Statistical comparison of observed temperature and rainfall extremes with climate model outputs in the Mediterranean region

D. Tsaknias[1, †*], D. Bouziotas[1, 2], D. Koutsoyiannis[1]

[1] Department of Water Resources and Environmental Engineering, School of Civil Engineering, National Technical University of Athens, Greece
[2] Delft University of Technology, Netherlands
*Corresponding author: Dimosthenis Tsaknias (dtsaknias@gmail.com)

Abstract

During the recent decades, climate model outputs have been widely used to support decision making for social and financial policies, with special focus on extreme events. Moreover, a general perception has been developed that extreme events will be more frequent in the future. To evaluate whether climate models provide a credible basis for predictions of extremes, their ability to reproduce annual extreme values of daily temperature and precipitation is studied, as well as a set of climate indices which are used to investigate the occurrence of droughts, heat waves and floods. Comparisons of climate model outputs with observed data are made in terms of probability distributions of extreme events. The case study focuses on the Mediterranean area, which is regarded to be one of the most vulnerable areas to climate change. The statistical comparison indicates that the observed extremes are not simulated sufficiently by climate models. Therefore, serious concerns are raised about the usefulness of climate models in hydrological applications.

[This research study has been submitted in several journals and was rejected; review comments and authors' replies to them are included as an Appendix]

Keywords: climate models; climate change; extreme events; Mediterranean; climatic indices

---

[†] Currently working at Risk Management Solutions (RMS), London

1. Introduction

In the recent decades, future climate projections have attracted the interest of the entire scientific community. Assumptions about shifts of hydrometeorological patterns, especially with regards to their impact on water resources, have been increasingly popular. As a result, there is an ongoing demand for 21st century predictions of climatic evolution on gradually finer scales. The outcome of modelling attempts to long-term climate forecasting are the General Circulation Models (GCMs), which are considered as sophisticated tools able to simulate past, present and future climatic conditions. In particular, it has become a common perception that their outputs constitute credible guides for the future climatic conditions and, therefore, they can be used by the hydrological and water management community to access climate impacts. The demand for such studies is not limited to general mitigation policies, but extends to more detailed level, i.e. local and site-specific water management decisions (Kundzewicz & Stakhiv, 2010), related to agriculture and irrigation policies, the design of hydraulic structures, land use and any other environmental aspect.

However, the credibility of GCMs in reproducing the observed climatic conditions has been questioned (e.g. Anagnostopoulos *et al.*, 2010, Koutsoyiannis *et al.*, 2008, Sakaguchi *et al.*, 2012). Consequently, it is debatable whether one should adapt the outcome of these models to hydrological studies, which demand a great level of accuracy at finer spatio-temporal scales. What is more, the way in which uncertainty is addressed is questionable, since climatic processes are very complex and undoubtedly they cannot be described only by deterministic approaches. Recent publications (Anagnostopoulos *et al.*, 2010, Kiem & Verdon-Kidd, 2011, Koutsoyiannis *et al.*, 2008) have triggered scepticism on their ability to simulate adequately the climate, advocating that a different approach based on a stochastic framework would not disregard system dynamics. Earlier, Koutsoyiannis *et al.* (2007) had proposed a framework for a stochastic approach to characterize future climate uncertainty and have shown that deterministic approaches based on climate model outputs underestimate uncertainty.

The uncertainty about future climate is more intense on finer scales and especially when it comes to extreme events (Blöschl *et al.*, 2007). According to commonly invoked projections, during the 21st century extreme events will be more frequent, more intense and with greater duration, rendering these natural hazards as a threat not only for hydraulic structures, but also for human lives at a global scale (IPCC, 2007). Areas that are expected to face such events will have to sustain dramatic changes in their economies, policies, environment and obviously, societies, as they will suffer from floods, droughts, or even desertification, which will provoke rise on food prices, pandemic diseases and waves of immigrants (European Environment Agency, 2010). Damage caused by extreme weather phenomena during 2014 triggered overall losses of 280 billion dollars, while the claims for insured properties were 70 billion dollars (Munich Re, 2011).

While their efficiency has been questioned, GCMs have been used as the main decision tool in order to assess risk and quantify projections in the aforementioned cases. It should be noted, though, that globally the IPCC projections for more frequent and more intense extreme events have not been proven as yet. For example, a recent comprehensive study of precipitation by Sun *et al*. (2012), which analyses observations of monthly precipitation (1940–2009) over the global land surface "unexpectedly" found a reduction in the variance of precipitation, so that "the dry became wetter while wet became drier".

The purpose of this study is to investigate whether it is justified to use GCMs as risk assessing tools. To this aim we make a statistical comparison of climate model outputs with observed maximum temperature and maximum precipitation in the Mediterranean area at a daily time scale. There are good reasons for the choice of the Mediterranean area as a study case as summarized below.

The climatic variability in the Mediterranean area has attracted the interest of the scientific community since the 18^{th} century. According to Luterbacher *et al.*, (2006), Mann (1790) based on written sources, concluded that temperatures were rising and the climate was getting drier due to deforestation. On the other hand, he was criticized that his sources were not reliable (Ideler, 1832). It seems, thus, that there is a long tradition on studies of climate variability in the Mediterranean area.

The Mediterranean area is classified as 'Csa' and 'Csb' in Köppen classification system, being a transitional climate zone between Central Europe and North Africa (Köppen, 1918). Therefore, even small changes in climatic patterns of the neighbouring areas can have a strong impact on the Mediterranean area, which is consequently considered as one of the most vulnerable areas to climate change (Giorgi and Lionello, 2008). For all these reasons, including the fact that great changes were observed also in the past, the Mediterranean has been characterized as a 'hot-spot' of climate change (Giorgi, 2006). It is projected that extreme temperatures may rise by 6 ^{o}C in some areas and precipitation will follow a dominant negative trend, reduced by up to 30% in the winter (Alpert *et al*., 2008; García-Ruiz *et al*., 2011) causing important shifts on the annual water balance. Shohami *et al.* (2011) analysed trends for the years 1964-2003 in daily precipitation and temperature and supported that there is an increase in warming conditions especially during the summer and less precipitation-generation processes will occur during the winter months in the Eastern Mediterranean. Drier conditions are projected by both A2 and B2 scenarios of IPCC (see below), and Beniston *et al.* (2007) claimed that precipitation intensity will be reduced with droughts occuring earlier in the year and with a longer duration than in the current climatic conditions.

2. Data used

Three GCMs are used and they are run with the 20^{th} Century Climate in Coupled Models (20C3M) and Special Report on Emissions Scenarios (SRES) A2 scenario of the IPCC Fourth Assessment Report (AR4; IPCC 2007) on the 20^{th} century and for the first decade of the 21st century (IPCC, 2007), respectively (Table 1). The availability of daily maximum temperature and precipitation data on the internet, as well as the coverage of past periods with adequate length of time series have been the criteria for the selection of GCMs. The GCM outputs were downloaded from https://esg.llnl.gov:8443/index.jsp and http://www.cccma.ec.gc.ca/data/cgcm3/cgcm3.shtml. The fact that climate models simulate past and present climate makes our comparison feasible. The daily annual maxima of temperature and precipitation are investigated, as well as some climate indices for extreme events, as they have been defined in previous studies (Alexander *et al*., 2006; Frich *et al*., 2002).

Table 1: GCMs used for the statistical comparison.

Name	Institution	Resolution (°)	Grid points (Latitude × Longitude)	Time span	Scenario	
					20C3M	SRES A2
ECHAM5	Max-Planck Institute for Meteorology & Deutsches Klimarechenzentrum	1.88 ×1.88	96×192	1860-2000	✓	×
CSIRO	Australia's Commonwealth Scientific and Industrial Research Organisation	1.88 ×1.88	96×192	1871-2010	✓	✓
CGCM3-T63	Canadian Centre for Climate Modelling and Analysis	2.81 × 2.81	64×128	1850-2010	✓	✓

Table 2: Stations used in the statistical comparison.

ID	Station	Country	Coordinates	# of years with temperature data	# of years with precipitation data
1	Bologna	Italy	44.50N, 11.35E, 53.0 m	194	194
2	Hvar	Croatia	43.17N, 16.45E, 20.0 m	116	152
3	Perpignan	France	42.74N, 2.87E, 42.0 m	109	99
4	Tortosa	Spain	40.82N, 0.49E, 44.0 m	70	89
5	Istanbul	Turkey	40.97N, 29.08E, 33.0 m	77	75
6	Torrevieja	Spain	37.98N, -0.71E, 1.0 m	73	73
7	Albacete - Los Llanos	Spain	38.95N, -1.86E, 704.0 m	107	N/A
8	San Sebastian	Spain	43.31N, -2.04E, 251.0 m	81	N/A
9	Dar-El-Beida	Algeria	36.72N, 3.25E, 24.0 m	72	71
10	Valencia	Spain	39.48N, -0.37E, 11.0 m	65	N/A
11	Madrid	Spain	40.41N, -3.68E, 667.0 m	90	N/A
12	Sarajevo	Bosnia	43.87N, 18.43E, 630.0 m	105	105
13	Athens	Greece	37.58N, 23.43E, 104 m	N/A	110
14	Larnaca	Cyprus	34.88N, 33.63E, 1.0 m	N/A	94
15	Palermo	Italy	38.11N, 13.35E, 37.0 m	N/A	157
16	Tel Aviv	Israel	32.10N, 34.78E, 4.0 m	N/A	67
17	Malaga	Spain	36.67N, -4.49E, 7.0 m	N/A	70
18	Belgrade	Serbia	44.80N, 20.47E, 132.0 m	N/A	75
19	Nimes	France	43.86N, 4.41E, 59.0 m	N/A	83
20	Pesaro	Italy	43.91N, 12.90E, 11.0 m	N/A	138

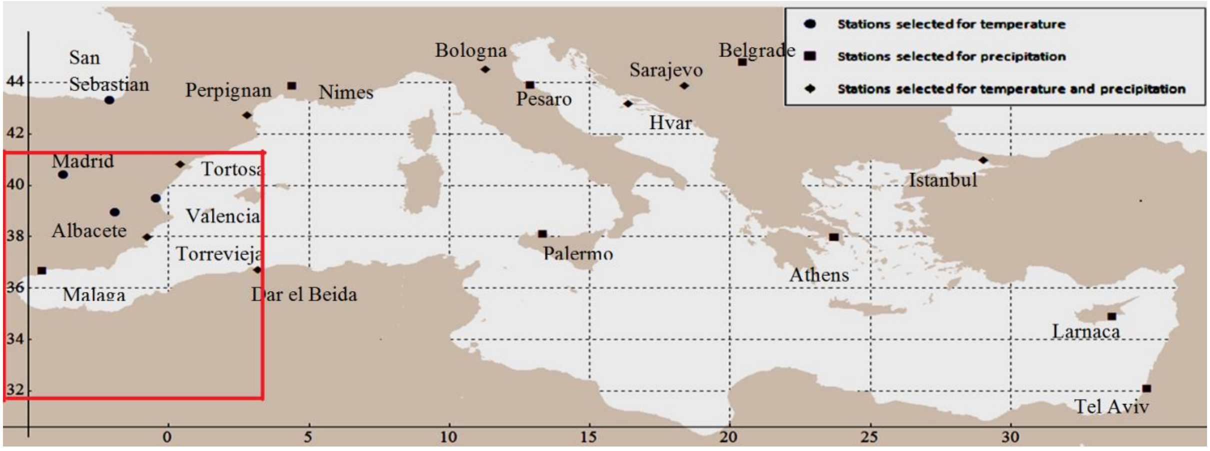

Figure 1: Map of the Mediterranean along with the stations used.

The statistical comparison is made at several stations (according to the methodology described in the next section) and 16 and 12 stations are used for the precipitation and the maximum temperature, respectively (Table 2, Figure 1). These stations were selected from those available online based on a first criterion of a minimum time length of observations, i.e. 65 years with as few missing values as possible. Although, even geographical distribution is desirable, this criterion was not completely fulfilled, as Spain and France have more meteorological stations. The databases used are those of the Royal Netherlands Meteorological Institute (http://climexp.knmi.nl/start.cgi?someone@somewhere) and the National Observatory of Athens.

3. Methodology

In order to compare the observed time series with the climate model outputs, the four grid points closest to the station of interest were statistically interpolated (Figure 2), through a non-linear conversion of the Best Linear Unbiased Estimation (BLUE) method. The use of the four neighbouring grid cells is favourable for the GCMs, since it has been found that the highest correlations between observed and modelled variables do not always occur within the same cells (Wilby & Wigley, 2000). The BLUE method (Kitanidis, 1993) has been successfully used in previous analyses (Anagnostopoulos *et al*., 2010; Koutsoyiannis *et al*., 2008); however, it is suitable only for large time scales (monthly and beyond), as its performance degrades for daily gridded time series. In addition, it smooths out the extreme values. Thus, a nonlinear conversion of the BLUE method is used.

The nonlinear conversion of BLUE is as follows. Let $\underline{z}_i$ be a non-negative stochastic process at a point denoted by $i = 1, \ldots, k$. Let $\underline{z}'_i$ a non-linear transformation of z_i defined by (Papalexiou and Koutsoyiannis, 2011):

$$\underline{z}'_i = p\left\{\ln\left[1+\left(\frac{\underline{z}_i}{p}\right)^q\right]\right\}^{1/q} \quad (1)$$

where $\underline{z}_i$ represent the GCM outputs for the four nearest grid points, and p is a scale parameter with units identical to $\underline{z}_i$, and q is a non-negative exponent.

