Reply to review comments R. T. Clarke:

We would like to thank Professor Clarke for his review comments which helped us improve the quality of our manuscript. We realize that we needed to add some comments and justifications for the applied statistical analyses. Some of the analyses have been removed from the manuscript as the criteria for a Gaussian distribution and sample independency could not be met. Please find our specific comments below.

General comment

This is a very substantial piece of work which uses the meteorological time series outputs from no less than 12 GCMs as inputs to the distributed global hydrological model PCR-GLOBWB, run on a daily time-step with spatial resolution 0.5o. Two sets of meteorological time series from each GCM were input to the hydrological model: those from (i) the 20C3M experiment for the period 1971–1990 ("past"), and (ii) the SRES scenarios A1B for the period 2081–2100 ("future"). From the runoff generated by the hydrological model, six flow statistics were calculated (min , mean, max, peak, Var, RC) as defined in the paper's Table 2, averaged over the two 20-year periods past and future. For each of the 6 statistics and 12 GCMs, relative changes of the type _Qij = (ij, future - ij,past) / ij,past were computed for each 0.5o grid-square, where i=1. . .6 denotes the statistic and j=1. . .12 the GCM. Finally, unweighted means i over the 12 GCMs were calculated for each of the six derived statistics, and these means were then tested to assess whether they differed significantly from zero, using paired-sample t-tests. It is understood that the test was made for each 0.5o grid-square. For each GCM individually, past and future means were also compared by t-tests, after reducing the number of years in each of the two 20-year periods to an "effective sample size" to allow for serial correlation between years in each of the two periods. The authors find "a consistent decrease in runoff for southern Europe, southern Australia, the south and north of Africa and southwestern South America. Significant discharge decreases are also projected for most African rivers, for the Murray and for the Danube. Runoff increases are projected for sub-Arctic and Arctic regions and an advance in phase in the annual cycle is projected for the sub-Arctic regions." Although these conclusions are said to be similar to those reported by others, new aspects of the work are "(1) the comparison of spatial patterns of regime changes and (2) the quantification of consistent significant change calculated relative to both the natural variability and the inter-model spread."

The scope of the authors' study is global, with emphasis on 19 of the world's major river basins, and results are clearly presented. However this Reviewer must take issue with the authors on a number of points, as set out in the following section.

Specific comments

1.Noting that "Projections of different GCMs diverge widely", underscoring "the need [to use] a multi-model ensemble", the Authors use the meteorological time series produced by 12 GCMs for input to the single global hydrological model PCR-GLOBWB, a model "showing similar performance to other global hydrological models". However its parameterization "is based on best available global datasets and so far the model has not been calibrated." Like GCMs, however, hydrological models can also produce widely varying outputs even when calibrated, and the Authors have touched on this point in their earlier paper (Hydrol. Earth Syst. Sci., 14, 1595–1621, 2010, section 2.1). To be fully credible, therefore, the Authors' results need to be confirmed by other global hydrological models; when they argue for the use of a multi-model ensemble of GCM outputs, the same argument surely applies to global hydrological models (recognizing, however, that the computational effort that this would require is formidable).

We do understand the reviewers concern here. Indeed the hydrological model introduces uncertainty as well and results may be different when other hydrological models are used. Yet, unfortunately a multiple hydrological and general circulation model analysis is out of our reach. As the study of Gosling et al. also concluded, the spread in projections obtained from multiple GCMs is in general larger than the spread obtained from multiple hydrological models, therefore we here focus on the GCM uncertainty. We added a comment on the relevance of hydrological model uncertainty to the Synthesis section.

2. The Authors state that one of the two new aspects of their work is "the quantification of consistent significant change calculated relative to both the natural variability and the inter-model spread" and it is with the "significance" of the changes, and in particular their use of t-tests to assess significance, that the Reviewer wishes to take particular issue. Beginning first with the most trivial point, Section 2.3.3 of the paper says "significance is tested for a significance level of 95 %." Here there is the common confusion between significance level and confidence probability found in much climate-related literature, as reported long ago by von Storch (1995) and von Storch and Zwiers (1999). The Authors' "significance level of 95%" should be corrected to "significance level of 5%"; the smaller the significance probability, the stronger is the conclusion that an observed difference is "real" and not due to random variability.

