statistiques développées à la Direction de la Climatologie, permettent la détection et la correction des ruptures. Ils peuvent être employés pour d'autres jeux de données que des séries climatologiques. On dispose maintenant de 70 séries mensuelles de températures démarrant avant 1900 et de 226 séries mensuelles de cumuls de précipitation. Ces séries sont utilisées pour la détection des changements climatiques et la validation des modèles régionaux de climat.
Ce type de données apparaît comme une **archive vivante** qui s'améliorera au rythme de la recherche en nouvelles données anciennes et de l'amélioration des méthodes d'homogénéisation.

Abstract - The **historical data rescue program** engaged since 1994 by *Météo-France* has allowed the enhancement of French climatological heritage, especially for monthly averages of daily minimal and maximal temperature and monthly rainfall. The digitisation effort was mainly devoted to the 1880-1950 period, until then poor in data. Climate change study using raw long-term data is hazardous due to many breaks caused by displacement of meteorological stations, replacement of sensors, modifications of the local environment, etc. Long-term data **homogenization** appeared as an imperative step prior to calculating long-term trends. Homogenization tools, based on statistical method developed at the Climatology department of *Météo-France* (*Direction de la Climatologie*), allow the detection and the correction of breaks. Their purpose may be applied to other kinds of datasets. There are now 70 monthly temperature series and 226 monthly rainfall series beginning before 1900 available. These series are useful for climate change studies and regional climate model validation.
This kind of data can be seen as a **living archive** that will be enhanced with new raw data discovery or homogenization techniques improvements.

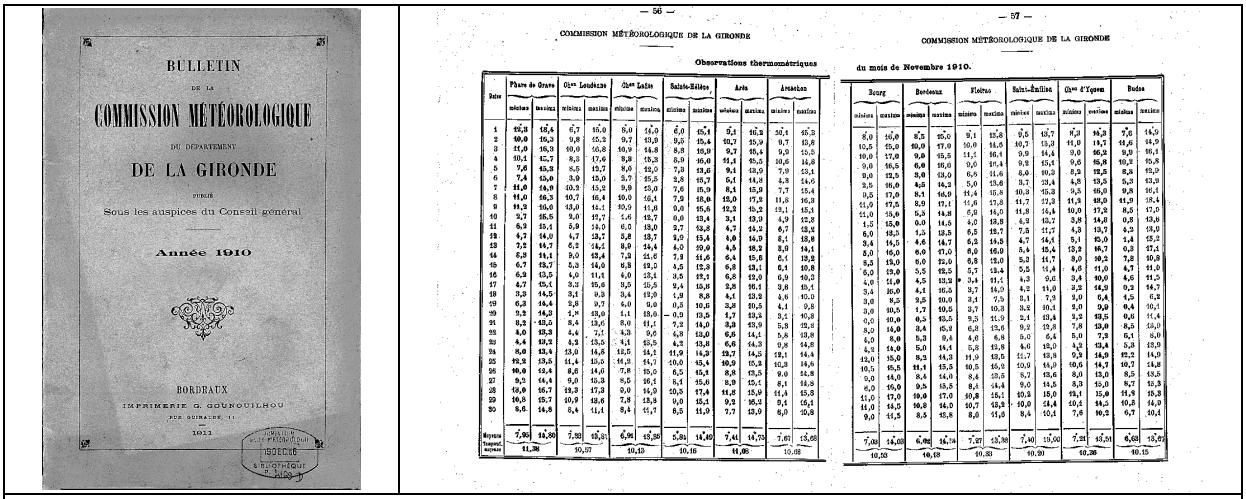
Introduction

Identification and digitisation of historical climatological data contained in old documents is a project by itself. The need of meteorological data covering a large period is now increasing with climate change : the more we can describe past changes, the more we can validate climate models and trust models simulations for 21th century. Many climatologists have noticed that many factors are likely to introduce homogeneity breaks into long-term climatological series. Breaks in long-term series is a general problem, not limited to climatology. Historical data rescue program and homogeneity issues are described in first chapter.
Detection and correction of breaks is a complex statistical problem. A new method is now used by *Météo-France*. This general method can be applied in many domains, but we will illustrate it mainly with long-term climatological series. Method and applications are described in second chapter.