We interpolate at a point with the transformed process being:

$$\underline{x}' = \sum_{j=1}^{4} \lambda_j \cdot \underline{z}'_j \tag{2}$$

where λ_j are weighting coefficients, non-negative for physical consistency and $\sum_{j=1}^{4} \lambda_j = 1$ for unbiasedness. Notice that we use the so-called Dutch notational convention, i.e. we underline the symbol of a random variable whereas the same symbol not underlined denotes a value (realization) of the variable. Based on (1), the back-transformation of the interpolated time series x' will be:

$$\underline{x} = p\left\{\exp\left[\left(\frac{\underline{x}'}{p}\right)^q\right] - 1\right\}^{1/q} \tag{3}$$

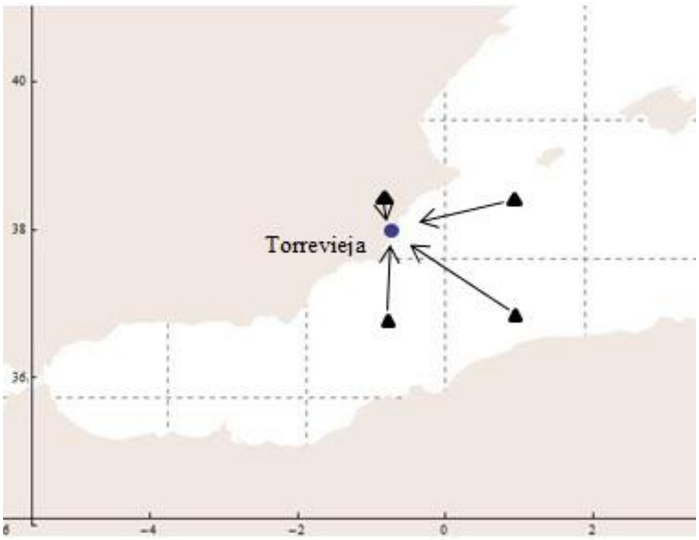

Figure 2: Explanation sketch of the interpolation method: Torrevieja station and the four neighbouring grid cells.

Generally, interpolation tends to reduce the variance and higher moments of order r. To see this, we can assume that z_i are independent to each other, that they have identical distributions (so that $\mathrm{E}[(z'_j)^r] = \mathrm{E}[(z')^r]$, i.e. same for all j) and that their expected values are zero ($\mathrm{E}[z'_j] = 0$; if not, we can subtract the mean value $\mathrm{E}[z'_j]$ from each variable before raising to power r), thus

$$\mathrm{E}[(\underline{x}')^r] = \left(\sum_{j=1}^{4} \lambda_j^r\right) \cdot \mathrm{E}[(\underline{z}_j')^r] \tag{4}$$

Unless $\lambda_i = 1$, for some i and $\lambda_j = 0$, for every $i \neq j$ we get $\mathrm{E}[(\underline{x}')^r] < \mathrm{E}[(\underline{z}')^r]$ because $\sum_{j=1}^{4} \lambda_j^r < 1$ for $r > 1$. It is also expected that $\mathrm{E}[\underline{x}^r] < \mathrm{E}[\underline{z}^r]$. Nevertheless, by appropriate choice of p and q we may considerably increase $\mathrm{E}[\underline{x}^r]$. Here we choose:

$$\frac{\left\{\mathrm{E}[\underline{x}^r]\right\}^{1/r}}{\mathrm{E}[\underline{x}]} = \sum_{j=1}^{4} \lambda_j \frac{\left\{\mathrm{E}\left[\underline{z}_j^{\,r}\right]\right\}^{1/r}}{\mathrm{E}[\underline{z}_j]} \tag{5}$$

A natural choice of the order of moment would be $r = 2$, in order to avoid bias in the variance; however, since we deal with maxima, a higher order of moment would be more favourable for the climate model outputs; thus we choose $r = 4$.

The optimized coefficients p, q, λ_j are derived after minimizing the error function:

$$\mathrm{Er} = \mathrm{MSE} + w \cdot e \tag{6}$$

where MSE is the mean square error between the interpolated and the historical time series and e is the error in (5) (the square of the difference between the two sides), which is penalized by a large number w.

In rainfall time series, to make the optimal interpolated values comparable to the observed ones, we must have in mind the substantial spatial variability of a rainfall field. We assume that the interpolated values correspond to an area as defined by the size of the grid, while clearly the observations refer to a point. To take this into account, in design rainfall studies a point maximum is multiplied by the so-called areal reduction factor (φ). Here we followed a similar approach, dividing the gridded model output series by φ, in order to transform them to point rainfall of the same duration and return period. According to Koutsoyiannis and Xanthopoulos (1999), φ can be approximated by the following equation:

$$\varphi = 1 - \frac{0.048 \cdot A^{0.36 - 0.01 \cdot \ln A}}{d^{0.35}} \geq 0.25 \tag{7}$$

where d is the time scale in h and A is the area of each cell in km^2. This was empirically derived from tabulated data provided by the National Environmental Research Council (1975). As the grid cells are on the order of 200 km x 200 km, we estimate $\varphi \approx 0.76$.

4. Justification of the methodology

The observed, the modelled, in four GCM grid points, and the extremes derived from the optimized time series at one of the stations (Torrevieja) for the CSIRO model are depicted in Figure 3. While the methodology is the same in both variables, decent fitting can be achieved only in temperature, where the results of non-linear transformation are satisfactory. On the contrary, the extreme daily rainfall is systematically underestimated by the models and the results of the optimization still diverge, despite the fact that they reproduce the reality better than the four gridded time series of the station.

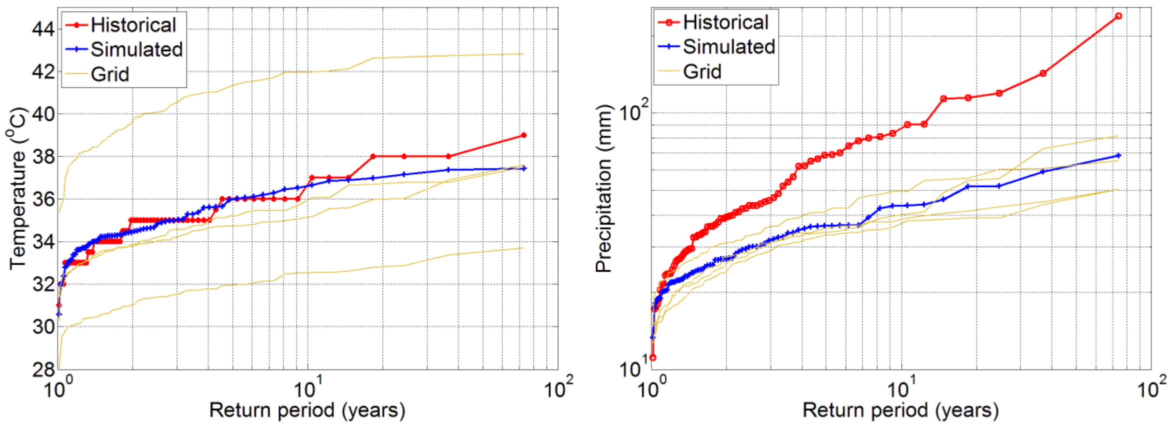

Figure 3: (a) Annual maximum daily high temperature and (b) precipitation at Torrevieja station for the CSIRO model.

Koutsoyiannis *et al*. (2008) claim that the BLUE method is appropriate not only for temperature and precipitation fields at the climatic scale, but even for rough random fields. They tried to estimate the efficiency of the BLUE method by assuming that they know the values of three out of four neighboring grid cells and the fourth was considered as unknown. Using the BLUE technique they tried to estimate the time series of the unknown grid cell with three weights. Their results were particularly good, as the Nash-Sutcliffe efficiency coefficient was 0.91 and 0.99 for precipitation and temperature, respectively. It is clear then, that with the use of 4 weights the results should be improved.

Moreover, Anagnostopoulos *et al.* (2010) reconstructed time series of precipitation and temperature from point measurements for the whole area of USA and compared them to GCM outputs. They found out that large scale results are worse compared to the point comparison after using the BLUE method which allows the modelled time series to fit the observed ones. Hence, the BLUE algorithm is more favourable to GCMs and allows them to avoid local influences. In addition, the non-linear conversion of BLUE method introduced here, explicitly avoids bias at the first and fourth moments implying small biases for the second and third moments.

With regard to the specific study area, it should be noted that extreme storm and flood events in the Mediterranean area usually occur during the winter period, driven mostly by frontal events (large scale precipitation). On the contrary, summer floods, which are caused by local convective precipitation events, are rare in the study area. Furthermore, as observed in Table 2, most of the stations are located at low altitudes, and the terrain of neighbouring areas is relatively flat, thus the effects of orographic precipitation do not influence the measurements. We believe, therefore, that the use of a constant areal reduction factor is justified and the spatial variability on the daily time scale does not have a significant impact on our analysis.

One could argue that we compare unlike data, since the observed time series refer to a point, whereas the model outputs are gridded with the resolution presented in Table 1.

Likewise, precedent studies (Anagnostopoulos *et al.*, 2010, Koutsoyiannis *et al.*, 2008) have been criticized on the basis that a comparison in different scales is meaningless without downscaling. However, the goal of our study is not to project future climate patterns in the Mediterranean. Our scope is to identify elements of reality in the climate model outputs. We think that before downscaling, one should verify whether model outputs reflect observed patterns. If it is proved that model outputs have sufficient performance, then one can proceed with downscaling. Otherwise, applying downscaling techniques is just a mathematical exercise that will make a curve fit better to a given problematic dataset (cf. Kundzewicz, 2010, from Koutsoyiannis *et al*., 2011).

With reference to downscaling, while it has been a very common technique, we believe that its use has never been justified. The rest of this paper provides further evidence against its use, but doubts about it have been already been cast by several studies. The lack of falsifiability studies may lead to overestimation of model predictive skill based on the fact that some of their properties have been tuned in order to match the known state of Earth's climate (Mauritsen *et al.*, 2012). Similarly, Dawson *et al.* (2012) found that low resolution GCMs are not capable of simulating regimes observed by reanalysis data, while high resolution RCMs simulated regimes described by non-Gaussian probability distributions accurately. They deduce that the credibility of climate projections is questionable, since the boundary conditions provided to RCMs through downscaling may err systematically. In the same spirit, Xu & Yang (2012) showed that the traditional dynamic downscaling is not adequate to simulate reality, as precipitation is overestimated, while extreme temperatures are overestimated by 2-6 °C and in order to achieve good results one should tune the model using an improved dynamical downscaling technique.

A skilful model should not produce results that are significantly different from the observed ones. However, Fyfe *et al.* (2011) claim that if the model does not have a satisfactory performance in reproducing long-term trends, a linear trend correction along with a systematic long-term bias correction should be applied. We believe that in the latter case the model results are just tuned to fit observations and the authors rather prove that multi-decadal properties are not captured by climate model outputs. Besides this, it has been demonstrated that the assumption of invariance in climate change projections can lead to overestimation of warming if temperatures during the warmest months exceed by more than 4-6 °C current climatic conditions (Christensen *et al.*, 2008). Bias may be increased in the case of precipitation during the wettest months as well.

Finally, Pielke Sr & Wilby, (2012) discuss the cases when regional dynamic downscaling is not skilful. Among other things, they state that dynamic downscaling dependent on the lateral boundary conditions from a multi-decadal Earth system model has practical value only for sensitivity experiments and not for climate predictions. If GCMs do not accurately predict large-scale circulation phenomena, such as El Niño and La Niña, then they will not be able to provide accurate lateral boundary conditions. They deduce that CMIP5 model results are not skilful enough to assess impacts on communities.

5. Extreme Value Distribution and Climate Indices

For completeness, to each of the observed annual maximum series we also fit the Generalized Extreme Value (GEV) distribution. Its probability distribution function is:

$$F(x) = \exp\left(-\left(1+\frac{\kappa\cdot x}{\lambda}-\kappa\cdot\psi\right)^{-1/\kappa}\right) \qquad \kappa\cdot x \geq \kappa\cdot\lambda\cdot\psi-\lambda \tag{8}$$

where κ, ψ, λ are the shape, location and scale parameters.

Contrary to Kharin *et al.* (2007), who had short samples and used the L-moments, our samples are considerably larger and allow us to use the maximum likelihood which can provide better estimates for the three parameters (and especially the shape parameter).

In rainfall series, one of the most important statistical characteristics which should be preserved by a model is intermittency. For numerical reasons, climate models outputs are a set of non-zero daily precipitation values. Many of these values are close to zero, so in order to retain intermittency we characterize a day as wet when its precipitation depth is larger than 1 mm.

Previous studies of extreme events have suggested a set of climate indices which apply on daily time series (Zhang *et al*., 2011). Such indices have been used both on observed data (Alexander *et al*., 2006; Frich *et al*., 2002) and on climate model outputs (Sillmann and Roeckner, 2008). Regarding rainfall the following indices are examined:

1. Average days per year with daily precipitation larger than the threshold of 10, 20 and 30 mm (R10mm, R20mm and R30mm, respectively).
2. Maximum number of consecutive dry days (CDD).
3. Average daily precipitation of the wet days (simple precipitation intensity index, SDII).
4. 90th percentile of wet days (Prec90p).

For temperature we use the following indices:

1. 90th percentile of maximum temperature.
2. Average number of warm spell periods per year (warm spell duration index, WSDI). A warm spell is defined as the occurrence of at least six consecutive days with maximum temperature larger than the one corresponding to the 90th percentile.

6. Results

5.1. Rainfall

For each study station we evaluated the extreme events of precipitation and maximum temperature, along with the climate indices, as they were defined in the previous section. We present the results in two characteristic stations for precipitation (Perpignan and Torrevieja) and in another two for maximum temperature (Bologna and Tortosa). In the case of precipitation, the empirical distributions of daily annual maxima and their corresponding time series are plotted, along with the annual time series and the 10-year moving average, the probability density functions of precipitation aggregated on the monthly time scale, bar charts with the climate indices (R10mm, R20mm, R30mm) as well as the five greatest CDD.

In the case of rainfall (Figure 4), the discrepancy of the quantitative results between observed and modelled daily annual maximum precipitation is apparent, as the GCMs consistently underestimate the extreme events, especially for large return periods. One can observe that at Perpignan the historical maximum is about 220 mm, while the GCMs predict record maxima of about 70-95 mm. Similar observations can be made at Torrevieja, as the record maximum is 240 mm, whereas according to GCMs this is 70-80 mm. Besides the presented two stations, there was not any location where the results of maxima could be considered as decent. Kharin *et al.*, (2007) found that precipitation corresponding to 20-yr return period is underestimated by around 20% for temperate regions, while results deteriorate for the tropical areas, where convective events are dominant. Figure 4 shows that the underestimation in the Mediterranean area is much worse (exceeding 50%).

