Testing the shape of distributions of weather data

Ana L P Baccon\textsuperscript{a} and José T Lunardi\textsuperscript{a,b}

\textsuperscript{a} Department of Mathematics and Statistics, State University of Ponta Grossa, Brazil
\textsuperscript{b} School of Physics and Astronomy, University of Glasgow, UK

E-mail: jttlunardi@uepg.br (JTL)

Abstract. The characterization of the statistical distributions of observed weather data is of crucial importance both for the construction and for the validation of weather models, such as weather generators (WG’s). An important class of WG’s (e.g., the Richardson-type generators) reduce the time series of each variable to a time series of its residual elements, and the residuals are often assumed to be normally distributed. In this work we propose an approach to investigate if the shape assumed for the distribution of residuals is consistent or not with the observed data of a given site. Specifically, this procedure tests if the same distribution shape for the residuals noise is maintained along the time. The proposed approach is an adaptation to climate time series of a procedure first introduced to test the shapes of distributions of growth rates of business firms aggregated in large panels of short time series. We illustrate the procedure by applying it to the residuals time series of maximum temperature in a given location, and investigate the empirical consistency of two assumptions, namely i) the most common assumption that the distribution of the residuals is Gaussian and ii) that the residuals noise has a time invariant shape which coincides with the empirical distribution of all the residuals noise of the whole time series pooled together.

1. Introduction

The statistical characterization of the distributions of the relevant stochastic variables is of crucial importance in the construction and empirical validation of stochastic models, such as climate models. In particular, weather generators (WG’s) play an important role in agrometeorological studies, such as in crop models, as well as in studies of hydrological changes, risk analysis, and as part of climate changes models [1, 2, 3, 4, 5, 6]. The main goal of a WG is to produce synthetic time series for weather variables that statistically resemble the actual ones from which the generator is calibrated. Several WG’s were proposed in the literature, and several studies were performed to investigate their adequacy to describe weather variables in several specific sites (see, for instance, [3, 6, 7, 8, 9]).

An important class of WG’s are the Richardson-type generators [1], which describes the time series of a weather variable as an autoregressive process for its residuals [10], which are obtained after the removal of the annual course of the mean and standard deviation of the observed values. Autocorrelations and cross-correlations among residuals of different variables are taken into account, and the stochastic terms associated with each variable is often assumed to be independent gaussian shocks (white noise). Besides, each variable is modeled conditionally to the precipitation status of each day (dry or wet). The parameters of the model, such as the distribution parameters for precipitation and for the lengths of dry and wet periods, as well as the cross and autocorrelations of the autoregressive process for the several variables must be...
estimated from the aggregated historical data. It is generally assumed in WG’s that all years are statistically similar, and then several years of data (typically of the order of 20-30 years) for a given site are pooled together to the aim of parameter estimations (see [3] for an outline of Richardson-type generators and a description of the parameter estimation procedures).

An important issue in the process of validation of a WG is to investigate if the observed residuals are in fact distributed according to the assumptions of the underlying model. For instance, if the residuals are not normally distributed then a WG which rely on the assumption of normality may not be able to accurately describe extreme events in the corresponding variable. Statistical tests for normality consider in general data aggregated from several years [11]. However, aggregation of data within large periods may hide temporal changes in the residuals distribution, and the distribution of the pooled annual data (or data corresponding to a given month but with several years pooled together) may not represent the actual distribution of the residuals if its parameters, or the shape of this distribution, change with time.

In this work we are interested in testing the empirical consistency of the assumption that the shape of the residuals distribution of a weather variable is constant in time. To this aim we adapt a procedure first introduced in reference [12], which was originally designed to test the assumption that (generally short) time series corresponding to the growth rates of business firms aggregated in a large panel of firms can be considered as being withdrawn from distributions which share the same shape, even if some of their parameters (namely the mean and variance) may vary among the firms. In Section 2 we present the basic ideas underlying the test procedure and its adaptation to climate time series. In Section 3 we illustrate the application of the procedure to a simple model for just one weather variable, namely the maximum temperature in a given site, conditioned on the precipitation status (wet or dry) of each day. In the last Section we discuss our results and present our conclusions.

2. The test procedure

In reference [12] a procedure was proposed to test the assumption that the (generally short) time series of growth rates of business firms aggregated in a large balanced panel of firms all share the same shape for the growth rate distribution. Here we present a brief outline of the basic ideas of such a procedure in the original context, and after we show how to adapt such a procedure to a climatet time series, which we are interest in in this work. For deeper details about the test and its power we refer the reader to [12].