1. Historical data rescue program at *Météo-France*

1.1. Program description

Since 1994, the Climatology department of *Météo-France* (*Direction de la Climatologie*, Toulouse) has put emphasis on his historical data rescue program. Various types of documents have been used: meteorological station documents, regional or national synthetic documents. An example of original document is shown in FIG. 1. These documents were found in various places: *Centre des Archives Contemporaines* (Fontainebleau), local meteorological centres (through regular queries), Paris-Montsouris centre (closely involved in French, and specially Parisian, climate History), etc.


BULLETIN
DE LA
COMMISSION MÉTÉOROLOGIQUE
DU DÉPARTEMENT
DE LA GIRONDE
PUBLIÉ
Sous les auspices du Conseil général
Année 1910
BORDEAUX
IMPRIMERIE G. GOUNOUILHOU
1911

— 56 —
COMMISSION MÉTÉOROLOGIQUE DE LA GIRONDE
Observations thermométriques

— 57 —
COMMISSION MÉTÉOROLOGIQUE DE LA GIRONDE
du mois de Novembre 1910.

FIG. 1: Example of regional archive document from which values have been keyed

Special emphasis has been put on temperature and precipitations monthly data during the 20th century:

- This action is closely linked to climate change problems
- This hierarchy of parameters and periods allows to reach a French territory coverage of this essential long-term data.

Original documents are firstly checked, to give an idea of quality of data (in terms of missing data, keyboarding easiness, length of series).
Millions of data have been keyed and these data are now available in *Météo-France* operational climatological database: this action allows both long-term preservation of this data and access to end-user. It is an on-going action: other periods, parameters, have to be looked for. For temperature and precipitations data, the goal is now to complete critical periods (such as World Wars I and II), to extend the research and preservation Program to 19th century data and to increase the amount of daily data (there is now strong needs related with variability studies). Large amounts of easily recoverable and readable documents have been used but it remains to deal with less readable documents (such as diagrams) or scattered documents. For monthly data of temperature, we only plan very specific insertions. Other parameters are also in progress: old pressure data have been keyed.
The amount of data keyed is illustrated by FIG. 2. Very old data are episodic (as were the data prior to 1950 before the data rescue program). Nowadays, the amount of raw data for the first half of the 20th century is

less than one tenth of the volume of the 1950-2000 periods. We consider that this ratio is a good result. One can see also the impact of World Wars I and II on the amount of data. There are more meteorological stations measuring rainfall than temperature.