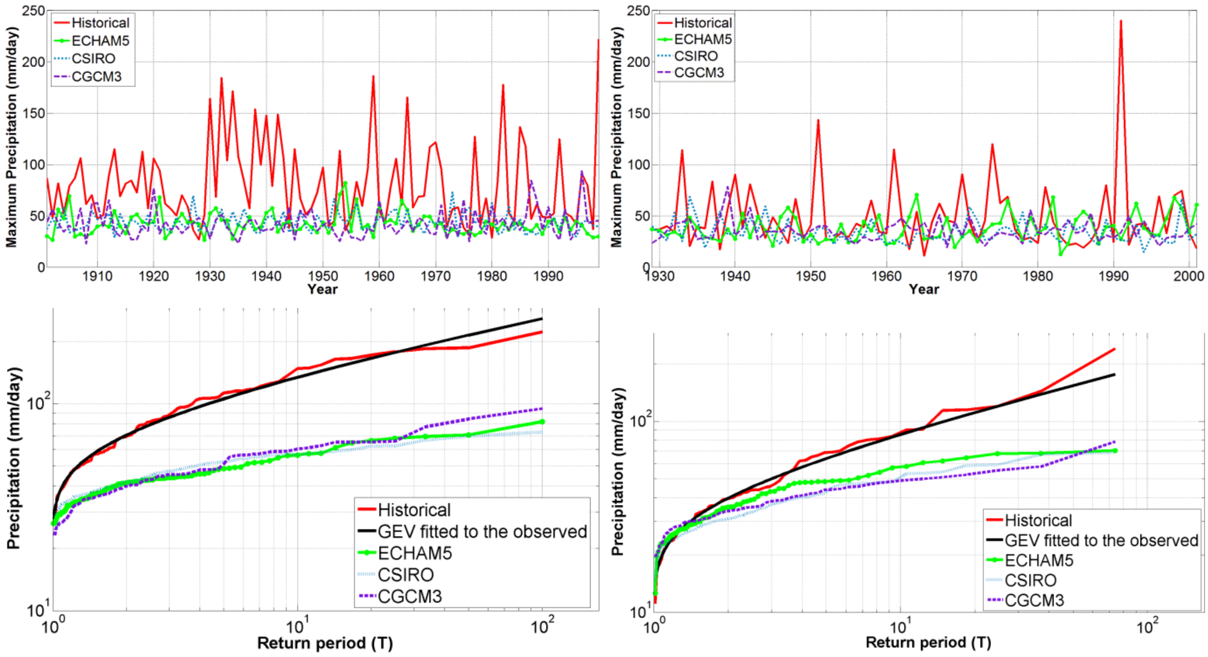

Figure 4: (a) Daily annual maximum precipitation in Perpignan and (b) Torrevieja and (c) empirical distributions in Perpignan and (d) in Torrevieja.

We also tested whether climate models qualitatively capture rising or falling climatic trends of extreme rainfall, despite the fact that their overall fit to the data is not satisfactory. An example is given in Figure 5 for the maximum precipitation in Perpignan. The entire time series is partitioned into three periods, each one with length of 33 years, a length that in climate research is regarded long enough to form a consistent picture of climate. The plots of the empirical distributions in Figure 5 indicate that the prominent discrepancy of the model results from historical records extends in all three periods. Moreover, it can be seen that in the second period the observed extremes were increased, in comparison to the first period; for example, the rainfall depth of 100 mm in the upper graph has return period of about 5 years, whereas in the middle graph its return period lowers to 3 years. However, the CSIRO and CGCM3 models do not capture this increasing trend. Furthermore, in the third period the highest rainfall events increased more, with the record rainfall now exceeding 200 mm. The ECHAM model, which behaved better when we moved from the first to the second period, now behaves in a worse way, suggesting a significant decrease of the most extreme rainfall instead of an increase.

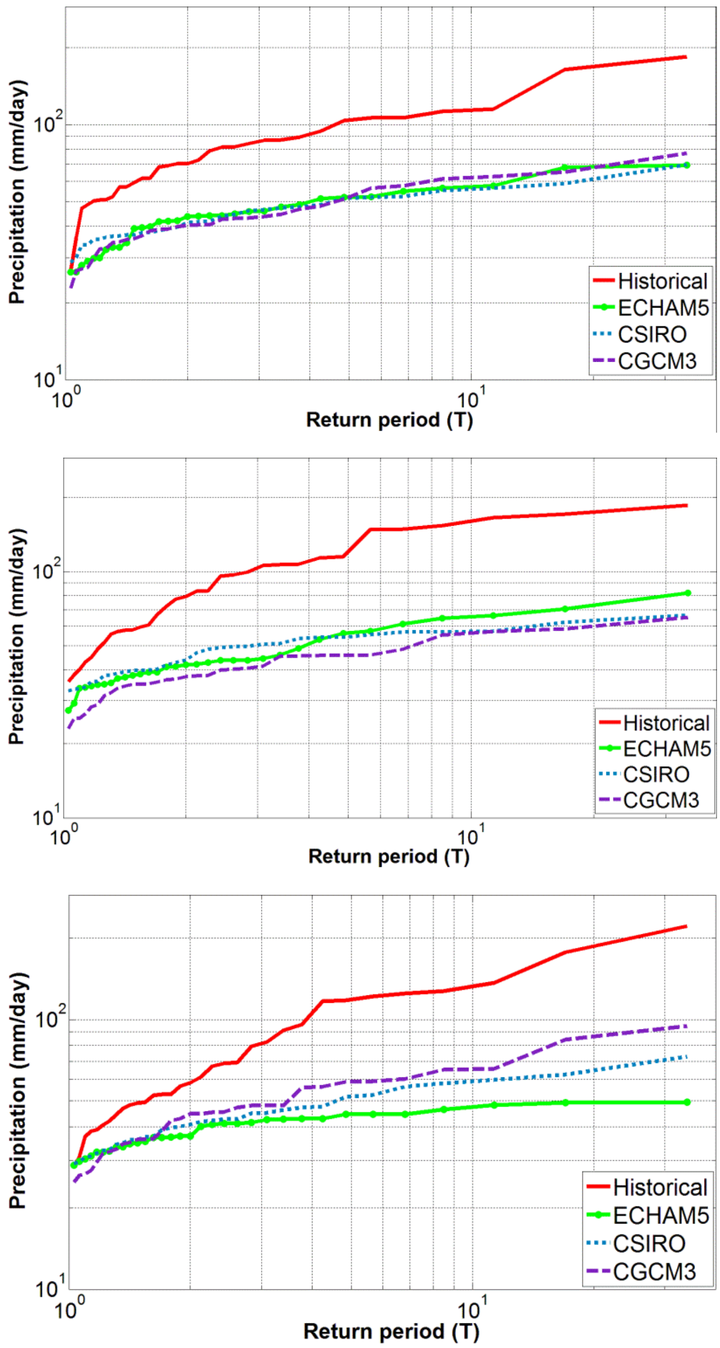

Figure 5: Empirical distribution of daily annual maximum precipitation in Perpignan for the sub-periods (a) 1901-1933, (b) 1934-1966 and (c) 1967-1999.

The values of correlation coefficient and Nash-Sutcliffe efficiency (NSE) are close to zero and negative, respectively, in daily time series of precipitation (Table 3). Consequently, the historical time series are uncorrelated with the simulated ones, leading to the conclusion that GCMs cannot reproduce the seasonality of wet and dry periods. The negative values of the NSE indicate that their predictive skill is worse than using the average. It should be mentioned, though, that GCMs do not produce results that can be used as weather forecast, namely the observed and modelled time series cannot be compared on a daily basis. Thus, one should not expect very high values of correlation coefficient and NSE. On the other hand, the results indicate that even the seasonality of rainfall cannot be captured by the models and this may perhaps be regarded as a cautionary note against their use in water resources management studies.

Figure 6 gives additional information on the performance of models at aggregated scales. Specifically, the modelled daily time series aggregated on an annual scale, as well as their 10-year moving averages, disagree with the observed data; neither of them being able to simulate the actual fluctuations. The probability density functions of monthly precipitation (Figure 7) indicate that the skewness is greater in the observed precipitation compared to the simulated. These findings are consistent with previous studies on aggregated time scales, which reported poor performance of GCMs (Koutsoyiannis *et al*. 2008, Anagnostopoulos *et al*. 2010).

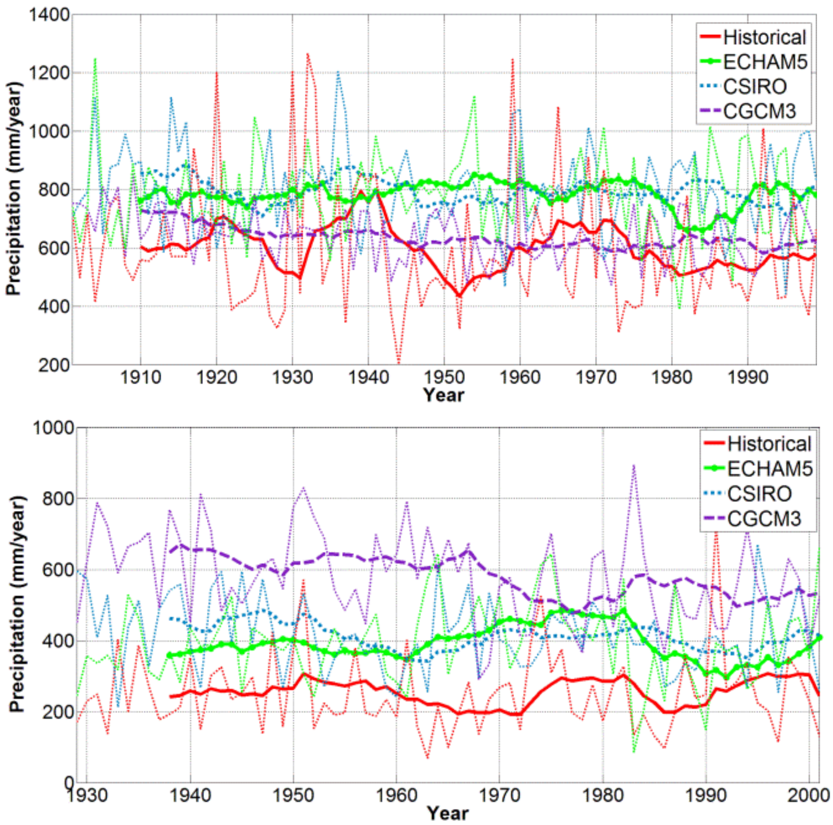

Figure 6: Annual and moving 10-year average precipitation in (a) Perpignan and (b) Torrevieja.

When we examine the occurrence of consecutive dry days (Figure 8) we observe that the results vary according to the GCM used; namely ECHAM5 can simulate them with a slight overestimation, while CGCM3 systematically underestimates such events. From the bar charts of daily precipitation thresholds (Figure 9) we can confirm that GCMs substantially underestimate events which can lead to a flood. This finding is further analyzed in Figure 10, Figure 11 and Figure 12 which depict the probability of wet days, the average rainfall depth of wet days and events with very heavy precipitation, respectively. For the probability wet we considered thresholds of 1 mm and 0.2 mm; note that in Istanbul the accuracy of measurements is 1 mm, therefore we calculate the probability wet only for the threshold of 1 mm/d.

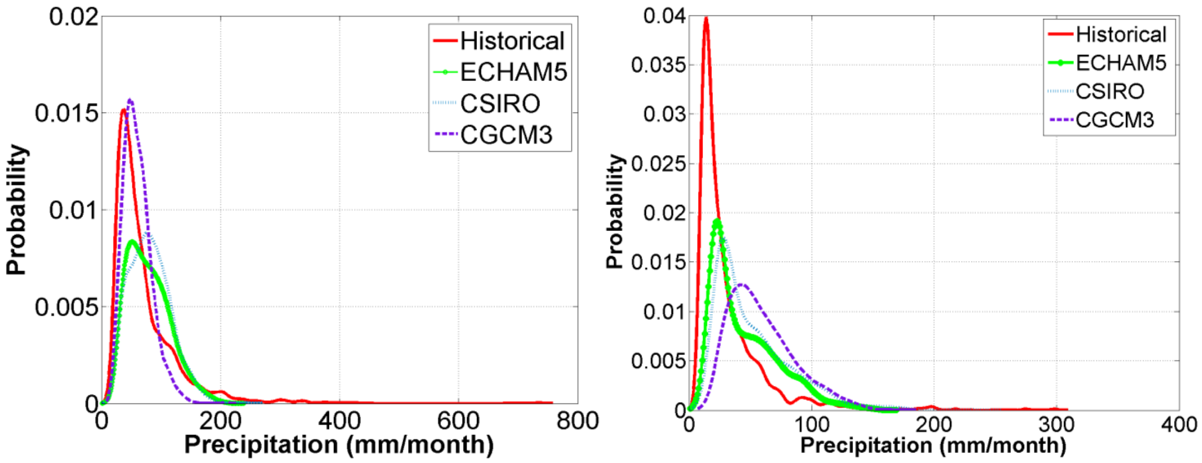

Figure 7: Probability density function of monthly precipitation in (a) Perpignan and (b) Torrevieja.

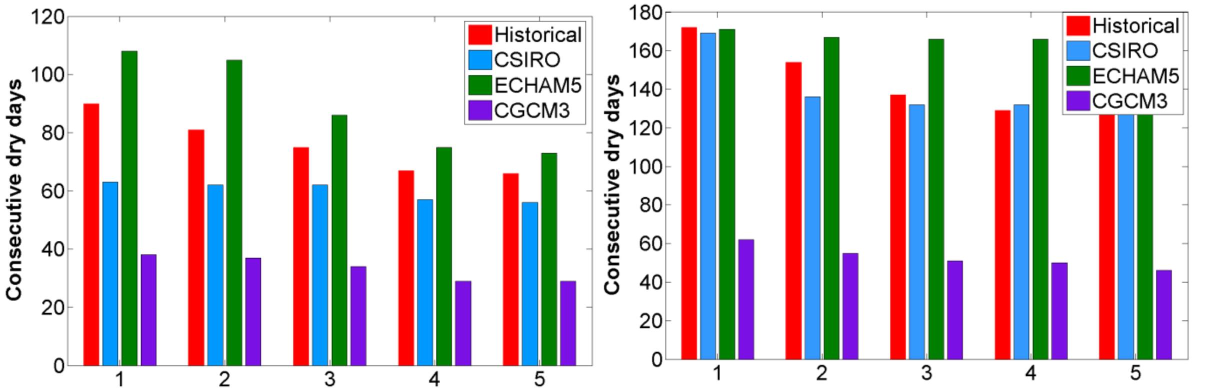

Figure 8: The five largest CDD in (a) Perpignan and (b) Torrevieja.

