Journal de la Société Française de Statistique
Vol. 156 No. 1 (2015)

Modeling of air temperatures: preprocessing and trends, reduced stationary process, extremes, simulation

Titre: Modélisation des séries de température de l'air : prétraitement, réduction à un processus stationnaire, extrêmes, simulation.

Didier Dacunha-Castelle[1], Thi Thu Huong Hoang[2] and Sylvie Parey[2]

Abstract: Our first goal is to give a complete methodology to get a simulation model valid for a very large number of air temperature series. Existing simulators have quite good properties for the bulk of the distribution but they are unable to reproduce extreme values as well as cold or hot waves. Thus we focus on this aspect in order to cover both the bulk of the data and extreme values. First we give some new results on preprocessing (the phase during which trends and seasonalities are removed). In this context, the choice of non parametric trends requires some new results on cross validation. Once additive (mean) and multiplicative (variance) trends and seasonality are suppressed, we test the cyclo-stationarity of the reduced series. In almost all cases, there does not remain any trend even for extremes values. To model the observations, we start from non parametric estimates of the expectation and variance from a day conditional to the previous one and from studies of correlations. The best class of model seems the seasonal FARCH class. Then for mathematical improvements supported by physical ones, FARCH process is interpreted as the Euler scheme of continuous-time diffusions and thus as misspecified models and modified by adaptive truncation of innovations using new results of extreme theory on diffusions. These results are plugged in the likelihood of FARCH processes to get a convenient estimation. We detail some results of the application to about 200 stations from Eurasia and United States, including the validation procedures for the model. The comparison to some existing models is also considered.

Résumé : Notre but premier est de donner une méthodologie complète pour construire un simulateur valide pour un très grand nombre de séries de température de l'air. Les simulateurs existants ont un comportement plutôt bon pour le cœur de la distribution mais sont incapables de reproduire les valeurs extrêmes comme les vagues de chaud ou de froid. Nous nous focalisons particulièrement sur cet aspect de qualité tant pour les valeurs ordinaires que pour les valeurs plus extrêmes Nous donnons d'abord des résultats nouveaux sur le prétraitement des séries visant à extraire les saisonnalités et les tendances. Nous obtenons alors un processus réduit cyclostationnaire, la plupart du temps sans tendances résiduelles même pour les extrêmes. Pour bâtir le modèle, nous partons d'une étude non paramétrique des espérances conditionnelles et des variances conditionnelles d'un jour sur les jours précédents. Nous construisions alors un modèle FARCH cyclostationnaire sur l'année. Pour améliorer l'estimation, à partir de considérations physiques faites en temps petit, nous interprétons la chaine FARCH comme le schéma d'Euler d'une diffusion cyclostationnaire. Pour cette diffusion nous obtenons des résultats nouveaux liant le coefficient de diffusion et les paramètres d'extrêmes. Par plug-in dans la vraisemblance du modèle FARCH nous obtenons une bonne estimation de ce coefficient de diffusion. Le plug-in est justifié par l'ergodicité géométrique du FARCH vu comme une chaîne de Markov squelette de la diffusion (bien que le coefficient de diffusion ne soit pas lipchitzien au voisinage des frontières). Nous illustrons les résultats sur 200 stations en Eurasie et aux USA.

[1] Université Paris Sud 11
E-mail: didier.dacunha-castelle@math.u-psud.fr
[2] EDF R& D
E-mail: thi-thu-huong.hoang@edf.fr and E-mail: sylvie.parey@edf.fr

Keywords: temperature, time series, preprocessing, cyclo-stationarity, extremes, diffusion, simulation model
Mots-clés : température, séries chronologiques, prétraitement, cyclostationnarité, extrêmes, diffusion, modèle de simulation
AMS 2000 subject classifications: 62-02, 62P12

1. Introduction

Statistics are of crucial importance in climate studies as recent debates about data treatments have emphasized, for example around the well-known "Hockey stick". Nevertheless, collaborations with statisticians remain unusual. Climate results can strongly depend on the considered period or geographical area, and some arbitrary choices like linear trends for non linear phenomena or the use of statistical tests without any power computation are still common practice. However climate data strongly differ from econometric ones and very specific tools are needed.

In this work we focus on air temperatures with a particular goal: we try to give a treatment of these series as complete as possible in order:

- To give an accurate data analysis of the data,
- To design a simulation model that can reproduce the bulk of the distributions as well as the extremes values. The simulation model allows estimating the probability of rare events like very hot waves.

Stochastic models for simulation of air temperatures are more frequently used as part of pricing weather derivatives, especially in the framework of energy prices, than as tools for climate studies. They are now commonly used in financial studies as a way of simulating key properties of observed meteorological parameters. This necessitates some knowledges on this Şweather noiseŤ over space and time, which motivated the development of such stochastic models (Campell and Diebold, 2005, Mraoua and Bari, 2007, Benth and Benth, 2011).

The general principle of such models consists in modeling the temperature (daily maximum or minimum or mean for instance) as the summation of a deterministic part and a stochastic process, designed to represent the random fluctuations: $X(t) = \Lambda(t) + \Phi(t)Y(t)$, where $\Lambda(t)$ and $\Phi(t)$ the deterministic parts and $Y(t)$ the stochastic one. $\Lambda(t)$ contains at least a seasonal component, and usually also a trend (low frequency) component. The stochastic part generally presents a stationary autoregressive structure, more or less sophisticated: from an AR(1) to an AR(3) for the linear part and from a constant to a GARCH model for the innovation's variance. For some models, the standard deviation is considered as seasonal (Richardson, 1981).

We address in this paper the behavior of the extremes in order to study rare events as heat waves. The interest in the extremes further motivates the use of such models; however they generally must be improved to adequately reproduce extreme events (Furrer and Katz, 2008) mainly because the conditional variance is not well reproduced. In fact, all studies performed with observations, say during the last 60 years, show that temperature distributions are bounded thus with a negative shape parameter (Katz, 2011). This result does not depend on the method used to study extremes:

GEV, POT, Hill estimator for rough or standardized data. This point is decisive and explains why linear Gaussian models or GARCH processes with heavy tails give poor representations of the extremes events if the bulk of the distributions is quite well fitted. Then what kind of stochastic processes have to be selected to model air temperatures?

Let us explain now how to choose a model for the stochastic part. The first step is a non parametric qualitative analysis of $E(Y_t/Y_{t-1})$ and $Var(Y_t/Y_{t-1})$. If the expectation is clearly a linear function by, the variance $a(y)$ decreases or increases in the bulk of the distribution of Y_{t-1} following the season and it diminishes drastically out of a compact interval. It seems difficult to get such a behavior using a GARCH or bilinear model as they have been used in the previous models. At the opposite a FARCH model given by:

$$Y_t = b(Y_{t-1}) + a(Y_{t-1})\varepsilon_t \tag{1}$$

with $a = 0$ out of an interval seems able to support the qualitative behavior of the conditional variance.

However if ε_t is Gaussian, the distribution of Y_t is unbounded but the observation gives a negative shape parameter strictly negative thus associated with a bounded distribution. Thus (1) is not satisfactory. Then, another manner to consider a FARCH model is to take it as the first Euler scheme of a continuous time diffusion whose discrete observations Z_t define a Markov chain M, the skeleton of the diffusion Z given by:

$$dZ_t = b(Z_t)dt + a(Z_t)dW_t \tag{2}$$

where W is a Brownian motion.

The transition of M is too complicated (see Dacunha-Castelle and Florens-Zmirou, 1986 to use it as a simulation model and more to estimate a and b). It is why often the FARCH approximation is used for this statistical goal. Nevertheless if we consider M as the Ştrue modelŤ we obtain a satisfactory estimate of the extremes parameters and more using the embedding in a continuous time diffusion and a new theoretical result, we can not only estimate the support of the distribution but also, this is a difficult point, the behavior of a near the boundary.

Once this work done we come back to the FARCH likelihood in order to estimate a and b but using the constraint obtained on the support and the behavior of a at the boundary in a Şplug-inŤ step.

The last step is to modify the FARCH model in order to get bounded distributions. This is done taking for the residues a truncated distribution with adaptative (depending on the data Y_{t-1}) bounds. The new chain is easy to simulate.

We do not enter in this paper in physical considerations. Temperature is a continuous time process with continuous trajectories, like wind although generally smoother. Arguments for diffusions can be found for instance in Sura (2012) starting from physical models of non linearity.

From a purely mathematical point of view, to support the hypothesis of diffusive phenomena, we have to test if the process is of a Markov type; the Markov property implies the diffusive character. Markov property can be tested only on discrete observations with a fixed interval between observations. Thus the Markov property can be checked only for some "scale" associated to the lag between measures. We can test by a non parametric method the Markov property associated to the data interval (minute, hour, day). We can also test directly the goodness of fit of a parametric class of Markov processes. Both methods have been used in our work and lead to a diffusion (with inaccessible boundaries) of the form (2).

The use of purely stochastic models such as diffusions with time depending coefficient is not convenient because of the existence of very low frequency and seasonal dynamics in the temperature. Thus the first intend is to get stationary data from the rough data $X(t)$ by removing the additive and multiplicative estimations of trends and seasonality to obtain the "reduced series" $\hat{Y}_t = \frac{X_t - \hat{u}_t - \hat{m}_t}{\hat{v}_t \hat{s}_t}$ where $\hat{u}, \hat{v}^2$ are respectively the estimated seasonalites in mean and in variance and $\hat{m}, \hat{s}^2$ are the estimated trends in mean and in variance.

The reduced series is potentially stationary but it is not because the short term correlations, the higher order moments (skewness, kurtosis), remain seasonal despite the first preprocessing. Thus we have to work with a non stationary, but instead a cyclo-stationary process whose distribution is invariant by time shifts $kA, k \in \mathbb{Z}$, A length of the period, here a year. We test the cyclo-stationarity of the reduced series.

After this preprocessing step, we model the reduced series $\hat{Y}(t)$ by the model explained previously.

Our work presents four main steps (including applications) now detailed.

1- **Preprocessing**: we go from the rough data X to the reduced data $\hat{Y}$, and then check the cyclo-stationarity of $\hat{Y}$. Preprocessing is here difficult because of trends. Trends are often modeled, without any discussion, by a linear function. This implies a loss of information. It can be seen that the slope depends highly on the period of observation. We choose a non parametric estimation, which needs the choice of a smoothing parameter here difficult because the observations are highly dependent. Moreover, it is necessary to consider multivariate trends: trends in variance are crucial for climate. Once trends and seasonalities have been removed, nothing proves that there does not exist any other trends, for instance for short interval correlation. A more difficult question concerns the trends in extremes: are they correctly explained as a non random function of the trends in mean and variance? If it is not the case, it is almost impossible to get a simulation model valid for extremes. For the geographical area studied here the answer is yes (see Parey et al., 2013 for details).

Once detrending and deseasonalisation have been done, we have to check the cyclo-stationarity of the reduced series. We suggest a new test of stationarity which is not only adapted to climatic data but also useful in many other practical situations.

2- **Estimation**: we estimate a and b taking into account specific results for extreme values of a diffusion. Then we have to work with a cyclo-stationary diffusion discretely observed. From the geometric ergodicity of the diffusion we are able to estimate the extreme parameters and the bounds using only this discrete skeleton. We give here only the main mathematical ideas. These informations are then plug-in as constraints in the likelihood of the Euler scheme to estimate coefficients a and b. This technical point is new and allows getting coefficient giving a good representation of the extreme behavior.

3- **Improve the Euler scheme** whose extremes cannot be bounded and thus which is not adapted to our purpose.

