Hydrological Sciences Journal
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t911751996

A comparison of local and aggregated climate model outputs with observed data

G. G. Anagnostopoulos[a]; D. Koutsoyiannis[a]; A. Christofides[a]; A. Efstratiadis[a]; N. Mamassis[a]
[a] Department of Water Resources, Faculty of Civil Engineering, National Technical University of Athens, Heroon Polytechneiou 5, GR 157 80, Zographou, Greece

Online publication date: 13 October 2010

To cite this Article Anagnostopoulos, G. G. , Koutsoyiannis, D. , Christofides, A. , Efstratiadis, A. and Mamassis, N.(2010) 'A comparison of local and aggregated climate model outputs with observed data', Hydrological Sciences Journal, 55: 7, 1094 — 1110

To link to this Article: DOI: 10.1080/02626667.2010.513518

URL: http://dx.doi.org/10.1080/02626667.2010.513518

A comparison of local and aggregated climate model outputs with observed data

G. G. Anagnostopoulos, D. Koutsoyiannis, A. Christofides, A. Efstratiadis & N. Mamassis

Department of Water Resources, Faculty of Civil Engineering, National Technical University of Athens, Heroon Polytechneiou 5, GR 157 80 Zographou, Greece
a.christofides@itia.ntua.gr

Received 10 April 2009; accepted 10 May 2010; open for discussion until 1 April 2011

Citation 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.

Abstract We compare the output of various climate models to temperature and precipitation observations at 55 points around the globe. We also spatially aggregate model output and observations over the contiguous USA using data from 70 stations, and we perform comparison at several temporal scales, including a climatic (30-year) scale. Besides confirming the findings of a previous assessment study that model projections at point scale are poor, results show that the spatially integrated projections are also poor.

Key words climate models; general circulation models; climate change; Hurst-Kolmogorov climate

Comparaison de sorties locales et agrégées de modèles climatiques avec des données observées

Résumé Nous comparons les résultats de plusieurs modèles climatiques avec les observations de température et de précipitation en 55 points du globe. De plus, nous agrégeons spatialement les sorties de modèles et les observations couvrant les Etats-Unis d'Amérique à partir des données de 70 stations, et nous procédons à une comparaison à plusieurs échelles temporelles, y compris à l'échelle climatique (30 ans). Les résultats sont non seulement cohérents avec ceux d'une évaluation antérieure pour conclure que les projections par modélisation à l'échelle ponctuelle sont pauvres, mais montrent aussi que les projections intégrées dans l'espace sont également pauvres.

Mots clefs modèles climatiques; modèles de circulation générale; changement climatique; climat de Hurst-Kolmogorov

INTRODUCTION

According to the Intergovernmental Panel on Climate Change (IPCC), global circulation models (GCM) are able to "reproduce features of the past climates and climate changes" (Randall *et al*., 2007, p. 601). Here we test whether this is indeed the case. We examine how well several model outputs fit measured temperature and rainfall in many stations around the globe. We also integrate measurements and model outputs over a large part of a continent, the contiguous USA (the USA excluding islands and Alaska), and examine the extent to which models can reproduce the past climate there. We will be referring to this as "comparison at a large scale".

This paper is a continuation and expansion of Koutsoyiannis *et al.* (2008). The differences are that (a) Koutsoyiannis *et al.* (2008) had tested only eight points, whereas here we test 55 points for each variable; (b) we examine more variables in addition to mean temperature and precipitation; and (c) we compare at a large scale in addition to point scale. The comparison methodology is presented in the next section.

While the study of Koutsoyiannis *et al.* (2008) was not challenged by any formal discussion papers, or any other peer-reviewed papers, criticism appeared in science blogs (e.g. Schmidt, 2008). Similar criticism has been received by two reviewers of the first draft of this paper, hereinafter referred to as critics. In both cases, it was only our methodology that was challenged and not our results. Therefore, after presenting the methodology below, we include a section "Justification of the methodology", in which we discuss all the critical comments, and explain why we disagree and why we think that our methodology is appropriate. Following that, we present the results and offer some concluding remarks.

ISSN 0262-6667 print/ISSN 2150-3435 online
© 2010 IAHS Press
doi: 10.1080/02626667.2010.513518
http://www.informaworld.com

METHODOLOGY AND DATA

Comparison at point basis

For the first part, that is, for comparison at point basis, we employed the same methodology as Koutsoyiannis *et al.* (2008). We compared at 55 points worldwide, selecting them with the following criteria: (a) distribution in all continents and in different types of climate; (b) availability of data on the Internet at a monthly time scale; and (c) existence of long data series (>100 years) without, or with very few, missing data. In total we used 84 stations, of which 29 for temperature, 29 for precipitation, and 26 for both. The stations are shown in Fig. 1 and are listed by Anagnostopoulos (2009, pp. 8–10). The data were downloaded from the web site of the Royal Netherlands Meteorological Institute (http://climexp.knmi.nl).

The next step was to retrieve a number of climatic model outputs for historical periods. We picked exactly the same outputs as Koutsoyiannis *et al.* (2008). The models used are shown in Table 1, which is reproduced from Koutsoyiannis *et al.* (2008). For TAR models, we used the runs for scenario SRES A2, except for ECHAM, for which we used IS92a. Since these runs are based on historical GCM input information prior to 1989, and extended using scenarios for 1990 and beyond, the choice of scenario is actually irrelevant for test periods up to 1989, whereas for later periods there is no significant difference between different scenarios for the same model. For AR4 models we

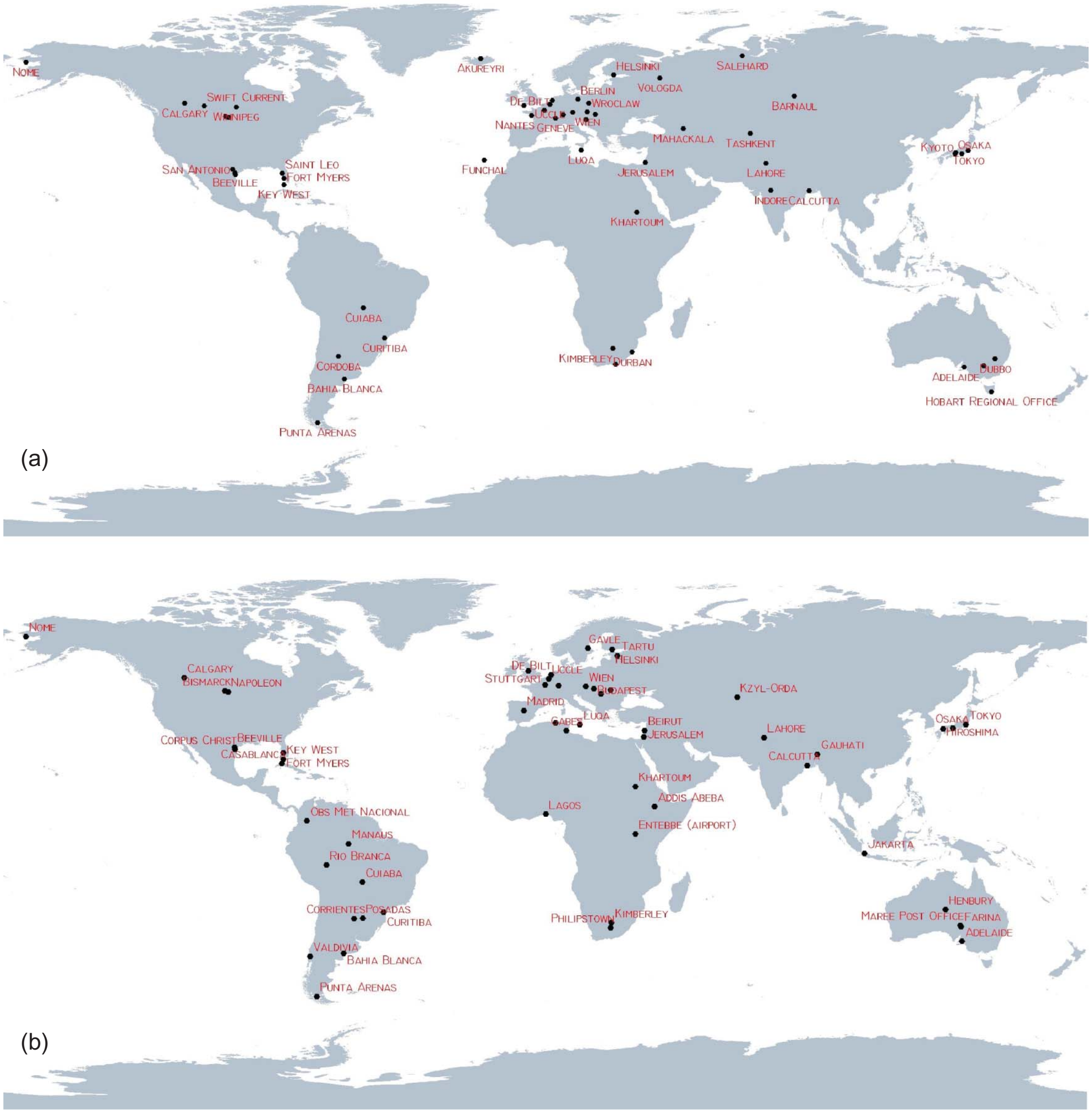

Fig. 1 Stations selected for (a) temperature and (b) precipitation.

Table 1 Models used in the study (reproduced from Koutsoyiannis *et al.*, 2008).

