RAPID COMMUNICATION

On the credibility of climate predictions

D. KOUTSOYIANNIS, A. EFSTRATIADIS, N. MAMASSIS & A. CHRISTOFIDES

Department of Water Resources, Faculty of Civil Engineering, National Technical University of Athens, Heroon Polytechneiou 5, GR-157 80 Zographou, Greece

dk@itia.ntua.gr

Abstract Geographically distributed predictions of future climate, obtained through climate models, are widely used in hydrology and many other disciplines, typically without assessing their reliability. Here we compare the output of various models to temperature and precipitation observations from eight stations with long (over 100 years) records from around the globe. The results show that models perform poorly, even at a climatic (30-year) scale. Thus local model projections cannot be credible, whereas a common argument that models can perform better at larger spatial scales is unsupported.

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

De la crédibilité des prévisions climatiques

Résumé Des prévisions distribuées dans l'espace du climat futur, obtenues à l'aide de modèles climatiques, sont largement utilisées en hydrologie et dans de nombreuses autres disciplines, en général sans évaluation de leur confiance. Nous comparons ici les sorties de plusieurs modèles aux observations de température et de précipitation de huit stations réparties sur la planète qui disposent de longues chroniques (plus de 100 ans). Les résultats montrent que les modèles ont de faibles performances, y compris à une échelle climatique (30 ans). Les projections locales des modélisations ne peuvent donc pas être crédibles, alors que l'argument courant selon lequel les modèles ont de meilleures performances à des échelles spatiales plus larges n'est pas vérifié.

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

INTRODUCTION

Hydrologists are very attentive in the development and use of mathematical models for hydrological processes, particularly if these models are to be applied for prediction of future events. They use several indices to assess the prediction skill of their models, and they evaluate the indices not only in the model calibration period, but also in a separate validation period, whose data were not used in the calibration (the split-sample technique, Klemeš, 1986). Long prediction horizons, such as 50 or 100 years, are very common in engineering hydrological applications, as these are the lifetime periods of major engineering works. Traditionally, in such cases, deterministic models and approaches, which are good for prediction horizons of some hours to a few days, are replaced by probabilistic and stochastic approaches. To the authors' knowledge, no attempt to cast long-term hydrological predictions based on deterministic hydrological approaches has ever been made.

On the other hand, in recent decades, numerous hydrological studies have attempted to cast projections of the impacts of hypothesized anthropogenic climate change on freshwater resources and their management, adaptation and vulnerabilities (Kundzewicz *et al*., 2008). All these studies are essentially based on the explicit or tacit assumptions that climate is deterministically predictable in the long term and that the climate models (or general circulation models, GCMs) can give credible predictions of future climate for horizons of 50, 100 or more years (e.g. Alcamo *et al*., 2007). Less effort has been put into falsifying or verifying such assumptions. However, the widespread use of statistical downscaling methods in hydrological studies may be viewed as an indirect falsification of the reliability of climatic models: for this downscaling refers in essence to techniques that modify the climate model outputs in an area of interest in order to reduce their large departures from historical observations in the area, rather than techniques to scale down the coarse-gridded GCM outputs to finer scales.

As falsifiability is an essential element of science (Popper, 1983), the scientific basis of climatic predictions may be disputed on the grounds that they are not falsifiable or verifiable at present. Such a critique may arise from the argument that we need to wait several decades before we will know how reliable the predictions may be. However, we maintain that elements of falsifiability already exist. These should be traced in at least two directions: in the very structure and core hypotheses of GCMs and the related modelling practices, and in the agreement of model results in past periods with reality (hindcasting or retrodiction).

In terms of the first direction, several studies have tried to shed light on GCM inconsistencies and the resulting uncertainty (the most recent being Frank, 2008). However, we think that the following passage from Kerr (2008) is indicative of the state of the art in GCMs. Kerr's article refers to the study by Keenlyside *et al*. (2008), who explained the observed constancy (or slight decrease) of the global temperature (instead of the projected global warming) in the last decade, using real sea-surface temperatures as initial conditions (and with this they predict that in the next decade European and North American surface temperatures will decrease slightly):

"*To take account of such ocean-driven natural variability, Keenlyside and his colleagues began their model's forecasting runs by giving the model's oceans the actual sea surface temperatures measured in the starting year of a simulation. Providing the initial state of the ocean doesn't make much difference when forecasting out a century, so long-range forecasters don't usually bother. But an initial state gives the model a starting point from which to calculate what the oceans will be doing a decade hence and therefore what future natural variability might be like*".

This reveals a culture in the climatological community that is very different from that in the hydrological community. In hydrology and water resources engineering, in real-time simulations that are used for future projections in transient systems (in contrast to steady-state simulations), it is inconceivable to neglect the initial conditions; likewise, it is inconceivable to claim that a model has good prediction skill for half a century ahead but not for a decade ahead.

Furthermore, the climatological community focuses on theories and models, whereas the hydrological community has greater trust in data. In this respect, here, we focus on the second falsifiability path, the testing of GCM performance in reproducing observed past climatic behaviours, a path also explored by others (e.g. Douglass *et al*., 2008). The IPCC Third Assessment Report (TAR; IPCC, 2001) contains comparisons at the global scale, and the IPCC Fourth Assessment Report (AR4; IPCC, 2007) extends these comparisons of observed and simulated climate to the continental scale; however, these do not offer validation, in the sense described above (and are not presented as validation results by IPCC). In our falsifiability framework, the TAR models provide a good basis because they have projected future climate starting from 1990. Thus, there is an 18-year period for which comparison of model outputs with reality is possible. Besides, several TAR model runs include longer past periods with historical inputs. The situation is different with AR4 models, but again comparisons with past climate are possible as detailed below. The comparisons are done using existing long records of the past. The use of climatic records was also recommended by the US National Research Council (2005) in a different context (to investigate relationships between regional radiative forcing and climate response). As climate is defined to be a long-term average of processes, long time series are necessary to investigate whether or not observed climatic trends (more precisely, long-term fluctuations) are captured by climatic models. The hydrometeorological time series with the longest observation periods, i.e. temperature and precipitation, which also happen to be the most important to hydrology, were chosen for the comparison. Records of 100 years or more were retrieved from eight locations belonging to different climates worldwide.

METHODOLOGY

The methodology we employed in this study is very simple. We decided to use eight stations (this number was dictated by time and resource limitations—the research is not funded). Our criteria for the selection of the eight locations were: (a) the distribution of stations in all continents and in dif-

ferent types of climate; (b) the availability of data on the Internet at a monthly time scale; and (c) the existence of long data series (>100 years for both temperature and precipitation) without missing data (or with very few missing data, which were filled in with average monthly values). In Australia, we were not able to find a station satisfying the data size criterion for both temperature and precipitation and we accepted a shorter length of the precipitation record. The study locations are shown in Table 1, and their characteristic climatic diagrams are presented in Koutsoyiannis *et al*. (2008b; this is an on-line report with all detailed information supplementary to this paper). An additional location (Aliartos, Greece) had been studied previously (Koutsoyiannis *et al*., 2007, 2008a).

The next step was to retrieve a number of climatic model outputs for historical periods (available on the Internet), which are shown in Table 2. We picked three TAR and three AR4 models, and one simulation for each model. The selection of simulation runs, which is presented in

Table 1 Study locations and their characteristics.