Once the statistical work achieved, we build the simulation model as a Markov chain.

$$Z_t = c([t], Z_{t-1}) + a([t], Z_{t-1})\eta_t(Z_{t-1}) \tag{3}$$

where $\eta_t(Z_{t-1})$ is a particular Markov chain subordinated to Z, the distributions of η_t being adaptively truncated in order to satisfy the extreme constraints. The coefficients c and a depend on the time $[t]$, the days of year, which allows taking into account the remained seasonality in the reduced series.

4- **Applications** We give a complete treatment of air temperature series. The quality of simulations is tested for different characteristics as marginal distributions, correlations, but also, of course, for extreme properties.

In the bulk of the distribution, many models, which are essentially linear models with a Gaussian noise (so non bounded), are close to our model in terms of performance. Nevertheless, our model seems to take into account seasonalities and trends in a better way. For the high quantiles or the extremes, our model clearly shows a better performance compared with the existing simulation models. This is the result of our better representation of the shape parameter : our model takes into account the boundedness of the temperature.

2. Preprocessing

2.1. The model for preprocessing

Temperature series at a daily time scale exhibit clearly a seasonal cycle and trends. Thus we consider the model:

$$X_t = m_t + u_t + s_t v_t Y_t, t \in [1, T] \subset \mathbb{N} \tag{4}$$

m and s are to be looked as trends, they take into account slow dynamics (low frequency). u and v are periodic functions with period A (A=365 days in our applications). t means the day. Y is a centered and normed process, $EY_t = 0, EY_t^2 = 1$. We do not suppose that Y is a white noise, it can be non stationary.

The model (4) is identifiable if the following conditions are verified:

$$\sum_{t=1}^{T} m(t) = 0, \qquad \sum_{t=1}^{T} s(t)^2 = 1 \tag{5}$$

The trends are estimated by Loess (locally weighted scatterplot smoothing, Cleveland, 1979) and the seasonalities are estimated using the trigonometric functions. In practice, when the smoothing parameter (large enough) of Loess, or any other non parametric method, is fixed, the estimation order of m and u has no importance. In our case, the bandwidth which is used for detrending gives a low frequency corresponding to a window of 10 to 15 years large enough with respect to one year, corresponding to the frequency of the seasonality. Here we estimate the seasonality first because the cross-validation algorithm to choose the bandwidth for the trend works better on the series without seasonality.

In our study, we estimate the parameters of (4) in the following way:

- estimate u from the series X_t,
- estimate m from the series $X_t - \hat{u}_t = \hat{X}_t^{(1)}$,
- estimate v^2 from the series $(X_t - \hat{u}_t - \hat{m}_t)^2 = \hat{X}_t^{(2)}$,
- estimate s^2 from the series $\frac{(X_t - \hat{u}_t - \hat{m}_t)^2}{\hat{v}_t^2} = \hat{X}_t^{(3)}$,
- series $\hat{Y}_t = \frac{X_t - \hat{u}_t - \hat{m}_t}{\hat{v}_t \hat{s}_t}$ is called the "reduced series".

Remark 2.1. - *For the purpose of constructing simulation models for the temperature, many authors model the trend m by a line. According to our studies, this linear trend shows its insufficiency and inadequacy. Even when a seasonally linear trend (with periodic coefficients α and β) $\alpha(t)t + \beta(t)$ is used as an estimator, the trend in mean is not completely detected. The result of this is that $\hat{Y}_t$ still has a significant trend, which can be seen by simple graphics.*

- In Parey et al. (2009), we show that the (increasing) trends for mean and variance are very similar in many situations for all stations in Europe for instance, except for Iberian Perninsula where the evolutions of mean and variance are opposite.

2.1.1. Estimation of seasonalitites and trends

The additive u and multiplicative v seasonal effects are estimated by trigonometric polynomials of the form:

$$P_p(t) = \theta_0 + \sum_{k=1}^{p} \theta_{k,1} \cos \frac{2\pi tk}{365} + \theta_{k,2} \sin \frac{2\pi tk}{365} \tag{6}$$

The degree p is chosen through the use of an Akaike criterion. This parametric estimation has been compared to the non parametric STL (Seasonal-Trend decomposition based on Loess smoothing) method (Cleveland et al., 1990) and the two approaches have been found very similar for temperature series.

For asymptotics of the estimation of m and s, we consider as usually the continuous functions m and s defined on $[0,1]$, then $m(t) = m\left(\frac{t}{T}\right), s(t) = s\left(\frac{t}{T}\right)$ with $t = 1, \ldots, T$. The details can be found in Hoang (2010), we just mention the main results.

Once the seasonalites u_t estimated, to estimate m, we consider the model:

$$\hat{X}_t^1 = m_t + \tilde{Y}_t \tag{7}$$

where $\hat{X}_t^1 = X_t - \hat{u}_t$.

We do not suppose Y stationary.

To get asymptotic results, we need to control the non stationarity, this is done in the following way, using only the covariance function:

Let W be a stochastic process, $t \in \mathbb{N}, \sup EW_t^2 < \infty$. We suppose that W follows the following condition:

$$\sup_T C_T(W) = C(W) < \infty \tag{8}$$

with

$$C_T(W) = \frac{1}{T} \sum_{i=1}^{T} \sum_{j=1}^{T} \mid Cov(W_i, W_j) \mid \tag{9}$$

The condition (9) for a stationary process implies L^2 weak-dependence. For non stationary case, it controls the degree of non stationarity, which allows to apply the framework valid for iid observations to dependent and non-stationary data. Using (9), we obtains an asymptotic result close to that of an ordinary regression.

In order to apply Loess, let K be a compact support kernel, a bandwidth is a positive sequence h_T such that $h_T \to 0$ and $Th_T \to \infty$ as $T \to \infty$.

Theorem 2.1. *(m estimate, see Hoang, 2010)*

Let $X_t - \hat{u}_t = m_t + \tilde{Y}_t$. Suppose $m \in C^2$ and (8) verified for $\tilde{Y}$. Let $\hat{m}_T$ be the Loess estimator of m. It means that $\hat{m}(x) = \hat{\beta}_0(x)$ where $\hat{\beta}_0(x)$ is obtained by minimizing:

$$\sum_{i=1}^{T} \left(\left[X_t - \beta_0 - \beta_1 \left(x - \frac{t}{T} \right) \right]^2 K_T \left(x - \frac{t}{T} \right) \right) \tag{10}$$

with $K_t(.) = \frac{1}{h_T} K(./h_T)$. Then we have:

$$\begin{aligned} E(\hat{m}_T(x) - m(x))^2 = & \left(\frac{1}{2} h_T^2 m^{(2)}(x) \mu_2^1 \right)^2 \\ & + \frac{1}{Th_T} \mu_0^2 C_Y(T) + o\left(\frac{1}{Th_T} + h_T^4 \right) \end{aligned} \tag{11}$$

where μ_j^r is the moment of order j of K^r.

The proof of this theorem is shown in Hoang (2010), as an extension of Ruppert and Wand (1994) .

For the estimation of s^2, consider now:

$$\hat{X}_t^2 = \frac{(X_t - \hat{m}_t - \hat{u}_t)^2}{\hat{v}_t} = s_t^2 + s_t^2(Y_t^2 - 1) = s_t^2 + R_t \tag{12}$$

Suppose that the stochastic process R_t satisfies the condition (9) (which is equivalent to a condition on fourth order cumulants of Y).

Theorem 2.2. *(estimation of s^2 (see Hoang, 2010))*
Let k_T be a bandwidth for the kernel K used to estimate s, suppose that $k_T^4 + (Tk_T)^{-1} = o(h_T^2)$ then

$$\begin{aligned} E(\widehat{s^2}(x) - s^2(x)) &= \frac{k_T^2}{2}\mu_2^1 (s^2)''(x) + o(k_T^2) \\ Var(\widehat{s^2}(x)) &\leq \frac{1}{Tk_T^2}\mu_0^1 C_T(\eta) + o\left(\frac{1}{Tk_T}\right) \end{aligned} \tag{13}$$

Remark 2.2. *: 1- The speed of the estimation of s^2 does not depend on the knowledge of m. Optimal choice of h_T using theorem 2.1 is difficult, we do not know the constants like $m^{(2)}(t), C_Y(t)$. We come back below to this choice.*

2- We can use block-bootstrap to compute confidence intervals for m and s if Y is stationary. In the case of temperature series, we can use 10 days as the size of blocks of bootstrap as we show later.

Now we address the choice of the bandwidth h_T and k_T. To our knowledge, it does not exist today any theoretical cross-validation (CV) result under the hypothesis of dependence and non stationarity of Y_t.

In Hoang (2010), a new algorithm of CV is suggested and called MPCV (modified partitioned cross-validation), a modified version of PCV (partitioned cross-validation) of Marron (1987). Through simulations, we show that MPCV gives better results than PCV but also than other algorithms as MCV (Chu and Marron, 1991), plug-in CV (Francisco-Fernández and Vilar-Fernández, 2005), block bootstrap (Hall et al., 1995). Details on MPCV can be found in Hoang (2010).

In our study, we take $k_T = h_T$.

3. Stationarity of the reduced series

Let $\hat{Y}_t = \frac{X_t - \hat{u}_t - \hat{m}_t}{\hat{v}_t \hat{s}_t}$ the reduced series.

If we neglect estimation errors, we can suppose that:

$$E\hat{Y}_t = 0, E\hat{Y}_t^2 = 1, 1 \le t \le T \tag{14}$$

In application to temperature series $T \sim 10^4$ days, we have enough data to test condition (14) without any problem using the test theory developed in this part.

We want to test H_0: "$\hat{Y}_t$ is a stationary process" or H_0': "$\hat{Y}_t$ is a cyclo stationary process" against H_1: "$\hat{Y}_t$ is non stationary" or H_1' for non cyclo-stationarity.

Stationarity means that all multidimensional distribution of $\hat{Y}$ are invariant for every time translation and cyclo-stationarity of a period A means invariance for the group of translations of $kA, k \in \mathbb{Z}$.

We treat jointly both stationarity and cyclo-stationarity. We take the following convention: *Every time that we talk about a constant estimation for stationarity, we can take place by a periodic function estimation for cyclo stationarity.*

H_0 and H_1 are too large to have some practical meaning. In fact, we can first informally limit H_1 to "$\hat{Y}$ have a low frequency trend for some important characteristics" and H_0 can be changed into an hypothesis which can be statistically tested using simple and meaningful statistics rather than checking an infinity of conditions.

For T fixed and for every t $(1 \le t \le T)$, consider a finite set

$\Phi(t) = (\varphi_1(t), ..., \varphi_k(t))$ of real functionals of $(\hat{Y}_1, ..., \hat{Y}_T)$.

We distinguish now some kinds of functionals of $\hat{Y}$: the moments, the density and the characteristics of extremes.

Moments

Univariate moments: $\varphi_j^{mom}(t) = E\hat{Y}_t^j, j \ge 1$.

Bivariate moments: $\varphi_h^{cor}(t) = E(\hat{Y}_t - E\hat{Y}_t)E(\hat{Y}_{t+h} - E\hat{Y}_{t+h})$.

Density: $\varphi^{dens}(t) = f_t(\hat{Y})$.

Extremes:

Let $N = \oplus_{j=1}^{\infty} A_j$ where A_j are disjoint blocks of $\hat{Y}$ of size $\| A \|$. Let $N(T)$ the number of blocks belonging to $[0,T]$. Let $M_j = \max_{t \in A_j} Z_t$. We consider:

$\varphi_1^{ext}(t) = EM_j$ for $t \in A_j$

$\varphi_2^{ext}(t) = Var(M_j)$ for $t \in A_j$

If we suppose that the distribution of M_j is approximated, for $\| A \|$ large enough, by a GEV distribution with parameters (μ_j, σ_j, ξ_j). In this case, we will consider: $\varphi_1^{GEV}(t) = \mu(t)$, $\varphi_2^{GEV}(t) = \sigma(t)$, $\varphi_3^{GEV}(t) = \xi(t)$.