IPCC report	Name	Developed by	Resolution (°) in latitude and longitude	Grid points, latitude × longitude
TAR	ECHAM4/OPYC3	Max-Planck-Institute for Meteorology & Deutsches Klimarechenzentrum, Hamburg, Germany	2.8 × 2.8	64 × 128
TAR	CGCM2	Canadian Centre for Climate Modeling and Analysis	3.7 × 3.7	48 × 96
TAR	HADCM3	Hadley Centre for Climate Prediction and Research	2.5 × 3.7	73 × 96
AR4	CGCM3-T47	Canadian Centre for Climate (as above)	3.7 × 3.7	48 × 96
AR4	ECHAM5-OM	Max-Planck-Institute (as above)	1.9 × 1.9	96 × 192
AR4	PCM	National Centre for Atmospheric Research, USA	2.8 × 2.8	64 × 128

Sources: cera-http://www.dkrz.de/IPCC_DDC/IS92a/Max-Planck-Institut/echam4opyc3.html; Flato & Boer, 2001; Gordon *et al.*, 2000; www.mad.zmaw.de/IPCC_DDC/html/SRES_AR4/index.html.

used the runs for scenario 20C3M, which is the only AR4 scenario relevant to this study, as the other scenarios either concern the future (SRES, COMMIT) or are not supposed to represent historical reality (1%-2X, 1%-4X, PI-cntrl). More details on the choice of scenarios can be found in Koutsoyiannis *et al.* (2008).

To compare model outputs to historical time series, we used the best linear unbiased estimation (BLUE) technique in order to fit each historical time series to a linear combination of the GCM outputs at the four nearest grid points. Specifically, we optimize the weight coefficients λ_1, λ_2, λ_3, λ_4 (assuming positive values for physical consistency and $\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1$) in a linear relationship $\tilde{x} = \lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 + \lambda_4 x_4$, where $\tilde{x}$ is the best linear estimate of the historical value x (i.e. $\tilde{x} - x$ is the estimation error), and x_1, x_2, x_3, x_4 are the model outputs for the four nearest grid points. Optimization is done on the basis of the coefficient of efficiency, defined as Eff $= 1 - e^2/\sigma^2$, where e^2 is the mean square error in estimation and σ^2 is the variance of the historical series. The methodology is the same as in Koutsoyiannis *et al.* (2008), but here we examine more time series for each station; specifically, for temperature we examine: (a) annual average, (b) minimum monthly, (c) maximum monthly, (d) annual amplitude, (e) seasonal DJF (December-January-February), and (f) seasonal JJA (June-July-August); for precipitation, we examine: (a) total annual, (b) minimum monthly, (c) maximum monthly, (d) seasonal DJF, and (e) seasonal JJA. The comparison is made at three different time scales: monthly, annual and climatic (30-year moving average).

We used two main statistical indices (as in Koutsoyiannis *et al.*, 2008): the correlation coefficient (between $\tilde{x}$ and x) and the coefficient of efficiency. Several other statistical parameters were examined, depending on time scale. The average and the standard deviation were examined at all scales. At the annual time scale, the first-order autocorrelation coefficient and the Hurst coefficient were examined. The latter is a well-recognized metric of long-term fluctuations (also known as Hurst-Kolmogorov behaviour; cf. Kolmogorov, 1940; Hurst, 1951; Koutsoyiannis, 2010). The Hurst coefficient has a value of 0.50 for time-independent processes and 1.0 for fully dependent ones (Koutsoyiannis, 2003). Finally, at the climatic scale, the following metrics were additionally examined: (a) the change of 30-year moving average temperature or precipitation in the 20th century; (b) the change between the first and last values of each climatic time series; and (c) the maximum climatic fluctuation across the entire period. We calculate the change in moving average as the difference of 30-year moving averages centred at 1985 and 1915 (the 20th century is the common simulation period for all models except HadCM3 on scenario SRES A2); and we define the maximum fluctuation as the difference of maximum minus minimum observed or simulated climatic values, where a positive sign indicates that the minimum value precedes (in time) the maximum (positive trend), and a negative sign indicates the opposite.

Comparison at a large scale

We collected long time series of temperature and precipitation for 70 stations in the USA (five were also used in the comparison at the point basis). Again the data were downloaded from the web site of the Royal Netherlands Meteorological Institute (http://climexp.knmi.nl). The stations were selected so that they are geographically distributed throughout the contiguous USA. We selected this region because of the good coverage of data series satisfying the criteria discussed above.

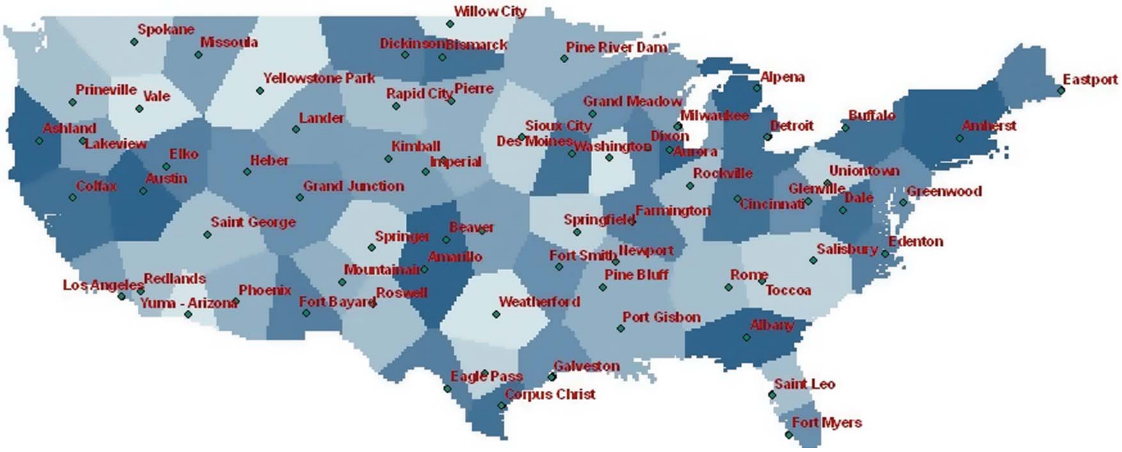

Fig. 2 Stations selected for areal integration and their contribution areas (Thiessen polygons).

The stations selected are shown in Fig. 2 and are listed by Anagnostopoulos (2009, pp. 12–13).

In order to produce an areal time series we used the method of Thiessen polygons (also known as Voronoi cells), which assigns weights to each point measurement that are proportional to the area of influence; the weights are the "Thiessen coefficients". The Thiessen polygons for the selected stations of the USA are shown in Fig. 2.

The annual average temperature of the contiguous USA was initially computed as the weighted average of the mean annual temperature at each station, using the station's Thiessen coefficient as weight. The weighted average elevation of the stations (computed by multiplying the elevation of each station with the Thiessen coefficient) is $H_m = 668.7$ m and the average elevation of the contiguous USA (computed as the weighted average of the elevation of each state, using the area of each state as weight) is $H = 746.8$ m. By plotting the average temperature of each station against elevation and fitting a straight line, we determined a temperature gradient $\theta = -0.0038$°C/m, which implies a correction of the annual average areal temperature $\theta(H - H_m) = -0.3$°C.

The annual average precipitation of the contiguous USA was calculated simply as the weighted sum of the total annual precipitation at each station, using the station's Thiessen coefficient as weight, without any other correction, since no significant correlation could be determined between elevation and precipitation for the specific time series examined.

We verified the resulting areal time series using data from other organizations. Two organizations provide areal data for the USA: the National Oceanic and Atmospheric Administration (NOAA) and the National Aeronautics and Space Administration (NASA). Both organizations have modified the original data by making several adjustments and using homogenization methods. The time series of the two organizations have noticeable differences, probably because they used different processing methods. The reason for calculating our own areal time series is that we wanted to avoid any comparisons with modified data. As shown in Fig. 3, the temperature time series we calculated with the method described above are almost identical to the time series of NOAA, whereas in precipitation there is an almost constant difference of 40 mm per year.

Determining the areal time series from the climate model outputs is straightforward: we simply computed a weighted average of the time series of the grid points situated within the geographical boundaries of the contiguous USA. The influence area of each grid point is a rectangle whose "vertical" (perpendicular to the equator) side is $(\varphi_2 - \varphi_1)/2$ and its "horizontal" side is proportional to $\cos\varphi$, where φ is the latitude of each grid point, and φ_2 and φ_1 are the latitudes of the adjacent "horizontal" grid lines. The weights used were thus $\cos\varphi(\varphi_2 - \varphi_1)$; where grid latitudes are evenly spaced, the weights are simply $\cos\varphi$.

JUSTIFICATION OF THE METHODOLOGY

Scale of comparison

The study of Koutsoyiannis *et al.* (2008) has been criticized (see Introduction) on the grounds that the comparison at point scale is meaningless. The critics also pointed out that the natural variability makes evaluations meaningless if year-to-year values are

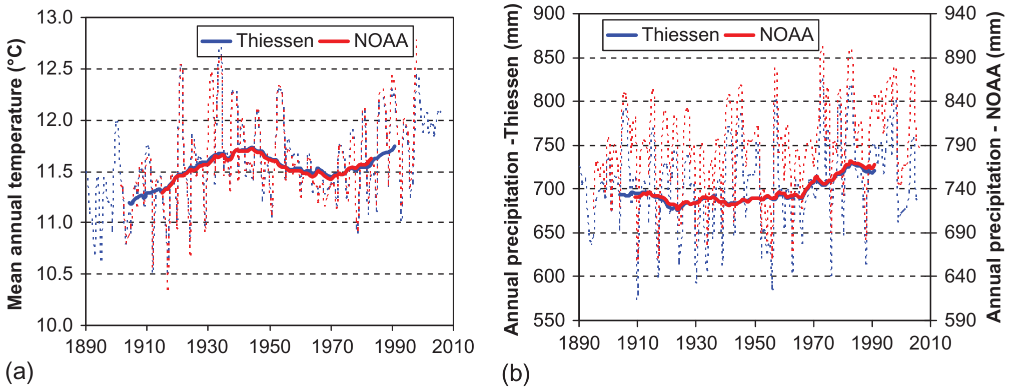

Fig. 3 Comparison between areal (over the USA) time series of NOAA (downloaded from http://www.ncdc.noaa.gov/oa/climate/research/cag3/cag3.html) and areal time series derived through the Thiessen method; for (a) mean annual temperature (adjusted for elevation), and (b) annual precipitation.

compared, especially since model runs are not initialized with real initial conditions. Both objections are essentially the same, and are related to scale, spatial and temporal.