Station	Climate	Latitude (°)	Longitude (°)	Altitude (m)	Data source
Albany (USA)	Sub-tropical	31.53N	84.13W	60	www.ncdc.noaa.gov/oa/climate/
Athens (Greece)	Mediterranean	37.97N	23.72E	107	www.knmi.nl
Alice Springs (Australia)	Semi-arid	23.80S	133.88E	547	www.knmi.nl
Colfax (USA)	Mountainous	39.11N	120.95W	735	www.ncdc.noaa.gov/oa/climate/
Khartoum (Sudan)	Arid	15.60N	32.50E	380	www.knmi.nl
Manaus (Brasil)	Tropical	3.17S	60.00W	60	www.knmi.nl
Masumoto (Japan)	Marine	36.20N	138.00E	611	www.knmi.nl
Vancouver (USA)	Mild	45.63N	122.68W	10	www.ncdc.noaa.gov/oa/climate/

Table 2 Main characteristics of the GCMs used in the study.

IPCC report	Name	Developed by	Resolution (°) in latitude and longitude	Grid points, latitudes × longitudes
TAR	ECHAM4/OPYC3	Max-Planck-Institute for Meteorology & Deutsches Klimarechenzentrum, Hamburg, Germany	2.8 × 2.8	64 × 128
TAR	CGCM2	Canadian Centre for Climate Modelling 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 Center for Atmospheric Research, USA	2.8 × 2.8	64 × 128

Sources: www.mad.zmaw.de/IPCC_DDC/html/SRES_TAR/; www.mad.zmaw.de/IPCC_DDC/html/SRES_AR4/.

Table 3 IPCC scenarios and their relevance to the study.

	Scenario	Characteristics	Reason for being appropriate or inappropriate
TAR	✓ SRES, IS92a	Many runs are based on historical GCM input information prior to 1989 and extended using scenarios for 1990 and beyond.	For such runs, choice of scenario is irrelevant for test periods up to 1989; for later periods, there is no significant difference between different scenarios for the same model.
AR4	✗ SRES	Various hypothetical scenarios for the future.	Runs start in the 21st century (out of study period).
	✗ COMMIT	Greenhouse gases fixed at year 2000 levels.	Runs start in the 21st century (out of study period).
	✗ 1%-2X, 1%-4X	Assume a 1%-per-year increase in CO_2, usually starting at year 1850.	Results in CO_2 being 570 cm^3/m^3 (ppm) already in 1920, when in fact it was 379 cm^3/m^3 in 2005. Actual 20th century concentrations are required.
	✗ PI-cntrl	Uses pre-industrial greenhouse gas concentrations.	Actual 20th century concentrations are required.
	✓ 20C3M	Generated from output of late 19th & 20th century simulations from coupled ocean–atmosphere models, to help assess past climate change.	This is the only AR4 scenario relevant to this study, and therefore the outputs of model runs on this scenario were used.

Sources: Leggett *et al*. (1992); Nakicenovic & Swart (1999); Carter *et al*. (1999); Hegerl *et al*. (2003); www.mad.zmaw.de/IPCC_DDC/html/SRES_AR4/.

Table 3, was based on the criterion that it cover past periods, rather than merely referring to future. As Table 3 indicates, the selection of scenario does not matter for TAR models; therefore, we randomly chose SRES A2 and simulations covering the 20th century with historical GCM input information prior to 1989; there is an exception in ECHAM, whose simulations with SRES do not cover the 20th century but for which a run is nevertheless available for the scenario IS92a starting at 1860. Of the AR4 scenarios, only 20C3M is appropriate.

The next step was to extract the monthly time series for the four grid points closest to each of the eight examined stations (the specific grid depends on the model) and to estimate a modelled time series for each station. A technique for making inferences at small regional scales from coarser climate model scales based on grid points nearest to the area of interest was proposed by Georgakakos (2003). However, this technique aims to downscale model information to finer spatial scales, which is different from the falsification/validation scope of the present study, as explained below. Nonetheless, we kept the idea of using the nearest grid points for making inferences on the location of interest. Specifically, we used the time series of all four nearest grid points and produced the modelled time series for the station location based on the best linear unbiased estimation (BLUE; e.g. Kitanidis, 1993), i.e. by optimizing 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 prediction error), and x_1, x_2, x_3, x_4 are the model outputs for the four closest grid points. Optimization is done on the basis of the coefficient of efficiency, which is the main comparison statistic in this study, defined as $\mathrm{Eff} = 1 - e^2/\sigma^2$, where e^2 is the mean square error in prediction and σ^2 is the variance of historical series. Given that at each point the historical variance is constant, maximizing efficiency is precisely equivalent to minimizing the mean square error (as in the standard BLUE method). Thus, we let the modelled time series fit the historical monthly time series as closely as possible. Other techniques with fixed weights, based for instance on the distances of the study location to each grid point, would obviously produce larger prediction errors and smaller efficiencies than the BLUE technique. Interestingly, in some of the cases (particularly in temperature), the resulted weights were zero for three out of the four grid points and one for the fourth point, which means that one grid point (i.e. four-point combination with weights 0, 0, 0, 1) was more representative for the study location than any linear combination of the four grid points.

The final step was the comparison of the historical with modelled time series. The comparisons were made graphically and using various statistical indicators, as detailed below. Three scales of comparison were used: the monthly, the annual and the climatic, where the latter was assumed to be the 30-year (moving) average.

JUSTIFICATION OF THE METHODOLOGY

An essential element of our approach is the use of directly observed time series instead of any type of processed (e.g. gridded) data sets. This is justified by the falsifiability scope of the study, according to which, one of the two objects to be compared should be an observable quantity and the other a modelled quantity. Thus, we preclude any type of manipulation of observed data that could lead to an artificial agreement with models. Unavoidably, this leads to comparisons on a point basis rather than on an areal basis, which requires the estimation of a modelled time series at the chosen point from a gridded output data set. This is much more feasible and natural than the opposite way (estimating an "observed" time series at a GCM's grid point), or than an areal comparison (aggregating observed time series to produce an areal average and then aggregating GCM gridded time series on the same area), for it is difficult or infeasible to find long historical time series at many locations on the area, in order to combine them to produce either a spatial interpolation to a GCM grid point or a spatial integration at a given area. The difficulty in doing such spatial interpolations or integrations from observed data is reflected, for instance, in the fact that the Climatic Research Unit (CRU), which systematically monitors global temperature on a monthly basis using observations from 3000 stations (www.cru.uea.ac.uk/cru/ data/ temperature/),

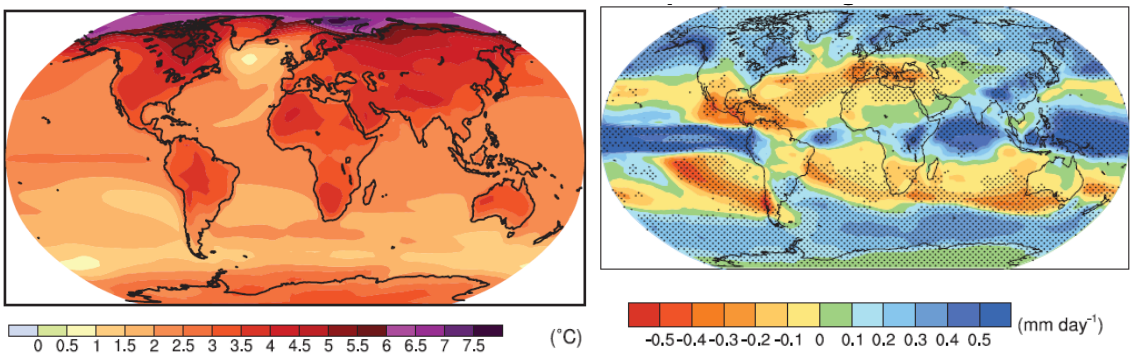

Fig. 1 IPCC projected changes in (left) temperature and (right) precipitation, for 2080–2099 relative to 1980–1999 for the SRES scenario A1B (Figs 10.8, middle right panel, and 10.12, upper left panel, respectively, of Meehl *et al*., 2007; stippled regions in the precipitation map indicate consistency of at least 80% of models in the sign of change).