Definition 3.1. *Z is said Φ-stationary if Φ is constant.*

Let H_0^{Φ} be the hypothesis "Φ is constant". Rather than testing H_0^{Φ} against H_1^{Φ}: "Φ is not constant", we test H_0^{Φ} against H_{LF}^{Φ}: "Φ has a low frequency non zero trend".

3.1. Theory of test of stationarity and cyclo-stationarity

The main idea of our test bases on the distance between the constant estimator and the non parametric one. Let $\Delta(\varphi, \psi)$ be the distance between two sequences on $[1, T]$ defined by:

$$\Delta^2(\varphi, \psi) = \frac{1}{T} \sum_{t=1}^{T} \mid \varphi(t) - \psi(t) \mid^2 \tag{15}$$

and $\Delta^2(\Phi, \Psi)$ be the distance between two sets of functionals:

$$\Delta^2(\Phi, \Psi) = \sum_{j=1}^{k} \Delta^2(\varphi_j, \psi_j)\delta_j \tag{16}$$

where $\Phi(t) = (\varphi_1(t), ..., \varphi_k(t))$, $\Psi(t) = (\psi_1(t), ..., \psi_k(t))$ and δ_j is the weight linked to $\Delta^2(\varphi_j, \psi_j)$ in the set (Φ, Ψ).

Now let $\hat{\Phi}^T$ be a non parametric estimator of Φ and $\bar{\Phi}^T$ an estimator of Φ by a constant (stationary case) or a p-periodic function (p = 365) (cyclo stationary case).

The theoretical support to our test is developed below. Suppose that $\varphi^T(.) = \varphi\left(\frac{\cdot}{T}\right)$ where φ is the "true" function defined on $[0, 1]$. The following conditions need to be verified $\forall j = 1, ..., k$:
i/ $\Delta_T(\hat{\varphi}_j^T, \varphi_j^T) \to 0$ for H_0 as for H_1
ii/ Under H_0, $\Delta_T(\bar{\varphi}_j^T, \varphi_j^T) \to 0$.
iii/ Under H_1, $\Delta_T(\bar{\varphi}_j^T, \varphi_j^T) > a > 0$ for a independent of T.

Let us remark that if φ_j^T is non constant, then:

$$a \leq \int_0^1 \mid \varphi_j^T(t) - c(t) \mid^2 dt \tag{17}$$

for every constant (stationary case) or every periodic function c (cyclo stationary case).

So to get an asymptotic result, we need to know the behavior of a "good" estimation of φ by a constant even under H_1.

If the conditions i/, i// and iii/ are verified, then:
Under H_0: $\Delta_T(\hat{\varphi}_j^T, \bar{\varphi}_j^T) \to 0$ as $T \to \infty$
Under H_1: $\exists a > 0$ so that $\Delta_T(\hat{\varphi}_j^T, \bar{\varphi}_j^T) > a/2 > 0$ for every T large enough.

Then a test of level α of H_0 can be associated to a rejected region:
$\Big(\Delta(\hat{\varphi}_j^T, \bar{\varphi}_j^T) > c_\alpha\Big) = \upsilon_\alpha$ with $P_{H_0}(\upsilon_\alpha) = \alpha$. The test is asymptotically consistent.

In practice, depending on the situation, we use two ways to estimate the distribution of $\Delta(\hat{\varphi}_j^T, \bar{\varphi}_j^T)$. When we do not know the true distribution of the considered functional (moments, density), we can have a good approximation of the distribution of Δ using a group of permutations. The principle of this method can be summarized as follows: Let $(P_1, ..., P_k)$ be a partition of $[0, T]$ in k disjoint intervals of the same length (for instant years) and $\pi \in \Pi$ a permutation of $(P_1, ..., P_k)$. Thus π gives also a permutation of the data. We note $\pi\varphi_j$ for $\varphi_j \circ \pi$. We have then $\pi\bar{\varphi}_j^T = \bar{\varphi}_j^T$ and $\{\Delta(\pi\hat{\varphi}_j^T, \pi\bar{\varphi}_j^T) = \Delta(\pi\hat{\varphi}_j^T, \bar{\varphi}_j^T), \pi \in \Pi\}$ gives an approximation of the distribution of Δ under H_0. When we know the true distribution of the considered functional (GEV for the maxima), the distribution of Δ can be approximated through simulation. We simulate the trajectories of $\varphi(t)$ with parameter $\bar{\varphi}_j^T$. For each trajectory, we calculate two estimators: constant and non parametric. The distance of these two estimators gives an approximation of the distribution of Δ under H_0.

Now in order to calculate the power of test for a function φ_j, we should know the distribution of $\Delta(\hat{\varphi}_j^T, \bar{\varphi}_j^T)$ under H_1. We can have an approximation of this distribution by using the bootstrap technique or by simulation.

In the case of functionals of moments, when we do not know the true distribution of the considered functional, let R_t the residuals of the estimator $\hat{\varphi}_j^T$. Let now R_t^* be a block bootstrap sample of R_t, by adding the estimator $\hat{\varphi}_j^T$ to this block bootstrap sample, we obtain a block bootstrap sample of the observed sample. By estimating the constant estimator and the non parametric estimator from this bootstrap sample, we obtain an approximation of $\Delta(\hat{\varphi}_j^T, \bar{\varphi}_j^T)$ under H_1. Thus we can compute $P_{H_1}(\Delta(\hat{\varphi}_j^T, \bar{\varphi}_j^T) > \zeta)$ which is the power of the test.

When we know the true distribution of the considered functional, we can obtain an approximation of $\Delta(\hat{\varphi}_j^T, \bar{\varphi}_j^T)$ from the simulated sample with parameter $\hat{\varphi}^T$.

4. Complete model and model for simulation

4.1. Continuous time process and extremes

Considering air temperatures, many statistical problems, linked with the mean, the extreme values or the construction of simulation models, require for physical coherence to consider the observations as a discrete sample of a continuous time process. Air temperatures are measured at different fixed lag Δ, from 10 minutes to 24 hours. Tests can be made in order to check the Markov property for these discrete time sampled data. If the Markovianity is not rejected, the continuous time process of temperatures, whose trajectories are continuous, can be considered as a diffusion process. Deeper physical considerations on the diffusive character of climate temperatures can be found in Sura (2012).

In this part, we work with reduced variables, tested as being cyclo-stationary. Thus if Markovianity is accepted as a physical hypothesis, we can consider that discrete reduced observations are provided by a cyclo stationary diffusion. For applications to temperatures, for different theoretical points, we suppose first that the reduced continuous time process $Z(t)$ is a stationary diffusion with inaccessible boundaries given by the stochastic differential equation:

$$dZ_t = b(Z_t)dt + a(Z_t)dW_t \tag{18}$$

where b is the drift, a the diffusion coefficient and W_t a Brownian motion. We choose $a \geq 0$ in order to have an identifiable model. Let ν be the invariant probability of $Z(t)$ given by:

$$\nu(x) = \frac{e^{\int^x 2\frac{b(v)}{a^2(v)}dv}}{a^2(x)} \tag{19}$$

then this means that there is no seasonality in the coefficients a and b. The cyclo-stationarity is considered in a second time.

We first address the behavior of the extreme values of Z_t, more specifically for the case where Z_t is bounded.

The seminal theory on the behavior of extremes is due to Berman (Berman, 1973) and reformulated by Davis (Davis, 1982). In the general case, Davis has shown that for any GEV distributions G and H, it is possible to find c_T and d_T such that the limit distribution of the centered and normed (by c_T and d_T) maximum M_T of the diffusion observed on $(0,T)$ tends to H as $t \to \infty$, although ν is in the domain of max-attraction of G.

From now on, we are concerned by a specific situation. The support of Z, say (r_1, r_2), is bounded so r_1, r_2 are inaccessible boundary points. We suppose that at the boundary, the following hypothesis are verified:

j/ a and b are defined and continuous on $[r_1, r_2]$

jj/ $b(r_1)b(r_2) \neq 0$ (checked in the applications)

We prove that in this situation, if there exists a limit distribution for suitably centered and normalized maxima $M_T = \{\max Z_t, 0 \leq t \leq T\}$ then the same is true for an iid sample of ν and the two limit distributions are the same GEV one.

Now consider the theorem 3.2 in Davis (1982).

Theorem 4.1. *Let s be the scale function of the process* (18)*:*

$$s(x) = \int^x e^{\int^u -2\frac{b(v)}{a^2(v)}dv} du \tag{20}$$

Define the function F by:

$$F(x) = exp(-1/s(x)) \tag{21}$$

If there exist some functions $T \to A_T \in \mathbb{R}^*$ and $T \to B_T \in \mathbb{R}$ and a proper distribution G such that

$$\frac{M_T - B_T}{A_T} \to G \text{ in distribution}$$

then G is a GEV distribution such that F is in the extreme domain of attraction of G.

We now prove the following theorem and its corollary which is very useful in our applications:

Theorem 4.2. *Under hypothesis **j/** and **jj/**:*
1- If the distribution of the maximum of the diffusion is in the domain of attraction of a GEV distribution with $\xi < 0$ then the marginal distribution ν of the diffusion is in the same domain.
2- Suppose that F is in the extreme domain of attraction of some GEV distribution G with shape parameter $\xi < 0$, let r_2 be the common upper bound of F and G.
We have the following behavior of a as $x \to r_2$:

$$a^2(x) = -2b(r_2)\xi'(r_2 - x) + o(r_2 - x) \tag{22}$$

with $\frac{1}{\xi} + \frac{1}{\xi'} = 1$

The proof is given in the Appendix.

Corollary 4.1. *as $x \to r_2$:*

$$\nu(x) \simeq \frac{-(r_2 - x)^{-\frac{1}{\xi}}}{2b(r_2)\xi'} \tag{23}$$

4.2. Discrete observations: statistical application of geometric ergodicity and extremes

The Markov chain Z_n of discrete observations is geometrically ergodic with invariant measure ν. In Dacunha-Castelle and Florens-Zmirou (1986), it is proved that the chain Z_n is L^2-positively contractive and thus geometrically ergodic. This implies that if $M_n = \max(Z_1, ..., Z_n)$ and $M_T = \max(Z_t, 0 \leq t \leq T)$, if there exist sequences A_n, B_n and functions A_T, B_T such that:

$$\frac{M_n - B_n}{A_n} \to G_1, \frac{M_T - B_T}{A_T} \to G_2 \tag{24}$$

then G_1, G_2 have the same shape coefficient ξ and an iid sample of ν satisfies a limit theorem for its extremes, with ξ as shape coefficient. This result is of course very useful for statistics, the data being discrete.

The behavior of ν given by (23) implies that the extremes of an iid sample of ν, suitably normalized, converges to a GEV distribution with shape coefficient ξ.

The lower and upper bounds of ν can then be estimated from the discrete data. Let $\hat{\mu}_n, \hat{\sigma}_n, \hat{\xi}_n$ be the estimators of the parameters of GEV associated to Z_n. $\hat{r}_{2,n} = \hat{\mu}_n - \hat{\sigma}_n/\hat{\xi}_n$ is the estimator of the upper bound of ν. We have an equivalent formula for the lower bound.