2.1. Outline of the original test

Consider a large balanced panel (a matrix) formed by $N$ rows and $T$ columns. The $i$-th row contains $T$ time records (annual growth rates) for the firm $i$ ($i = 1, ..., N$). The null hypothesis is that all the firms in the panel share the same shape for their growth rate distributions, without necessarily sharing the same parameters. Thus, for example, a typical null could assert that all the time series in the panel were withdrawn from normal distributions, which may have different means and variances. The basic assumptions behind the null hypothesis are the following (below $R_i$ denotes the random variable (r.v.) associated with the growth rates of the firm $i$).

A1. the time records $\left\{r_{ij}\right\}_{j=1}^{T}$ of firm $i$ are $T$ independent outcomes of the same r.v. $R_i$.
A2. different firms grow independently of each other.
A3. all the time series are stationary.

1 More precisely, in the general case we must refer to the distribution of the residuals noise [see Eq. (3)]. When the residuals noise distribution is stable, as it is the case of Gaussian noise, the residuals and the residuals noise will belong to the same family of distributions and, therefore, will share the same shape.

2 The notation used here is slightly different from that used in [12].
A4. each variable $R_i$ may have an idiosyncratic mean $\mu_i$ and standard deviation $\sigma_i$, and it is assumed that these are the only possible differences between the $R_i$’s. Therefore, according to this assumption the standardized r.v.

$$X_i = \frac{R_i - \mu_i}{\sigma_i}$$  \hspace{1cm} (1)

will be the same for all $i$. Let us denote this variable by $X$ and its distribution function by $F_X$. It is precisely this distribution which characterize the null about the common shape assumed to be shared by all the distributions of the firms (rows) in the panel.

Now, if one knows the actual values of all the idiosyncratic parameters $\mu_i$ and $\sigma_i$, then by simply pooling together all the standardized outcomes $\{x_{ij}\}_{i=1}^{N} \times_{j=1}^{T}$, the resulting aggregated distribution would resemble the distribution of $X$. In order to test the null $F_X$ one could simply apply a goodness-of-fit test to compare this distribution with that of the aggregated observed data standardized by rows. However, in general we do not know the actual idiosyncratic parameters $\mu_i$ and $\sigma_i$, and at the best these should be estimated from each row in the panel; when $T$ is large enough such estimates will be accurate, and the lack of knowledge about the exact idiosyncratic parameters would not be an obstacle to do the standardization and perform the suggested test. However, as it was shown in [12], when $T$ is not large (roughly $T \lesssim 50$) such estimates may be very rough, and even if the null is correct the aggregate distribution of the outcomes standardized with respect to these roughly estimated parameters may deviate significantly from $F_X$. In [12] it was also shown that

$$Z_i = \frac{R_i - \hat{\mu}_i}{\hat{\sigma}_i} = \frac{X_i - \bar{X}_i}{s_{X_i}} = \frac{X - \bar{X}}{s_{X}}$$  \hspace{1cm} (2)

where $\hat{\mu}_i$ and $\hat{\sigma}_i$ are the (generally innacurate) time series estimates for $\mu_i$ and $\sigma_i$, and $\bar{X}_i$, $s_{X_i}$ are the (unknown) time series estimates for the mean and standard deviation of the standard r.v. (1), and similarly for $\bar{X}$ and $s_{X}$. From now on we will reserve the word “standardization” for standardization with respect to the actual parameters $\mu_i$ and $\sigma_i$, and will use the expression “z-transformation” for the standardization with respect to the corresponding time series estimates $\hat{\mu}_i$ and $s_{X_i}$.

Equation (2) is central to designing a test procedure able to deal with the issue of short time series in the panel. It states that under the null hypothesis the variables $Z_i$ obtained by z-transforming the time series variables $R_i$ must be the same for all firms and is distributed identically to the variable $\bar{Z} \equiv \frac{X - \bar{X}}{s_{X}}$, where $X$ is the r.v. corresponding to the null $F_X$. Therefore, what we should to do to test the null is simply to compare (by using a goodness-of-fit test) the distribution of all the z-transformed growth rates pooled together with the modified null distribution $F_{\bar{Z}}$ (for details of how to obtain this distribution by a Monte Carlo simulation see [12]).