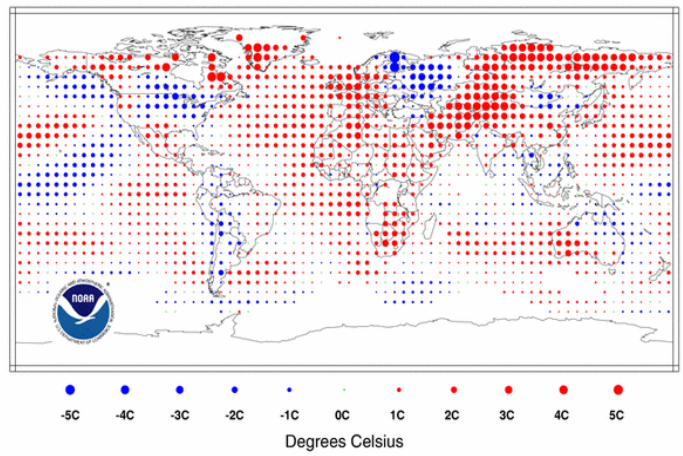

Fig. 2 Land and sea (blended) temperature departures from the 1961–1990 mean in May 2008 (US National Climatic Data Center, www.ncdc.noaa.gov/oa/climate/research/2008/may/global.html).

has avoided providing temperature time series on any spatial basis and, instead, provides only the temperature departures from the 1961–1990 mean (more commonly referred to as "anomalies").

It is very easy and natural to perform spatial interpolation of the GCM outputs. For this purpose we have chosen the BLUE technique as it is the one that makes the error (departure of observation from model) as small as possible. This technique is commonly used even in "rough" random fields and, apparently, it is even more appropriate for the rather smooth temperature and precipitation fields of GCM outputs, particularly at the climatic time scale. Figure 1 shows that GCM output temperature and precipitation fields at the climatic scales are quite smooth in space. This smoothness extends also to observed temperature fields at time scales as fine as monthly (Fig. 2). Apparently, the BLUE interpolation is more accurate for sites in flat terrain, whereas in mountainous terrain the method cannot capture variations due to orographic effects. Even in the latter cases, BLUE is appropriate at least for the climatic scales, where random errors should be smoothed out. While BLUE is by definition unbiased, the contingency of bias cannot be excluded, for instance in a location at a very high altitude with decreased temperature due to altitude. For that reason, the comparisons of observed and modelled series are done not only in terms of the coefficient of efficiency, which is affected by the presence of bias, but also in terms of the correlation coefficient, which by definition removes the effect of bias.

Another possible objection could be that the performance of GCMs on the point basis is expected to be low because of the "noise" of local (spatial or temporal) weather conditions,

whereas the performance should be higher at large spatial scales. This is implied, for instance, in Randall *et al*. (2007), where it is stated that GCMs provide credible quantitative estimates of future climate change, particularly at continental scales and above, and that the estimates are more confident for temperature than for other climatic variables (e.g. precipitation). However (see Fig. 1 and IPCC, 2007), geographically distributed projections are provided, and not only continental or global-scale projections. These geographically distributed data are then used, after downscaling, by many scientists from many disciplines to project the impacts of climate change onto any type of natural process and human activity, and at any spatial scale (not only at continental scales and above). Two questions then arise: (1) Can the continental or global climatic projections be credible if the distributed information, from which the aggregated information is derived, is not? (2) Are geographically distributed projections credible enough to be used in further studies?

The common answer to the first question is positive. The logic behind it is a premise that atmospheric phenomena consist of two components: "climate", i.e. a low frequency component, and a high frequency "noise". Likewise, in space, there is "climate" and local "micro-climate". The argument is then that GCMs can faithfully represent the climate at very large scales, although they may fail on "noise" and "micro-climate", despite the fact that GCMs primarily model the "noise", since they are numerical prediction models integrating differential equations at small time steps (Edwards, 2001). Koutsoyiannis & Montanari (2007) criticized this dichotomous logic, giving emphasis to the continuity of stochastic descriptions on a spectrum of scales with no specific scale dominating. This is manifested by the Hurst behaviour of hydroclimatic processes, herein referred to as Hurst-Kolmogorov (HK) climate, after Hurst (1951) who studied it in natural processes and Kolmogorov (1940) who devised its stochastic representation as a mathematical tool for the research of turbulence (see also Shiryaev, 1989; Koutsoyiannis & Cohn, 2008). Koutsoyiannis & Montanari, as well as Koutsoyiannis (2003, 2005, 2006a,b) and Koutsoyiannis *et al*. (2007) have provided empirical evidence and theoretical support of the HK behaviour of climate, and have shown that climatic variability is very much higher in HK processes than in purely random or Markov-type processes. Specifically, in a HK climate, the uncertainty at a climatic (30-year) scale proves to be only slightly lower than that at the annual one (Koutsoyiannis *et al*., 2007), in contrast to the classical approach, which yields significant reduction as we proceed from the annual to the climatic scale and justifies different perception of climatic and finer scale views of processes. Furthermore, Koutsoyiannis (2006b) has demonstrated, using a toy model with fully known simplified deterministic dynamics capable of producing a HK climate, that even slight perturbations in initial conditions produce very high departures, not only at a fine time scale but also (and mainly) at the climatic time scale. Such a result is in line with Collins (2002), who used a GCM (HadCM3) and, assuming this to be a "perfect" model, concluded that the climate predictability is likely to be severely limited by chaotic error growth. Thus, we think that the commonly assumed positive answer to the first question is unsupported.

Yet, the focus of this paper is the second question, namely the credibility of the geographically distributed representation of climate by GCMs. We must clarify that the methodology we propose for this testing is not "downscaling" in the sense that is commonly attributed to this term. Statistical downscaling is in fact a technique to utilize the available local data series to adapt (correct) climate model outputs so as to agree with observations. This typically involves the establishment (fitting) of a statistical relationship between GCM outputs (predictors) and observational data (e.g. Heyen *et al*., 1996; Wetterhall *et al*., 2005).