The coefficients a and b of the diffusion can be estimated directly by using the sequence (Z_n) but computations are very complicated in general(see Dacunha-Castelle and Florens-Zmirou, 1986). So we need an approximation.

The first order Euler scheme approximation is given by:

$$Z_{n\Delta} = c(Z_{(n-1)\Delta}) + a(Z_{(n-1)\Delta})\varepsilon_n \tag{25}$$

where ε_n is a white noise or more precisely a sequence of iid random variables with distribution E and Δ is the time interval between observations, we choose here $\Delta = 1$.

There is an important literature about these processes. The most important results concern stationarity, geometric ergodicity and the existence of an invariant density λ such that the extremes of an iid sample of λ have the same asymptotic properties than the extremes of the chain Z_n. Some important cases and tools can be found in Doukhan (1994) where various criteria are given for the geometric ergodicity. Some papers (Chen and An, 1999 , Wang, 2008) give more specific conditions. If $c(x) = cx$ with $|c| < 1$, the main conditions are: $a > 0$, a is bounded on $\mathbb{R}$ and the density of ε is almost everywhere positive. So bounded supported noises are excluded. Now if a is uniformly Lipschitzian, the existence of a unique stationary solution as the geometrical decrease of the covariance and strong mixing properties are proven. But in our application, from the previous results of the extreme theory, a is not Lipschitzian at the boundary for $\xi < 0$. Moreover, a is zero out of a bounded interval $I = (r_1, r_2)$.

We have to look for a model with bounded marginals, but unfortunately we have the following almost evident (and thus given without proof) negative result.

Lemma 4.1. *If E has a non bounded support (for instance E Gaussian) then there are not bounded supported stationary solutions of* (25).

It is thus impossible to have such an Euler first order scheme with the following wanted properties: it is stationary with bounded inaccessible boundary and Gaussian innovation. Then what are the possible improvements for this situation?

First we try keeping at least in the first step a Gaussian noise. We prove results for stationarity and geometric ergodicity adapted to this situation.

We give only one of the results obtained by applying the work of Cline and Pu (1999).

Theorem 4.3. *If E is a Gaussian distribution, $c(x) = cx$, $|c| < 1$ and $a(x)$ is continuous and null outside a bounded interval then the process Z is geometrically ergodic.*

These results will be used only to justify the simulation process.

We now modify the Euler scheme in order to obtain a bounded process. We define a new Markov chain $\tilde{Z}_n$ by:

$$\tilde{Z}_n = b\tilde{Z}_{n-1} + a(\tilde{Z}_{n-1})\eta_n(\tilde{Z}_{n-1}) \tag{26}$$

where η_n is conditionally independent of σ-algebra generated by $\left(\tilde{Z}_{n-2}, ..., \tilde{Z}_{n-k}, ..., \eta_{n-1}, ..., \eta_{n-k}\right)$ given $\tilde{Z}_{n-1}$.

Let $\psi_2(x) = \frac{r_2 - bx}{a(x)}$, $\psi_1(x) = \frac{bx - r_1}{a(x)}$, $\Phi(x)$ the Gaussian cumulative distribution function, $R(x) = \Phi(\psi_2(x)) - \Phi(\psi_1(x))$, $N_\psi(x)$ the normal distribution truncated at $\psi_1(x)$ and $\psi_2(x)$ whose density is $f(\eta) = \frac{1}{R(x)\sqrt{(2\pi)}} \exp^{-\eta^2/2} \mathbf{1}_{\psi_1(x) < \eta < \psi_2(x)}$.

4.2.1. Estimation procedure with constraints for extremes

We start from:

$$Z_{n+1} = c(Z_n) + a(Z_n)\varepsilon_n, \quad \varepsilon_n \sim N(0,1) \tag{27}$$

The informations of the theorem (4.2) on the behavior of a^2 in the continuous-time process is used in a plug-in manner: the likelihood of the Euler scheme is considered as a Gaussian contrast function. As shown in part 4.2.1, we can estimate the lower and upper bounds $\hat{r}_1, \hat{r}_2$ by applying the GEV method to the block maxima of the reduced series. It remains now the estimation of a and b with bound constraints.

There is a very large literature about the estimation of parameters a and b from a discrete observation, such as Florens-Zmirou (1989), Hansen et al. (1998), Kessler and Sorensen (1999) and Gobet et al. (2004). We can separate the results into:

1. non parametric results, considering c and a respectively as the conditional mean $E(Y_{n\Delta}|Y_{(n-1)\Delta})$ and conditional variance $a^2(Z_{(n-1)\Delta}) = E(Z^2_{n\Delta}|Z_{(n-1)\Delta}) - \left[E(Z_{n\Delta}|Z_{(n-1)\Delta})\right]^2$. In the following, we will use these methods only to give preliminary results.

2. Methods working with a fixed Δ using the likelihood of the first order Euler scheme associated to the diffusion (Florens-Zmirou, 1989).

For caution, we first estimate both the drift coefficient and the diffusion coefficient by nonparametric methods. Knowing the distribution of ε, we can estimate at the same time a^2 and c by the cubic spline smoothing by maximizing the penalized likelihood.

The penalized likelihood is then given as a function of a^2 and c by:

$$L(N, a^2, c) = \sum_{i=2}^{N} \log p(\tilde{Z}_{n-1}, \tilde{Z}_n, a, c) - \lambda \int (a''(x))^2 dx - \mu \int (c''(x))^2 dx \tag{28}$$

The regularization parameters λ, μ can be chosen by modified generalized cross-validation in order to take into account the dependence. The method is shown in Hoang (2010) (chapter 5). However, to avoid a too heavy numerical computations, we estimate a^2 and c separately. We estimate first the drift coefficient c in the supplied model by local smoothing Loess with c being the conditional expectation $E(Z_n/Z_{n-1})$. Then with given $\hat{c}$, we estimate a^2 by maximizing the penalized likelihood for a^2.

Our main interest is to find a parametric simulation model. Both coefficients are then estimated parametrically. In general for temperatures, we can check that the drift c is linear. The drift function has a quite clear physical meaning as the elastic part of the basic oscillator, which is the deterministic justification of the diffusion. We can write then $\hat{c}(x) = cx$, with $c < 0$. The shape of the diffusion coefficient a can be quite complicated. From previous non parametric studies and the property of a in the continuous-time diffusion, we suggest a strictly positive a in the interval (r_1, r_2) and zero outside.

To choose a parametric model which satisfies the constraints given by extreme theory, we model a in the following way

$$a^2(x,\beta) = (x-r_1)(r_2-x)\,|P_k(\beta,x)|^2\,\mathbf{1}_{r_1<x<r_2} \tag{29}$$

$P_k(\beta,x)$ is a trigonometric polynomial of order k with parameters β, k chosen using AIC, submitted to the following quadratic constraints provided by the results in section 4.1:

$$(a^2)'(\hat{r}_2,\beta) = 2(1-c)(\hat{r}_2)\xi'$$

so we plug in $\hat{r}_2$ and we consider the constraints given by the a behavior as coefficient of the diffusion:

$$(c-1)r_1\xi' = (r_2-r_1)\,|P_k(\beta,r_1)|^2 \;\text{ and }\; (1-c)r_2\xi' = (r_2-r_1)\,|P_k(\beta,r_2)|^2 \tag{30}$$

Once r_1 and r_2 are estimated , we use a Gaussian constrast (which is not the likelihood of the observed Markov chain) defined by:

$$L(N,k,c,\beta) = \sum_{i=1}^{N}\left[-\frac{1}{2}\frac{(\tilde{Z}_i - c(\tilde{Z}_{i-1})^2}{\left|P_k(\tilde{Z}^j_{i-1},\beta)\right|^2} - \log\left|P_k(\tilde{Z}^j_{i-1},\beta)\right|\right] \tag{31}$$

Estimating a^2 and c turns now to maximize (31) with boundary constraints. In practice, we first estimate c by least squares, then we estimate a from the series $(Z_t - \hat{c}(Z_{t-1}))^2$ with constraints. The estimation of a becomes an optimization problem with equality and inequality constraints. We first estimate a by least squares with constraints, using the algorithm in Golfarb and Idnani (1982) and Godfarb and Idnani (1983). Then we use the results of least square estimation as the initial values for the parameters in the likelihood maximization. Using the algorithm of Nelder and Mead (1965), we obtain the estimation of a^2.

Finally, outside the intervals $(\hat{r}_1, \hat{r}_2)$, $\widehat{a^2}$ is fixed as zero.

5. Applications to temperatures and simulations

5.1. Data

The validation of the model is conducted for different climates in Eurasia and in the United-States. For Eurasia, weather station time series of minimum daily temperature (TN) and maximum daily temperature (TX) are obtained from the ECA&D project database. The project gives indications of homogeneity through the results of different break identification techniques (A.M.G., 2002). First, the series which could be considered as homogenous (stated as ŞusefulŤ in the database) over the period 1950-2009 have been selected for both TN and TX. Then, only the time series with less than 5% missing data are kept, leading to 106 series for TX and 120 for TN (many TX series, mostly in Russia, have missing values from 2007 onward whereas the corresponding TN ones have missing values only in 2009). For the United States, weather station TX and TN time series are obtained from the Global Historical Climatology Network Ű Daily Database (GHCN daily) (Menne et al., 2012). A similar selection procedure left us with 86 series for TX and 85 for TN. The studied stations are shown in the figure 1.

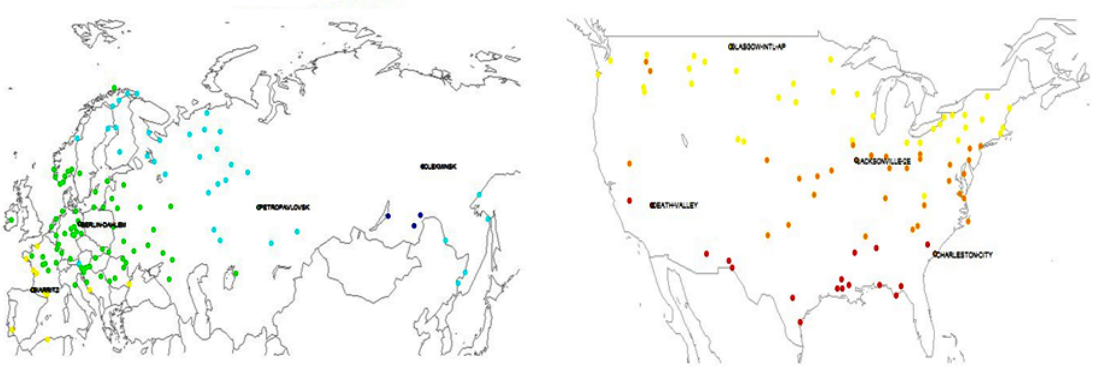

FIGURE 1. *Studied stations in Eurasia and USA*

We will show just the results of some stations for example. To have a complete panorama, all the results from the preprocessing to the simulation of the reduced series will be shown.

5.2. From the preprocessing to the reduced series

We give an example of the preprocessing applied to the minimum daily temperatures (Tmin) in Berlin (Germany). Trends and seasonalities, are estimated by the procedure explained in the section 2.