We think that the criticism about temporal scale is evidently unjustified because one of our comparison time scales is the climatic, 30-year, scale. In order to address the objection concerning the spatial scale, we will first attempt to clear up some confusion in the literature. Von Storch *et al.* (1993) introduced the notion of the "skillful scale" of the GCM and mentioned that the skillful scale is "likely" at least eight grid point distances, but we cannot see in what way this conclusion can be inferred from Grotch & MacCracken (1991), which they cited. Today the notion of skillful scale is not being used any more, and, instead, Randall *et al.* (2007, p. 600) claim that "there is considerable confidence that climate models provide credible quantitative estimates of future climate change, particularly at continental scales and above." Despite being vague in several points (e.g. "considerable", "credible", "continental scales"), the statement seems to imply that GCMs are not very skillful at a single grid cell scale. However, Christensen *et al.* (2007, p. 852) mention that "providing information at finer scales [than the GCM computational grid] can be achieved through using high resolution in dynamical models or empirical statistical downscaling", and imply that GCMs are skillful even at single grid cells, as is also assumed by most literature on downscaling (e.g. Kotroni *et al*., 2008).

There are, therefore, two ways in which the spatial scale objection can be interpreted. The first alternative way is that GCMs are not skillful at single grid cells, and therefore assessing their performance at single grid cells is not meaningful; this criticism might follow from von Storch *et al.* (1993) and Randall *et al.* (2007), and it appears that, if true, it would automatically invalidate most literature on downscaling. The second alternative way is that, although GCMs are allegedly skillful even at single grid cells, it is not meaningful to compare a grid cell directly to a single point; this is the criticism by Schmidt (2008), and could also follow from Christensen *et al.* (2007).

Our comparison for the contiguous USA addresses both these arguments empirically. If the models produce "credible" results, "particularly at continental scales and above", this should be visible when comparing model outputs to reality at the scale of the contiguous USA, an area of 8 × 10^6 km^2 (similar to Australia, which is 7.7 × 10^6 km^2), and at the climatic scale, that is, on the 30-year moving average, which should lessen the effect of variability.

Except for this empirical addressing of the problem, the arguments can also be addressed theoretically. We will first deal with the first possible argument, that GCMs are not skillful at single grid cells whereas they are at larger scales. Climate models make local simulations, and the global value is derived from the local results. Can the global estimate be credible, therefore, if the local estimate, from which the global is derived, is not? The general belief is that the answer to this question is positive, but this should be justified (or negated) using probabilistic reasoning. As an example, in statistical thermodynamics, one can correctly estimate aggregate macroscopic quantities (such as pressure, temperature, etc. of a gas volume) even though microscopic quantities (such as position and momentum of each molecule) are incorrect or not known. The "correctness" of the macroscopic quantities has a proof based on concepts of probability theory, such as the law of large numbers, the central limit theorem or the principle of maximum entropy. Such a proof cannot be extended from a large thermodynamic system, say a mole of a gas, to climate, and a simple parallelism does not suffice. We recall that in a mole of a gas the macroscopic quantities are averages of a number, $N_A = 6.022 \times 10^{23}$, of molecules, and that independence between different molecules can be assumed (e.g. Stowe, 2007), so that after simple calculations using classical statistics we can see that the uncertainty at the single molecule level is multiplied by $1/\sqrt{N_A} = 1.3 \times 10^{-12}$ when macroscopic properties are calculated. This tiny number explains why we do not need an accurate representation at the molecular level to have the macroscopic quantity correct.

In contrast, spatial aggregation of climate model outputs on a continent is calculated as the average of, say, $n = 300$ grid point values and, apparently, n is much less than N_A. Moreover, the classical statistical inverse-square-root law is not valid because independence does not hold. Rather, a Hurst-Kolmogorov dependence is more plausible in hydrometeorological processes both in time and in space (Koutsoyiannis & Langousis, 2011). As shown by Koutsoyiannis & Montanari (2007) using temperature proxy records, under this dependence an averaging size of $n = 1000$ to 2000 is equivalent to the averaging of two to three statistically independent points, which does not give any hope for elimination of "local noise" when averaging model outputs even at a global scale.

Although in this analysis we used "local" and "global", which have a spatial meaning, the same argumentation applies to temporality if we substitute "short-term" and "long-term" for "local" and "global". Climate models perform their simulations at a high temporal resolution. If the annual estimates are not credible, then, for exactly the same reasons as those presented above, the 30-year moving average, which is derived from the annual, cannot be credible either. The implications of the Hurst-Kolmogorov dependence in predictability of climatic type (long-term average of future values *vs* single future values) has been elaborated in Koutsoyiannis (2010).

Finally, we will address theoretically the second possible version of the spatial scale objection, that it is not meaningful to compare point observations to the value of a grid cell. The argument is that the grid cell, which is of the order of 200 km × 200 km, provides the general climate of the area, whereas a point is affected by local factors. Daily temperatures may indeed differ significantly at a distance of 200 km, but the maximum temporal resolution we use is monthly; and departures from the mean in monthly temperature at points that close will be almost indentical. The same applies to precipitation at over-year scales (see also further demonstration and justification in Koutsoyiannis *et al*., 2008 – Figs 2 and 3 in particular). There could, however, be a systematic bias; for example, a point being consistently 1°C higher than a nearby one or than the grid-cell average. This bias may show in our comparison, since we make unbiased estimation ($\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1$), but this is not a problem, since we also use other metrics, such as the correlation coefficient, which ignores bias.

Comparison of actual values rather than departures from the mean

A common practice of climate modellers and the IPCC is to make comparisons in terms of departures from the mean (also called "anomalies"), rather than actual values. Some of the metrics we use, i.e. the correlation coefficients, are actually based on departures from means, i.e. they neglect bias. However, we think that actual values are important; as we will see, some model outputs have enormous differences from reality (up to 6°C in temperature and 300 mm in annual precipitation at the climatic time scale), which are not visible when differences from the mean are taken.

The GCMs calculate radiation balances, and, according to the Stefan-Boltzmann law, the amount of energy radiated is proportional to the fourth power of the absolute temperature. This means that the ratio of energy radiated at 15°C to the energy radiated at 12°C is $(273+15)^4/(273+12)^4 = 1.043$. For a difference of 0.5°C instead of 3°C, the ratio becomes 1.007. The question then arises on what grounds a model that errs by 3°C in the 20th century, that is, by 4.3% in radiative energy, could detect a future decadal trend of 0.7%, that is, six times lower. The same argument applies to precipitation and is even more important in hydrology: the actual value of over-year precipitation is an important factor determining the hydrological regime of a specific area (for instance an area with 400 mm of precipitation has a different hydrological regime from one receiving 600 mm of precipitation).

One of the earliest uses of departures from the mean was by Kim *et al.* (1984), who wrote that "the reason for considering the anomaly as an input rather than the mean value itself reflects our belief that while the long-term climate of a GCM may be different from reality, the GCM's anomaly with respect to its climate may be realistic." They did not further justify this belief. The National Climatic Data Center (NCDC, 2010) mentions that global mean temperatures are difficult to determine for two reasons: "Some regions have few temperature measurement stations (e.g., the Sahara Desert) and interpolation must be made over large, data-sparse regions. In mountainous areas, most observations come from the inhabited valleys, so the effect of elevation on a region's average temperature must be considered as well." They then argue that using departures from the mean addresses these two problems. However, for the first of these problems, that is, data scarcity in some areas, there is no explanation in what way using departures from the mean could help. In any case, we do not have this problem for the contiguous USA. The second problem can also be easily addressed by making proper adjustments, e.g. using the temperature gradient. In general, departures from the mean discard valuable information and ignore the facts described above about the Stefan-Boltzmann law as far as temperature is concerned, whereas in precipitation we also do not see any reason for using departures from means.

Alternative evaluation methods

The critics and Schmidt (2008) mentioned alternative evaluation methods which, in their opinion, might be better than our method. These methods are model inter-comparison, perturbed physics ensembles, and comparison of statistical indices of model outputs to statistical indices of measurements.

In model inter-comparison methods (e.g. Johnson & Sharma, 2009), all model outputs could inter-compare perfectly, but still they could all be wrong. Similarly, perturbed physics ensembles (e.g. Murphy *et al*., 2007) are a kind of sensitivity analysis, where the response of a given GCM to modifications of its parameters or inputs is investigated. Although we have no reason to doubt that model inter-comparison and perturbed ensembles could provide useful clues to modellers concerning the behaviour of the models, we do not see any way in which they could help assess whether model outputs can be a "credible" estimation of future climate.

The comparison of statistical indices of model outputs to those of measurements can also be useful, and for this reason our study includes such comparisons. However, we note that, whereas the failure of a model to capture certain statistical parameters can disprove the model, the opposite does not hold; if a model correctly reproduces statistical parameters, this does not necessarily mean that the model is a valid means of deterministic prediction of the future. For example, Koutsoyiannis (2006) presented a toy model that simulates hydroclimatic processes and correctly reproduces the standard deviation, skewness, and Hurst coefficient, and even reproduces the (known) past evolution, but this model clearly cannot (and is not intended to) provide a deterministic prediction of the future (see also Koutsoyiannis *et al*., 2007).

RESULTS

Comparison at point basis

The results of the point comparison confirm the findings of Koutsoyiannis *et al.* (2008). The conclusions here are safer because the sample is much larger (55 *vs* 8 stations), and the selection of time series examined is wider (Koutsoyiannis *et al*., 2008, examined only the annual mean temperature and annual precipitation).

The results vary depending on the time scale. At the monthly time scale the models generally reproduce the sequence of cold-warm and wet-dry periods at all stations examined. The average correlation coefficient (for all stations and all models) is 0.909 for the temperature and 0.256 for the precipitation. The average coefficient of efficiency is 0.721 for the temperature and –0.433 for the precipitation.

The statistics are dramatically different at the annual time scale. The average correlation coefficient drops remarkably and the average coefficient of efficiency is negative regardless of the time series examined (Table 2). Furthermore, the Hurst coefficient and the standard deviation are systematically underestimated (Table 3).