The precipitation frequency is systematically overestimated by model outputs, with the exception of some stations for CSIRO and ECHAM5 where the actual data are reproduced, while CGCM3 diverges from the observed data and simulates events which are not representative of the Mediterranean climate. Additionally, the mean values of wet days along with their 90% quantiles are underestimated in all locations.

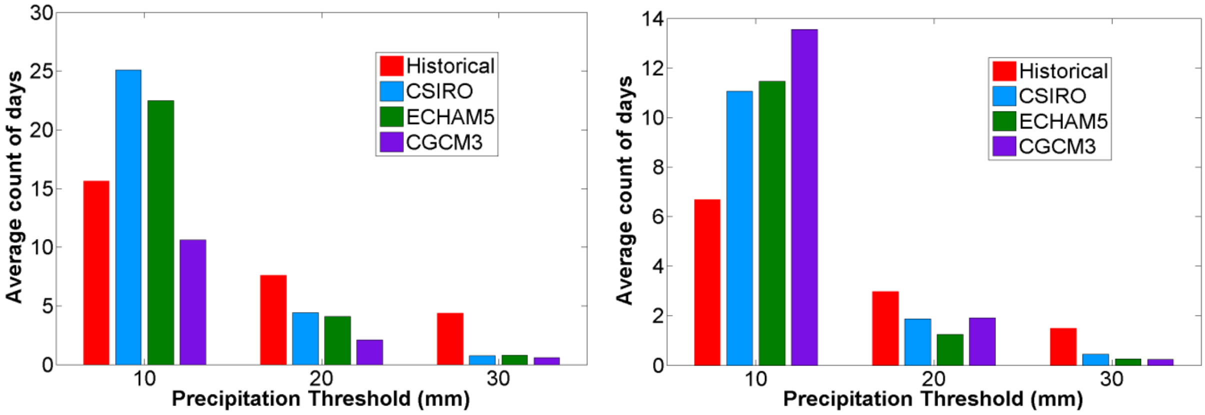

Figure 9: Precipitation thresholds (R10mm, R20mm, R30mm) in (a) Perpignan and (b) Torrevieja.

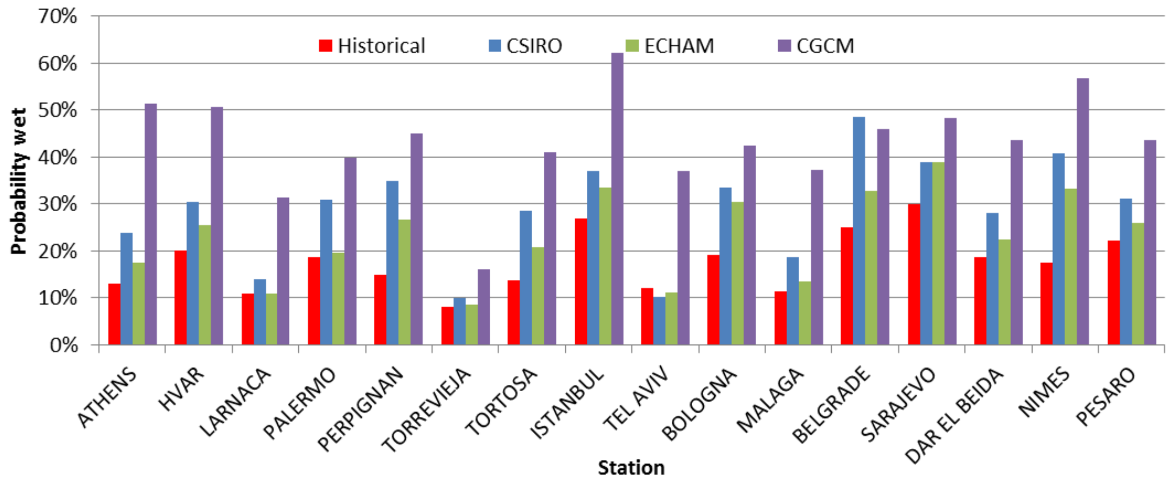

Figure 10: Probability of wet days in the complete dataset of stations with a threshold of (a) 1 mm and (b) 0.2 mm.

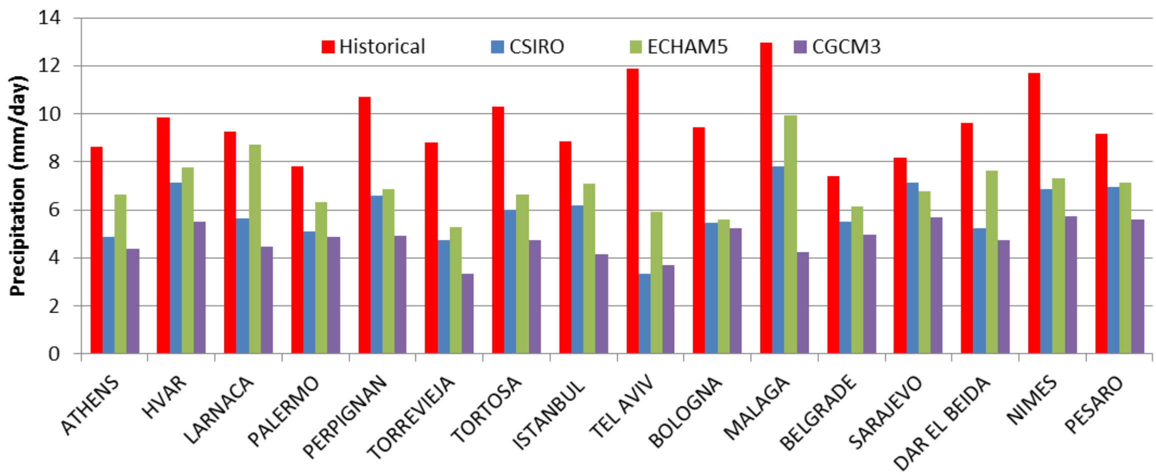

Figure 11: Mean precipitation depth of wet days (SDII) in the complete dataset of stations.

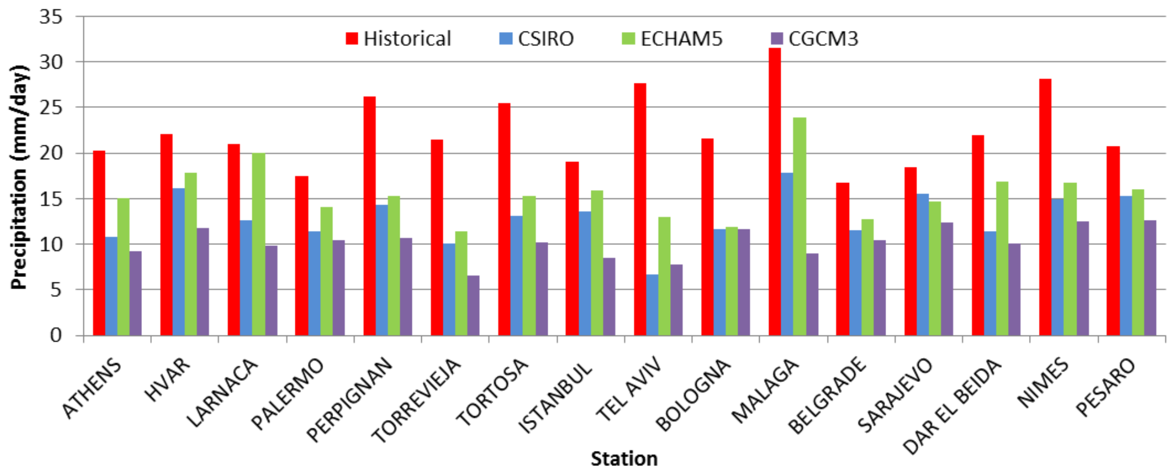

Figure 12: 90% quantile of the wet days (Prec90p) in the complete dataset of stations.

Table 3: Mean values of correlation and Nash-Sutcliffe efficiency coefficients in all studied locations for the daily and annual maximum time series of precipitation.

	Precipitation			**Temperature**		
	CSIRO	ECHAM5	CGCM3	CSIRO	ECHAM5	CGCM3
Correlation	0.05	0.06	0.04	0.65	0.58	0.69
NSE coefficient	-0.34	-0.44	-0.34	0.32	0.24	0.21

This has also been found by Dai (2006) who found that the probability of dry days is 45%-60% according to climate models, while TRMM data show that in reality it is 71% on the global scale. Climate models produce rainfall mostly in the interval 1-10 mm, while they generate daily precipitation larger than 20 mm too rarely. This shows that even if climate models could reproduce the reality at larger time scales, this would be a result of aggregating many days with very low rainfall depths, rather than simulating the reality.

5.2. Temperature

In the case of daily annual maximum temperature, the distributions of model interpolated time series fit the historical maxima in the majority of stations, not only for small return periods, but also for the largest ones (Figure 13). Similarly, Kharin *et al.*, (2007) found that warm temperature extremes are plausibly simulated by the IPCC AR4 models, since the global mean of averaged 20-year return period of maximum temperature was close to the one of NCEP2 and ERA-40 reanalyses datasets. GCMs are not supposed to forecast the weather, therefore relatively low values of the correlation and NSE coefficients of daily time series are expected. On the other hand, if the models are able to simulate the seasonality, then the time series should not be totally uncorrelated or random. In the case of temperature, both NSE and the correlation coefficient are rather satisfactory which indicates that the models capture the annual seasonality of temperature variation (Table 3).

In terms of the climate indices, the temperatures which correspond to the 90%-quantile fit well the observed ones, while the frequency of heat waves (WSDI) is overestimated by climate models (Figure 14).

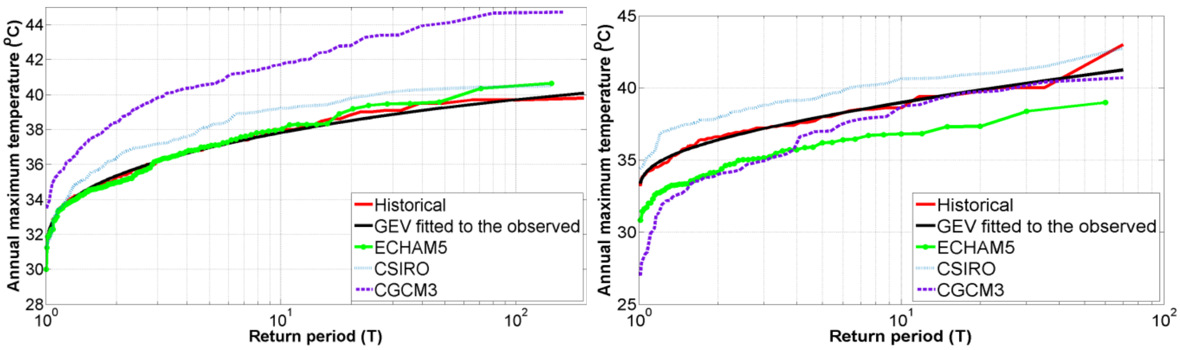

Figure 13: Empirical distribution of daily annual maximum temperatures in (a) Bologna and (b) Tortosa.

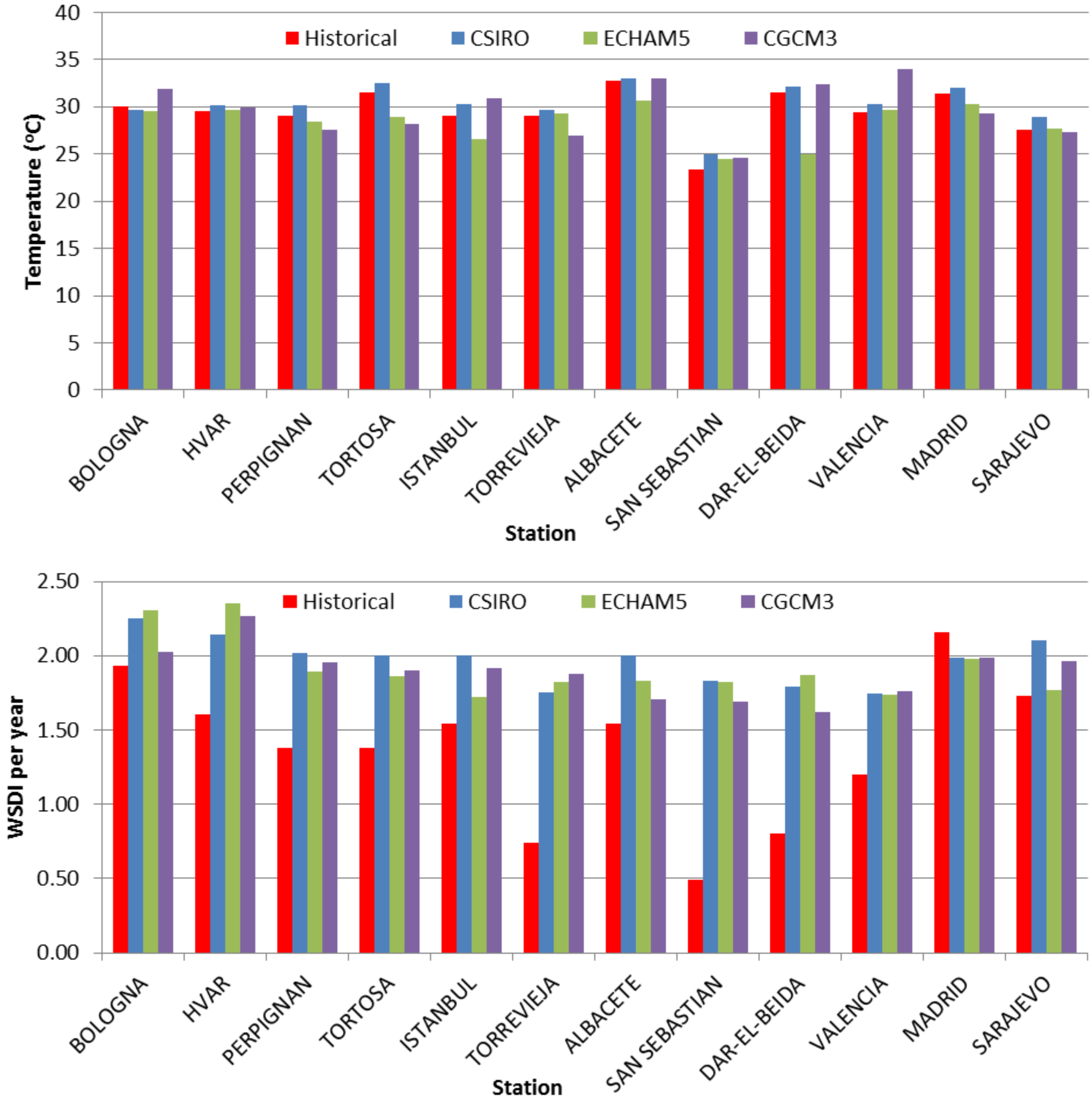

Figure 14: (a) 90% quantile of maximum temperature and (b) WSDI in the complete dataset of stations.