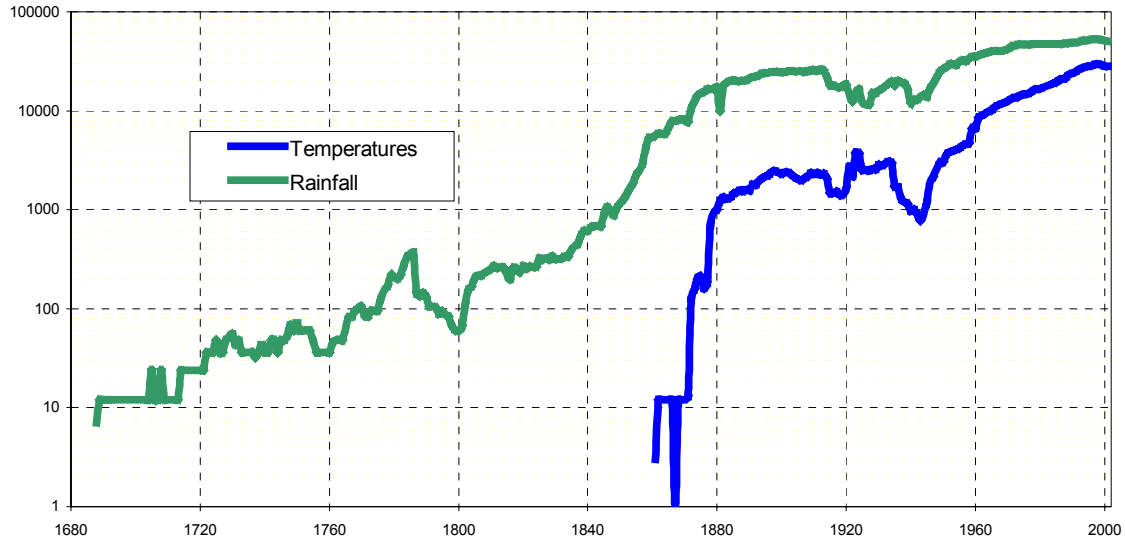


FIG. 2: Volume of raw data in *Météo-France* operational climatological database. X-axis: date, Y-Axis: annual amount of monthly data.

This old data rescue concern is shared by all National Meteorological Service. An international project, DARE (Data Rescue) of WMO (World Meteorological Organization) aims at assisting countries in the management, preservation and use of climatic data over their own territories.

1.2. Difficulties to use raw data for climate change detection

After insertion in *Météo-France* operational database (*BDCLIM*), new data go through a set of quality control and labelling procedures (the same set as recent data): range control, space and time coherency. Monthly "raw data" are calculated from daily raw data.

First goals of old data are climate change studies. To study regional warming, for example, we have to deal with different kind of problems:

- Measurement location continuity: there are only few measurement made at the same location over a long period. Relocations are numerous. Mixing data from different locations is necessary in order to build up long-term series (**concatenation**).
- Even for a single location, many changes can affect the **homogeneity** of data: sensor change, displacement, environment of measurement (obstacle), etc. Moreover, spurious observations are frequent.

So the real climatological series result from the superposition of two signals (FIG.3): the climatological one (the one we want to study) and station effect (we want to remove).

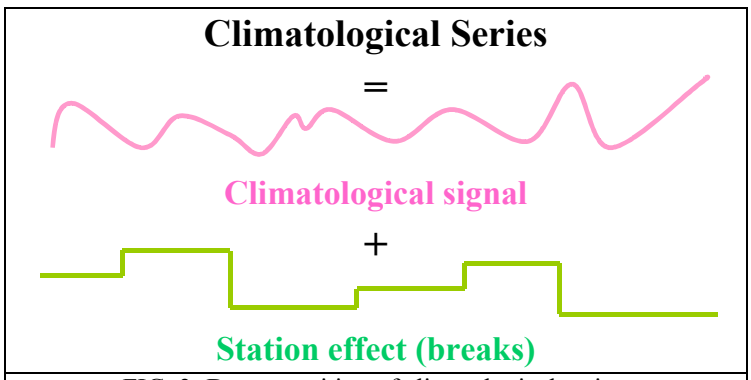


FIG. 3: Decomposition of climatological series.

Measurement conditions have changed a lot during 19th century, as shown in FIG. 4.

FIG. 4: Meteorological measurement stations used to build up the raw data long-term series of La Rochelle. Left: *Ecole Normale* (1910), right: place known as *Le Bout Blanc*

Nowadays conditions of measurement are well defined by WMO: height, type of sensors, shelter (colour, type, exposition), fetch. Long ago, conditions of measurement were not so standardized. One can notice for example, on the left side of FIG.4 a stony ground, which can influence the afternoon temperatures especially in summer (overestimation) and presence of obstacles (low wall, trees) and possible heat sources.
Amplitude of breaks often has the same magnitude as the embedded climate signals such as long-term variations trends or cycle. An imperative step before any serious climate change study is the **detection** and then the **correction** of homogeneity breaks.