Our falsification/validation framework merely involves spatial interpolation of the GCM output fields to infer their values at the points of interest. For this interpolation we used the local observations only to estimate the optimal weights. We do not propose this technique as a downscaling approach. Rather, it is a step prior to downscaling, necessary to justify whether downscaling is meaningful or pointless. In the case that the GCM results are invalidated against observations at the climatic scale, there is no meaning to proceeding with further analyses. As discussed above, the smoothness of the given fields justifies the reliability of interpolation. This is further demonstrated with an example, which shows that four neighbouring grid points (or even three) are

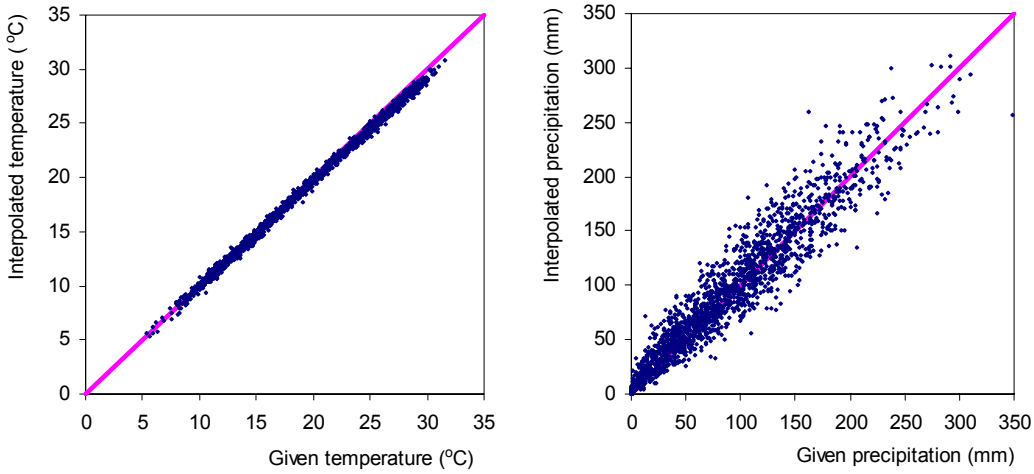

Fig. 3 Interpolated *vs* given GCM temperature and precipitation at a grid point (see text).

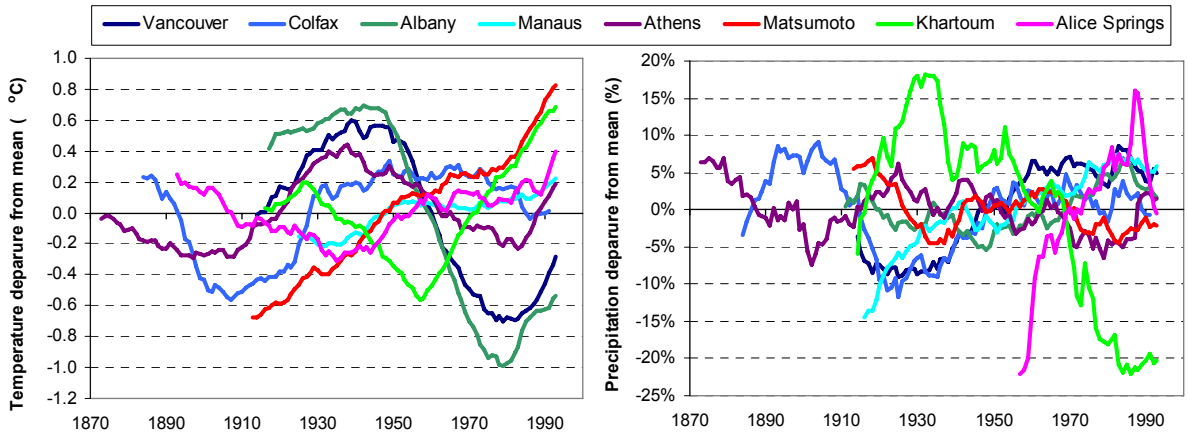

Fig. 4 Plots of 30-year moving average time series for temperature (departures from historical mean; left) and precipitation (relative departures from historical mean; right).

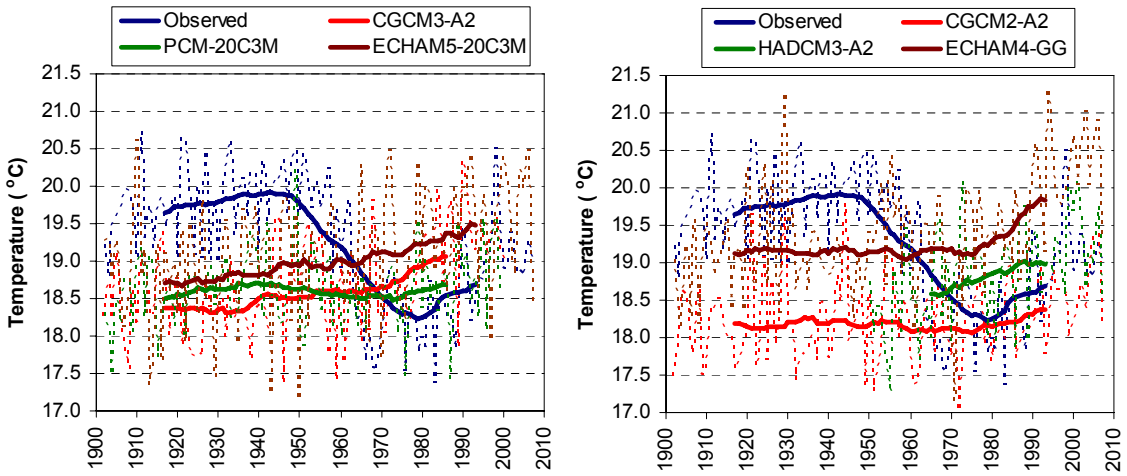

Fig. 5 Plots of observed and modelled annual (doted lines) and 30-year moving average (continuous lines) temperature time series at Albany (left: AR4 models; right: TAR models).

enough to provide a sufficient interpolation. The example refers to the four grid points closest to the station Albany for model ECHAM5-OM and scenario 20C3M, whose data were used next in the comparisons with observed data. The temperature and precipitation time series of one of the four points were assumed unknown and estimated from the time series of the other three points, using the BLUE technique described above (but with three weights). As shown in Fig. 3, the agreement of interpolated with given time series is impressively good, particularly for temperature (coefficients of efficiency 0.99 and 0.91 for temperature and precipitation, respectively). *A fortiori*, with four points instead of three, the agreement would be even better.

LONG-TERM VARIABILITY OF HISTORICAL TIME SERIES

All examined long records exhibit large over-year variability (i.e. long-term fluctuations), with no systematic signatures across the different locations/climates. This is clearly illustrated in Fig. 4, where long excursions from overall historical means are observed in almost all climatic (30-year moving average) series. The 30-year average temperature departs up to ±1°C from the mean of the entire historical period, while the 30-year average precipitation departs up to about ±20% from the historical mean. These excursions demonstrate the large variability at long time scales, which is an essential element for our testing framework: indeed, it is vital to evaluate whether the modelled series are able to reproduce the observed over-year "trends", which would be a powerful indicator of their potential to generate realistic climatic behaviours in their future projections.

A well-recognized metric of long-term fluctuations is the Hurst coefficient (H), which is estimated for the annual time scale and longer (Koutsoyiannis, 2003). All records but two have Hurst coefficients larger than 0.50, i.e. 0.72–0.93 for temperature and 0.56–0.86 for precipitation (see Fig. 6 and detailed statistics in Koutsoyiannis *et al*., 2008). The two exceptions are the precipitation time series of Albany and Athens, in which the Hurst coefficient is around 0.50. One is reminded that the value 0.50 corresponds to time independent processes and the value 1.0 to fully dependent ones. These results, based on long observed time series, do not confirm earlier studies claiming that the Hurst coefficient of temperature in continental areas is 0.5, which would imply temporal independence of continental climatic processes (Blender & Fraedrich, 2003; Fraedrich & Blender, 2003); this was also disputed in a discussion by Bunde *et al*. (2004), who found values 0.6–0.7, whereas, more recently, Alvarez-Ramirez *et al*. (2008) estimated even higher values in some periods. The three continental stations in our study, i.e. Alice Springs, Manaus and Khartoum, gave Hurst coefficients for temperature equal to 0.72, 0.89 and 0.90, respectively.