7. Conclusions

We compared observed time series of maximum temperature and precipitation to those produced by GCMs, in order to evaluate the credibility of the latter in reproducing the extremes. The comparison is made in the Mediterranean region, which is regarded to be one of the most vulnerable areas to climate change. The quantitative results of our analysis depict the GCMs do not reproduce the actual temporal variation of rainfall and temperature adequately, and in particular the occurrence of extreme events.

More specifically, in the case of temperature, models reproduce the seasonality and some statistical characteristics of maxima. The 90%-quantile of maximum temperature is simulated successfully, but the frequency of heat waves is overestimated by climate models.

In rainfall extremes, it is concluded that there is a systematic overestimation of the rainfall frequency, along with a severe underestimation of the rainfall intensity in all studied locations. Moreover, neither the statistical characteristics of daily maxima nor the succession of wet and dry periods can be reproduced, while the results for climate indices vary. Models consistently err in all time scales, from the daily to the 10-year moving average and they cannot capture the observed long-term trends. Our results are in accordance with other studies (e.g. Stephens *et al.*, 2010) who found that projected rainfall is far too light, but with a double frequency compared to the observed. These systematic errors imply different mechanisms of precipitation from the real Earth system. Therefore, downscaling GCM outputs of rainfall at individual grid cells is not meaningful. The problematic simulation of precipitation is also underlined in other studies which however find satisfactory results for other variables such as the surface specific humidity, surface pressure and surface air temperature (Johnson & Sharma, 2009; Wilby & Wigley, 2000). Furthermore, Beniston *et al.* (2007) found that precipitation extremes projected by seven RCMs can substantially vary, even though the forcings are provided by only two GCMs.

We can deduce that the discrepancy between the modelled outputs and the observed time series does not allow GCMs to provide the hydrological community with the required level of credibility in studying extreme events. The key issue of our study is that climate models are not ready for “prime time” in water resources management. Therefore, they should not be used injudiciously for hydrological applications, despite their constant promotion and the resulting shift of many water scientists in favour of their use. Model results may be helpful to enhance knowledge on some climate processes, but they are inappropriate for risk assessments having an impact on societies or the environment (Pielke Sr, 2012). Sakaguchi *et al.*, (2012) found that GCMs are not skilful at the regional (5^o x 5^o – 20^o x 20^o grid) and at the decadal scales, hence direct use of GCM outputs of long-term simulations is not recommended. Even if agreement between observed and modelled trends can be achieved in the global scale, this does not apply at the local scale, especially at areas with high-frequency variability (Oldenborgh *et al.*, 2012). Major concerns have been also raised by previous studies (Anagnostopoulos *et al.*, 2010; Kiem and Verdon-Kidd, 2011; Koutsoyiannis *et al.*, 2008; see also Kundzewicz and Stakhiv, 2010).

Changes in such a complex system, as is climate, can be triggered either by positive or negative feedbacks which compete each other and whose balance points are unknown constituting any multi-decadal climate predictions impractical (Rial *et al.*, 2004). Besides being a system where interactions are not easily identified, we are confident that the defects of current climate models are not only a result of biases and errors. On the contrary, they may reflect the intrinsically unpredictable character of climate and, in particular, the emergence of extreme events. Even if progress is made in comprehending climate, which is extremely complicated, the uncertainty does not necessarily decrease and predictability in the long term may not be achieved (Koutsoyiannis, 2010). Instead of trying to make accurate climate predictions, we believe that climate policies should emphasize on the sustainability and more specifically on societal and environmental resources affecting humans, ecosystems, water, energy and nutrition (Pielke Sr. *et al.*, 2009).

Acknowledgements: The comments of Associate Editor A. Sharma and three anonymous reviewers are gratefully appreciated. Their both encouraging and negative feedback helped us to improve the presentation of our findings and express them in a clearer way. We also thank Andreas Efstratiadis and Antonis Christofides for their contribution in a preliminary study presented in the EGU 2011 conference and F. Pierros of the National Observatory of Athens for allowing our access to rainfall data of Athens.

Funding Information: The authors received no funding for this research, which was performed for scientific curiosity.

Disclaimer: The authors declare that there is no conflict of interest regarding the publication of this paper.

References

Alexander, L.V. *et al*., 2006. Global observed changes in daily climate extremes of temperature and precipitation. J Geophys Res-Atmos, 111(D5).

Alpert, P., Krichak, S.O., Shafir, H., Haim, D., Osetinsky, I., 2008. Climatic trends to extremes employing regional modeling and statistical interpretation over the E. Mediterranean. Global and Planetary Change, 63(2-3): 163-170.

Anagnostopoulos, G. G., Koutsoyiannis, D., Christofides, A., Efstratiadis, A. & Mamassis, N. (2010) A comparison of local and aggregated climate model outputs with observed data. Hydrol. Sci. J. 55(7), 1094–1110. doi:10.1080/02626667.2010.513518

Beniston, M., Stephenson, D. B., Christensen, O. B., Ferro, C. a. T., Frei, C., Goyette, S., Halsnaes, K., *et al*. (2007) Future extreme events in European climate: an exploration of regional climate model projections. Clim. Change 81(S1), 71–95. doi:10.1007/s10584-006-9226-z

Blöschl, G., Ardoin-Bardin, S., Bonell, M., Dorninger, M., Goodrich, D., Gutknecht, D., Matamoros, D., *et al*. (2007) At what scales do climate variability and land cover change impact on flooding and low flows? 1241–1247. doi:10.1002/hyp

Christensen, J. H., Boberg, F., Christensen, O. B. & Lucas-Picher, P. (2008) On the need for bias correction of regional climate change projections of temperature and precipitation. Geophys. Res. Lett. 35(20), L20709. doi:10.1029/2008GL035694

Dai, A. (2006) Precipitation Characteristics in Eighteen Coupled Climate Models. J. Clim. 19(18), 4605–4630. doi:10.1175/JCLI3884.1

Dawson, A., Palmer, T. & Corti, S. (2012) Simulating regime structures in weather and climate prediction models. Geophys. Res. Lett. 39. doi:10.1029/2012GL053284

European Environment Agency. (2010) The European Environment, State and Outlook, Adapting to Climate Change.

Fyfe, J., Merryfield, W., Kharin, V. V., Boer, J., See, W. & Salzen, K. von. (2011) Skillful predictions of decadal trends in global mean surface temperature. Geophys. Res. Lett. 38. doi:10.1029/2011GL049508

IPCC. (2007) . Summary for Policymakers, Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge.

Johnson, F. & Sharma, A. (2009) Measurement of GCM Skill in Predicting Variables Relevant for Hydroclimatological Assessments. J. Clim. 22(16), 4373–4382. doi:10.1175/2009JCLI2681.1

Kharin, V. V., Zwiers, F. W., Zhang, X. & Hegerl, G. C. (2007) Changes in Temperature and Precipitation Extremes in the IPCC Ensemble of Global Coupled Model Simulations. J. Clim. 20(8), 1419–1444. doi:10.1175/JCLI4066.1

Kiem, A. S. & Verdon-Kidd, D. C. (2011) Steps toward “useful” hydroclimatic scenarios for water resource management in the Murray-Darling Basin. Water Resour. Res. 47(July 2010), 1–14. doi:10.1029/2010WR009803

Koutsoyiannis, D., Efstratiadis, A., Mamassis, N. & Christofides, A. (2008) On the credibility of climate predictions. Hydrol. Sci. J. 53(August 2008), 671–684. Retrieved from http://www.tandfonline.com/doi/abs/10.1623/hysj.53.4.671

Kundzewicz, Z. W. (2010) Comments on the paper “Credibility of climate predictions revisited.”

Kundzewicz, Z. W. & Stakhiv, E. Z. (2010) Are climate models "ready for prime time" in water resources management applications, or is more research needed? Hydrol. Sci. J. 55(7), 1085–1089. doi:10.1080/02626667.2010.513211

Luterbacher, J., Xoplaki, E., Casty, C. & Wanner, H. (2006) Mediterranean climate variability over the last centuries: a review. Dev. Earth. Retrieved from http://www.sciencedirect.com/science/article/pii/S1571919706800042

Mauritsen, T., Stevens, B., Roeckner, E., Crueger, T., Esch, M., Giorgetta, M., Haak, H., *et al*. (2012) Tuning the climate of a global model. J. Adv. Model. Earth Syst. 4(3). doi:10.1029/2012MS000154

Munich Re. (2011) Great natural catastrophes worldwide 1950 – 2011.

National Environmental Research Council (1975), Flood Studies Report, Institute of Hydrology, Wallingford, UK.

Oldenborgh, G. J., Doblas-Reyes, F. J., Wouters, B. & Hazeleger, W. (2012) Decadal prediction skill in a multi-model ensemble. Clim. Dyn., Vol. 38, 1263–1280. doi:10.1007/s00382-012-1313-4

Pielke Sr, R. (2012) Climate Science Malpractice – The Promotion Of Multi-Decadal Regional Climate Model Projections As Skillful. Retrieved from http://pielkeclimatesci.wordpress.com/2012/03/14/climate-science-malpractice-the-promotion-of-multi-decadal-regional-climate-model-projections-as-skillfull/

Pielke Sr, R. & Wilby, R. L. (2012) Regional Climate Downscaling: What's the Point? Eos Forum 93(5), 52–53. doi:10.1029/2012EO050008

Pielke Sr., R., Beven, K., Brasseur, G., Calvert, J., Chahine, M., Dickerson, R., Entekhabi, D., *et al*. (2009) Climate Change : The Need to Consider Human Forcings Besides Greenhouse Gases. Eos (Washington. DC). 90(45). doi:10.1007/s10533-

Rial, J. a., Pielke Sr., R., Beniston, M., Claussen, M., Canadell, J., Cox, P., Held, H., *et al*. (2004) Nonlinearities, Feedbacks and Critical Thresholds within the Earth's Climate System. Clim. Change 65(1/2), 11–38. doi:10.1023/B:CLIM.0000037493.89489.3f

Sakaguchi, K., Zeng, X. & Brunke, M. a. (2012) The hindcast skill of the CMIP ensembles for the surface air temperature trend. J. Geophys. Res. 117(D16), D16113. doi:10.1029/2012JD017765

Shohami, D., Dayan, U. & Morin, E. (2011) Warming and drying of the eastern Mediterranean: Additional evidence from trend analysis. J. Geophys. Res. Atmos. 116(D22), n/a–n/a. doi:10.1029/2011JD016004

Stephens, G. L., L'Ecuyer, T., Forbes, R., Gettlemen, A., Golaz, J.-C., Bodas-Salcedo, A., Suzuki, K., *et al*. (2010) Dreary state of precipitation in global models. J. Geophys. Res. 115(D24), 1–13. doi:10.1029/2010JD014532

Sun, F., Roderick, M., & Farquhar, G., (2012) Changes in the variability of global land precipitation, Geophysical Research Letters, vol. 39, L19402, doi:10.1029/2012GL053369, 2012

Wilby, R. L. & Wigley, T. M. L. (2000) Precipitation predictors for downscaling: observed and general circulation model relationships. Int. J. Climatol. 20(6), 641–661. doi:10.1002/(SICI)1097-0088(200005)20:6<641::AID-JOC501>3.3.CO;2-T

Xu, Z. & Yang, Z. (2012) An improved dynamical downscaling method with GCM bias corrections and its validation with 30 years of climate simulations. J. Clim.

Appendix: Submissions to Journals

This research study has been submitted in several journals and was rejected; review comments and authors' replies to them (in red font, excluding those which are related to typos or minor corrections) are included below.

Journal of Hydrology, First Submission (2012/03/21)

From: Konstantine Georgakakos

Date: 2012/06/02
Subject: HYDROL12807: Editor's decision

To: dtsaknias@gmail.com

Cc: hydrolkpg@xxx.xxx, a.sharma@xxx.xxx

Ref.: "Statistical comparison of observed temperature and rainfall extremes with climate model outputs in the Mediterranean region" (Mr Dimosthenis Tsaknias)

Dear Mr Tsaknias,

I very much regret to have to tell you that publication in our journal is not recommended.

We would, however, consider as a new submission for review a substantially revised version of this paper that addresses all of the reviewers' comments. Should you choose to submit such a revised manuscript please refer to the present manuscript number, provide a detailed point-by-point reply to all of the reviewers' comments, and state how the revised manuscript addresses these.

An explanation for this decision is given in the attached review reports (and on http://ees.elsevier.com/hydrol/). I hope that the comments contained therein will be of use to you.

Thank you for your interest in our journal.