The results for the standard deviation are better than those for the Hurst coefficient (Figs 4 and 5). This is expected because the models can reproduce relatively well the seasonal variations in temperature and precipitation, and the variations in these variables with respect to latitude (Fig. 6). Moreover, the standard deviation varies with latitude (and also with the proximity of the station to the sea). The reproduction of dependence on latitude (and, in general, on the

Table 2 Average correlation coefficient and average coefficient of efficiency at the annual time scale for temperature and precipitation.

	Temperature		Precipitation	
	Average correlation coefficient	Average efficiency coefficient	Average correlation coefficient	Average efficiency coefficient
Annual mean/total	0.122	−5.157	0.003	−3.008
Max monthly	0.062	−5.254	0.007	−1.266
Min monthly	0.033	−3.748	0.004	−167.368
Annual amplitude	0.008	−4.068		
Seasonal DJF	0.051	−3.865	0.002	−3.750
Seasonal JJA	0.073	−7.495	−0.001	−12.168

Table 3 Percentage of the stations in which the Hurst coefficient and the standard deviation are underestimated.

	Temperature		Precipitation	
	Hurst	St Dev	Hurst	St Dev
Mean annual or total	74%	70%	79%	89%
Max monthly	72%	60%	67%	95%
Min monthly	66%	72%	68%	35%
Annual amplitude	69%	68%		
Seasonal DJF	69%	70%	67%	77%
Seasonal JJA	77%	58%	59%	84%

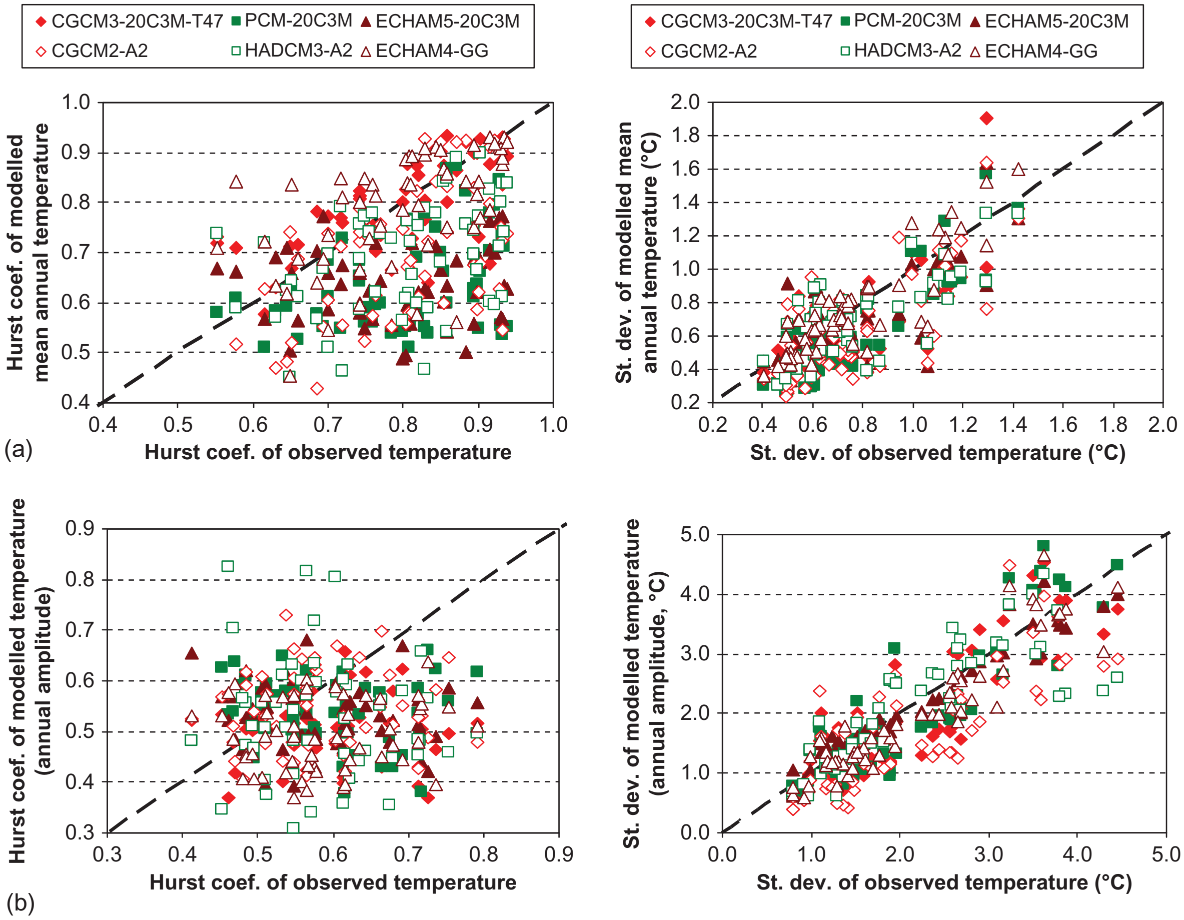

Fig. 4 Hurst coefficients (left) and standard deviations (right) of observed and modelled series for (a) mean annual temperature, and (b) annual temperature amplitude.

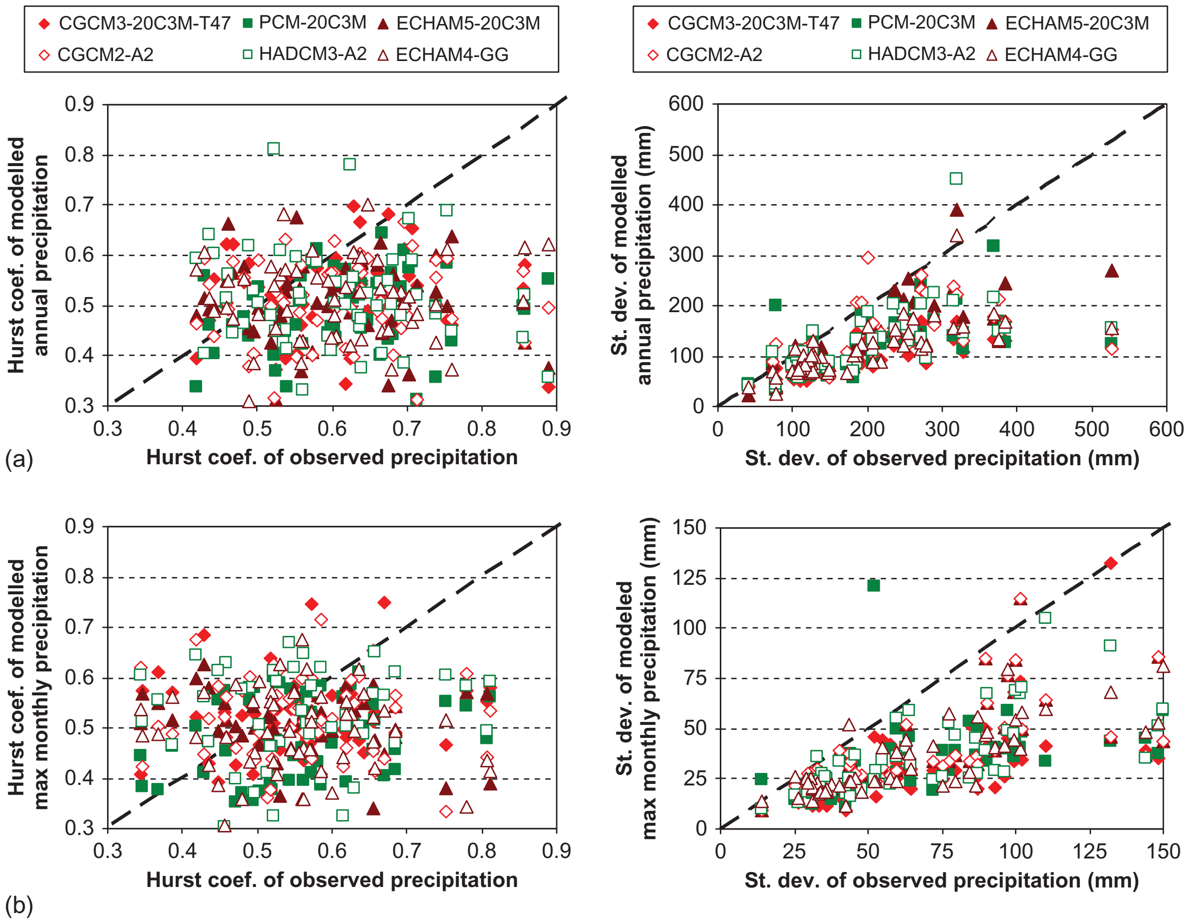

Fig. 5 Hurst coefficients (left) and standard deviations (right) of observed and modelled series for (a) annual precipitation, and (b) maximum monthly precipitation.