We realize we applied the t-statistics here while not all required criteria were met. As a reply to the reviewer's comments below we removed these significance values (which are calculated relative to the ensemble spread by the changes obtained from the 12 GCMs) from the manuscript. This because the, for the t-statistics required normal distribution, can not be guaranteed for the changes derived from the12 GCMs.

Even allowing for this change, however, the reader is confronted with other difficulties when attempting to interpret the Authors' Table 4 which shows percentages of change in the six flow statistics in 19 of the world's major drainage basins; the legend to this table says "If applicable the significance level (sig) for which change is significant is given as well." Quite a number of the entries in columns headed "sig" are missing, whilst others are reported as 80%, 70%, 60% and even 50%. Even allowing for the fact that the "significance level" shown should be subtracted from 100, what would "sig" values of 20%, 30%, 40% and even 50% mean? And why are there missing values in both the "sig" columns, and the _Qmin column?

As stated before, these significance values have been removed from the manuscript.
We added a note underneath the table. The minimum flow becomes zero for several GCMs for the Zambezi and Niger, therefore % changes could not be derived and is not included in the table.

3. A more serious comment concerns the use of t-tests to establish the "significance" of the differences ΔQ (i.e., the means of QGCM fut,j −QGCM past,j, j=1. . .12, in the Authors' notation) between flow statistics for the A1B scenario and the 20C3M control period, averaged over the 12 GCMs. Use of a t-test for this purpose requires that (i) the 12 ΔQ values are a random sample from a population of ΔQ values (taken from a hypothetical population of GCMs); (ii) each ΔQ has the linear structure $\Delta Qi = \mu + "j$ (j = 1. . .12) where μ is the mean of the hypothetical population of GCMs, and "j is a Normally-distributed random variable with constant variance _2. The t-tests used in the paper then test the hypothesis that $\mu=0$, for each of the six derived statistics. Some of the six statistics (mean, max, min....) in the Authors' Table 2 will be at least approximately Normally distributed, since they are means. Others, such as Var and the ratios in the table, will not be: nor will the relative changes (future - past) / past. It is not obvious that these assumptions - random sampling from a hypothetical population of GCMs; μ a constant; "j a Normally-distributed random variable with zero mean, constant variance _2) are justified, and if

they are not, conclusions regarding the "significance" shown by t-tests must be open to doubt. There is, of course, no problem with calculating t-statistics, such as the t=xd/[sdpM] in the Authors' equation (5), provided that these are regarded merely as indices of change; but the problem arises when such t-statistics are used to test for "statistical significance" under conditions for which such tests are not justified.

In the revised manuscript we only apply the t-statistics for the calculation of the significance of change obtained from a single GCM relative to its own inter-annual variability to identify notable changes. As indeed it can not be assumed that the 12 GCMs are completely independent due to similar parameterizations and use of the same numerical methods.
The remaining significance calculations are restricted to the mean, minimum and maximum discharge. For these variables we added a statement on the use of the t-test as the distribution of the 20 annual average discharge values may indeed not be normal. See section 2.3.3 (former 2.4.3) where the t-statistics are introduced.