2. Adding Value to data: homogenization

2.1. Break Detection: Detecting multiple change-points and outliers

2.1.1. Formulation

Climatic signal being mostly undetermined and non-stationary, it has to be removed as far as possible to put into evidence outliers or changes in measurement conditions. The principle of relative homogeneity in climatology [1] relies on the assumption that the difference series between the data at the tested station and at an assumed homogeneous neighbouring station (reference series) is fairly constant in time, up to the perturbations to be detected. It is also assumed that the distribution of the difference series is normal, and that most of the breaks are step-like changes that typically alter the average value only, usually leaving the higher moments unchanged. The difference series behaves like a Gaussian sample whose mean varies from sub-period to sub-period. Detecting such changes becomes challenging when the number of breaks is unknown.
When it is assumed that there is at most one change-point in a normal linear model, many procedures are based on the likelihood ratio test. Hawkins [4], Worsley [13] test for no change-in-mean versus existence of one change. When the number of possible change-points is known in advance, say k, all the previous procedures can be straightforwardly generalized to the choice between no change and k changes. But detecting an **unknown** number of change-points in a normal linear model is more difficult because the various hypotheses do not have the same dimension. The more general problem of determining a normal linear model with an unknown number of change-points and outliers has been studied by Caussinus and Lyazrhi [2]. They formulate it as a problem of testing multiple hypotheses and provide a multi-decision rule.
Let us give now the formulation of this procedure in the case of a normal sample. We consider n normal random variables Y_i (i=1,...,n) and let Y denote the column vector of the Y_i's. We assume that the probability distribution of Y is n-dimensional normal, with covariance matrix I_n (identity matrix of order n×n) up to the unknown variance σ^2.
Let k be the number of change-points and ℓ the number of outliers, let $\tau_1,\tau_2,\ldots,\tau_k$ be the positions of the k change-points, and let $\delta_1,\delta_2,\ldots,\delta_\ell$ be the positions of the ℓ outliers. Let $K=(\{\tau_1,\ldots,\tau_k\},\{\delta_1,\delta_2,\ldots,\delta_\ell\})$ be the set of change-points and outliers. To simplify the notation, we will set $\tau_o=0$, and $\tau_{k+1}=n$. Finally, let

$\Delta=\{\delta_1,\delta_2,\ldots,\delta_\ell\}$ and $n_j = \tau_j-\tau_{j-1}-\text{Card}[\{\tau_{j-1}+1,\tau_{j-1}+2,\ldots,\tau_j\}\cap\Delta]$, *i.e.* n_j is equal to the length of the period $[\tau_{j-1}+1,\tau_j]$ minus the number of outliers within this period.

We denote: $\overline{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i$ $\overline{Y}_j=\frac{1}{n_j}\sum_{\substack{i=\tau_{j-1}+1\\ i\notin\Delta}}^{\tau_j}Y_i$ for j=1,…,k+1

Let:

$$C_\varnothing(Y)=0$$

$$C_K(Y)=\ln\left[1-\frac{\sum_{j=1}^{k+1}n_j\left(\overline{Y}_j-\overline{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}\right]+\frac{2(k+\ell)}{n-1}\ln(n) \qquad (1)$$

The penalized log-likelihood procedure proposed by Caussinus and Lyazrhi [2] is:

$$\textbf{select } H_{K^*} \textbf{ such that } K^*=\text{Argmin}_K(C_K(Y)) \qquad (2)$$

The variance σ^2 is estimated by $\frac{1}{n-k-\ell-1}\sum_{j=1}^{k+1}\sum_{\substack{i=\tau_{j-1}+1\\ i\notin\Delta}}^{\tau_j}\left(Y_i-\overline{Y}_j\right)^2$ where the number and positions of outliers and change-points are those given by (2).

Procedure (2) has been proved to be asymptotically Bayes invariant optimal under a set of assumptions which turn out to be realistic in the problem we are dealing with [8]. For the particular problem of change-points in a Gaussian sample, the chosen penalty term gives much better results than Akaike's or Schwartz's criteria.

2.1.2. The step by step versions

The natural way to compute the procedure is to calculate $C_K(Y)$ for every possible hypothesis H_K (complete procedure). Nevertheless, this approach suffers from a major drawback: the number of hypotheses to examine rises very fast with n (length of the series) and $k+\ell$ the number of accidents to be detected. When detection is only performed for change-points, a dynamic programming algorithm can be used [5, 7]. Computation time then becomes only linear in k, and quadratic in n. To enable the detection of outliers as well, at a reasonable computing cost, a slightly different algorithm is used [8].

At each step, **one or two** more change-points are added to the previous selected hypothesis. Analytical studies shows that this **double step** procedure gives better detection results than single step procedure for up-and-down change points (and without significant improvement for staircase configuration). Furthermore, a triple step procedure, much more greedy in terms of computation time, leads to small improvement.