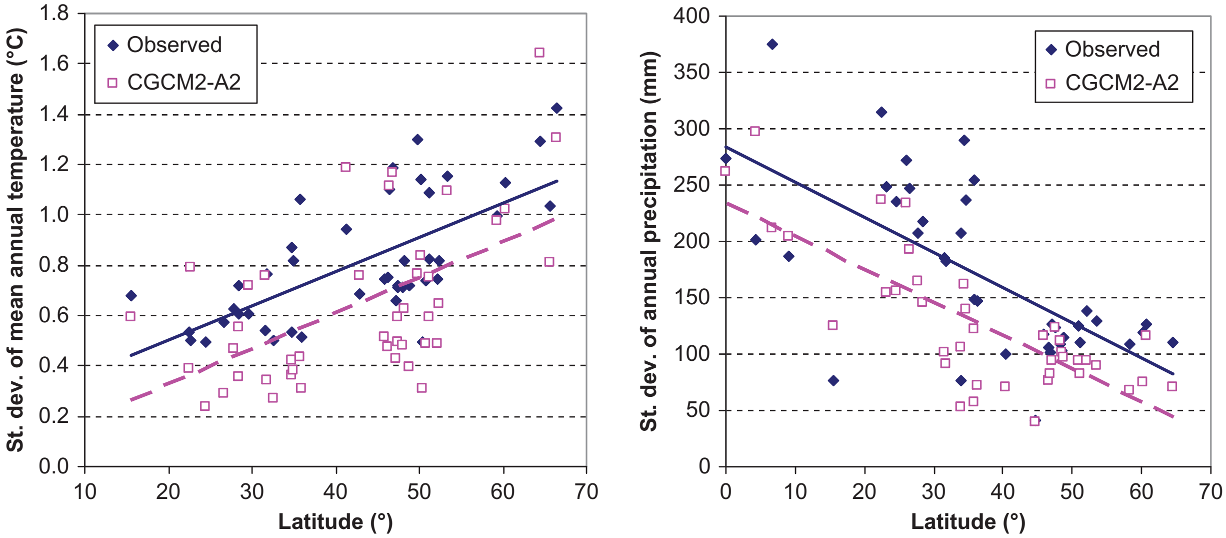

Fig. 6 Standard deviations of (a) mean annual temperature, and (b) annual precipitation, with respect to latitude (for the Northern Hemisphere).

location on the globe) is satisfactory, and since standard deviation depends on latitude, it is not surprising that it is also reproduced.

At the climatic (30-year moving average) time scale, the correlation coefficient increases slightly for temperature and precipitation in all the time series examined, in contrast to the coefficient of efficiency, which has, in most cases, large negative values (Table 4, Fig. 7). However, in a large number of stations, the correlation coefficient has low or even negative values for both temperature and precipitation. Therefore, large-scale fluctuations, and hence the Hurst-Kolmogorov behaviour, are not reproduced.

Results vary with stations. At the De Bilt station, The Netherlands (Fig. 8), the annual mean temperature is reproduced relatively well in the TAR model runs (but less so in AR4). While the models reproduce the annual mean temperature quite well, they do not reproduce the time series of the other variables examined (e.g. maximum and minimum monthly temperature, annual temperature amplitude). In addition, the relatively good fit is present only for the version of the time series which is reported to have had some homogeneity and other errors corrected; the model outputs do not fit the original time series (which is also shown in Fig. 8). At the Durban station, South Africa (Fig. 8), not a single model output shows the 1.5°C fall in mean annual temperature during 1920–1960; instead, all model outputs show a constant increase. At all stations examined, there is not a single model run that successfully reproduces the time series of all variables examined.

Model outputs are also in disagreement with actual data in the change in the 30-year moving average temperature and precipitation through the 20th century, as well as in the maximum fluctuation of these variables across the entire period of study (Fig. 9). In many cases, the model outputs show a temperature rise when the temperature actually falls. The differences in the direction of change are more marked in precipitation. An interesting example is the station of Valdivia. The observed change of the precipitation 30-year moving average during the 20th century is −853.3 mm and the maximum fluctuation is −877.4, whereas the results in the model outputs vary from −106.5 to 24.3 for the change and from −118.8 to 74.8 for the maximum fluctuation. Most outliers in the lower panel of Fig. 9 are because of Valdivia.

Comparison at a large scale

As in point basis, results vary depending on the time scale examined. At the monthly time scale the models reproduce the sequence of warm-cold and wet-dry periods. The highest (among the different models) correlation coefficient is 0.984 for temperature and 0.287 for precipitation and the highest coefficient of efficiency is 0.950 for temperature and −0.776 for precipitation.

At the annual time scale, the correlation coefficient has much lower values, sometimes near zero or even negative, in all time series examined, while the coefficient of efficiency is negative in all cases for both temperature and precipitation. The Hurst coefficient and the standard deviation are systematically underestimated in the majority of model outputs (Table 5, Fig. 10).

At the climatic (30-year) time scale, the correlation coefficient has slightly higher values than at the annual time scale, but the coefficient of efficiency has strongly negative values. In addition, model outputs disagree with observed values in the fluctuation of the 30-year moving average during the 20th century and

Table 4 Average correlation coefficient and average coefficient of efficiency at climatic (30-year) scale for temperature and precipitation.

	Temperature		Precipitation	
	Average correlation coefficient	Average efficiency coefficient	Average correlation coefficient	Average efficiency coefficient
Mean annual or total	0.328	−89.0	0.020	−125.9
Max monthly	0.207	−118.5	−0.024	−51.4
Min monthly	0.177	−117.4	0.006	−5456.7
Annual amplitude	0.027	−107.4		
Seasonal DJF	0.243	−92.0	0.053	−208.0
Seasonal JJA	0.208	−180.4	−0.041	−1064.1

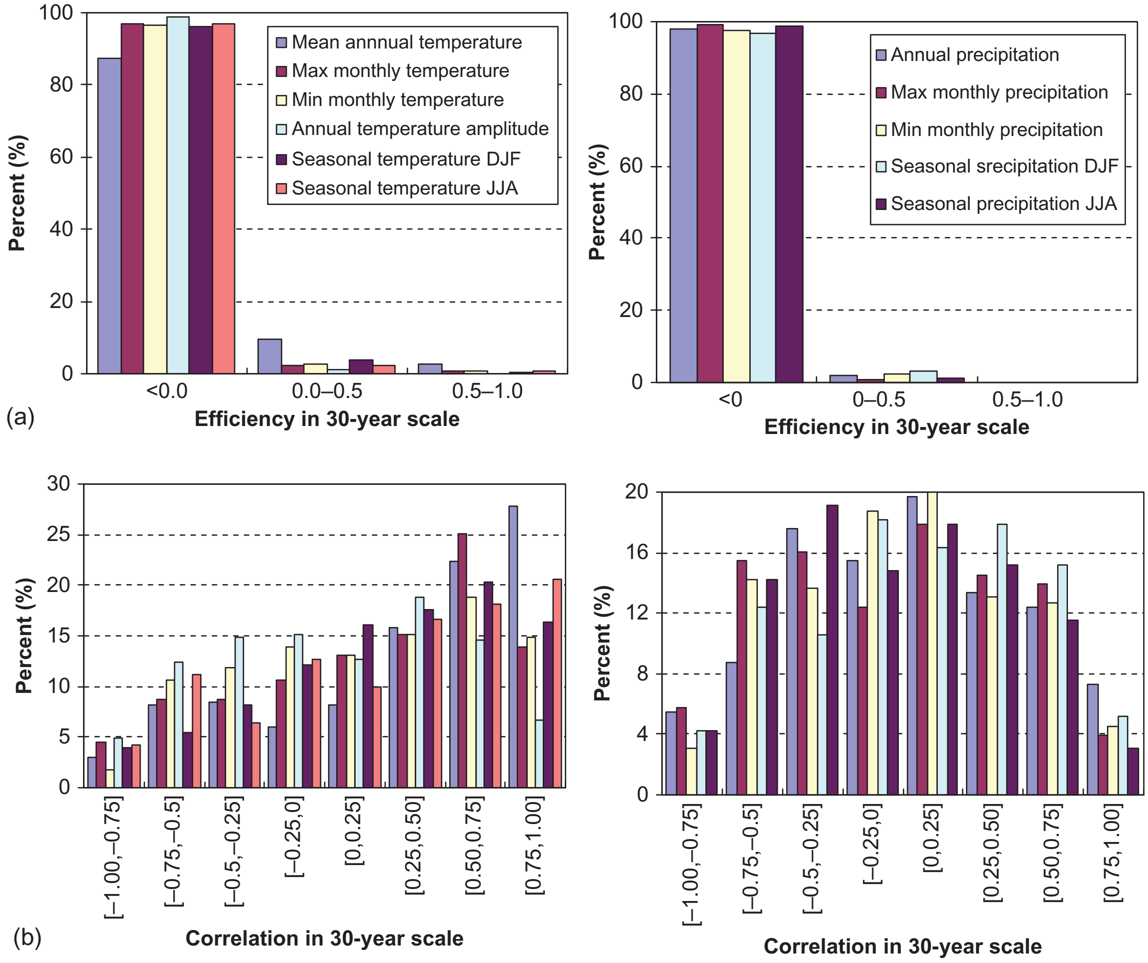

Fig. 7 Frequency distribution of (a) coefficient of efficiency and (b) correlation coefficient for temperature (left) and precipitation (right) at the climatic (30-year) scale.

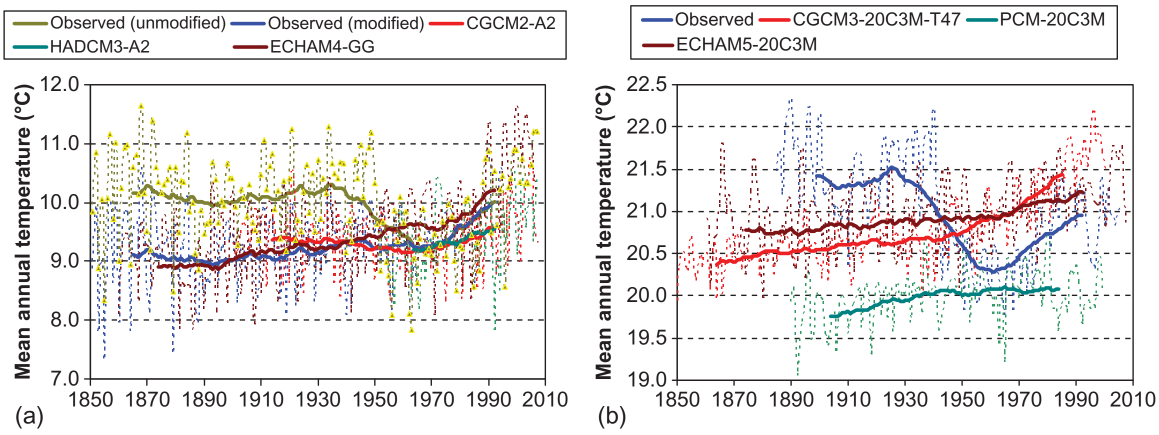

Fig. 8 Mean annual and 30-yearly temperature at (a) De Bilt and (b) Durban.

the maximum fluctuation during the period examined (Table 6, Fig. 11), and sometimes they indicate a rise when there was actually a drop, and *vice versa*.