With kind regards,

Konstantine P. Georgakakos, Sc. D.
Editor
Journal of Hydrology

COMMENTS FROM EDITORS AND REVIEWERS

Associate Editor: This paper presented a method for transforming GCM precipitation and temperature to point scales, and then a comparison of this method with observed data for the Mediterranean region. The paper was reviewed by 3 reviewers and by me, and all except one of the reviewers have significant concerns about what is reported. The biggest concerns are laid out by reviewer 1, who states that the problem in comparing observations to the downscaled values lies not with the GCMs but with the downscaling method that is used. Additional problems are noted by reviewer 2, who quite rightly states that the comparison being performed is not appropriate, unless the authors use a downscaling model that corrects for the known biases associated with GCM simulations of hydrological variables (and especially rainfall). In my own reading of the paper, I tended to agree with these two reviewers, and had a few additional comments that I am writing below:

l137 - If I understand correctly, what has been done here is to present a way to estimate p and q in the transformation, such that the interpolated values exhibit moments that are consistent with observations. This part of the presentation is quite confusing - at the very least authors should use different notation for the observations and the transformed series - and since this is probably the key contribution thay are making here, this needs a much better explained presentation.

l140 - This doesn't make sense - are the authors attempting to form error measures between coincident raw (areally corrected) observations and GCM downscaled rainfall and temperature? They should know that because the GCM is initialised at a common point in time and then set to evolve based on atmospheric concentrations and extraterrestrial radiations, one cannot compare values for the same days/times with observations. Instead comparisons are made with reference to overall probability distributions for a much larger reference time span. What the GCM gives us better than other alternatives are estimates of likely changes due to changed atmospheric forcings. It is for this reason that sophisticated downscaling alternatives are used and more trust is attached to certain variables simulated by GCMs (such as upper level temperature and pressures - see [Johnson, F. and A. Sharma (2009). "Measurement of GCM Skill in Predicting Variables Relevant for Hydroclimatological Assessments." Journal of Climate 22(16): 4373-4382.]).

No, we do not try to preserve the moments from the observed time series. After interpolation, the moments have smaller absolute values than the raw climate model outputs for each grid cell. Therefore, with eq. (5) we capture the moments of the raw data.

The role of observed time series is introduced in eq. (6), where the mean square error is minimized. It is well-known that GCMs do not produce results that can be compared with the observed time series in a straightforward way and that the reference should be the overall statistical behaviour. However, with our methodology we reduce the bias and allow GCMs to capture observations as close as possible. We believe that after changing the notation, the presentation of our algorithm is clear.

Based on this, I feel that considerable work is needed in revising the paper before it can be suitable for publication. At the minimum, the authors do need to assess the viability of their proposed downscaling logic by contrasting it with other downscaling metrics. They also need to take into consideration the relatively large uncertainty of raw GCM rainfall, and use this in developing the model they are coming up with. They also need to rethink what exactly is the focus and novelty of the paper - I assumed this was the downscaling metric, but get a feeling from the writing at many stages that it is the comparison of GCMs and observations. Given the considerable work that is needed in doing so, I am recommending the paper be rejected in its current state, but the authors invited to submit a marked revised version if they feel they have been able to address the concerns that were stated here.

The novelty of our paper is the identification of elements of reality at GCMs outputs at the daily time scale. If this assumption is verified, then proceeding with techniques of the climatological toolbox can be meaningful.

We have added a new section (Justification of methodology) where we explain why one should be cautious with downscaling. We had erroneously written that our methodology is downscaling, which does not hold. In fact, what we did is an interpolation where climate model outputs are compared with observed time series. If the interpolation verifies the credibility of GCMs, then one could go to the next step which is downscaling, in order to derive results at the scale and area of interest.

Reviewer #1: This paper investigates whether GCMs can reproduce extremes in observed daily rainfall and temperature at point locations. It is motivated by the fact that hydrological and water resource models need climate inputs at spatial scales much smaller than those used in GCMs. The authors show that GCM grid outputs mapped to point locations using their nonlinear BLUE transformation cannot adequately reproduce extremes observed at the point scale. They therefore conclude that the usefulness of GCMs in hydrological applications is questionable.

The finding that the BLUE transformation on GCM outputs fails to reproduce extreme characteristics at the point scale is well supported by the evidence presented in the paper. However, it does not follow from this finding that the usefulness of GCMs is questionable. It may be that the BLUE model given by eqns (1) to (3) is inadequate.

Indeed the authors admit on p. 7 that "decent fitting can be achieved only in temperature. On the contrary, the extreme daily rainfall are systematically underestimated by the models and the results of the optimization still diverge." The main problem appears to be that the BLUE model is deterministic and, for that reason, may be intrinsically incapable of reproducing extremes. A similar situation arises with deterministic rainfall-runoff models. They underestimate the actual variability in runoff and are unable to simulate the flood frequency curve.

So it appears that the main conclusion is that the authors were unable to satisfactorily transform GCM rainfall outputs to point rainfall using their BLUE approach. It does not follow that use of GCMs in hydrological applications is of questionable value. There may be other approaches for transforming GCM rainfall to point rainfall that work better. It is surprising that the authors did not consider the extensive literature on downscaling given that the objective of downscaling is similar to the authors' BLUE approach. Without this perspective there seems little value in this failed attempt to transform GCM outputs to point rainfall.

First of all, the BLUE technique is not a deterministic method but a statistical method. As its name says, it provides the best linear unbiased estimate of an unknown random variable based on correlations with other random variables whose values are known. Second, the problems indicated by the reviewer are tackled in the proposed modification of the blue method. We show that with this variant the extremes can be reproduced by increasing the value of the adjustable parameter r.

Therefore we do not agree with the criticism of BLUE. The reviewer may also see Koutsoyiannis *et al*., (2008) who show that BLUE is appropriate not only for temperature and precipitation fields at the climatic scale, but even for rough random fields. They tried to estimate the efficiency of the BLUE method by assuming that they know the values of three out of four neighboring grid cells and the fourth was considered as unknown. Using the BLUE technique they tried to estimate the time series of the unknown grid cell with three weights. Their results where particularly good, as the Nash-Sutcliffe efficiency coefficient was 0.91 and 0.99 for precipitation and temperature, respectively. It is clear then, that with the use of 4 weights the results should be improved.

Even though bias is minimized, it is not completely removed. Therefore, we examine not only the Nash-Sutcliffe efficiency coefficient, but also the correlation coefficient that neglects any systematic bias.

Apparently, its use is justified at flat terrains, where the variations due to orographic effects are not significant. However, almost all the stations used in our analysis are located at low elevation as can be seen in Table 2.

Moreover, we added to the manuscript the following paragraph: “Anagnostopoulos *et al*., (2010) reconstructed time series of precipitation and temperature from point measurements for the whole area of USA and compared them with GCM outputs. They found out that large scale results are worse compared to the point comparison after using the BLUE method which allows the modelled time series to fit the observed ones. Hence, our algorithm is more favourable to GCMs and allows them to avoid local influences. In addition, the non-linear conversion of BLUE method, avoids any bias in the first four orders of moments.”

Finally we do not propose our methodology as not another downscaling technique. In the new section “Justification of methodology” we explain why downscaling should be the step after the use of our interpolation technique, where we attempt to assess whether the error is larger than the signal.

The calibration scheme for the nonlinear BLUE model deserves further comment. Eqn (6) is a weighted combination of two criteria, MSE and "e". The text preceding eqn (6) states that the weight "w" is calibrated along with the parameters "p" and "q". It is difficult to understand what such a calibration would produce as "w" controls the trade-off between MSE and "e". It is unclear which value of "w" would be best in terms of bias and reproduction of variability. The absence of any analysis on this issue adds further doubt about the value of the nonlinear BLUE approach to transform GCM rainfall to point locations.

Please see our reply in editor's comments about the methodology and the equations used.

Reviewer #2: Review of manuscript number HYDROL12807 titled "Statistical comparison of observed temperature and rainfall with climate model outputs in the Mediterranean region" by D. Tsaknias et al.

The authors compared the statistics of climate model data and observed station data are compared in order to study whether climate models qualify as adequate with respect to the prediction of extremes, i.e. maxima of temperature and precipitation. The authors find that the observed time series are not simulated sufficiently by the models. They conclude that climate models may not be useful in hydrological applications.

The deficiencies of climate models, which are addressed in this paper, have been studied in the past and it is well-known that many climate models tend to produce too many "drizzle" days with weak but rather persistent precipitation intensity, a model artifact that relates to cloud precipitation and the process of generating precipitation, e. g. when a humidity threshold is exceeded. These artifacts are often compensated - to some extent - by the use subsequent statistical procedures, when climate models are to be used as input fields to hydrological applications. The literature on these problems is rapidly growing and models are still far from sufficient as far as proper precipitation simulation is concerned.

Some references:

- Dai, Aiguo, 2006: Precipitation Characteristics in Eighteen Coupled Climate Models. J. Climate, 19, 4605-4630.
- Kharin, Viatcheslav V., Francis W. Zwiers, Xuebin Zhang, Gabriele C. Hegerl, 2007: Changes in Temperature and Precipitation Extremes in the IPCC Ensemble of Global Coupled Model Simulations. J. Climate, 20, 1419-1444.
- Martin Beniston, David B. Stephenson, Ole B. Christensen, Christopher A. T. Ferro, Christoph Frei, Stéphane Goyette, Kirsten Halsnaes, Tom Holt, Kirsti Jylhä and Brigitte Koffi, et al. Future extreme events in European climate: an exploration of regional climate model projections. Climatic Change, Volume 81, Supplement 1 (2007), 71-95

Concerning bias correction, e.g.:

Jens H. Christensen et al. On the need for bias correction of regional climate change projections of temperature and precipitation. GEOPHYSICAL RESEARCH LETTERS, VOL. 35, L20709, 6 PP., 2008

Against this background, the study of Tsaknias et al. lacks a thorough discussion of the literature on the topic and it is not clear how the current study adds to that.

The fact that our results are in accordance with the literature is reassuring for our study. However, literature is dominated by publications which injudiciously use climate models to project future climate drawing conclusions - or even defining policies. We believe that the credibility of climate models on the multi-decadal time scale has not been sufficiently evaluated yet.

We believe that with our study we enhance literature about the falsifiability of GCMs with a focus on the Mediterranean, which is a “hot-spot” of climate models as we explain in the manuscript. As we support in the section “Justification of methodology” downscaling can be implemented only after verifying that climate model outputs show some skill to simulate current climate conditions. In addition, we propose a non-linear transformation of the BLUE technique which allows the interpolation of time series at fine time scales.

A further point that needs to be motivated more is the comparison made between point rainfall data (stations) and climate model data. The task of converting extremes between these two statistical representations is cumbersome and a great deal of the paper hinges on equation (7), namely a formula for the "reduction factor" that is said to convert between point maxima and gridded model output. This formula needs to be explained and motivated better. The reference to Koutsoyiannis and Xanthopoulos (1999) is incomplete, therefore the derivation of this formula is hard to trace. Because of its central role in this paper, this formula needs to be re-derived in the body of the article.

At any rate, it is questionable how this formula - a function only of area and time scale - can possibly work for all precipitation events studied in this article (e.g. convective events are known to occur on very small spatial scales compared to large-scale events are both are strongly dependent on temperature). The authors estimate the "reduction factor" to be ~.76, does this mean, point rainfall and 200 x 200 km^2 rainfall are almost the same?

The formula for the estimation of the areal reduction factor is empirical and has not been derived analytically. The book containing it (Koutsoyiannis and Xanthopoulos, 1999, http://itia.ntua.gr/getfile/115/3/documents/1999EngHydroChap2.pdf p.154) is written in Greek. We will hereby briefly explain how it has been derived.

There have been extensive empirical studies about the changes of the areal reduction factor depending on the area and the event duration for the USA (Hershfield, 1961; Miller, 1973; Dingman 1994; Viessman *et al*., 1989) and the UK (National Environmental Research Council, 1975; Wilson, 1990; Shaw, 1994). The most complete dataset was provided by the National Environmental Research Council, hence the function presented in eq. (7) has been fitted on it. The equation was further compared with the data provided by the U.S. Weather Bureau for the Western States and exhibited a very good performance. Although there were some obvious differences, they were not of significant importance. Thus, we believe that the aforementioned relationship can be used in our study without introducing errors that affect our analysis.

Reviewer #3: This paper continues a sequence of studies by Koutsoyianni and colleagues of the predictive skill (using hindcast runs) of multi-decadal global model predictions. In this study the focus is on the Mediterranean region.

The paper certainly should be published as such assessments of regional predictive skill of climate statistics on multi-decadal time periods is mostly lacking.

I do have several suggested added papers that the authors might choose to include and comment on [two are just using global models and two refer to regional dynamic downscaling from the global models]. They are:

1. Fyfe, J. C., W. J. Merryfield, V. Kharin, G. J. Boer, W.-S. Lee, and K. von Salzen (2011),Skillful predictions of decadal trends in global mean surface temperature, Geophys. Res. Lett.,38, L22801, doi:10.1029/2011GL049508

who concluded that "....for longer term decadal hindcasts a linear trend correction may be required if the model does not reproduce long-term trends. For this reason, we correct for systematic long-term trend biases."

2. Xu, Zhongfeng and Zong-Liang Yang, 2012: An improved dynamical downscaling method with GCM bias corrections and its validation with 30 years of climate simulations. Journal of Climate 2012 doi: http://dx.doi.org/10.1175/JCLI-D-12-00005.1

who find that without tunlng from real world observations, the model predictions are in significant error. For example, they found that "...the tradiional dynamic downscaling (TDD) [i.e. without tuning) overestimates precipitation by 0.5-1.5 mm d-1.....The 2-year return level of summer daily maximum temperature simulated by the TDD is underestimated by 2-6°C over the central United States-Canada region."

3. van Oldenborgh, G.J., F.J. Doblas-Reyes, B. Wouters, W. Hazeleger (2012): Decadal prediction skill in a multi-model ensemble. Clim.Dyn. doi:10.1007/s00382-012-1313-4
who report quite limited predictive skill when they conclude that "A 4-model 12-member ensemble of 10-yr hindcasts has been analysed for skill in SST, 2m temperature and precipitation. The main source of skill in temperature is the trend, which is primarily forced by greenhouse gases and aerosols. This trend contributes almost everywhere to the skill. Variation in the global mean temperature around the trend do not have any skill beyond the first year. However, regionally there appears to be skill beyond the trend in the two areas of well-known low-frequency variability: SST in parts of the North Atlantic and Pacific Oceans is predicted better than persistence. A comparison with the CMIP3 ensemble shows that the skill in the northern North Atlantic and eastern Pacific is most likely due to the initialisation, whereas the skill in the subtropical North Atlantic and western North Pacific are probably due to the forcing."