RESULTS

For each study location, we evaluated the proximity of the modelled and observed time series by means of various statistical indicators at the monthly, annual and 30-year moving average (climatic) time scales, as well as through graphical inspections. Two global performance indices were used for all scales, namely the coefficient of efficiency and the correlation coefficient (where the latter removes the effect of bias). Moreover, at all scales, we provided comparisons between the observed and modelled average and standard deviation (the former is obviously the same at all scales). In addition, in the annual time series, we calculated and compared the first-order auto-correlation coefficient and the Hurst coefficient, whereas, at the climatic scale, we also compared three fluctuation indices. These refer to: (a) the change of 30-year moving average temperature or precipitation in the 20th century (which is the common simulation period for all models except HadCM3-A2), calculated as the difference of 30-year moving averages centred at 1985 and 1915; (b) the change between the first and last year of each time series; and (c) the maximum fluctuation across the entire period. The latter is defined to be 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. For

the TAR models and the annual scale, we used three periods of comparison, i.e. the entire period, the period before 1989 (with historical GCM inputs) and the period after 1989 (with projected inputs and outputs). In addition, we plotted the annual and 30-year moving average time series, separately for the AR4 and TAR models. A representative example of comparisons of the temperature series at Albany is shown in Tables 4–6 and Fig. 5, while the results of all analyses (tabulated and graphical) are given in Koutsoyiannis *et al*. (2008).

The performance of the modelled against the observed time series varies across the different time scales. At the monthly scale, GCMs generally reproduce the broad climatic behaviours at the different locations and the sequence of wet/dry or warm/cold periods, although bias may be marked. The average (over the eight sites and six models) correlation coefficient between the modelled and observed time series reaches 0.848 for temperature but only 0.331 for precipitation.

Table 4 Detailed statistics for observed and modelled temperature time series at Albany at the monthly scale.

	Period	Average (°C)	St. dev. (°C)	Correlation	Efficiency
Observed	1899–2007	19.24	6.78		
CGCM3-A2	1899–2000	18.65	6.31	0.921	0.840
PCM-20C3M	1899–1999	18.60	6.40	0.922	0.840
ECHAM5-20C3M	1899–2007	19.03	6.31	0.913	0.829
CGCM2-A2	1900–2007	18.21	5.41	0.912	0.795
HadCM3-A2	1950–2007	18.80	7.15	0.930	0.853
ECHAM4-GG	1899–2007	19.31	6.50	0.931	0.866

Table 5 Detailed statistics for observed and modelled temperature time series at Albany at the annual scale.

	Period	Average (°C)	St. dev. (°C)	Correlation	Efficiency	Autocorrel.	Hurst coeff.
Observed	1899–2007	19.24	0.83			0.660	0.930
CGCM3-A2	1899–2000	18.65	0.63	–0.215	–1.368	0.306	0.802
PCM-20C3M	1899–1999	18.60	0.48	0.076	–0.828	0.005	0.511
ECHAM5-20C3M	1899–2007	19.03	0.77	–0.154	–1.214	0.078	0.577
CGCM2-A2	1900–2007	18.21	0.54	–0.012	–1.999	0.167	0.562
	1900–1989	18.15	0.54	0.016	–2.013	0.091	
	1989–2007	18.52	0.45	–0.016	–3.115	0.213	
HadCM3-A2	1950–2007	18.80	0.65	0.053	–0.905	0.232	0.713
	1950–1989	18.62	0.64	–0.319	–1.631	0.121	
	1989–2007	19.18	0.48	0.142	–1.238	0.045	
ECHAM4-GG	1899–2007	19.31	0.74	0.151	–1.001	0.279	0.803
	1899–1989	19.12	0.63	–0.072	–0.642	–0.023	
	1989–2007	20.23	0.57	0.031	–9.561	–0.222	

Table 6 Detailed statistics for observed and modelled temperature time series at Albany on climatic scale.

	Period	St. dev. (°C)	Correla-tion	Efficiency	DT (20th century) (°C)	DT (all data) (°C)	max DT (°C)
Observed	1899–2007	0.62			–1.09	–0.95	–1.68
CGCM3-A2	1899–2000	0.21	–0.856	–2.018	0.69	0.69	0.76
PCM-20C3M	1899–1999	0.06	0.455	–1.149	0.19	0.19	0.22
ECHAM5-20C3M	1899–2007	0.21	–0.832	–0.833	0.66	0.76	0.84
CGCM2-A2	1900–2007	0.07	0.110	–2.844	0.03	0.18	0.31
HadCM3-A2	1950–2007	0.15	–0.107	–4.443		0.41	0.45
ECHAM4-GG	1899–2007	0.19	–0.452	–0.359	0.40	0.71	0.80

DT: change of 30-year moving average in indicated period; max DT: difference of maximum minus minimum observed or simulated climatic value, with positive (negative) sign for positive (negative) trend.

The average efficiency is negative both for precipitation (–0.285) and temperature (–0.116), but without Manaus the latter value rises to 0.679. Even in the latter case, the performance is not very satisfactory if compared to that of even elementary statistical predictions. For instance, in Albany (Table 4), the highest efficiency, provided by ECHAM4-GG, is 0.866. Yet, if we replaced the modelled time series with a series of monthly averages (same for all years), the resulting efficiency would be 0.930, i.e. considerably higher. The corresponding results for precipitation in Albany are –0.168 for the best model (CGCM3-A2) and 0.140 for the elementary statistical method.

The performance of GCM time series degrades significantly when moving from the monthly to the annual time scale. At the annual scale, the average correlation coefficient between models and observations drops to only 0.112 for temperature and to –0.010 for precipitation, while the average efficiency is substantially negative for both processes (–8.829 for temperature, –2.043 for precipitation). In addition, GCMs underestimate the observed variability (expressed by the standard deviation for temperature and the coefficient of variation for precipitation) in 73% of cases for temperature and 90% for precipitation (Fig. 6), and the observed Hurst coefficient in 75% of cases for temperature and 83% for precipitation (Fig. 7).

Finally, at the 30-year climatic time scale, the average correlation coefficient rises slightly to 0.237 for temperature and remains slightly negative (–0.046) for precipitation; however, the average efficiency values become tremendously negative, –81.6 for temperature and –49.5 for

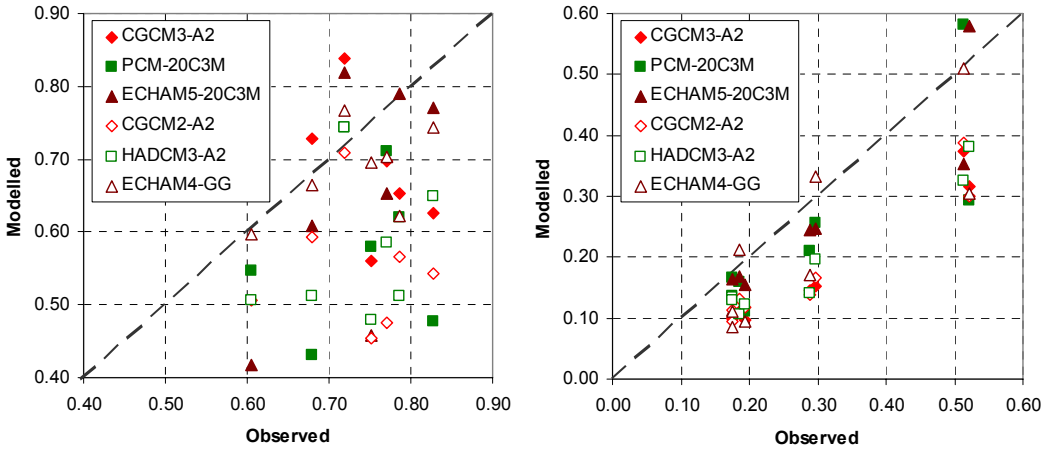

Fig. 6 Modelled *vs* observed standard deviations for temperature (left) and coefficients of variation for precipitation (right) time series.