In general, the results differ substantially from the observed time series (Fig. 12). The observed annual mean temperature of the USA gradually rose between

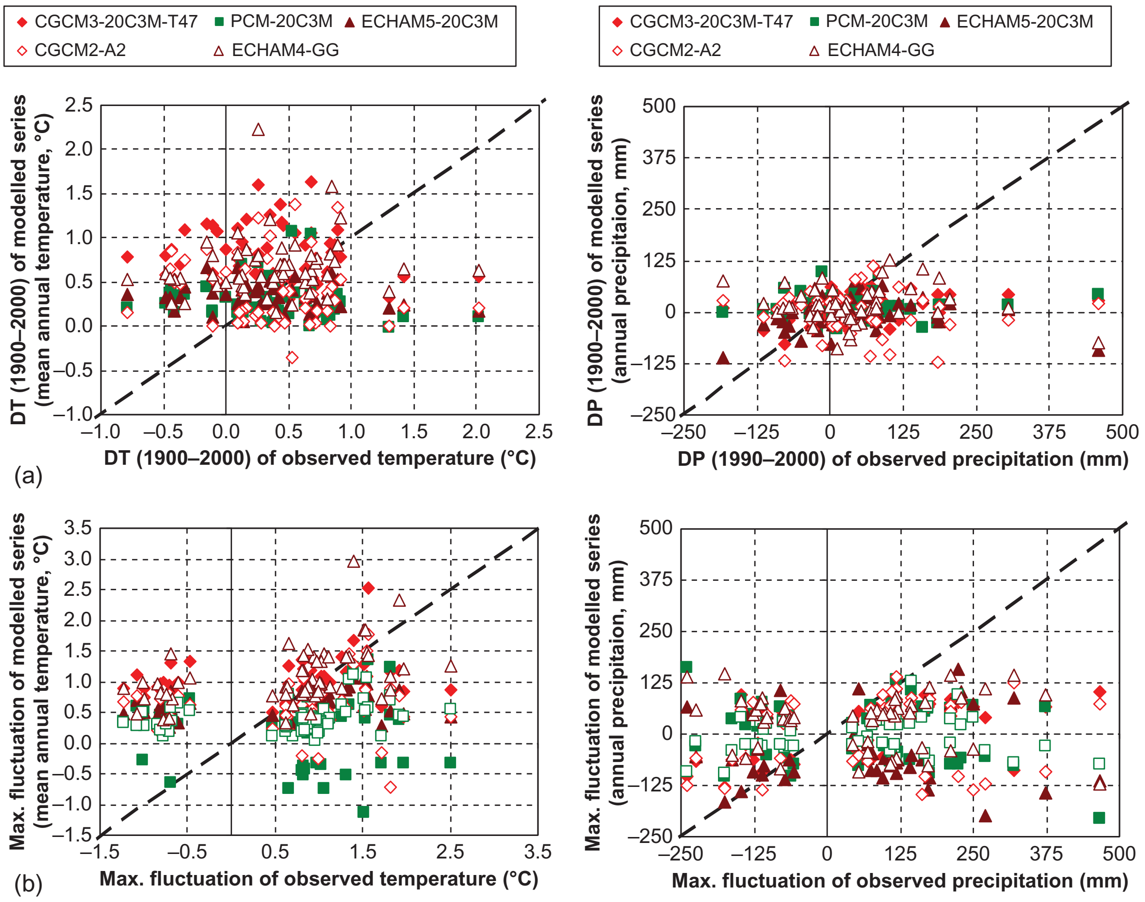

Fig. 9 (a) Change of 30-year moving average in the 20th century, and (b) maximum fluctuation across the entire period for temperature (left) and precipitation (right).

Table 5 Hurst coefficient and standard deviation of the observed areal time series.

	Temperature		Precipitation	
	Hurst	St Dev (°C)	Hurst	St Dev (mm)
Mean annual or total	0.765	0.442	0.628	52.175
Max monthly	0.762	0.621	0.420	8.897
Min monthly	0.515	1.450	0.469	7.252
Annual amplitude	0.555	1.519		
Seasonal DJF	0.662	0.990	0.412	20.666
Seasonal JJA	0.787	0.500	0.522	19.414

Table 6 Change of 30-year moving average in the 20th century (DT, DP) and maximum fluctuation (maxDT, maxDP) of the observed areal time series.

	Temperature (°C)		Precipitation (mm)	
	DT	maxDT	DP	maxDP
Mean annual or total	0.32	0.56	34.68	53.72
Max monthly	0.37	0.73	3.96	4.89
Min monthly	0.19	–0.81	1.34	3.70
Seasonal DJF	0.45	0.89	1.41	–8.75
Seasonal JJA	0.43	0.58	5.51	13.63

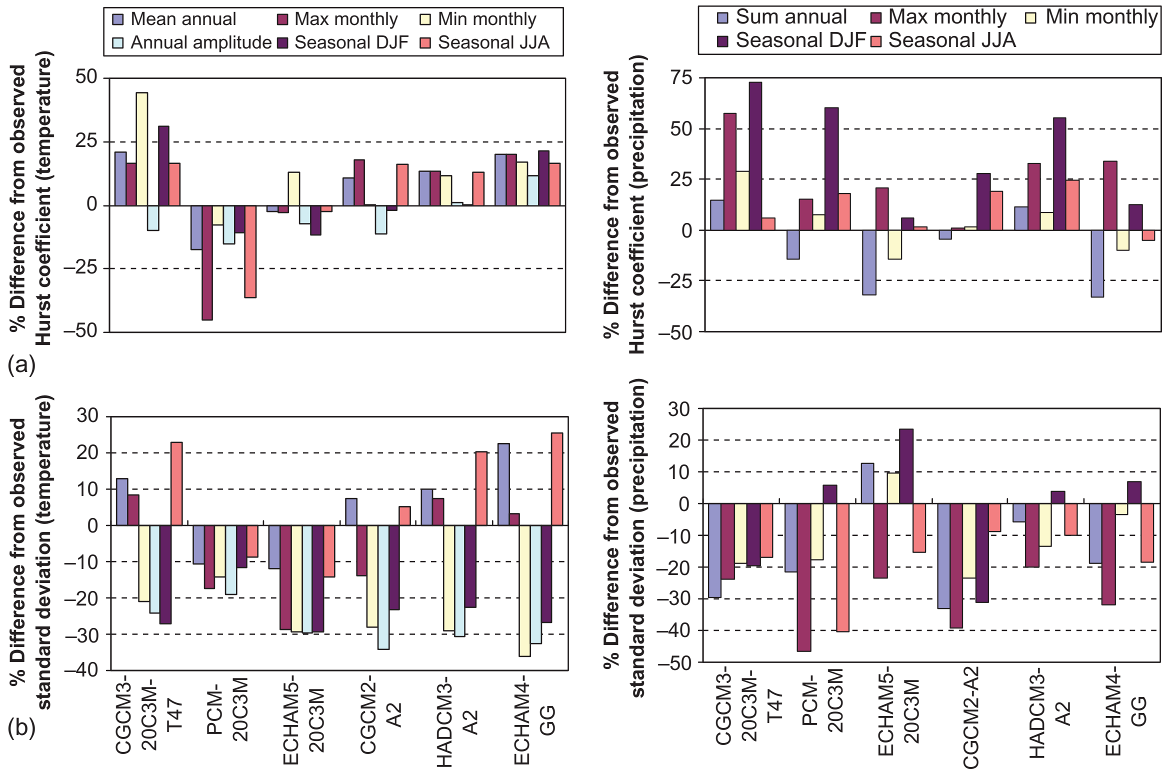

Fig. 10 Difference between (a) observed and modelled Hurst coefficient, and (b) standard deviation of the areal temperature (left) and precipitation (right) time series.

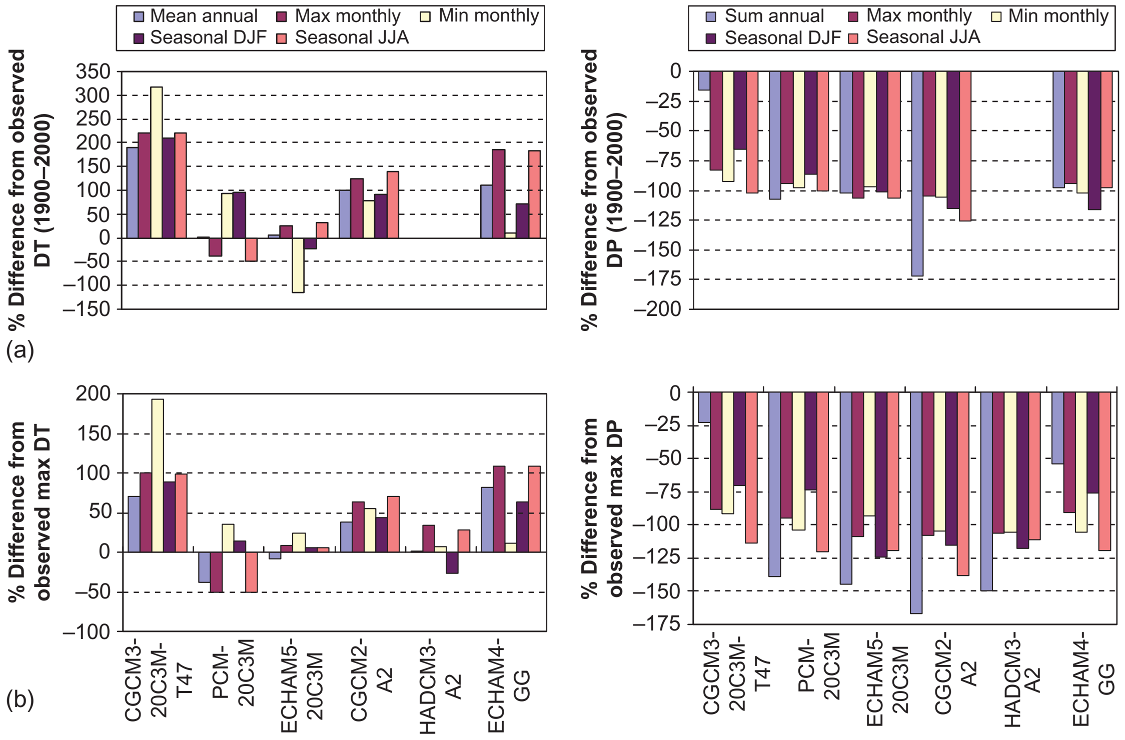

Fig. 11 (a) Difference between observed and modelled change of 30-year moving average in the 20th century, and (b) the maximum fluctuation, of the areal temperature (left) and precipitation (right) time series.