Mestre Method, with double step procedure, is now the **standard detection part of the homogenization method used by *Météo-France***.

2.2. Break correction

Knowledge of break positions can be a very interesting aspect for some users. For many applications (such as climate change studies) it is the half part of the problem. The other one, described below, is the break correction.

A two factors linear model is proposed for correction purposes. The series within the same climatic area are considered to be affected by the same climatic signal factor at each time, while the station factor remains constant between two breaks.

The model is applied after break detection. It provides the correction coefficient of a set of non-homogeneous series, through weighted least-squares estimation of the parameters. Weighted least squares allows to correct series with missing data. It also allows the weighting of the data, according to their supposed quality, which can be estimated for example with the correlation between the stations.

The formulation is equivalent to an exact modelling of the relative homogeneity principle. Given a set of non-homogeneous instrumental series, it allows unbiased estimations of the breaks affecting these series.

This method does not require to compute regional reference series, and is now **the standard correction part of the homogenization method used by *Météo-France***.

2.3. Homogenization of long-term climatological series

2.3.1. Way of working

Common conditions of detection and correction is the need of correlated series and homogenization only concerns monthly data. At a lower time scale, correlations are too weak.

Homogenization is performed on a set of about 20 series merged with geographical criteria. The first step is to perform quality control of series. Afterwards differences (for temperatures) or ratios (for precipitations) are computed and tested to put into evidence outliers or breaks.

Usual homogenization techniques are based on the assumption that climatic variations affect in the same way an homogeneous regional reference series and the candidate meteorological station. But this method rests on the existence of a homogeneous regional reference series whose reliability cannot be proved. The different methods for creating such series [1, 3, 12] do not guarantee their perfect homogeneity.

There is an easy way to get round the reference series. It is based on the simple statement that **between two change-points a series is reliable** (by definition), so these sections can be used as reference series. Each single series is compared to others within the same climatic area by making a series of differences. These difference series are then tested for discontinuities.

At this stage, we do not know which individual series is the cause of a shift detected on a difference (or ratios) series. However, if a detected change-point remains constant throughout the set of comparisons of a candidate station with its neighbours, it can be attributed to this candidate station. The detection of the outliers follows the same principle.

The procedure is not automatic and the **role of expert is essential**. The expert validates the break keeping in mind statistical and climatological issues. Metadata are the "data that describe the data": location, sensor-type, etc. Metadata are very helpful to confirm a break or to accurately locate a break. Obviously, metadata are not always available: there are more breaks than metadata.

It is impossible to locate straightaway all breaks: pronounced breaks hide smaller one's. Thus, the procedure is iterative. Iteration ends when all break risk is gone (Fig. 5). The whole processes can take a long time. Series are said "homogeneous", a relative notion linked to the Amplitude of Residual Breaks (ARB), which is estimated by the method.

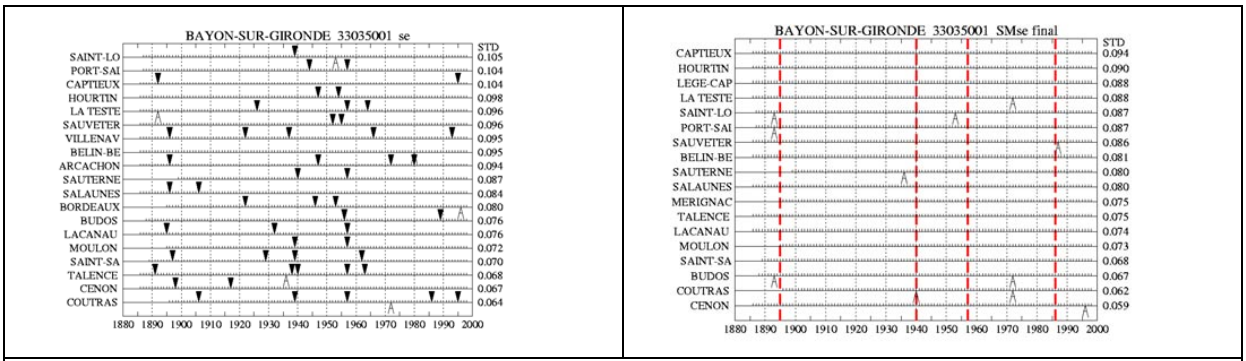


FIG. 5: Homogenization of rainfall monthly data of Bayon-sur-Gironde
Left: first step of homogenization. Right: end of homogenization
▼=break between Bayon-sur-Gironde and a neighbouring series, A=outlier, |=validated break, and then corrected