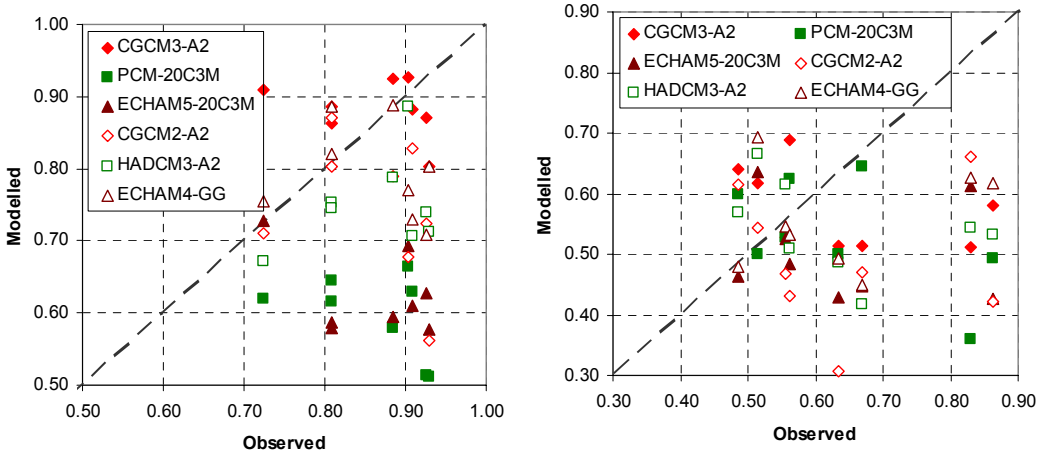

Fig. 7 Modelled *vs* observed Hurst coefficients for temperature (left) and precipitation (right) time series.

precipitation. This clearly shows that GCMs totally fail to represent the HK-type climate of the past 100–140 years, which is characterized by large-scale over-year fluctuations (i.e. successions of negative and positive "trends") that are very different from the monotonic trend of climatic models. In addition, they fail to reproduce the long-term changes in temperature and precipitation (Fig. 8). Remarkably, during the observation period, the 30-year temperature at Vancouver and Albany decreased by about 1.5°C, while all models produce an increase of about 0.5°C (Fig. 8, lower left). With regard to precipitation, the natural fluctuations are far beyond ranges of the modelled time series in the majority of cases (Fig. 8, lower right).

The systematically unsatisfactory agreement of modelled and observed time series can have four interpretations: (1) the models are poor; (2) the data are poor; (3) the modelled and observed time series are not comparable to each other (e.g. there should not be a direct link between observations at a point and model outputs at neighbouring grid cells); (4) our calculations and comparisons are wrong. The last possibility cannot be excluded in principle, but since all data series we used (observations and models) are available on the Internet, possible errors will be spotted. Interpretation (3) is not plausible in our opinion, for reasons explained in the section "Justification of the methodology"; in other words, we think that this interpretation is more or less a specific case of (1). Interpretation (2) has its merit: observations may be "contaminated", either by random and systematic errors or by changes in local conditions. The latter is expected particularly in stations in urban areas, where in recent decades the increasing heat island effect may have distorted the natural character of the time series and introduced artificial increasing trends in both precipitation and temperature (Huang *et al.*, 2008). However, the heat island effect, if present, would in fact improve rather than worsen the agreement between models and observations

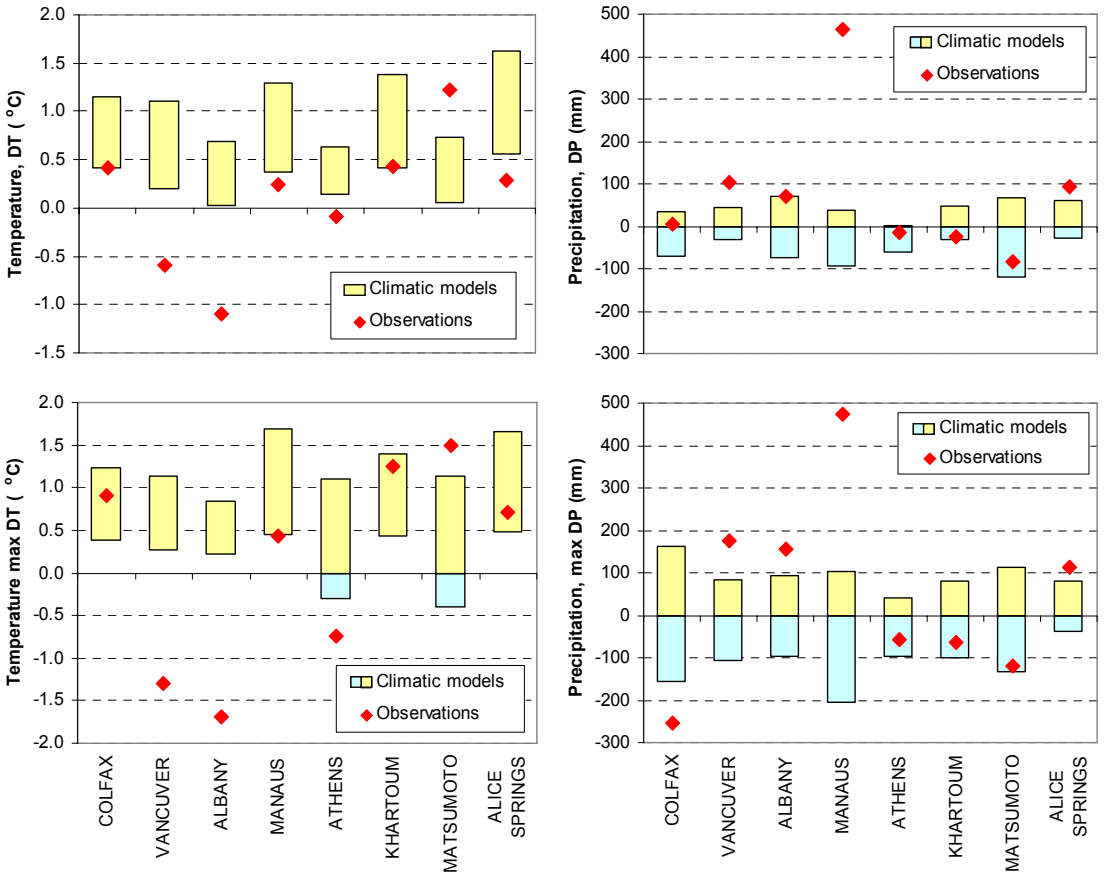

Fig. 8 Change of 30-year moving average temperature (upper left) and precipitation (upper right) in the 20th century and maximum fluctuation of 30-year moving average temperature (lower left), and precipitation (lower right) within the entire period.