Figure 2 shows the estimated seasonalites and trends for Tmin in Berlin. We observe an annual seasonality, for example the variance is more important in winter. For the estimation of trends, our cross validation algorithm gives 0.17 as the optimal value for the smoothing parameter. This parameter corresponds to a size of partitions g of 10 days. We remind that the main idea of this algorithm is to partition the data in $[n/g]$ sub-groups and in each sub-group, the observations of time lag of g days, and so we have the sub-groups of "sufficiently independent" data. The optimal

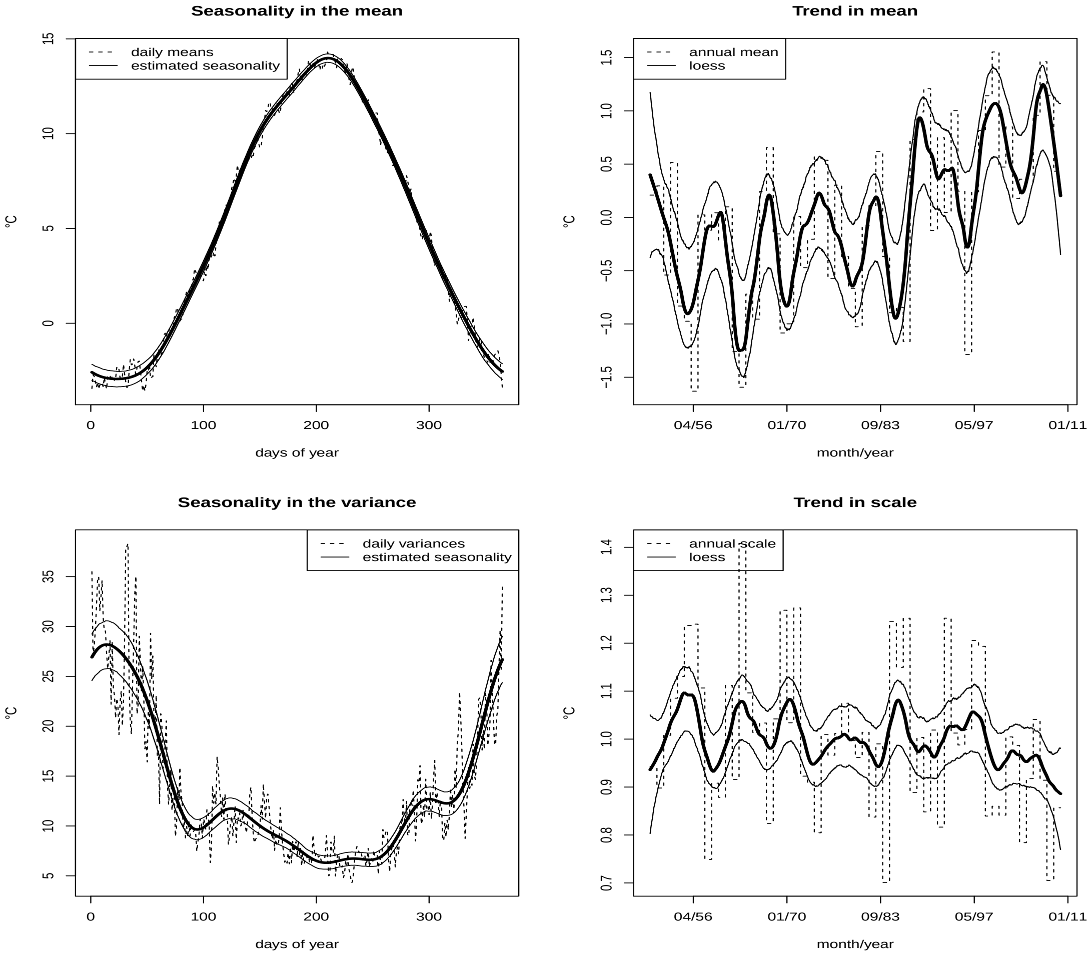

FIGURE 2. *Estimated seasonalites and trends in preprocessing for Tmin in Berlin (with 95% bootstrap confidence intervals)*

choice of the bandwidth h is predefined by the optimal choice of the size g which is also supplied in our algorithm.

Once the deterministic components removed, the stationarity or rather cyclo-stationarity of the reduced series will then be checked.

5.3. *Cyclo-stationarity of the reduced series*

The reduced series is in fact not stationary in the usual way, but in an extended sense: cyclic stationarity. As mentioned in Hoang (2010), the seasonality of the temperature has its own dynamic which cannot be removed: after removing the trend and seasonal components in mean and in variance from the observed series X_t, the seasonality still remains in the reduced series $\hat{Y}_t$.

This behavior can be firstly seen in the boxplot of $\hat{Y}_t$ for year (figure 3). The difference of the winter compared with the rest of the year can be detected.

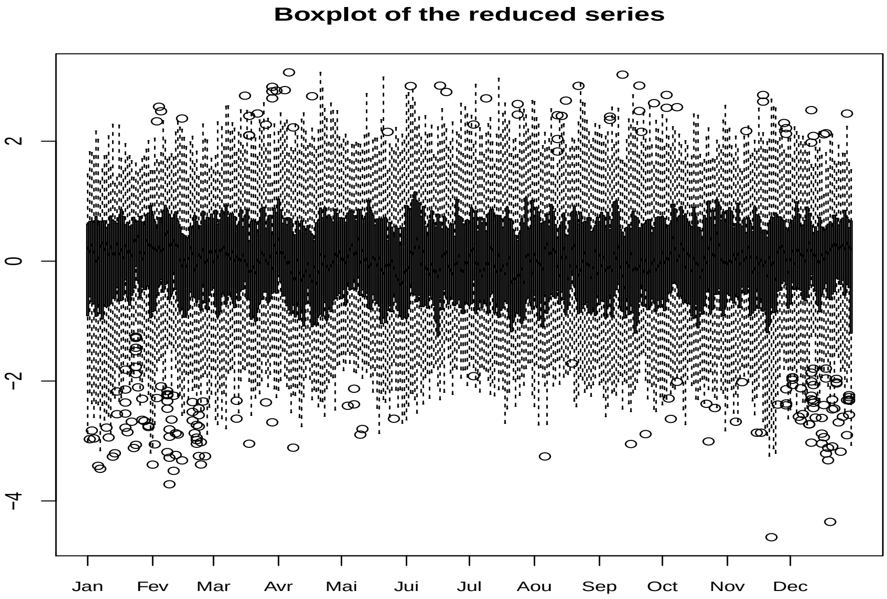

FIGURE 3. *Boxplot of the reduced series*

With a push-over analysis, it can be seen that the seasonality affects on different characteristics of the reduced series. Figures 4 and 5 which contain the optimal periodic approximations (by using the AIC criterion) for the moments and correlations of the reduced series show an example of this phenomenon. The seasonal effect is often found in skewness and kurtosis. In the correlations, the seasonality is always present. For this reason, the model that we will use is SFHAR (seasonal functional heteroscedastic autoregressive) model where both coefficients a and c are periodic.

The test of cyclo-stationarity presented in part 3 will be applied to different characteristics of $\hat{Y}$: mean, variance, skewness, kurtosis, correlations and the extremes. For moments, the distribution of the distance Δ is estimated using yearly permutations and for extremes, the empirical distribution of Δ is obtained by simulation using GEV distribution.

According to the results of the tests with 10% confidence level, the cyclo-stationarity of the reduced-series is not rejected for almost all the stations. In general, the non validation of the cyclo-stationarity comes from the reject of a constant trend for the extreme parameters. The constancy of the location parameter μ is rejected in 2% of the cases and for the scale parameter σ, that is about 10%. For both μ and σ, we have about 9% as the percentage of reject (for the details, see Parey et al., 2013). For a multidimensional test (describe in part 3) with a 10% confidence level, we can not reject the stationarity of the extremes. The validity of the cyclo-stationarity of the extremes is necessary to use the extreme theory of the diffusion process for modeling the extremes of the reduced series.

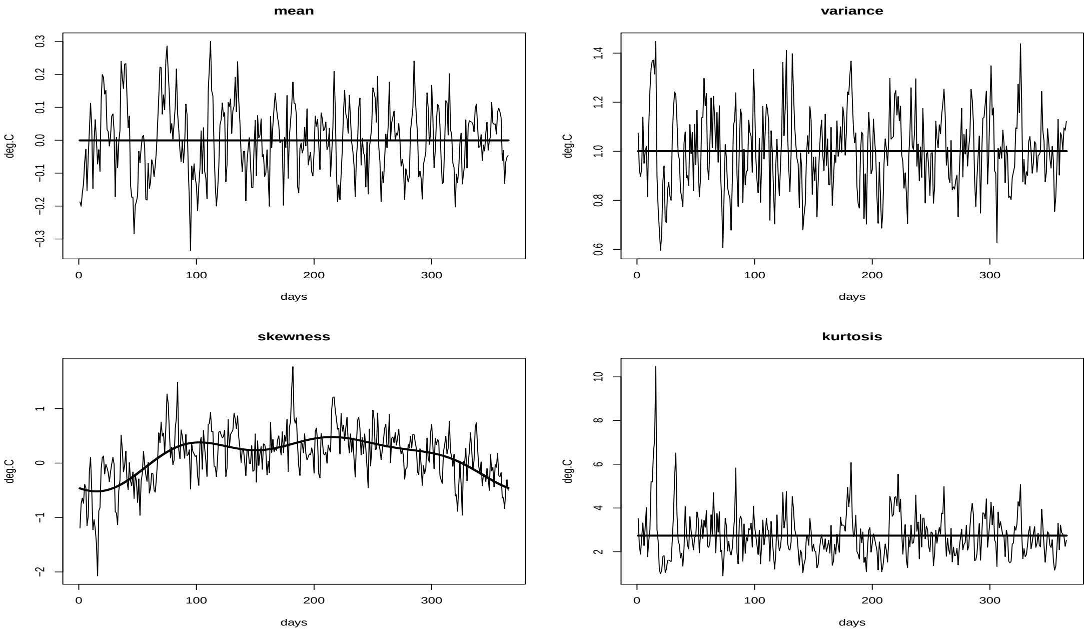

FIGURE 4. *Estimation of the seasonality in the moments of the reduced series: daily values and periodic estimation (bold line)*

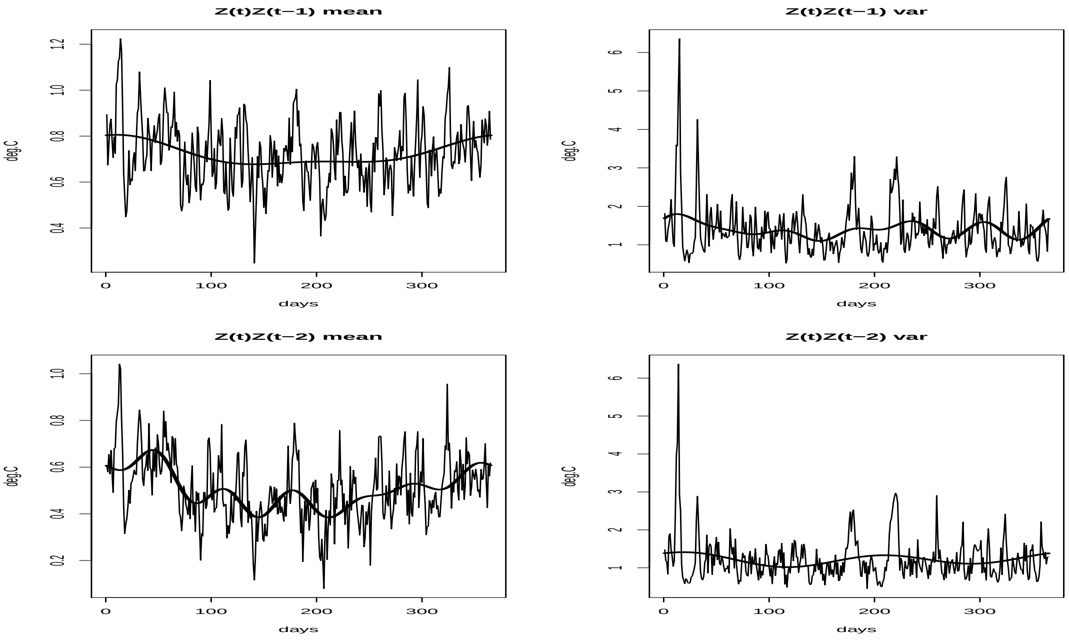

FIGURE 5. *Estimation of the seasonality in the correlations of the reduced series: daily values and periodic estimation (bold line)*

In order to compute the power of the test, we have to build the distribution of Δ under H_1 hypothesis. We compute then distances between constant and non-parametric estimates of the GEV parameters from simulated trajectories for non stationary GEV. We calculate next the percentage of values in the empirical table of Δ which do not fall in the rejection region. The results from different stations give 85% as the minimum value as the power of test.