The homogeneous series are inserted in *Météo-France* climatological database: 70 homogeneous monthly temperature series starting before 1900 and 226 rainfall series are now available. Temperature series signifies monthly means of daily minimum temperatures (Tn, also called end-night or morning temperatures) **and** monthly means of daily maximum temperatures (Tx, also called afternoon temperatures).

ARB is about 0.2°C for temperatures and 10% of the annual amount for rainfall. 382 breaks were detected and corrected for Tn series, 362 for Tx series and 977 for rainfall series. Tn and Tx homogenization have been carried out separately and Tn breaks and Tx breaks have not always the same location or magnitude.

Each new parameters treatment needs different tests and raises new kind of problems. For example, for wind speed, we do not know straightaway if break detection is better on difference series or quotient series. For Mean Sea Level Pressure (MSLP), we noticed breaks due to formulae (from station pressure to MSLP) changes.

2.3.2. Homogenization Impact

In the first part of the paper, we examined changes of measurement condition at La Rochelle. Compared to nowadays conditions, 1910 measurement conditions exhibit an underestimation of Tn (about 1,2°C) and overestimation of Tx (about 1°C). These breaks (and other ones) have naturally strong impact on 1901-2000 trends:

- +1,4°C/century with Tn homogeneous data against +2,1°C/century with raw data
- +0,6°C/century with Tx homogeneous data against -0,1°C/century with raw data

Of course, La Rochelle case is not an isolated case. Cartographies of 1901-2000 Tn or Tx trends calculated with raw data exhibit very **noisy** characteristics (FIG. 6). Isolated raw series generate many obvious bubbles. There is no spatial coherency. The general impression leads to warming for Tn (that would have shown a reference series). Use of raw data Tx is completely misleading. Homogeneous series exhibit coherent signals with strong regional patterns, such as smoothed gradients.