2.2. Adaptation to climate time series

For the sake of illustration of application of the above procedure to climate time series we consider a simple model consisting of a single variable $V$, whose daily time series is denoted by $\{V^j\}_{j=1}^{M}$, and $M$ is assumed to be large (in the next Section we consider 21 years of data, which correspond to more than 7 thousands of daily records). To obtain a residual series the daily observations are normalized with respect to the mean and standard deviation of the corresponding day, which are modeled by a truncated Fourier series with an annual period. Two Fourier models are
considered, depending on the wet or dry status of the day. After such a normalization is done, the resulting residuals $Y^j$ are assumed to follow an AR(1) process \[3, 10\]

$$Y^j = \rho Y^{j-1} + (1 - \rho^2) \epsilon^j, \quad j = 1, \ldots, M,$$  \hfill (3)

where the index $j$ labels the day in the whole time series, $\rho$ is the lag-1 autocorrelation, assumed to be constant within each calendar month, but which may differ among different calendar months (but not changing along the years), and $\epsilon^j$ are independent random shocks, whose distributions are often assumed in the literature to be a standard normal (white noise).

Our aim in this work is to adapt the procedure of the previous subsection to test the null hypothesis that a particular shape (e.g., Gaussian) of the distribution of the residuals noise $\epsilon^j$ is time invariant. The test will apply even if the mean $\mu^j$ and standard deviation $\sigma^j$ may change in the time; however, in order to apply the test outlined above we will make the assumption that within short time periods of length $T$ ($T \lesssim 30$) the variations of the noise’s mean and variance is negligible. Now, by splitting the whole time series in subperiods of length $T$ we can build a panel of $N$ rows and $T$ columns, where $N$ is the number of subperiods.\(^3\) Together with the null hypothesis we are interested in to test, it is straightforward to see that the set of assumptions made so far in our simple climate model (3) is compatible with the assumptions A1 – A4 of the previous subsection. Therefore, the test procedure outlined there can be straightforwardly applied to test the null hypothesis about the constancy of the shape of the residuals noise distribution in the model (3).

3. An illustrative example

To illustrate the application of the proposed procedure we considered a time series of daily data for the maximum temperature and precipitation for the locality of Chicago, USA (station “KMDW”), available in the software Wolfram Mathematica\(^@\). We considered data covering a period of 21 years, from January 1st, 1995 to December 31th, 2015, which resulted in a time series of length $M = 7,665$.

The time series of precipitations was used just to classify each day in the time series as wet or dry – we considered a day as wet if its accumulated precipitation was $\geq 0.1$ mm; otherwise, it was considered as dry. After, data on maximum temperature for each calendar day was split into two groups (according to the dry or wet status), and for each group it was computed the mean and standard deviation. The annual course of the means and the standard deviations were approximated separately by a third order Fourier fitting, one fitting for each group (dry or wet). After eliminating the annual course of the mean and standard deviation (by using the values for each calendar day as given by the third order Fourier fitting), the normalized residues $Y^j$ ($j = 1, \ldots, M$) were obtained.

To estimate the 1-lag autocorrelation $\rho$ for each calendar month we proceeded as follows. We used the well-known Yule-Walker (biased) estimator for the autocorrelation associated to each month in the time series, resulting in a set of 21 (since we have 21 years of observations) biased estimates for each calendar month \[13\]. Then, we taken the average of these 21 values and corrected for the bias, obtaining an estimate for the calendar month $m$ given by $\rho(m) = \frac{n^2 \bar{\rho} + 1}{n^2}$, where $n$ is the number of days of the calendar month $m$ and $\bar{\rho}$ is the mean of the set of 21 biased Yule-Walker estimates for that month.\(^4\) These estimates are shown in table 1.

Turning the estimated autocorrelations into (3) we computed the time series for the residual noise $\epsilon^j$. When data for maximum temperature or precipitation were lacking for a day (this occurred only for 263 days in 21 years of data) we filled that day in the noise time series by

\(^3\) $N = \lfloor M/T \rfloor$, where $\lfloor \rfloor$ denotes the integer part.

\(^4\) Due to the lacking of data in some days, this average was computed with approximately 16 values for every month.
Table 1. Estimates for the monthly autocorrelations in the model (3).

|    | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| ρ  | 0.75| 0.70| 0.69| 0.59| 0.63| 0.72| 0.59| 0.62| 0.68| 0.78| 0.63| 0.70|

Table 2. p-Values of the KS test for two nulls and several subperiods lengths.

| Length (days) | Gaussian | Empirical |
|---------------|----------|-----------|
| 7             | 0.024    | 0.838     |
| 14            | 0.008    | 0.937     |
| 21            | 0.030    | 0.871     |
| 28            | 0.002    | 0.594     |

randomly choosing a value from the remaining time series. Finally, we set the null hypothesis about the shape of the distributions of residual noises, split the whole time series in short subperiods of length $T$ and applied the adapted procedure of Subsection 2.2. As for the goodness-of-fit test, which was the last stage of our procedure, we used the Kolmogorov-Smirnov (KS) test available in the software Wolfram Mathematica®.