5.4. Model, simulation and validation

5.4.1. The model

In this section, we will show an extended version of the Euler scheme model which is used to fit the dynamics and different intrinsic characteristic of the reduced temperature.

Based on the previous preliminary study on the characteristics of the reduced series, we choose a model adapted to different characteristics (especially for the extremes) for the reduced temperature (in the following $P_q(t)$ is a trigonometric function of period $A = 365$ of order q):

$$\begin{gathered} Z_t = c(Z_{t-1}) + a(Z_{t-1})\eta_t(Z_{t-1}) \\ \text{where } c(Z_{t-1}) = P_p(t)Z(t-1) \\ \eta_t \text{ defined in section 4.2} \end{gathered} \tag{32}$$

The diffusion coefficient a is also seasonal and is constrained on the boundary. Then we define a in the following way:

$$\begin{aligned} & a^2(t, Z_{t-1}) = (\hat{r}_2 - t)(t - \hat{r}_1)\sum_{k=0}^{5} P_{p',k}(t)Z_{t-1}^k \\ & (a^2)'(r_1) = \frac{2(c(r_1) - r_1)}{1 - 1/\xi_1}, \quad (a^2)'(r_2) = \frac{2(c(r_2) - r_2)}{1 - 1/\xi_1} \\ & a^2(t) > 0 \ \forall t \end{aligned} \tag{33}$$

In practice, we first estimate the autoregressive part from the series Z_t, the numbers p, p' of cosinus and sinus terms in $P_p(t)$ and $P'_p(t)$ is chosen by the Akaike criterion. Then we estimate a from the series $(Z_t - \hat{c}(Z_{t-1}))^2$ using the maximum likelihood with constraints (described in part 4.2.1).

We observe that a is not constant but linear in the central part. For example, for the daily maximum temperature, a increase in summer and for the daily minimum temperature, the opposite effect is observed, a decrease in winter. In the following, we give an example (the minimum daily temperature in Berlin) of the form of estimated a (with constraints in the boundary) compared with the estimations without constraints for one day in summer and one day in winter. Without constraints in the boundary, the estimation of a for the high quantiles of Z has no sense because of the lack of data. Moreover, when we use a parametric estimation for a without boundary constraints, it tends to give high values near the boundaries for a and the results of this is the unrealistic values (tend to infinity) obtained by the simulation.

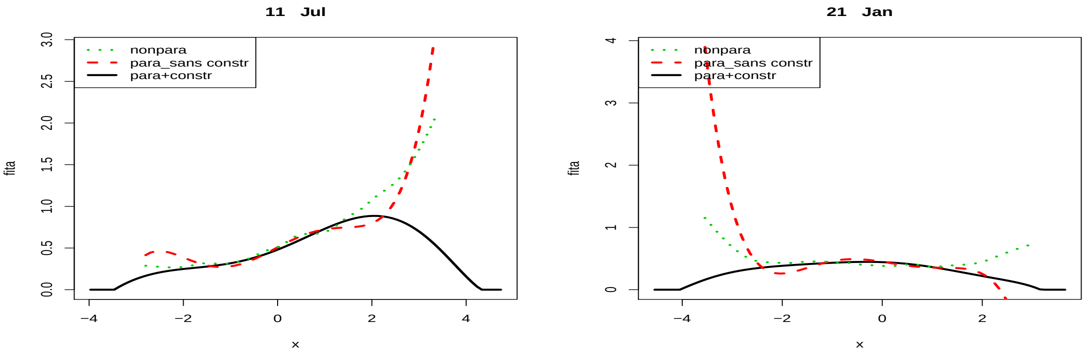

FIGURE 6. ***Left panel**. Estimated diffusion coefficients by different methods for 11th July. **Right panel**. Estimated diffusion coefficients by different methods for 21st January*

5.4.2. Simulation and validation

Usually, calibration is done on a sample used for learning and validation on a sample test to make predictions for instance. We do not detail here the predictive qualities. The main reason is that the model is a martingale and thus the non linearity does not change the predictor at least for prediction at one step.

We first consider the estimated residuals $\widehat{\varepsilon}_t$ of the model (32). $\widehat{\varepsilon}_t$ must be close to a stationary white noise. Most values of the autocorrelation function of the residuals and squared residuals are found in the confidence intervals of a white noise. The normality of the residuals is often rejected by Komogorov-Smirnov and Shapiro Wilk tests as expected from the modification of the Euler scheme for the tails.

In order to validate the model, we simulate 100 samples (with the same length of observed sample) based on the model above. We compare the following items for the series X_t and their simulated samples: moments, marginal distribution, quantiles, GEV parameters for Z_t and the hot and cold waves.

Compared to the other temperature generators found in the literature, our model differs in its bounded property. This property is expected to really improve the ability of the model at reproducing extremes. In order to have a better view on this point, we will compare the simulation results of our model to those of simpler models, where the extremes are not taken into account.

Following the estimation and simulation procedure:

- Preprocessing : estimate the mean function m of X, the scale function s of X, the additive and multiplicative seasonalities u and v
- Calculate $\hat{Y} = (X - \hat{u} - \hat{m})/(\hat{s}\hat{v})$ and apply model (32) to Z

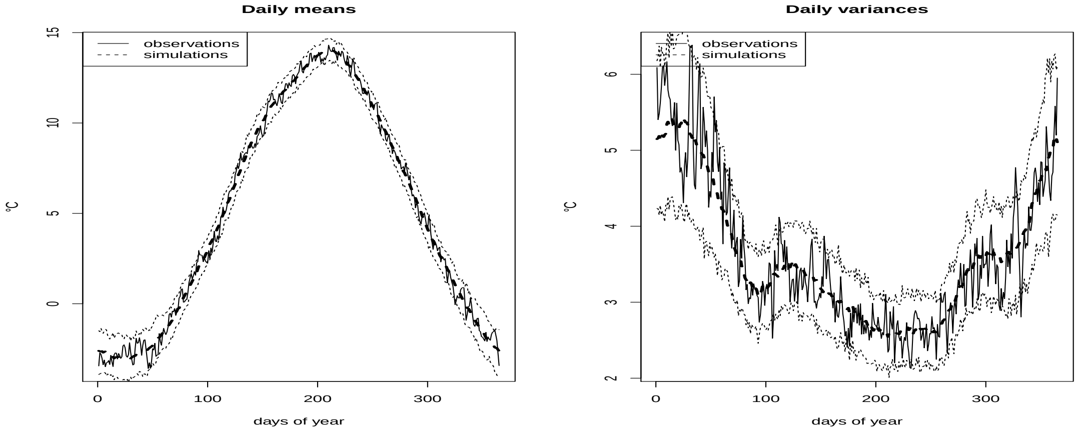

FIGURE 7. *Comparison of the observed and simulated (with 95% confidence intervals) daily means and variances*

• Estimate a and c with the method described in the previous section.

• Use these $\hat{a}$, $\hat{c}$ to create 100 simulated samples of Z.

• Deduce 100 samples of X by adding m,u and multiplying s,v : $X = svZ + m + u$

The results are quite better when the observations are not too asymmetric and their kurtosis is close to that of a Gaussian distribution.

The simulations represent correctly the daily characteristics:

- Daily distributions: 365 Komogorov-Smirnov tests are applied to test the homogeneity between the daily distributions of simulations and of observations. With a 5% confidence level, the homogeneity of simulated daily distributions with respect to observed ones is not rejected. We do not discuss here the problem of multi-hypothesis,
- Daily mean and variance: the simulations represent well the observed daily mean and variance (Figure 7 for Tmin in Fruholmen). It is normal that the estimators from simulations are smoother because for each day, we have 100 times more values than the observations,
- Daily skewness and kurtosis: the results are nevertheless less good for daily skewness and kurtosis (Figure 8 for Tmin in Fruholmen).

Remark 5.1. *In figure 8, the skewness and kurtosis corresponding to the 42th day and 355th day are remarkably low or high. The reason is that for these days, we have some very low temperatures: -21.6řC for the 42th day and -24.5řC for the 355th day.*

The quantiles between 1% and 99% are correctly represented: observed quantiles are found in the 95% confidence intervals of the simulations.

Let us now consider, at the same time, other models: constant a or a^2 is a trigonometric function

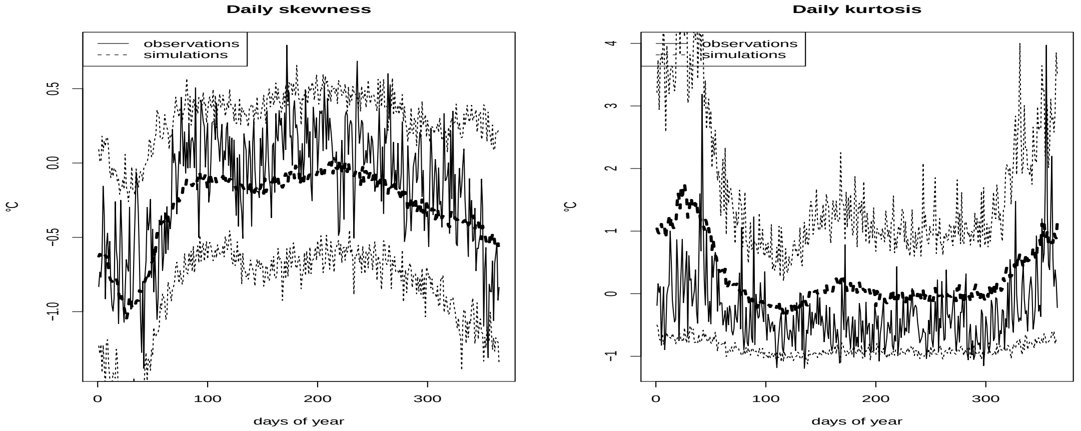

FIGURE 8. *Comparison of the observed and simulated (with 95% confidence intervals) daily skewness and kurtosis*

$f(t)$ which only depends on the dates, and not on the state $Z(t-1)$.

The example below (Figure 9) on Tmin in Berlin shows the better performance of our model for almost of quantiles deal with the others.

The construction of the model allows us to have bounded simulated trajectories, more concretely with the values remaining inside two estimated bounds. We then go further in this characteristics by considering the extreme events. For this, first the GEV parameters for the simulated reduced series are compared to those of the observed ones, both for the lowest and the highest extremes. The results show that the shape parameter, which determines the domain of max attraction, is better reproduced in the simulations than the location and scale parameters. We note that the estimation of the extreme parameters is very sensible to the size of block maxima.

According to the results, the simulations are less bounded. This fact is not surprising because the simulations produce 100 possibilities, among which higher or lower extremes could have been observed. Thus the model is not only able to reproduce extremes, but also to produce larger extremes than observed. Figure 10 shows, as an example, the distributions for each parameter (location μ, scale σ and shape ξ) obtained from the 100 simulated trajectories for the highest (warm) extremes (upper panels) and the lowest (cold) extremes (lower panel) together with the same parameters obtained from the observed reduced series (red line) for TN in Berlin and TX in Death Valley.