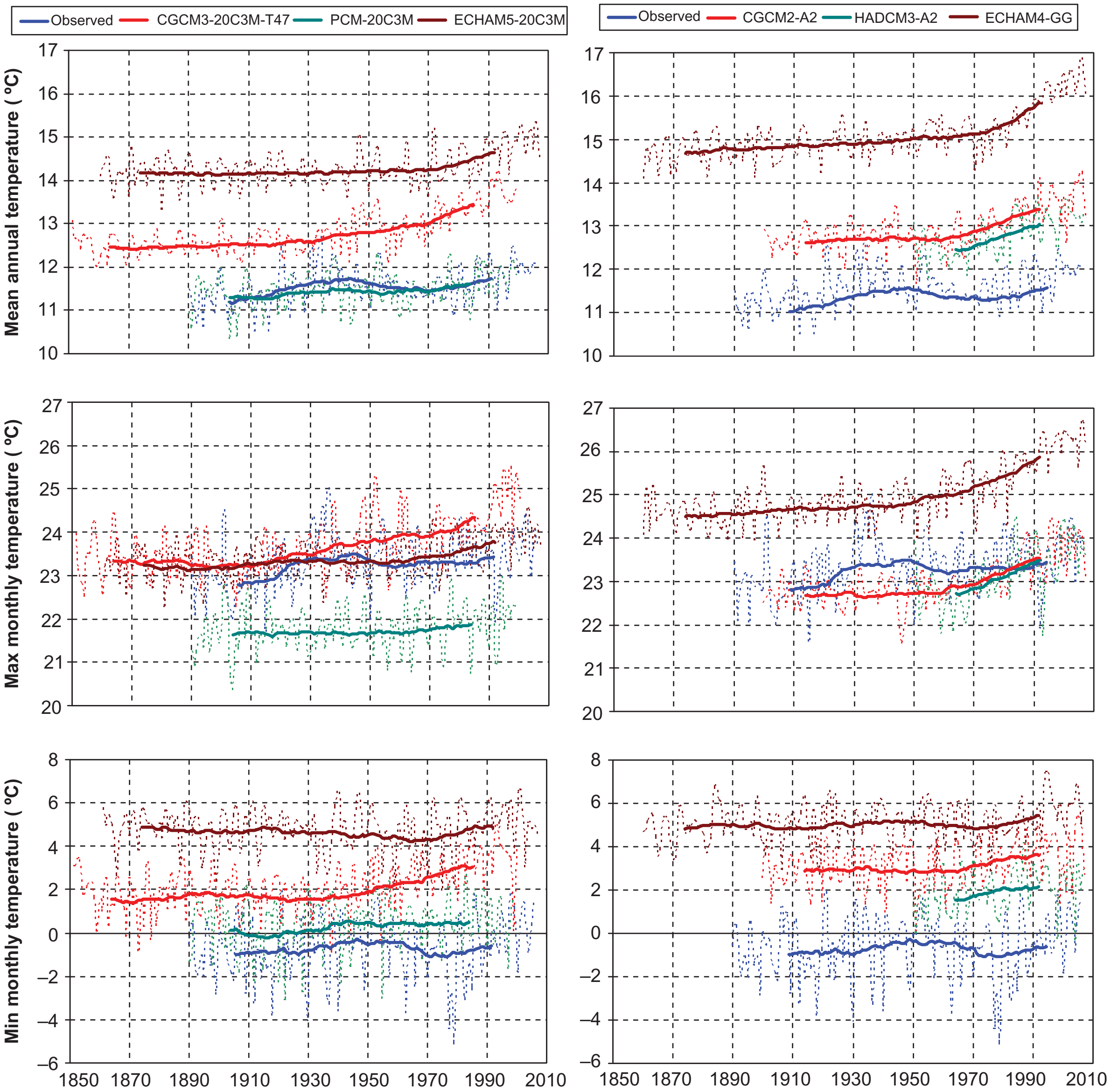

Fig. 12 Various temperature time series spatially integrated over the USA (mean annual, maximum monthly, minimum monthly), at annual and 30-year scales.

1890 and 1940, then had a falling trend until 1970, and from 1970 until today it had a slight upward trend. None of the model outputs fit these fluctuations of the annual mean temperature; most indicate a constant increase that becomes steeper in the last decades of the 20th century. The results closest to reality are the outputs of PCM-20C3M, but even these do not include the falling trend in 1940–1970 and have a very low coefficient of efficiency in 30-year time scale (only 0.05). However, this neutral performance of the annual mean temperature seems to result for the wrong reasons, as a result of averaging throughout a year, as indicated by examination of the maximum monthly temperature, minimum monthly temperature, the annual temperature amplitude, and the DJF and JJA seasonal temperature. Specifically, the maximum monthly temperature, the annual temperature amplitude and the JJA seasonal temperature are underestimated, while all the other time series are overestimated. The inter-annual fluctuations are not reproduced and, as a result, the Hurst coefficient is underestimated for all time series.

The results are worse for precipitation (Fig. 13). The annual precipitation is overestimated by up to 300 mm. In annual precipitation, the model outputs that are closer to reality (CGCM3-20C3M-T47) have a

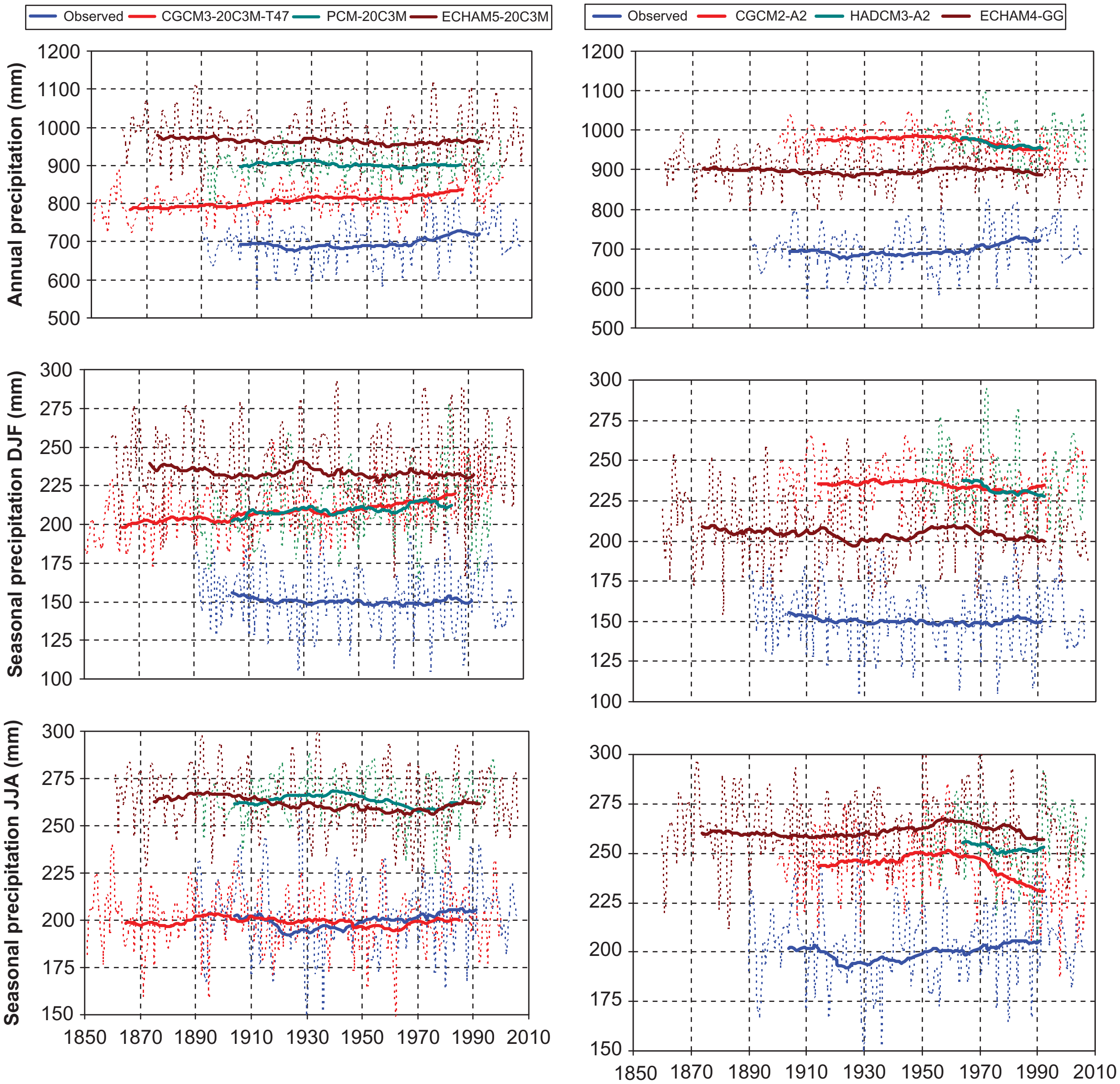

Fig. 13 Various precipitation time series spatially integrated over the USA (annual, seasonal DJF, seasonal JJA), at annual and 30-year scales.

coefficient of efficiency equal to −5.11 and a correlation coefficient of 0.171. The results are slightly better at the time series of maximum monthly temperature (coefficient of efficiency −0.57 and correlation coefficient 0.038) and JJA seasonal temperature (coefficient of efficiency −0.58 and correlation coefficient 0.057), but even there the best-fitting model outputs have a negative coefficient of efficiency and a correlation coefficient near zero.