3.1. Results
We considered two specific nulls. The first one consider that the invariant shape of the residuals noise distribution is Gaussian, which is the most often assumed shape in the literature. The second null we considered was that the invariant shape coincides with that empirically observed when we pool together all the observed residuals noise for the whole time series (this particular assumption implies that the parameters of this distribution does not change in the course of time as well). En passant, the last version of the widely used generator LARS-WG uses semi-empirical distributions for the noises and obtain better results in the simulation of maximum and minimum temperatures when compared with its previous versions which used Gaussian noises.[4].

We tested the above two nulls by partitioning the time series in subperiods of lengths $T = 7, 14, 21$ and 28 days. The results are shown in table 2. Considering a 5% significance level we observe that the Gaussian null should be rejected for all the lengths considered. On the other hand, at the same significance level the “empirical” null should not be rejected in any of the cases. It is not shown in the table, but the empirical null is consistently accepted also for larger subperiods of lengths 90, 180, 365, 1825 (5 years) and 7665 (21 years, the length of the whole time series). Figures 1 and 2 illustrates the goodness of fit of the two considered nulls for subperiods of lengths 7 and 28 days.

4. Discussions and conclusion
We proposed an adaptation of the testing procedure introduced in [12] to investigate the null hypothesis that the shape of the distribution of the residuals noise of a climate time series is time invariant, even if its parameters are not. We illustrated the application of the adapted procedure to a time series of a single weather variable (the maximum temperature) in a given site, and showed that it was very effective in testing for two specific nulls, one stating that the
time invariant shape is Gaussian and another stating that the empirically observed distribution for the residual noises is time invariant (time invariant shape and parameters). From the results we concluded that the Gaussian assumptions for the residuals does not describe well the observed data in the context of this single variable model; instead, in this context the assumption that the residuals noise maintain a time invariant distribution performs better. Further investigations in the context of this simple model may consider analytical assumptions to describe the empirical distribution. As such empirical distributions often present a non vanishing skewness (see, for example, [4, 11]; it seems to be the case also in the situations illustrated in figures 1 and 2) we could guess that members of the class of asymmetric exponential power distributions [14], which generalizes the Gaussian by changing the tails behavior and by introducing skewness, could be good candidates for it. It is worth to emphasize that the fact that the empirical distribution of the noises was not reject by the test is by no means obvious, because the resulting empirical distributions obtained by aggregating all the time series noises could be just an effect of aggregating different variables with different distributions – our test provided evidence that this is not the case for the data considered here.

Finally, the proposed procedure can be straightforwardly extended to deal with multivariate stochastic models presenting auto and cross correlations among the variables, and thus it can be an additional and useful tool in the studies of validation of actual multivariate WG’s for diverse sites. Such an extension, along with a wide study of validation of some WG’s for diverse locations is presently under our investigation. The results will appear elsewhere.

Acknowledgments
JTL thanks CNPq/Brazil for partial financial support (Grant PDE 233512/2014-4).

References
[1] Richardson C W 1981 Water Resources Research 17 182
[2] Kilsby C, Jones P, Burton A, Ford A, Fowler H, Harpham C, James P, Smith A and Wilby R 2007 Environmental Modelling & Software 22 1705
[3] Fodor N, Dobi I, Mika J and Szeidl L 2010 Meteorology and Atmospheric Physics 107 91
[4] Semenov M A and Strattonovitch P 2010 Climate Research 41 1
[5] Izumi T, Takayabu I, Dairaku K, Kusaka H, Nishimori M, Sakurai G, Ishizaki N N, Adachi S A and Semenov M A 2012 Journal of Geophysical Research: Atmospheres 117 D11114
[6] Chen J and Brissette F P 2014 International Journal of Climatology 34 3089
[7] Semenov M A, Brooks R J, Barrow E M and Richardson C W 1998 Climate Research 10 95
[8] Semenov M 2008 Climate Research 35 203
[9] Fodor N, Dobi I, Mika J and Szeidl L 2013 *The Scientific World Journal* **2013** 571367
[10] Mearns L O, Katz R W and Schneider S H 1984 *Journal of Climate and Applied Meteorology* **23** 1601
[11] Harmel R D, Richardson C W, Hanson C L and Johnson G L 2002 *Journal of Applied Meteorology* **41** 744
[12] Lunardi J T, Miccichè S, Lillo F, Mantegna R N and Gallegati M 2014 *Journal of Economic Dynamics and Control* **39** 140
[13] Krone T, Albers C and Timmerman M 2015 *Quality & Quantity* 1
[14] Bottazzi G and Secchi A 2011 *Industrial and Corporate Change* **20** 991–1030