(without it, the temperature in the most recent years would be lower, while GCMs predict a rise of temperature). Other random and systematic errors that may contaminate the measurements are not expected to be present in all examined stations. For all these reasons, we think that interpretation (1), i.e. that models are (intrinsically and perhaps inescapably) poor is the most plausible.

CONCLUDING REMARKS

The current scientific scene is dominated by the hypothesis that climate is deterministically predictable, combined with the belief that GCMs suitably implement this hypothesis and produce credible projections of future climate. As this hypothesis and this belief are widely accepted in a variety of scientific disciplines, including hydrology and water resources science, technology and management, and are used as a foundation upon which diverse impact studies are built, there is an urgent need to assess the credibility of climatic models. To date, the required attention has not been paid and many studies have built upon climatic projections without such prior assessment. This study compares observed, long climatic time series with GCM-produced time series in past periods in an attempt to trace elements of falsifiability, which is an important concept in science (according to Popper, 1983, "[a] statement (a theory, a conjecture) has the status of belonging to the empirical sciences if and only if it is falsifiable").

In all examined cases, GCMs generally reproduce the broad climatic behaviours at different geographical locations and the sequence of wet/dry or warm/cold periods at a monthly scale. Specifically, the correlation of modelled time series with historical ones is fair and the resulting coefficient of efficiency seems satisfactory. However, where tested, replacement of the modelled time series with a series of monthly averages (same for all years) resulted in higher efficiency.

At the annual and the climatic (30-year) scales, GCM interpolated series are irrelevant to reality. GCMs do not reproduce natural over-year fluctuations and, generally, underestimate the variance and the Hurst coefficient of the observed series. Even worse, when the GCM time series imply a Hurst coefficient greater than 0.5, this results from a monotonic trend, whereas in historical data the high values of the Hurst coefficient are a result of large-scale over-year fluctuations (i.e. successions of upward and downward "trends"). The huge negative values of coefficients of efficiency show that model predictions are much poorer than an elementary prediction based on the time average. This makes future climate projections at the examined locations not credible. Whether or not this conclusion extends to other locations requires expansion of the study, which we have planned. However, the poor GCM performance in all eight locations examined in this study allows little hope, if any. An argument that the poor performance applies merely to the point basis of our comparison, whereas aggregation at large spatial scales would show that GCM outputs are credible, is an unproved conjecture and, in our opinion, a false one. Our future plan also includes a study of this question after refinement and extension of our methodology.

None of the examined models proves to be systematically better than any other. In particular, AR4 models do not perform better than TAR ones, whereas the concept in AR4 of a scenario produced from the outputs of model runs for the 20th century (20C3M) does not serve well the requirement for falsifiability. In our opinion, however, the unsatisfactory state of the art in climatic (and hydrological) future projections does not reflect a general deadlock in related sciences, but only a wrong direction. Causality in climate and hydrology is not sequential and one-to-one but rather circular (due to feedbacks) and many-to-many (due to complexity). Such causality can be better described in probabilistic and stochastic terms (see Suppes, 1970), rather than in terms of the current deterministic climatic models and practices (see also Giorgi, 2005). Probabilistic and stochastic approaches should not be confused with current multi-model ensemble climate projections (e.g. Tebaldi & Knutti, 2007). A stochastic framework for future climatic uncertainty has been studied recently by Koutsoyiannis *et al*. (2007) in a stationary setting. Arguments that the increasing concentration of greenhouse gases causes nonstationarity (Milly *et al*., 2008) should not discourage stochastic descriptions: after all, nonstationarity is clearly a stochastic concept and, hence, stochastics is the proper mathematical tool to deal with it. For instance, the synchronized

palaeoclimatic data of atmospheric temperature and concentration of greenhouse gases (studied in a different context by Soon, 2007) can be utilized to establish a stochastic relationship between the two processes and test its significance. However, this will require great caution as it is well known that palaeoclimatic data often suffer from quality problems and perhaps unjustified interpretations, and thus involve great uncertainty. Possibly, deterministic climate models could also assist in establishing such a relationship, which, if proven to be significant, could be incorporated in a nonstationary stochastic framework of climatic uncertainty. It is noted, however, that the natural hydroclimatic variability (verified from long series of observations and seen in historical hydrology studies reviewed by Brázdil & Kundzewicz, 2006) is very large and underestimated by classical approaches. Thus, a consistent stochastic approach, under stationary conditions, yields uncertainty limits that well enclose current and future hydroclimatic trends projected by GCMs coupled with hydrological models (Koutsoyiannis *et al*., 2007).

Acknowledgements We are grateful to the Editor, Z. W. Kundzewicz, for his detailed and profound critique, and to R. A. Pielke Sr for his very encouraging and inspiring review. We also thank P. Frank and W. Soon for their informal reviews and L. Gould for a comment on the paper. A previous version of this study (Koutsoyiannis *et al*., 2008a) was thoroughly discussed in Internet forums (a full list is gratefully acknowledged in www.itia.ntua.gr/en/docinfo/850). We thank all discussers, anonymous and eponymous (including J. Crawford, P. Frank, C. Loehle, S. McIntyre, M. Morabito, S. Mosher, L. Ornstein, R. A. Pielke Sr, R. Pielke Jr, J. F. Pittman, G. Schmidt, G. Sherrington, D. Stockwell, J. Tofflemire and D. Wingo) for their comments, mostly positive and encouraging, and sometimes negative. All comments helped us to improve the study significantly.

REFERENCES

Alcamo, J., Flörke, M. & Märker, M. (2007) Future long-term changes in global water resources driven by socio-economic and climatic changes. *Hydrol. Sci. J.* **52**(2), 247–275.

Alvarez-Ramirez, J., Alvarez, J., Dagdug, L., Rodriguez, E. & Echeverria, J. C. (2008) Long-term memory dynamics of continental and oceanic monthly temperatures in the recent 125 years. *Physica A* **387**, 3629–3640.

Blender, R. & Fraedrich, K. (2003) Long time memory in global warming simulations. *Geophys. Res. Lett.* **30**(14), 1769.

Brázdil, R. & Kundzewicz, Z. W. (2006) Historical hydrology – Editorial. *Hydrol. Sci. J.* **51**(5), 733–738.

Bunde, A., Eichner, J. F., Havlin, S., Koscielny-Bunde, E., Schellnhuber, H. J. & Vyushin, D. (2004) Comment on "Scaling of atmosphere and ocean temperature correlations in observations and climate models". *Phys. Rev. Lett.* **92**(3), 39801.

Carter, T. R., Hulme, M. & Lal, M. (1999) Guidelines on the use of scenario data for climate impact and adaptation assessment, task group on scenarios for climate impact assessment. Intergovernmental Panel on Climate Change (ipcc-ddc.cru.uea.ac.uk/guidelines/ggm_no1_v1_12-1999.pdf).

Collins, M. (2002) Climate predictability on interannual to decadal time scales: the initial value problem. *Climate. Dyn.* **19**, 671–692.

Douglass, D. H., Christy, J. R., Pearson, B. D & Singer, S. F. (2008) A comparison of tropical temperature trends with model predictions. *Int. J. Climatol.* (in press, available on line, doi:10.1002/joc.1651).