The results for AR4 are no better than those for TAR. In some, the annual mean temperature of the USA is overestimated by about 4–5°C and the annual precipitation by about 300–400 mm.

In general, the results at the large scale are poorer than those of point comparison. One reason for this is probably the use of the BLUE technique. Specifically, in point comparison, rather than comparing each station to the nearest grid point, we compared it to the best-fitting of the four nearest grid points (more exactly, to the best-fitting unbiased linear combination of the four nearest grid points). This means that we made the comparison as generous as possible for the model output, in order to avoid influences of local factors. At the large scale, no such treatment of models was possible, or needed; instead, a simple areal integration of observed and modelled data was performed, and, therefore, the comparison is less forgiving.

CONCLUSIONS AND DISCUSSION

It is claimed that GCMs provide credible quantitative estimates of future climate change, particularly at continental scales and above. Examining the local performance of the models at 55 points, we found that local projections do not correlate well with observed measurements. Furthermore, we found that the correlation at a large spatial scale, i.e. the contiguous USA, is worse than at the local scale.

However, we think that the most important question is not whether GCMs can produce credible estimates of future climate, but whether climate is at all predictable in deterministic terms. Several publications, a typical example being Rial *et al.* (2004), point out the difficulties that the climate system complexity introduces when we attempt to make predictions. “Complexity” in this context usually refers to the fact that there are many parts comprising the system and many interactions among these parts. This observation is correct, but we take it a step further. We think that it is not merely a matter of high dimensionality, and that it can be misleading to assume that the uncertainty can be reduced if we analyse its “sources” as nonlinearities, feedbacks, thresholds, etc., and attempt to establish causality relationships. Koutsoyiannis (2010) created a toy model with simple, fully-known, deterministic dynamics, and with only two degrees of freedom (i.e. internal state variables or dimensions); but it exhibits extremely uncertain behaviour at all scales, including trends, fluctuations, and other features similar to those displayed by the climate. It does so with a constant external forcing, which means that there is no causality relationship between its state and the forcing. The fact that climate has many orders of magnitude more degrees of freedom certainly perplexes the situation further, but in the end it may be irrelevant; for, in the end, we do not have a predictable system hidden behind many layers of uncertainty which could be removed to some extent, but, rather, we have a system that is uncertain at its heart.

Do we have something better than GCMs when it comes to establishing policies for the future? Our answer is yes: we have stochastic approaches, and what is needed is a paradigm shift. We need to recognize the fact that the uncertainty is intrinsic, and shift our attention from reducing the uncertainty towards quantifying the uncertainty (see also Koutsoyiannis *et al*., 2009a). Obviously, in such a paradigm shift, stochastic descriptions of hydroclimatic processes should incorporate what is known about the driving physical mechanisms of the processes. Despite a common misconception of stochastics as black-box approaches whose blind use of data disregard the system dynamics, several celebrated examples, including statistical thermophysics and the modelling of turbulence, emphasize the opposite, i.e. the fact that stochastics is an indispensable, advanced and powerful part of physics. Other simpler examples (e.g. Koutsoyiannis, 2010) indicate how known deterministic dynamics can be fully incorporated in a stochastic framework and reconciled with the unavoidable emergence of uncertainty in predictions.

Acknowledgements We wish to thank the three anonymous reviewers, whose both strongly positive and strongly negative comments were important to us: the former for encouraging us and the latter for making us more confident that we did not err, as well as for forcing us to improve the presentation significantly. We also thank the editor Z. W. Kundzewicz for his decision to reconcile the polarized reviews by encouraging a public discussion of the different opinions, and for his own valuable comments on the paper. Finally, we wish to thank the colleagues who informally, during the first presentation of this study as a poster in the EGU 2009 Assembly, or later, offered us useful and constructive comments.

REFERENCES

Anagnostopoulos, G. G. (2009) Assessment of the reliability of climate models. Diploma thesis, Department of Water Resources and Environmental Engineering – National Technical University of Athens. Available at http://www.itia.ntua.gr/en/docinfo/893/.

Christensen, J. H., Hewitson, B., Busuioc, A., Chen, A., Gao, X., Held, I., Jones, R., Kolli, R. K., Kwon, W.-T., Laprise, R., Magaña Rueda, V., Mearns, L., Menéndez, C. G., Räisänen, J., Rinke, A., Sarr, A. & Whetton, P. (2007) Regional climate projections. In: *Climate Change 2007: The Physical Science Basis*. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change (S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor & H. L. Miller, eds), 847–940. Cambridge: Cambridge University Press.

Flato, G. M. & Boer, G. J. (2001) Warming asymmetry in climate change simulations. *Geophys. Res. Lett*. **28**, 195–198.

Gordon, C., Cooper, C., Senior, C. A., Banks, H., Gregory, J. M., Johns, T. C., Mitchell, J. F. B. & Wood, R. A. (2000) The simulation of SST, sea ice extents and ocean heat transports in aversion of the Hadley Centre Coupled Model without flux adjustments. *Climate Dyn*. **16**, 147–168.

Grotch, S. L. & MacCracken, M. C. (1991) The use of general circulation models to predict regional climatic change. *J. Climate* **4**(3), 286–303.

Hurst, H. E. (1951) Long term storage capacities of reservoirs. *Trans. Am. Soc. Civil Engrs* **116**, 776–808 (Published in 1950 as Proceedings Separate no. 11).

Johnson, F. & Sharma, A. (2009) Measurement of GCM skill in predicting variables relevant for hydroclimatological assessments. *J. Climate* **22**, 4373–4382.

Kim, J., Chang, J., Baker, N., Wilks, D. & Gates, W. (1984) The statistical problem of climate inversion: determination of the relationship between local and large-scale climate. *Monthly Weather Rev.* **112**(10), 2069–2077.

Kolmogorov, A. N. (1940) Wienersche Spiralen und einige andere interessante Kurven in Hilbertschen Raum. *Dokl. Akad. Nauk URSS* **26**, 115–118. (in German).

Kotroni, V., Lykoudis, S., Lagouvardos, K. & Lalas, D. (2008) A fine resolution regional climate change experiment for the eastern Mediterranean: analysis of the present climate simulations. *Global Planet. Change* **64**, 93–104.

Koutsoyiannis, D. (2003) Climate change, the Hurst phenomenon, and hydrological statistics. *Hydrol. Sci. J.* **48**(1), 3–24.

Koutsoyiannis, D. (2006) A toy model of climatic variability with scaling behaviour. *J. Hydrol.* **322**, 25–48.

Koutsoyiannis, D. (2010) A random walk on water. *Hydrol. Earth System Sci.* **14**, 585–601.

Koutsoyiannis, D., Efstratiadis, A. & Georgakakos, K. P. (2007) Uncertainty assessment of future hydroclimatic predictions: a comparison of probabilistic and scenario-based approaches. *J. Hydromet.* **8**(3), 261–281.

Koutsoyiannis, D., Efstratiadis, A., Mamassis, N. & Christofides, A. (2008) On the credibility of climate predictions. *Hydrol. Sci. J.* **53**(4), 671–684.

Koutsoyiannis, D. & Langousis, A. (2011) Precipitation, Chapter 27 . In: *Treatise on Water Science*. Elsevier (in press).

Koutsoyiannis, D., Makropoulos, C., Langousis, A., Baki, S., Efstratiadis, A., Christofides, A., Karavokiros, G. & Mamassis, N. (2009a) HESS Opinions "Climate, hydrology, energy, water: recognizing uncertainty and seeking sustainability". *Hydrol. Earth System Sci.* **13**, 247–257.

Koutsoyiannis, D., Montanari, A., Lins, H. F. & Cohn, T. A. (2009b) Climate, hydrology and freshwater: towards an interactive incorporation of hydrological experience into climate research—DISCUSSION of "The implications of projected climate change for freshwater resources and their management". *Hydrol. Sci. J.* **54**(2), 394–405.

Koutsoyiannis, D. & Montanari, A. (2007) Statistical analysis of hydroclimatic time series: uncertainty and insights. *Water Resour. Res.* **43**(5), W05429.1–W05429.9.

Murphy, J. M., Booth, B. B. B., Collins, M., Harris, G. R., Sexton, D. M. H. & Webb, M. J. (2007) A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles. *Phil. Trans. Roy. Soc. A* **365**, 1993–2028.

NCDC (National Climatic Data Center) (2010) Global surface temperature anomalies. Available from: http://www.ncdc.noaa.gov/cmb-faq/anomalies.html [Accessed 22 June 2010].

Randall, D. A., Wood, R. A., Bony, S., Colman, R., Fichefet, T., Fyfe, J., Kattsov, V., Pitman, A., Shukla, J., Srinivasan, J., Stouffer, R. J., Sumi, A. & Taylor, K. E. (2007) Climate models and their evaluation. In: *Climate Change 2007: The Physical Science Basis.* Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change (S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor & H. L. Miller, eds), 589–662. Cambridge: Cambridge University Press.

Rial, J. A., Pielke, R. A. Sr, Beniston, M., Claussen, M., Canadell, J., Cox, P., Held, H., de Noblet-Ducoudré, N., Prinn, R., Reynolds, J. F. & Salas, J. D. (2004) Nonlinearities, feedbacks and critical thresholds within the Earth's climate system. *Climate Change* **65**, 11–38.

Schmidt, G. (2008) Hypothesis testing and long range memory, blog post, Available from: http://www.realclimate.org/index.php/archives/2008/08/hypothesis-testing-and-long-term-memory/ [Accessed 22 August 2010].

Stowe, K. (2007) *Thermodynamics and Statistical Mechanics*, 2nd edn. Cambridge: Cambridge University Press.

von Storch, H., Zorita, E. & Cubasch, U. (1993) Downscaling of global climate change estimates to regional scales: an application to Iberian rainfall in wintertime. *J. Climate* **6**(6), 1161–1171.