4. Pielke Sr., R.A., and R.L. Wilby, 2012: Regional climate downscaling - what's the point? Eos Forum, 93, No. 5, 52-53, doi:10.1029/2012EO050008

who list a set of reasons that regional dynamic downscaling of multi-decadal are not skillful, and, at best state that "It is therefore inappropriate to present type 4 [i.e. multi-decadal climate predictions] results to the impacts community as reflecting more than a subset of possible future climate risks."

The current paper under review provides a framework that the authors could use to further discuss their study.

On the title I suggest changing to "Statistical comparison of observed temperature and rainfall extremes with multi-decadal climate model hindcasts in the Mediterranean region"

We appreciate the positive feedback from Reviewer #3. We included all the suggested publications apart from the title modification. We believe that the word "multi-decadal" is a pleonasm, since climate cannot be "intra-decadal". Moreover, "hindcasts" is not accurate enough, because some of the model outputs we use were forecasts when they were produced.

We further discuss the role of downscaling in climate projections, as this happens to be the main point of controversy over our study. We added a new section (*Justification of methodology*) where we explain that downscaling techniques are just mathematical exercises that fit model outputs to observed patterns.

Journal of Hydrology, Second Submission (2013/04/28)

From: Konstantine Georgakakos

Date: 2013/07/11

Subject: HYDROL14750: Editor's decision

To: dtsaknias@gmail.com

Cc: hydrolkpg@xxx.xxx, a.sharma@xxx.xxx

Ref.: "Statistical comparison of observed temperature and rainfall extremes with climate model outputs in the Mediterranean region" (Mr. Dimosthenis Tsaknias)

Dear Mr. Tsaknias,

I very much regret to have to tell you that publication in our journal is not recommended. An explanation for this decision is given in the attached review reports (and on http://ees.elsevier.com/hydrol/). I hope that the comments contained therein will be of use to you.

Thank you for your interest in our journal.

With kind regards,

Konstantine P. Georgakakos, Sc. D.
Editor
Journal of Hydrology

COMMENTS FROM EDITORS AND REVIEWERS

Associate Editor: This paper presents an assessment of GCM rainfall and temperature extremes in the Mediterranean region, and concludes that the GCM simulations are not of a high enough quality to warrant usage in planning and design. Two reviews of this paper have been received, one of which is from a new reviewer who is also an expert in this general area of research. Both reviewers are critical of this paper, arguing that what this paper presents (problems with raw GCM extremes) is widely accepted in the literature, and does not by itself present any innovation. They also go on to state that the BLUE approach presented here, by itself, does not offer a solution to the problems the authors have demonstrated exist, and hence the overall contribution is not high enough for the paper to merit publication in this journal. I am afraid I tended to agree with the assessments of these two reviewers, and felt that the authors should have enhanced the arguments related to BLUE logic, perhaps coupling it with a suitably formulated downscaling approach, to offer a novel solution to the problem they have raised. As such, this paper comes across as an assessment, rather than a contribution, and the assessment by itself does not have the novelty that is needed to merit publication in the journal. As a result, I am recommending that this paper be rejected from consideration in the journal in its current form. Having said so, I sincerely hope the comments that are presented here will assist the authors to further improve their work and publish in a suitable publication outlet. I am sorry I cannot be more positive in my assessment.

Reviewer #1: As mentioned in my previous review, this type of study is not new and the authors do not make additional attempts to justify its publication. Studies highlighting the insufficient ability of GCMs to describe rainfall extremes are abundant (I gave a number of references in my previous review). The authors do not properly address my concern from the previous review in this respect. I am left to guess what the novelty of the current study might be.

The authors' response "We believe that with our study we enhance literature about the falsifiability of GCMs with a focus on the Mediterranean, which is a "hot-spot" of climate models as we explain in the manuscript. As we support in the section "Justification of methodology" downscaling can be implemented only after verifying that climate model outputs show some skill to simulate current climate conditions. In addition, we propose a non-linear transformation of the BLUE technique which allows the interpolation of time series at fine time scales." is correct, but doesn't justify publication in Journal of Hydrology.

My second concern, regarding the formula for the "reduction factor" has also not been addressed satisfactorily. My concern was that a formula that is based only on area and time scale can be sufficient to convert between point maxima and gridded model output. I asked for a more detailed description of this formula but the explanation (and reference to a document in Greek) are again insufficient.

In conclusion, I do not feel that a convincing attempts has been made in supporting the claim of novelty or new insight into deficiencies of extreme event modeling in the Mediterranean region or elsewhere.

Reviewer #2: This manuscript tests the ability of GCMs to simulate extreme precipitation, and finds that the models do poorly. The authors find that "the observed time series are not sufficiently simulated by the climate models". which the authors say raises "serious concerns" about the "usefulness of climate models in hydrological applications."

I do not have any major problems with the results in the paper - by enlarge they are consistent with my experience of GCM's precipitation outputs. I would like the authors to clarify several aspects of their methodology (particularly adding dimensions to some of their variables), but this is a minor point. The authors should also provide a plot of monthly average intensity and occurrence statistics, to complement their NSE results that find that models do not properly simulate seasonality. In terms of the analysis/results, probably my major criticism of the paper is lack of originality: the fact that GCMs underestimate precipitation extremes in particular and precipitation intensity more generally - and overestimate occurrence statistics - is pretty well known. This was also mentioned by Reviewer #2.

However I do have a problem with how the authors contextualise their results. They have been highly selective in their literature review, taking an adversarial rather than a balanced approach in making their argument. The authors are adopting a 'straw man' argument in which they imply that it is standard (or even 'best') practice to use GCM outputs of extremes as a basis for making water management decisions, when this is not the case. Their arguments against downscaling are not convincing, given review papers by [Fowler et al., 2007; Maraun et al., 2010], and given that the paper does not evaluate downscaling techniques, they do not add to the debate on this subject. In Australia, a "multiple lines of evidence" approach to examining future changes to extreme precipitation is being taken, which includes (i) dynamically downscaling to ~2km grid resolution - generally regarded as the minimum resolution required to resolve convective processes needed to simulate extreme precipitation; (ii) looking at temperature scaling relationships between extreme precipitation and temperature; and (iii) historical analysis of systematic changes in extreme rainfall intensity and/or occurrence.

The other logical fallacy is that the authors are making conclusions on the general - "usefulness of climate models in hydrological applications" - from the specific [a single class of 'climate model' (GCM - see [McGuffie and Henderson-Sellers, 2005] for other types of model), a single region, and a focus on one measure of precipitation that is well-known to be particularly difficult to simulate (extremes)]. There are plenty of aspects of the planetary hydrological cycle (e.g. large-scale circulation, global water and energy budget calculations, temperature) that are reasonably well simulated by GCMs, and used in the right way I believe they can provide credible projections which can be a useful "line of evidence" for water resource managers.

I have therefore decided to reject the paper, for the following four reasons: (1) insufficiently original and 'new' for the Journal of Hydrology; (2) lack of balance in literature review; (3) use of the 'straw man' fallacy by implying that water resource managers would use GCM outputs of precipitation extremes, which in my experience is almost never the case; and (4) making very general conclusions from specific findings, with the findings not sufficient to substantiate the conclusions.

Specific comments:

Abstract: "Moreover, a general perception has been developed that extreme events will be more frequent in the future." The "perception" that the authors talk about is supported by significant evidence, and the authors should acknowledge this in the literature review. The authors need to provide a fair and balanced review of the literature on how this general perception has occurred. Particularly, the authors should review lines of evidence such as physical reasoning [e.g. the notion that the maximum intensity of preciptiation should somehow scale with the moisture holding capacity of the atmosphere: e.g. Trenberth et al., 2003], temperature scaling arguments [Berg et al., 2013; Lenderink and van Meijgaard, 2008], a large number of GCM and downscaled modelling experiments, idealised modelling experiments [Muller et al., 2011; Romps, 2011] and observations which at the global scale are consistently showing increases in daily extreme precipitation [Min et al., 2011; Westra et al., 2013]. Also IPCC reports such as the Special Report on Extremes [Seneviratne et al., 2012] which summarises many references. There are also quite a few references in climate journals over the last few months that have been written for the forthcoming IPCC.

Abstract: "Comparisons of climate model outputs with observed data are made in terms of probability distributions of extreme events. The statistical comparison indicates that the observed time series are not simulated sufficiently by the models. Therefore, serious concerns are raised by the usefulness of climate models in hydrological applications." This is making a general conclusion (re: "usefulness of climate models in hydrological applications") from a very specific result (performance in simulating extremes). "Hydrological applications" can mean anything from planetary-scale distribution of moisture, continental-scale studies of drought through to small urban drainage design.

Line 30: "In particular, it has become a common perception that [GCM] output constitute credible guides for future climate conditions, and, therefore, they can be used by the hydrological water management community to assess climate impacts." Please find reference to demonstrate the perception is "common". Also, please specify which aspects of climate impacts can be assessed using GCMs - my perception is they should not be used for catchment scale rainfall or extreme rainfall if the interest is in extremes.

Line 38: "However, the credibility of GCMs in reproducing the observed climatic conditions has been questioned." Please find references for this.

Line 52: "According to commonly invoked projections, during the 21st century extreme events will be more frequent, more intense and with greater duration." Please provide primary references for this (the IPCC just provides a synthesis). I have not seen projections that say extremes will be getting longer (in general) - rather the opposite. Also, I'm not sure whether extremes will become more frequent - only that they will be more intense when they occur. This of course depends on definitions (e.g. whether 'extremes' are defined above a threshold).

Line 63: "It should be noted, though, that globally the IPCC projections for more frequent and more intense extreme events have not been proven as yet." Although there is never such a thing as a complete 'proof' in this science, recent references by [Min et al., 2011; Westra et al., 2013] show reasonable evidence that the intensity and/or frequency of extremes is increasing.

Line 127: Define 'k'.

Line 131: Please specify length/dimension of zi.

Line 142: ".we can assume that zi are independent to each other." Please test this.

Line 173: ".decent fitting can be achieved.." Please define "decent".

Line 194: "With regard to the specific study area, it should be noted that extreme storm and flood events in the Mediterranean area usually occur during the winter period, driven mostly by frontal events (large scale precipitation). On the contrary, summer floods, which are caused by local convective precipitation vents, are rare in the study area." Please provide reference that shows the meteorological mechanisms causing floods in this area.

Line 214: "With reference to downscaling, while it has been a very common technique, we believe that its use has never been justified." What sort of downscaling? Statistical or dynamical, or both? If including dynamical, what resolution? There are increasing modelling efforts at the ~2km resolution, which is the minimum resolution needed to start resolving convective processes - are they part of the criticism? How about idealised models [Muller et al., 2011; Romps, 2011] that try to resolve specific processes?

Line 223: "In the same spirit, Xu and Yang (2012) showed that the traditional dynamic downscaling is not adequate to simulate reality." This is clearly a case of cherry picking the literature. There are plenty of papers that show dynamical downscaling can simulate certain processes quite well.

Line 387: "The key issue of our study is that climate models are not ready for 'prime time' in water resources management." Do the authors accept that 'stationarity is dead' [Milly et al., 2008]? If so, presumably the assumption that historical risk of extremes continues into the future is just as fraught as using GCMs. The point is that papers generally do not recommend using climate models (as discussed, usually downscaled rather than raw GCM output) because they are 'good', but rather because they are the 'least bad' option - particularly given the mounting evidence that extremes are already changing. The authors need to put their results in the proper context of alternative options.

Line 408: "Instead of trying to make accurate climate predictions, we believe that climate policies should emphasise on the sustainability and more specifically on societal and environmental resources affecting humans, ecosystems, water, energy and nutrition." This conclusion/recommendation is vague at best. On line 68, the authors say the purpose of the study is to "investigate whether it is justified to use GCMs as risk assessment tools." I understand that the authors are saying that GCMs are not suitable as risk assessment tools, but are the authors also saying that we should move away from a risk-based decision making framework to a more scenario-based framework? It is not clear to me from this sentence what the authors are recommending.

References

Berg, P., C. Moseley, and J. O. Haerter (2013), Strong increase in convective precipitation in response to higher temperatures, Nature Geoscience.

Fowler, H. J., S. Blenkinsop, and C. Tebaldi (2007), Linking climate change modelling to impact studies: Recent advances in downscaling techniques for hydrological modelling, International Journal of Climatology, 27, 1547-1578.

Lenderink, G., and E. van Meijgaard (2008), Increase in hourly precipitation extremes beyond expectations from temperature changes, Nature Geoscience, 1, 511-514.

Maraun, D., et al. (2010), Precipitation downscaling under climate change: recent developments to bridge the gap between dynamical models and the end user, Reviews of Geophysics, 48(RG3003).

McGuffie, K., and A. Henderson-Sellers (2005), A climate modelling primer.

Milly, P. C. D., J. Betancourt, M. Falkenmark, R. M. Hirsch, W. Zbigniew, Z. W. Kundzewicz, D. P. Lettenmaier, and R. J. Stouffer (2008), Stationarity is Dead: Whither Water Management?, Science, 319, 573-574.
Min, S. K., X. Zhang, F. Zwiers, and G. C. Hegerl (2011), Human contribution to more-intense precipitation extemes Nature, 470, 378-381.
Muller, C. J., P. A. O'Gorman, and L. E. Back (2011), Intensification of Precipitation Extremes with Warming in a Cloud Resolving Model, Journal of Climate, 24, 2784-2800.
Romps, D. M. (2011), Response of Tropical Precipitation to Global Warming, Journal of the Atmospheric Sciences, 68, 123-138.
Seneviratne, S. I., et al. (2012), Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation, in A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change (IPCC), edited by C. B. Field, et al., pp. 109-230, Cambridge, UK.
Trenberth, K. E., A. Dai, R. M. Rasmussen, and D. B. Parsons (2003), The changing character of precipitation, Bulletin of the American Meteorological Society, 84, 1205-1217.
Westra, S., L. V. Alexander, and F. W. Zwiers (2013), Global increasing trends in annual maximum daily precipitation Journal of Climate, in press.