Remark 5.2. *Sometimes, with some size of block, the shape parameter ξ estimated is very close to zero, the direct consequence is that the estimated bounds are not realistic. In this case, it is sufficient to change the size of blocks of maxima to obtain a reasonable value of ξ. One can also have a more stable estimation of ξ using a non parametric Hill estimator.*

Then the ability of the model to reproduce heat or cold waves has been investigated. Cold waves are defined as periods of consecutive days with daily minimum temperature lower than the 2nd percentile and heat waves as periods of consecutive days with daily maximum temperature

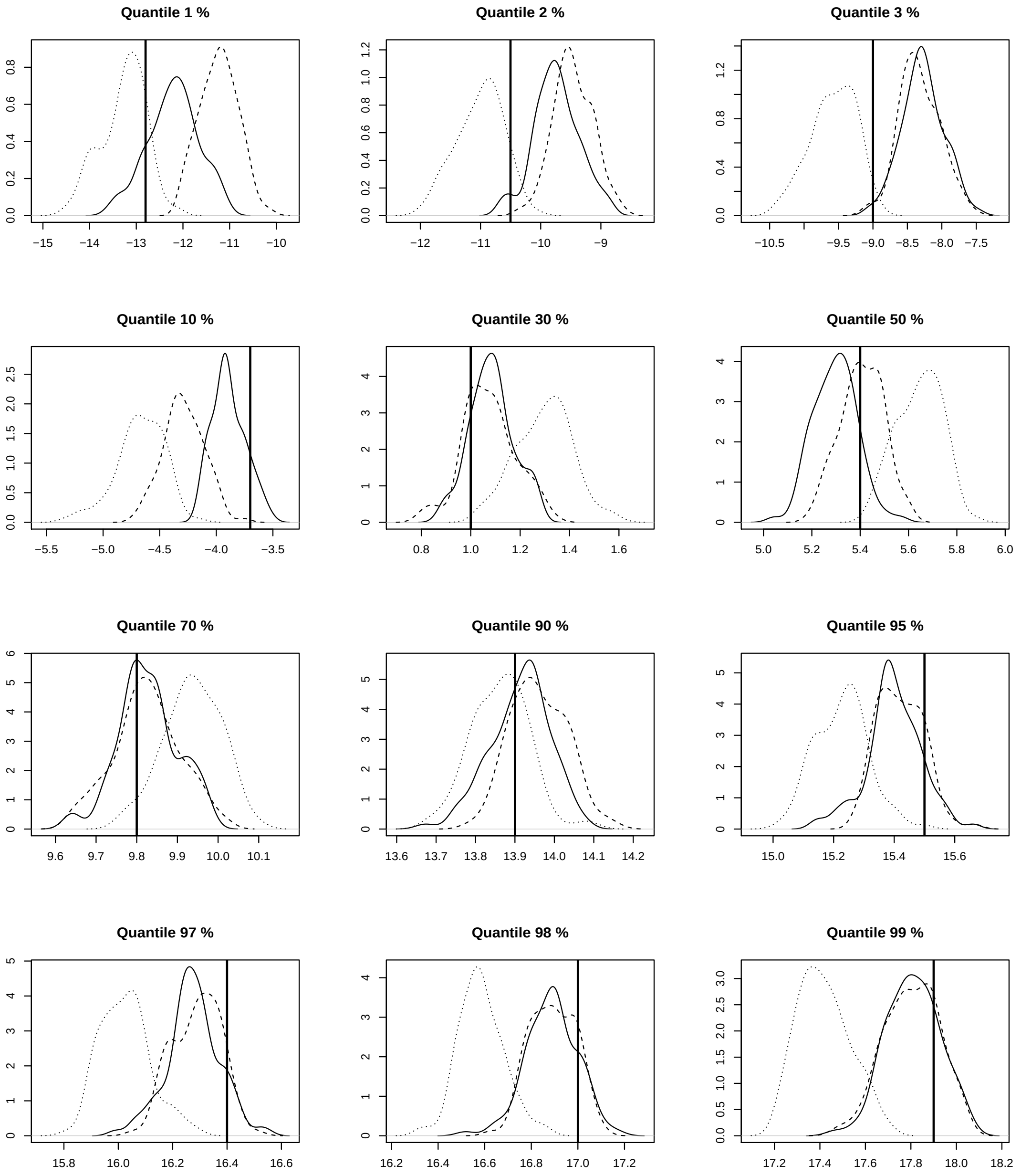

FIGURE 9. *Observed quantiles (vertical lines) for whole year of X_t and their distributions built from the simulations of different models: in black, model with constant a (dotted lines), models with $a^2 = f(t)$ (interrupted lines), model with $a(t, Z_{t-1})$ (continuous lines).*

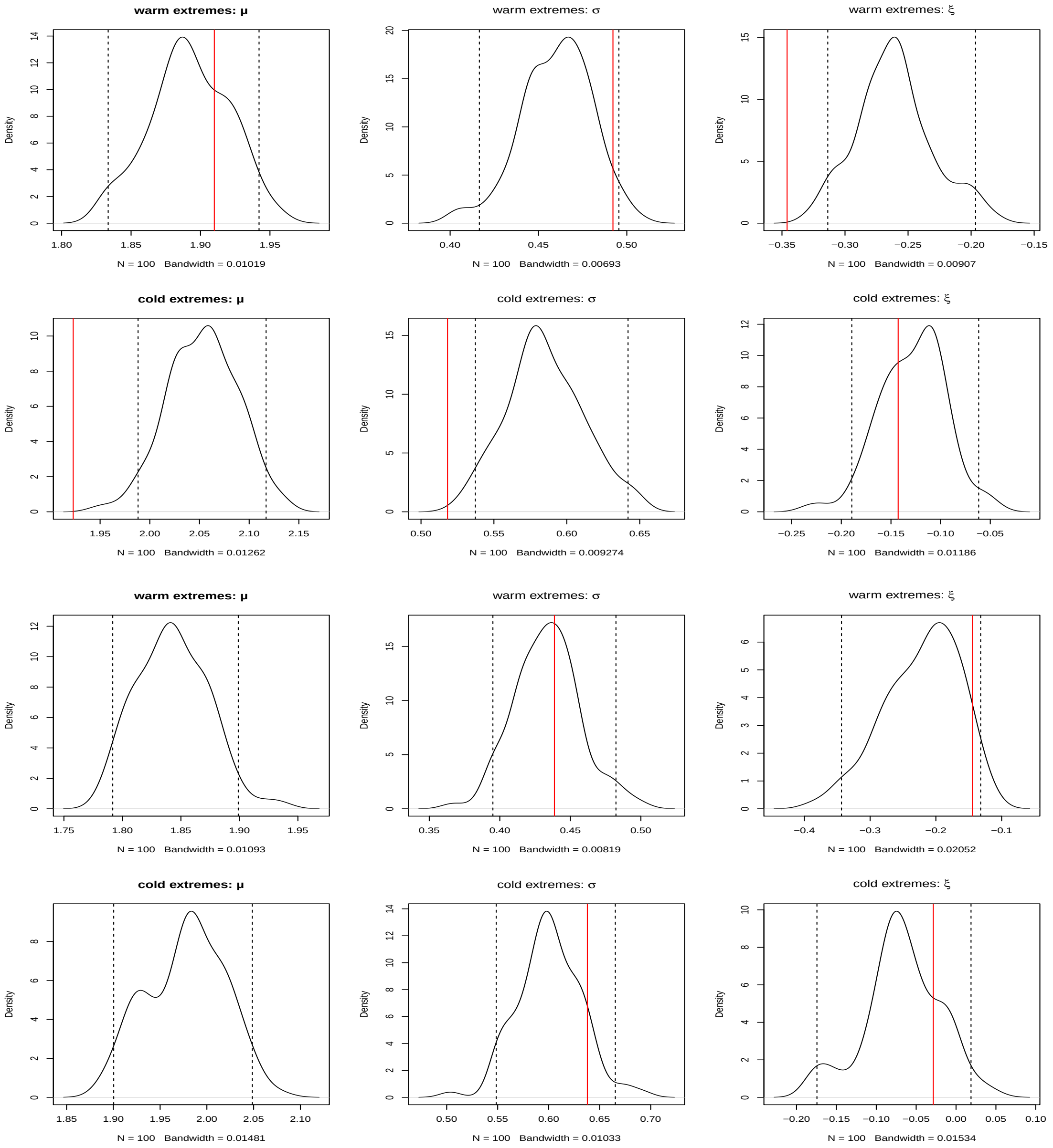

FIGURE 10. *Estimated extreme parameters (red vertical lines) from the observed reduced series and their distributions built from the simulations. Upper panel. Tn of Berlin. Lower panel. Tx of Death Valley*

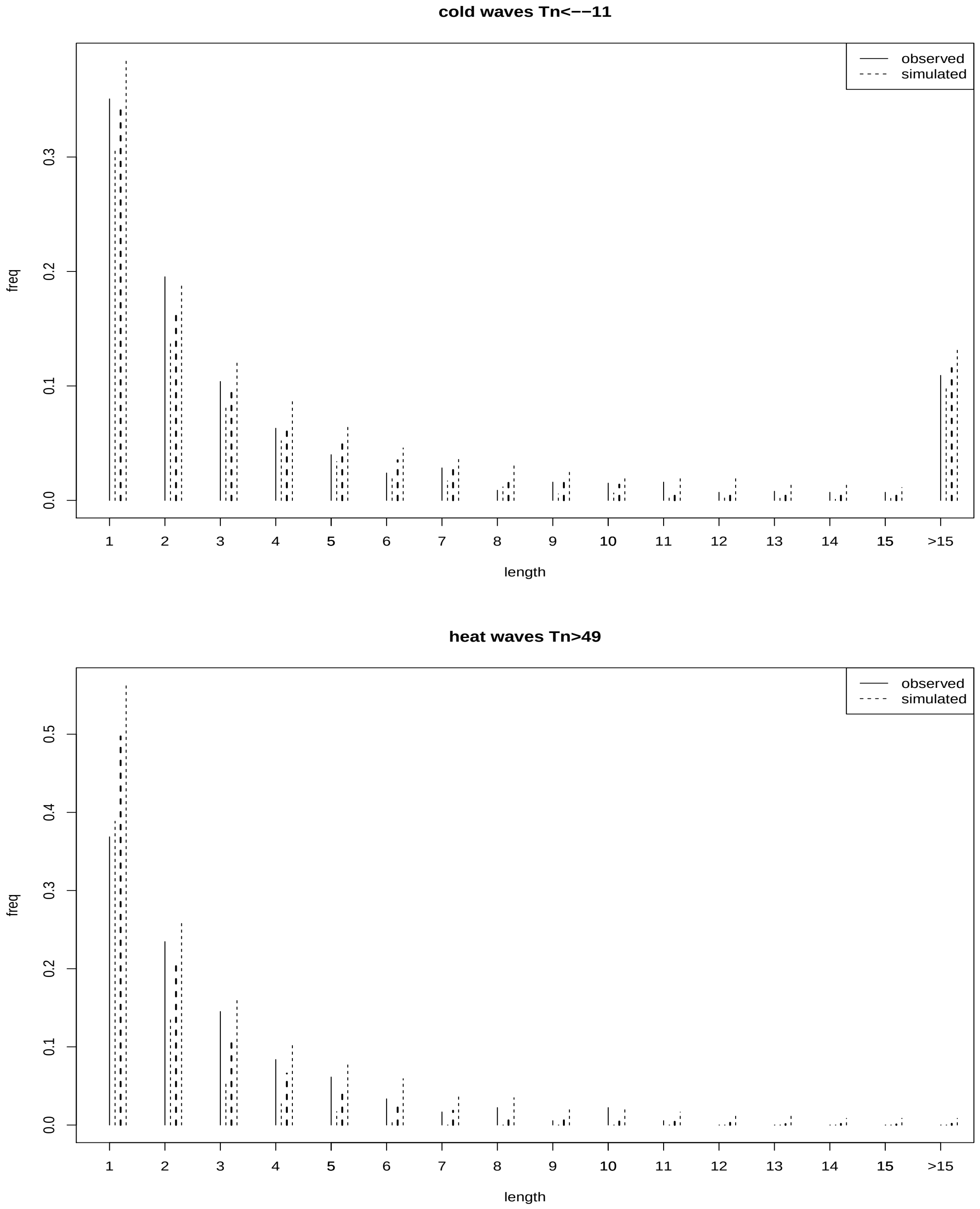

FIGURE 11. *Observed frequencies (red vertical lines) of the cold and hot waves and their distributions built from the simulations (in black).*

above the 98th percentile. The number of consecutive days varies between 1 and 15 days, the last class corresponding to the few episodes with more than 15 days, if any. Thus for each location the 2nd and 98th percentiles of the observed time series are computed and the distribution of episodes in the observed time series is compared to the minimum, maximum and mean frequencies of such a distribution in the 100 simulated trajectories. The results are in general good: even though the stochastic model sometimes tends to overestimate the proportion of 1-day cold excursions compared to the observations, it is still able to produce longer episodes in a reasonable proportion, even the longest ones. Figure 11 shows the results for cold waves in Berlin and heat waves in Death Valley.