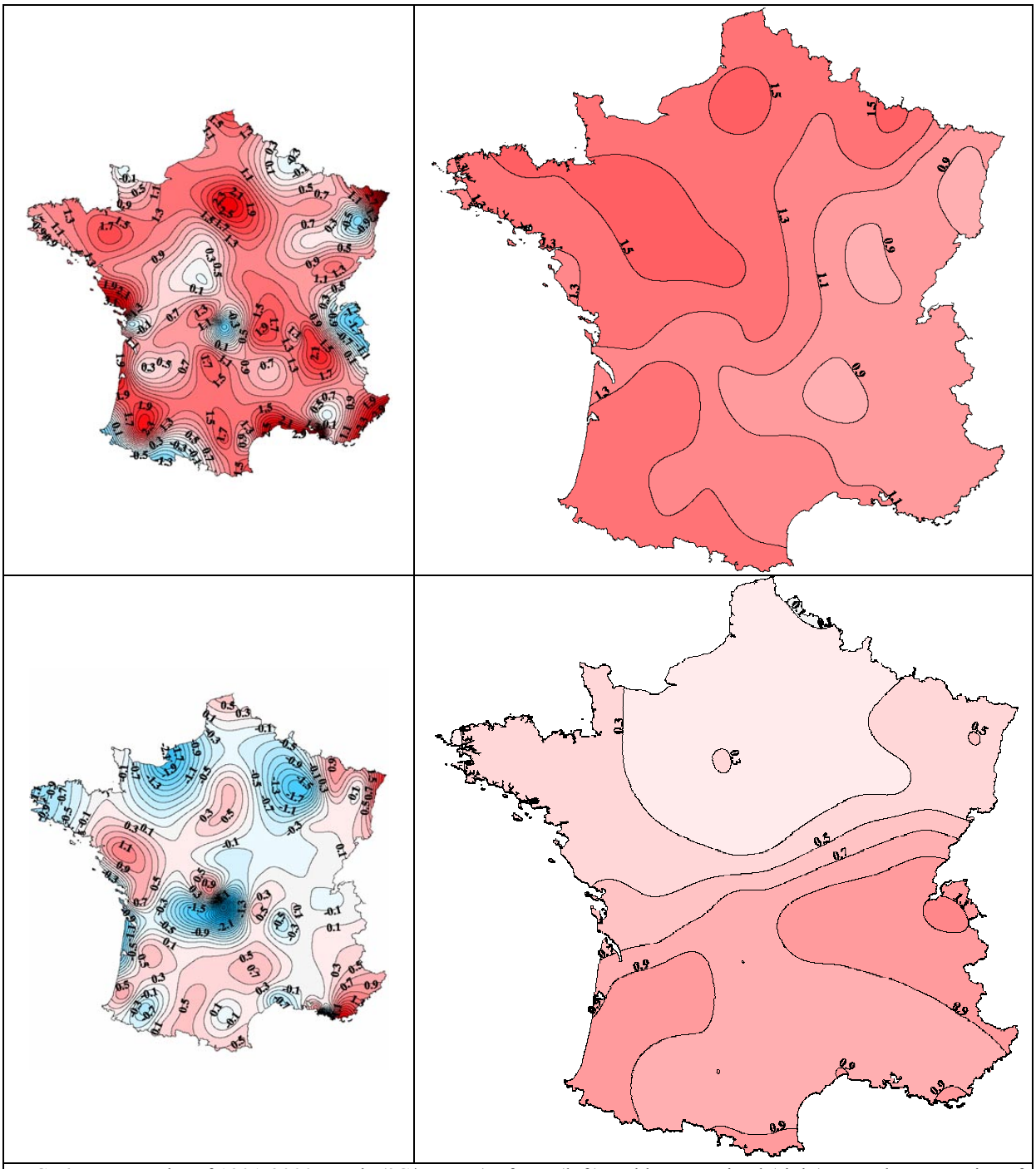


FIG. 6: cartography of 1901-2000 trends (°C/century) of raw (left) and homogenized (right) annual mean series of daily minimum temperature Tn (up) and maximum temperature Tx (down).

Comparisons of box-averages of series with French grid points of global climatologies (Jones for temperature and Hulme for rainfall) exhibit good correlations [9]. Studies are in progress to compare regional patterns of climate change to other kind of data.

2.3.3. Survey of climate change in France

Annual mean **temperature** (Tm, average of Tn and Tx) trends over France are **about 1°C for 1901-2000 period**. This value is higher than the diagnostic established by IPCC (Intergovernmental Panel on Climate Change) for the whole planet, 0,6±0,2°C [6]. The regional pattern of warming reveals more marked positive trends in Western part of France for Tn and a North-South gradient for Tx (FIG. 6). Of course, Tm warming reaches its highest point in South-West part of France (Tm trend of Périgueux: +1,22°C/century). Minimum temperatures rises are more pronounced than maximum temperatures one's. Summer is the season of maximal warming of Tn while autumn is the season of maximal warming of Tx [9].
Long-term homogenized rainfall series (226 series covering 20th century) are more numerous than temperature series (70 series covering 20th century). But we cannot have a complete view of rainfall changes, due to concentration of series in the same neighborhood and high spatial variability of this parameter. Annual **rainfall** exhibits contrasted but generally **positive trends**. Northern series increases are more pronounced than Southern one's. Rainfall decreases are frequent in Summer and there are hardly any decreases in Winter [10, 11].
Therefore, considering warming and rainfall increases, the hydrological cycle on France has sped-up during 20th century.

2.4. Other applications

Whole action "Research, Digitisation and Homogenization of Long-Term Climatological Data Series " aims climate change studies and validation of regional climate model but methods or results can be interesting for other purposes.
Methods and procedure can easily be used for other kinds of problem. Tests have been performed to study meteorological models performance scores. Breaks are closely linked to dates of major changes of the forecast systems.
Some users are only interested by break detection. Without correction, the period starting after the last break is homogeneous, it's for example a privileged period for statistical adaptation. If daily data are available, we plan to use such a period to study extreme events evolution (number of frost days by example).
The procedure calculates non-dimensional and unbiased variables and can be applied to heterogeneous set of variables.
Application for signals processing cannot be made directly because of computation-time considerations.