Reply by the last author

Subject: A personal comment on a JoH paper rejection

Date: Sat, 13 Jul 2013 11:47:05 +0300

From: Demetris Koutsoyiannis <dk@itia.ntua.gr>

To: a.sharma@xxx.xxx

CC: J Hydrology - Georgakakos, Dimosthenis Tsaknias, Dimitris Bouziotas

Dear Ashish,

(cc. Editor Kosta Georgakakos, Co-authors Dimosthenis Tsaknias & Demetris Bouziotas)

Dimosthenis Tsaknias, my coauthor in the paper “Statistical comparison of observed temperature and rainfall extremes with climate model outputs in the Mediterranean region”, forwarded to me the letter of rejection and the reviews for that paper (second review round). This is my personal reaction for the decision, not discussed or agreed with my coauthors. Note, I am not challenging the decision--just I wish to communicate my view.

I appreciate the time and effort you put on our paper and of course I understand that a rejection based on two negative reviews is justifiable. However, I would appreciate it more if either you had rejected outright the paper at the first submission or you followed an approach consistent for both submissions.

Of course I understand that there were difficulties for an outright rejection in the first round, given that one review looked very positive. By choosing the most critical of the three original referees and inviting a new one, I guess with a predictable verdict, things became simpler in the second round.

But this was not the only inconsistency between the first and the second round. In the second round you concluded that the paper:

> does not by itself present any innovation

Unless I did something wrong in my text searches, in the first-round reviews and your own assessment, the words “innovative”, “innovation”, “original”, “originality”, “new”, “novel” do not appear at all to indicate lack thereof, and the word “novelty” appears only once in your own assessment (not once in the reviews) in the phrase:

> They also need to rethink what exactly is the focus and novelty of the paper

So, if I understand correctly, you did not raise an issue of lack of originality, innovation and novelty in the first round—although you prompted us to rethink it (I guess with the meaning to better present the focus and the novelty).

Also, Reviewer #1 in the second review round writes:

> As mentioned in my previous review, this type of study is not new and the authors do not make additional attempts to justify its publication

but does not refer us to other instances of “this type of study”. In the first review round she (seems to be Reviewer #2 of the first round, right?) wrote:

> The deficiencies of climate models, which are addressed in this paper, have been studied in the past

but in my view this is different from charging the study with lack of novelty. Most interestingly, a few lines below she wrote:

> The literature on these problems is rapidly growing and models are still far from sufficient as far as proper precipitation simulation is concerned

Isn’t this an amazing contrast? Why is the literature rapidly growing if “the deficiencies … have been studied in the past” and “this type of study is not new”? Isn’t this a selective approach about what the “rapidly growing” literature should include?

With respect to the Reviewer #2, her impartial approach is evident from a single comment:

> Line 38: "However, the credibility of GCMs in reproducing the observed climatic conditions has been questioned." Please find references for this.

Had she read the entire paragraph that starts with this sentence, she would see not only the references but also summaries of what these references say. Overall, the feeling I got reading (partly at the moment) her review, is that this reviewer and I live in different planets. For example, she says:

> The authors are adopting a 'straw man' argument in which they imply that it is standard (or even 'best') practice to use GCM outputs of extremes as a basis for making water management decisions, when this is not the case.

In the planet I live, GCMs are used as a basis, on which basis other procedures such as downscaling are based, and myriads of impact studies on this basis have been produced and myriads of papers on this basis have been published.

This brings me to the last point I want to mention, which refers to the following phrase of your own assessment:

> that the BLUE approach presented here, by itself, does not offer a solution to the problems the authors have demonstrated exist

I really regret that, as evident from your statement, we were not able to make clear our intention and scope. For our intention was not to present any solution to the problems, which we indeed have demonstrated to exist. We introduced the non-linear conversion of the BLUE method not to solve the problems but to demonstrate how big they are, even in the most favourable (for GCM produced fields) case of interpolation, which we believe our proposed method is. Our scope is that, even using the most favourable interpolation, severe problems exist. Now, I generally believe if an approach does not give satisfactory results, then we do not try to rectify (call it downscale) the results obtained by this approach but we change the approach and the direction. Of course this paper does not describe which the new direction should be (this is the subject of other papers), but it underlines the need for another approach.

Concluding, I thank you for the lesson you provided to my youngster coauthors Dimosthenis and Demetris. They are very smart and I am sure that they understood that innovation and novelty, which produce publishable papers, are not to challenge existing approaches but to use them adding innovative copies suitably resembling the myriads of studies I mentioned above.

Sincerely,

Demetris

--
Demetris Koutsoyiannis
dk@ntua.gr - http://itia.ntua.gr/dk

Hydrology Research

2013/10/10 Ian G. Littlewood

Dear Mr Tsaknias,

I am sorry to tell you that your submission "Statistical comparison of observed temperature and rainfall extremes with climate model outputs in the Mediterranean region" has not been accepted for publication.

Thank you for submitting this manuscript with its history. I hope the following comments (and uploaded annotated file) will help you decide how to proceed.

I have sympathy with the some of the points you make about the use of current GCMs for hydrological aspects of water resource management. However, having looked at the Reviewers' comments arising from submission of this work to another journal, and having looked through the manuscript myself (fairly quickly, and I am not an expert in this area), I think there are problems of novelty and style (including its structure) for it to be a research paper in this journal. I know it will be disappointing for you to hear it but I would suggest that you re-think completely the way you present your work (your revision clearly did not work for the Reviewers). For example, in my opinion it would be better if you re-structured completely. I found too much 'discussion' text in the Introduction and interspersed with your specific research (description, results). Your 'Conclusions' largely comprise discussion of others' work (which should probably have been given in the Introduction). A section headed 'Conclusions' should not contain discussion. Perhaps you need 'Discussion' and 'Concluding remarks' sections? The written English needs improving in several places. I would encourage you to not use "we", "our", etc, and to use the third person narrative style instead.

Thank you for giving us the opportunity to consider your work. I hope that you will consider submitting to Hydrology Research in the future.

With Best Wishes

Ian G. Littlewood
Editor
Hydrology Research
IWA Publishing

Advances in Meteorology (submission 2015/12/31)

Subject: 7167280: Decision Finalized

Date: Wed, 24 Feb 2016 08:38:38 +0200

From: Amr Zedan

To: Demetris Koutsoyiannis

CC: Dimosthenis Tsaknias, Dimitris Bouziotas

Dear Dr. Koutsoyiannis,

Following the review of Research Article 7167280 titled "Statistical Comparison of Observed Temperature and Rainfall Extremes with Climate Model Outputs in the Mediterranean Region" by Dimosthenis Tsaknias, Dimitris Bouziotas and Demetris Koutsoyiannis, I regret to inform you that it was found to be unsuitable for publication in Advances in Meteorology. You may log in to the Manuscript Tracking System in order to read the review report(s) received for your manuscript.

Thank you for submitting your manuscript to Advances in Meteorology.

Best regards,

Amr

Amr Zedan
Editorial Office
Hindawi Publishing Corporation
http://www.hindawi.com

Review of "Statistical comparison of observed temperature and rainfall extremes with climate model outputs in the Mediterranean region" by Tsaknias, Bouziotas, and Koutsoyiannis.

The goal of this paper is to assess the ability of recent climate models to simulate extreme temperature and precipitation events using various statistical tools such as extreme value analysis. Such an assessment of model performance is important since one of the expected changes that will occur as a result of projected climate warming is an increase in the occurrence of extreme weather events. The authors approach this question by comparing the statistical properties of output of 20th century climate simulations (20C3M) from three different coupled global atmosphere-ocean models (ECHAM5, CSIRO, and CGCM3-T63) with meteorological records from selected stations (20 are listed but the main focus is on only two stations) around the Mediterranean basin. The Mediterranean region, which is the transition zone between the great deserts and the temperate mid-latitudes, is considered to be one of the regions that is highly vulnerable to the impacts of climate change. The main research question is indeed important, especially if one wishes to use the output of future climate model projections for downstream applications such as hydrological risk assessment. Their main conclusion is that the models fail to adequately reconstruct the variability and extreme statistics of the long term precipitation record at specific points (stations) while the temperature reconstructions exhibit some useful skill. This conclusion could have been anticipated from the outset even before conducting the analysis presented in the paper. The main shortcoming in their approach is that they focus exclusively on the statistical analysis while almost completely neglecting the meteorological processes, insight, and interpretation. Their expectation that a relatively coarse resolution model (grid spacing on the order of 200 km or more) should be able to faithfully reproduce the rainfall statistics of individual point measurements, especially over a geographically diverse region such as the Mediterranean, demonstrates a fundamental misunderstanding of the highly localized meteorological processes as well as the structure and capabilities of such models. Precipitation is very patchy and dependent upon smaller scale processes in contrast to temperature which is generally much smoother spatially.

One third of the text, 6 out of 18 pages, is devoted to a detailed description and justification of the rather complex method of interpolation of the gridded model output to the geographical locations of the specific stations at which historical data are available (Sections 4 and 5). Three pages are devoted to the results for the rainfall analysis while the temperature results only warrant slightly more than one half of a page. This seems rather disproportionate. On the one hand they appear to understand that conceptually, the values at a model grid point represent a spatial average of the fields over the relevant grid box and yet they seem to skip over the fact that this inherently means that the model grid values will be relatively smooth. Spatial interpolation of a smooth field, no matter how complex or precise the interpolation scheme, cannot reconstruct or restore the subgrid scale variability that is crucial here, especially for a patchy field such as precipitation which is strongly influenced by smaller scale, local effects such as land-sea contrast or topography. They provide a glaring example of this problem in Figure 2 where they show an example of the four model grid points surrounding Torrevieja used for the interpolation. Three of these are sea points while only one is a land point. Consequently the spatially interpolated model time series can be expected to be more similar to rainfall over the sea than over land, even the coastal region. But there is no discussion of the potential importance of these factors. The models are potentially skillful at reproducing the (smoothed) precipitation patterns over a region covering many grid points, but they are doomed to failure if one tries to extract individual point values.
On pages 10 and 11 the authors try to convince the reader that the alternative of dynamical downscaling using a high resolution regional climate model nested into a global model is not a sound approach if the parent model is inherently flawed. However the purpose of using a regional model is to overcome, at least partially, the shortcomings of the coarser resolution global model by accounting for the smaller scale physical and dynamical processes. Conceptually this is much more appealing that spatial interpolation alone which contains no

knowledge of the subgrid scale variability. Of course it is possible to design an interpolation scheme in which the weights contain some information about the smaller scale local processes or subgrid scale variability, similar to a MOS approach which is used in post processing model output for daily weather forecasting. But this of course would require determining appropriate weights for each location of interest. As an alternative, a much fairer assessment of the models would be to upscale the precipitation data to the model grid or compare the model results to a gridded historical precipitation data set such as CMAP, GPCC, or GPCP.

First of all, we thank the editor for his constructive comments which give us the opportunity to clarify some of our comments.

We agree that linear operations do perform smoothing. However, our equation (5) is non-linear just for the sake of avoiding smoothing. The reviewer may experiment with the equation to see that for r = 4, which we use, the outcome is close to the maximum value of the adjacent grid points. For increasing r tending to infinity, the outcome becomes exactly equal to that maximum in the adjacent points. This is a strongly non-linear operation that does not perform smoothing at all.

Regarding the comment that one should have anticipated the results even before performing any type of analysis, we totally agree, but at the same time we feel that these statements are addressed to the wrong recipient. Studies based on GCM outputs are used more and more frequently by policymakers, as there is a common perception that climate models can accurately simulate the climate in the future. As mentioned in out introductory section such studies are not used by just general studies, but can also drive local and site-specific decisions.

In addition to the general problems discussed above there are several specific points that should be addressed. Some examples are as follows:

1. In text on lines 94-95 and Table 1 they refer to data from the SRES A2 scenario simulations but these data are never used in the analysis. The authors even mention several times that their goal is to assess current climate simulations, not future projections. This should therefore be removed.

The data for the SRES A2 scenario (representing the 21st century) were available for the climate model outputs of the CSIRO and CGCM-T3 models. We are not sure why the editor believes we did not use these data, but we did. Where available, we used the observations up to 2010 and compared them with the corresponding years from the climate model outputs.

2. It would be very helpful to add a forth model to the analysis – a regional climate model for the Mediterranean region. Several are available from various projects such as CIRCE.

The focus of our analysis is the GCMs performance. We believe that if the GCMs are not good enough, then RCMs will be flawed as well, since thee downscaling procedure is originated by the GCMs outputs. In addition, as repeated in our manuscript Dawson *et al*. (2012) deduced that the credibility of climate projections is questionable, since the boundary conditions provided to RCMs through downscaling may err systematically. Similarly, Xu & Yang, (2012) showed that the traditional dynamic downscaling is not adequate to simulate reality, as precipitation is overestimated, while extreme temperatures are overestimated.

3. The results for the precipitation analysis focus almost exclusively on two stations only – Torrevieja and Perpignan out of the list of 20. Both of these stations are on the northern coast of the western Mediterranean and probably reflect similar geographic conditions. It is not clear why these two were selected. It would certainly be more convincing if several other stations in the central and eastern Mediterranean were also analyzed in detail.

Indeed, we chose to show the empirical distributions for only two stations in this publication, as they are representative of all the stations. This is also shown in the figures with the climate indices, where all the stations are shown. The readers can request for any stations of interest and we will share the corresponding figures.

4. In the discussion of Figure 5 (lines 294-305) they describe the changes in the GEV distribution between three consecutive 33 year periods. On lines 299-301 they refer to the historical 100 mm rainfall depth having a return period of 5 years in the upper graph and 2 years in the middle graph. According to my count it is 3 years in the middle graph (and 4 years in the third graph). They also mention a change in the most extreme rainfall increasing from under 200 mm in the first and second graphs to over 200 mm in the third graph. As statisticians they should know that the most extreme values at the tail of the GEV or any other extreme value distribution curve are highly unreliable and it is probably better not to mention them.

We agree that the return period of 100 mm/day in Perpignan is 3 years and not 2, as this was a typo. Regarding the most extreme precipitation values, we agree that one should be cautious when they perform exploratory data analysis at the tail of an extreme value distribution. However, our comment was not related to the absolute rainfall values in Perpignan; on the contrary our focus was the change of extreme values over time and the trends associated with them. One of the main reasons why climate models are extensively used is their ability to detect trends and climate signals; therefore we would expect that any change of extreme values over the years should be picked up by GCMs. Yet, our analysis suggests the CSIRO and ECHAM models do not capture the trends for all the three periods, while ECHAM captures the increasing trend from 1901-1933 to 1934-1966, but not from 1934-1966 to 1967-1999.