Edwards, P. N. (2001) Representing the global atmosphere: computer models, data, and knowledge about climate change. In: *Changing the Atmosphere: Expert Knowledge and Environmental Governance* (ed. by C. A. Miller & P. N. Edwards), 31–65. MIT Press, Cambridge, Massachusetts, USA.

Frank, P. (2008) A climate of belief. *Skeptic* **14**(1), 22–30.

Fraedrich, K. & Blender R. (2003) Scaling of atmosphere and ocean temperature correlations in observations and climate models. *Phys. Rev. Lett.* **90**, 108510.

Georgakakos, K. P. (2003) Probabilistic climate-model diagnostics for hydrologic and water resources impact studies. *J. Hydromet.* **4**, 92–105.

Giorgi, F. (2005) Climate change prediction. *Clim. Change* **73**, 239–265.

Hegerl, G., Meehl, G., Covey, C., Latif, M., McAvaney, B. & Stouffer, R. (2003) 20C3M: CMIP collecting data from 20th century coupled model simulations. *Exchanges* **26**, 8(1), Int. CLIVAR Project Office (www.clivar.org/publications/ exchanges/exchanges.php).

Heyen, H., Zorita, E. & Cubasch, U. (1996) Statistical downscaling of monthly mean North Atlantic air-pressure to sea level anomalies in the Baltic Sea. *Tellus* **48** A, 312–323.

Huang, S., Taniguchi, M., Yamano, M. & Wang, C.-H. (2008) Detecting urbanization effects on surface and subsurface thermal environment – a case study of Osaka. *Sci. Total Environ.* (in press, doi:10.1016/j.scitotenv.2008.04.019).

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).

IPCC (Intergovernmental Panel on Climate Change) (2001) *Climate Change 2001: The Scientific Basis*. Contribution of Working Group I to the Third Assessment Report of the IPCC (ed. by J. T. Houghton, Y. Ding, D. J. Griggs, M. Noguer, P. J. van der Linden, X. Dai, K. Maskell & C. A. Johnson). Cambridge University Press, Cambridge, UK.

IPCC (Intergovernmental Panel on Climate Change) (2007), Summary for Policymakers. In: *Climate Change 2007: The Physical Science Basis*, Contribution of Working Group I to the Fourth Assessment Report (AR4) of the IPCC (ed. by S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor & H. L. Miller). Cambridge University Press, Cambridge, UK.

Keenlyside, N. S., Latif, M., Jungclaus, J., Kornblueh, L. & Roeckner, E. (2008) Advancing decadal-scale climate prediction in the North Atlantic sector. *Nature* **453**(7191), 84–88.

Kerr, R. A. (2008) Mother Nature cools the greenhouse, but hotter times still lie ahead. *Science* **20**(5876), 595.

Kitanidis, P. K. (1993) Geostatistics. In: *Handbook of Hydrology* (ed. by D. R Maidment). McGraw-Hill, New York, USA.

Klemeš, V. (1986). Operational testing of hydrological simulation models. *Hydrol. Sci. J.* **31**(1), 13–24.

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

Koutsoyiannis, D. (2002) The Hurst phenomenon and fractional Gaussian noise made easy. *Hydrol. Sci. J.* **47**(4), 573–595.

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

Koutsoyiannis, D. (2005) Uncertainty, entropy, scaling and hydrological stochastics. 2, Time dependence of hydrological processes and time scaling. *Hydrol. Sci. J.* **50**(3), 405–426.

Koutsoyiannis, D. (2006a) Nonstationarity versus scaling in hydrology. *J. Hydrol.* **324**, 239–254.

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

Koutsoyiannis, D. & Cohn, T. A. (2008) The Hurst phenomenon and climate. *EGU General Assembly 2008, Geophys. Res. Abstracts*, vol. 10, Vienna, 11804. European Geosciences Union (www.itia.ntua.gr/en/docinfo/849/).

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

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

Koutsoyiannis, D., Mamassis, N., Christofides, A., Efstratiadis, A. & Papalexiou, S. M. (2008a) Assessment of the reliability of climate predictions based on comparisons with historical time series. *EGU General Assembly 2008, Geophys. Res. Abstracts*, vol. 10, Vienna, 09074. European Geosciences Union (www.itia.ntua.gr/en/docinfo/850/).

Koutsoyiannis, D., Efstratiadis, A., Mamassis N. & Christofides, A. (2008b) On the credibility of climatic predictions: Additional information. Report (www.itia.ntua.gr/e/docinfo/864/).

Kundzewicz, Z. W. Mata, L. J., Arnell, N. W., Döll, P., Jimenez, B., Miller, K., Oki, T., Sen, Z. & Shiklomanov, I. (2008) The implications of projected climate change for freshwater resources and their management. *Hydrol. Sci. J.* **53**(1), 3–10.

Leggett, J., Pepper, W. J. & Swart, R. J. (1992) Emissions scenarios for the IPCC: an update. In: *Climate Change 1992: The Supplementary Report to the IPCC Scientific Assessment* (ed. by J. T. Houghton, B. A. Callander & S. K. Varney), 75–95. Cambridge University Press, Cambridge, UK.

Meehl, G. A., Stocker, T. F., Collins, W. D., Friedlingstein, P., Gaye, A. T., Gregory, J. M., Kitoh, A., Knutti, R., Murphy, J. M., Noda, A., Raper, S. C. B., Watterson, I. G., Weaver, A. J. & Zhao, Z.-C. (2007) Global 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 (ed. by S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor & H. L. Miller), 747–845. Cambridge University Press, Cambridge, UK.

Milly, P. C. D., Betancourt, J., Falkenmark, M., Hirsch, R. M., Kundzewicz, Z. W., Lettenmaier, D. P. & Stouffer, R. J. (2008) Stationarity is dead: Whither water management? *Science* **319**(5863), 573–574.

Nakicenovic, N. & Swart, R. (eds) (1999) IPCC Special Report on Emissions Scenarios. Intergovernmental Panel on Climate Change. Available online at: http://www.grida.no/climate/ipcc/emission/.

Popper, K. (1983) *Realism and the Aim of Science. The Postscript to the Logic of Scientific Discovery,* 1. (ed. by W. W. Bartley III). Rowman & Littlefield, Totowa, New Jersey, USA.

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 (ed. by S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor & H. L. Miller), 589–662. Cambridge University Press, Cambridge, UK.

Shiryaev, A. N. (1989) Kolmogorov: Life and creative activities. *Annals Probab.* **17**(3), 866–944.

Soon, W. (2007) Implications of the secondary role of carbon dioxide and methane forcing in climate change: past, present, and future. *Phys. Geogr.* **28**(2), 97–125.

Suppes, P. (1970) *A Probabilistic Theory of Causality*. North-Holland, Amsterdam, The Netherlands.

Tebaldi, C. & Knutti, R. (2007) The use of the multi-model ensemble in probabilistic climate projections. *Phil. Trans. Roy. Soc. A* **365**, 2053–2075.

US National Research Council (2005) Radiative forcing of climate change: expanding the concept and addressing uncertainties. Committee on Radiative Forcing Effects on Climate Change, Climate Research Committee, Board on Atmospheric Sciences and Climate, Division on Earth and Life Studies. The National Academies Press, Washington DC, USA.

Wetterhall, F., Halldin, S. & Xu, C-Y. (2005) Statistical precipitation downscaling in central Sweden with the analogue method, *J. Hydrol.* **306**, 174–190.

Received 30 May 2008; accepted 10 June 2008