Another criteria as the return level can be considered, but this point is not discussed here, for more details, one can consult Parey et al. (2013).

With these results, we can conclude that the estimation of a combining the extreme theory in diffusion process is an interesting approach for both the non extreme and extreme parts of the maximum and minimum daily temperature, especially for the extremes.

6. Conclusion

The possible perspectives of our paper are some improvements of this model and its presentation for other scale of time as hour measurement. For other climate variables, the situation may be different. For wind, it should be interesting to develop similar ideas. Some similarities are found between wind and temperature. In general, wind speed is measured in continuous records or discrete records with very short time intervals. The inertia of the anemometers requires careful analysis but a simulator for temperatures similar to ours can be applied to wind taking into account the mentioned difficulty of measurement. However, this is quite different from the developments devoted to precipitation generators (see Wilks and Wilby, 1999 for a review) where both the occurrence and length of the precipitating events and the amount of precipitation must be adequately simulated. Most generators contain separate treatments for the precipitation occurrence and intensity processes. In fact, this kind of models of simulation can also be introduced as part of more global simulators, based on “weather type” and hidden Markov chains. The amount of statistical work should be more important.

7. Appendix

Proof of theorem 4.2

The proof is based on the proposition 0.7 in Resnick (2007). where RV_{φ} is the set of functions with φ regular variation.

Theorem 7.1. *(Resnick). Suppose $U : R^+ \to R^+$ is absolutely continuous with density u so that*

$$U(x) = \int_0^x u(t)du \tag{34}$$

If $U \in RV_\varphi$, that means U is regularly varying with index φ , $\varphi \in R$ and u is monotone then

$$\lim_{x \to \infty} \frac{xu(x)}{U(x)} = \varphi \tag{35}$$

and if $\varphi \neq 0$ then (sgn φ)$u(x) \in RV_{\varphi-1}$

Following the proposition 1.13 in Resnick (1987), if F is in the domain of attraction of the Weibull distribution with $\xi < 0$, then $1 - F(r_2 - x^{-1})$ is regularly varying with index $1/\xi$ when $x \to \infty$.

From the previous theorem of Davis, we have

$$1 - F(r_2 - x^{-1}) = 1 - exp(-1/s(r_2 - x^{-1})) \propto 1/s(r_2 - x^{-1}) \text{ when } x \to \infty$$

Then $1/s(r_2 - x^{-1})$ is regularly varying with index $1/\xi$, so $s(r_2 - x^{-1})$ is regularly varying with index $-1/\xi$.

Note $s^*(x) = s(r_2 - x^{-1})$, $s^*(x) \in RV_{-1/\xi}$, the first and second derivative of $s^*(x)$ are monotone. Applying twice the previous proposition of Resnick, we have:

$$\lim_{n \to \infty} \frac{xs^{*''}(x)}{s^{*'}(x)} = -\frac{1}{\xi} - 1 \Leftrightarrow \lim_{n \to \infty} \frac{x\left[s''(r_2 - x^{-1}\frac{1}{x^2} - \frac{2}{x}s'(r_2 - x^{-1}))\right]}{s'(r_2 - x^{-1})} = -\frac{1}{\xi} - 1$$
$$\Leftrightarrow \lim_{n \to \infty} \frac{1}{x} \frac{s''(r_2 - x^{-1}}{s(r_2 - x^{-1})} = 1 - \frac{1}{\xi}$$

We can then deduce the behavior of a near the upper bound with $t = r_2 - x^{-1}, x \to \infty$:

$$a^2(t) \approx \frac{-2b(t)(r_2 - t)}{1 - \frac{1}{\xi}} \tag{36}$$

From this lemma, we can deduce the value of the first derivative of a^2 at the upper bound r_2 :

$$(a^2)'(r_2) = \frac{2b(r_2)}{(1 - 1/\xi)} \tag{37}$$

and the same for the lower bound. From the lemma and the formula of ν, the elementary computations prove that the tail of $\nu(x)$ has the same behavior as the tail of F given by

$$F(x) = exp(-1/s(x))$$

In fact, from (36), we have when $u \to r_2$, $e^{\int^u 2\frac{b(v)}{a^2(v)}}$ is equivalent to $(r_2 - t)^{1-1/\xi}$ with $t \to 0$, then $\nu(v)$ is equivalent to $C(r_2 - t)^{-1/\xi}$ (C is an appropriate constant) which in turn proves that the extremes computed on a sample of i.i.d random variables of distribution $\nu(x)dx$ have the same shape coefficient as that of the diffusion.

So we have proven theorem 4.2.

References

A.M.G., K. T. e. a. (2002). Daily datasets of 20th-century surface air temperature and precipitation series for the european climate assessment. *Int. J. of Clim.*, 22:1441–1453.

Benth, J. and Benth, F. (2011). A critical view on temperature modelling for application in weather derivatives markets. *Energy Economics*, 34:592–602.

Berman, S. (1973). Maxima and large excursions of stationary gaussian processes. *Trans.Amer.Math.Soc.*, 160(67-85).

Campell, S. and Diebold, F. (2005). Weather forecasting for weather derivatives. *Journal of the American Statistical Association*, 100:6–16.

Chen, M. and An, H. (1999). The probabilistic properties of the nonlinear autoregressive model with conditional heteroskedasticity. *Acta Mathematicae applicatae sinica*, 15.

Chu, C. and Marron, J. (1991). Comparison of two bandwidth selectors with dependent errors. *Ann. Stat.*, 19:1906–1918.

Cleveland, R., Cleveland, W., Mcrae, J., and Terpenning, I. (1990). Stl: a seasonal-trend decompostion procedure based on loess (with discussion). *Journal of Official Statistics*, 6:3–73.

Cleveland, W. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74:829–836.

Cline, D. B. and Pu, H. (1999). Geometric ergodicity of on linear time series. *Statistica Sinica*, 9:1103–1118.

Dacunha-Castelle, D. and Florens-Zmirou, D. (1986). Estimation of the coefficients of a diffusion from discrete observations. *Stochastics*, 19:263–284.

Davis, R. (1982). Maximum and minimum of one-dimensional diffusions. *Stochastic Processes and their applications*, 13:1–9.

Doukhan, P. (1994). *Mixing : properties and examples*. Springer-Verlag.

Florens-Zmirou, D. (1989). Estimation de la variance d'une diffusion à partir d'une observation discrétisée. *C.R.A.S., t.309, Série I*, pages 195–200.

Francisco-Fernández, M. and Vilar-Fernández, J. (2005). Bandwidth selection for the local polynomial estimator under dependence: a simulation study. *Computational Statistics*, 20:539–558.

Furrer, E. and Katz, R. (2008). Improving the simulation of extreme precipitation events by stochastic weather generators. *Water Resources Research*, 44.

Gobet, E., Hoffmann, M., and Reib, M. (2004). Nonparametric estimation of scalar diffusions based on low frequency data. *Ann. Stat.*, 32:2223–2253.

Godfarb, D. and Idnani, A. (1983). A numerically stable dual method for solving strictly convex quadratic programs. *Mathematical Programming*, 27:1–33.

Golfarb, D. and Idnani, A. (1982). *Dual and Primal-Dual Methods for Solving Strictly Convex Quadratic Programs*. J. P. Hennart (ed.), Numerical Analysis, Springer-Verlag, Berlin.

Hall, P., Lahiri, S., and Polzehl, J. (1995). On bandwidth choice in nonparametric regression with both short and long-range dependent errors. *Ann. Stat.*, 23:1921–1936.

Hansen, L., Scheinkman, J., and Touzi, N. (1998). Indentification of scalar diffusions using eigenvectors. *Journal of econometrics*, 86:1–32.

Hoang, T. (2010). *Séries chronologiques non stationnaires non linéaires. Le cas des séries de températures en Europe.* PhD thesis, Université Paris Sud 11.

Katz, R. (2011). Overview of extreme value analysis under climate change. *US CLIVAR/NCAR ASP Researcher Colloquium on statistical assessment of extreme weather phenomena under climate change, NCAR, Boulder, CO.*

Kessler, M. and Sorensen, M. (1999). Estimating equations based on eigenfunctions for a discretely observed diffusion process. *Bernoulli*, 5(299-314).

Marron, J. (1987). Partiontioned cross-validation. *Econometric Rev.*, 6:271–284.

Menne, M., Durre, I., Vose, R., Gleason, B., and Houston, T. (2012). An overview of the global historical climatology network- daily database. *Journal of Atmospheric and Oceanic Technology*, 29:897–910.

Mraoua, M. and Bari, D. (2007). Temperature stochastic modeling and weather derivatives pricing: empirical study with moroccan data. *Afrika Statistika*, 2(1):22–43.

Nelder, J. and Mead, R. (1965). A simplex algorithm for function minimization. *Computer Journal*, 7:308–313.

Parey, S., Dacunha-Castelle, D., and Hoang, T. (2009). Mean and variance evolutions of the hot and cold temperatures in europe. *Climate Dynamics*.

Parey, S., Hoang, T., and Dacunha-Castelle, D. (2013). The role of variance in the evolution of observed temperature extremes in eurasia and in the united states. accepted by JPR.

Resnick, S. (1987). *Extreme values, regular variation and point processes*. Springer- Verlag.

Resnick, S. (2007). *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer- Verlag.

Richardson, C. (1981). Stochastic simulation of daily precipitation, temperature, and solar radiation. *Water Resources Research*, 17:182Ű190.

Ruppert, D. and Wand, M. (1994). Multivariate locally weighted least squares regression. *Ann. Stat.*, 22:1346–1370.

Sura, P. (2012). *Stochastic models of climate extremes: Theory and observations*. Springer-Verlag.

Wang, H. (2008). Nonlinear ARMA models with functional MA coefficients. *Journal of Time series analysis*, 29:1032–1056.

Wilks, D. and Wilby, R. (1999). The weather generation game: a review of stochastic weather models. *Progress in Physical Geography*, 23:329–357.