Conclusion

Analysis of climate change during past decades or centuries requires the availability of digitised historical data. The data rescue program going on since 1994 has made it possible.
The methodology for the relative homogeneity testing of climatological series and the model that allows the correction of non-homogeneous series (without requiring the computation of the so called "regional reference series") are essential tools for climate change studies. The whole action "Research, Digitisation, and Homogenization of Long-Term Climatological Data Series" combines historical, meteorological, statistical and instrumental issues.
Homogeneous series do not replace raw data. This series are the results of the use of processes (homogeneous) applied on data (raw data). Both can change with times. Homogenization techniques can be upgraded. Old data search can help to discover new raw data. These data can complete missing years or improve break detection or correction (whose quality depends on amount of series). Homogeneous data change the usual vision of historical climatological data.
Homogeneous series are inserted in the operational climatological database and part of the product catalogue of *Météo-France*. Users come mainly from research world but new kinds of users have appeared: underwriters for example for the climate derivative aspects.
Efforts put today on metadata will minimize and help tomorrow homogenization. Resorting to homogenization can be seen as a stopgap due to the lack of metadata. Efforts put on double measurement,

preservation of environmental conditions of measurement or at least complete description of measurement changes, will minimize or help a lot in future years homogenization of present series.

References

[1] H. Alexandersson. *A homogeneity test applied to precipitation data*. Int. Journal of Climatology, Vol. 6, 661-675, 1986

[2] H. Caussinus and F. Lyazrhi. *Choosing a linear model with a random number of change-points and outliers*. Ann. Inst. Statist. Math., 49, No. 4, 761-775, 1997

[3] E. J. Førland and I. Hanssen-Bauer. *Homogenizing long Norwegian precipitation series*. J. of Climate, 7, 1001-1013, 1994.

[4] D. M. Hawkins. *Testing a sequence of observations for a shift in location*. J. Amer. Statist. Assoc., 72, 180-6, 1977.

[5] D. M. Hawkins. *Fitting multiple change-points to data*. Comput. Statist. Data Anal. 37, 323-341, 2001.

[6] J. T. Houghton,Y. Ding, D.J. Griggs, M. Noguer, P. J.Van der Linden, D. Xiaosu. *Climate Change 2001: The Scientific Basis, Contribution of Working Group I to the Third Assessment Report of IPCC,* Cambridge University Press, UK, pp 944, 2001, (eds.).

[7] M. Lavielle. *Optimal segmentation of random processes*. IEEE Trans. on Signal Processing, vol. 46, no. 5, pp. 1365-1373, 1998.

[8] O. Mestre. *Méthodes statistiques pour l'homogénéisation de longues séries climatiques*, Thèse de doctorat de l'Université Paul Sabatier (Toulouse III), 2000.

[9] J.-M. Moisselin, M. Schneider, C. Canellas and O. Mestre. *Changements Climatiques en France au XXème siècle. Étude des longues séries de données homogénéisées françaises de précipitations et températures*, La Météorologie, Aug.2002.

[10] J.-M. Moisselin. *Les précipitations sur le XXème siècle en France*, Lettre PIGB-PMRC-France de Février 2002, disponible sur http://www.cnrs.fr/dossiers/dosclim/biblio/pigbsom.htm#som13

[11] J.-M. Moisselin and M. Schneider. *Homogénéisation des séries françaises de précipitations couvrant le XXème siècle*, à paraître dans La Houille Blanche.

[12] D. C. Peterson and D. R. Easterling. *Creation of homogeneous meteorological reference series*. Proc. Eight conf. on Appl. Climatology, Anaheim, Amer. Meteor. Soc., 31-34, 1993.

[13] K. J. Worsley. *On the likelihood ratio test for shift in location of normal population*. J. Amer. Statist. Ass., 74, 36-57, 1979.