4. One of the requirements for a t-test to be valid is that the quantities that constitute the random sample (of the 12 differences _Q, in the present context) should be statistically independent. When this is the case, the calculated t-statistic will be based on M-1=11 degrees of freedom (df), where M, the number of GCMs, is 12. If the test assumptions were valid, a mean difference would be judged "significantly different from zero" if the calculated t-statistic exceeded 2.201 in absolute value (for a two-sided t-test) or 1.796 (for a one-sided test: the paper does not state which was used). But it is not obvious whether the df are really 11, or some smaller number, because it is not certain that the 12 values of a statistic Q from the 20C3M experiment are statistically independent: similarly for the 12 values of Q from the A1B scenario.
The last paragraph of section 2.1 of the paper explains that "to overcome initialization problems, initial states [were] obtained for each GCM data-set individually. For the control climate experiment and the future scenario, PCR-GLOBWB was initialized with the first ten years of data starting with the 10 initial states obtained from a 30 yr run based on CRU TS2.1 monthly time series. . .. downscaled to daily values using ERA-40 re-analysis data. The end-states of the ten year during GCM runs are used as initial states for the 20 yr GCM scenario runs." The meaning of this is not absolutely clear, but the Reviewer's interpretation is that, although the hydrological model PCR-GLOBWB had been initialized differently when the meteorological time series from the 12 different GCMs were fed into it, these initial states had all been calculated originally from the same dataset (the downscaled CRU TS2.1). Thus the initial states for the 12 runs of PCR-GLOBWB appear to have been computed from the same data, and in theory will not be statistically independent; they will be inter-related, in some very complex way, though their mutual dependence on the data used to compute them. It may be that the degree of dependence between (say) the 12 values of a statistic Q from the 20C3M experiment is so slight that they can safely be assumed independent of their values under the A1B scenario, but this is not obvious.

Here we need to clarify the calculation of the initial states. These are generated in a two step approach. In the first step the hydrological model is spin-up with a 30 year run based on a combined dataset created from the CRU TS 2.1 and the ERA-interim datasets. The end-states of this run are used as initial states for the second step. In this second step, the hydrological model is run for a 10 year period with data from the specific GCM, to create independency between the runs with the individual GCMs. The end-states of these ten year runs are used as initial states for the hydrological model runs for the individual GCMs. In summary this means that each GCM based run has its own initial conditions which are derived from data of that specific GCM.

This text has been added to the manuscript at the end of section 2.1.

In the revised manuscript we only apply the t-statistics for the calculation of the significance of change obtained from a single GCM relative to its own inter-annual variability. As indeed it can not be assumed that the 12 GCMs are completely independent due to similar parameterizations and use of the same numerical methods.

The same issue of lack of independence arises if, as it appears from the text, t-tests for the statistical significance of were made in each 0.5o grid square, since the in adjacent (and not so adjacent) grid-squares will be spatially correlated.

We believe that within the calculation of the t-statistics the adjacent cells can be seen as independent. Within the t-test the cells are treated as independent systems and, information of adjacent cells is not included. We do agree that the discharge values in individual cells are correlated to their up- and downstream neighbors. Yet, if within the calculation of the t-statistics, only the discharge time-series of one specific cell are used, this can be interpreted as an independent system.

5. When comparing the mean value of a statistic Q calculated from the 20-year 20C3M past run with its value under the 20-year A1B future scenario for each GCM, the Authors used a modified form of t-test in which an "effective sample size" was calculated for each of the two 20-year sequences, to allow for the possibility of serial correlation between the values within them. The t-tests used in the paper to compare the past and future means of any flow statistic Q are based on similar assumptions to those mentioned above: the annual values (e.g., Qpast,j, j=1. . .20) are assumed to be a random sample from a population with mean μpast and variance _2, and the 20 annual values Qfut,j are regarded as a random sample with mean μfut and (the same) variance _2. The t-test shown in equations (7) and (8) of the paper, if it were valid, would then test the hypothesis μpast = μfut. The use of effective sample sizes nfut* , npast* as a procedure for allowing for serial correlation between annual values does not, in this Reviewer's opinion, adequately compensate for failure to satisfy the remaining assumptions.
Are there statistical procedures for comparing means which have less restrictive assumptions than Student's t-test? Yes, but they require the data values that make up the two groups to be statistically independent. Given independence, a permutation test shows whether the observed difference between two means lies in the tails of the empirical distribution of differences found be permuting the data, giving a measure of whether the observed difference could have arisen by chance. This would not require that the 12 GCMs be a random sample from a hypothetical population of GCMs; nor would it require the assumptions of Normality and homogeneous variances. To conclude, several papers have drawn attention to limitations of classical statistical methods for the analysis of hydrological and hydrometric data (Koutsoyiannis et al., 2007; Koutsoyiannis and Montanari, 2007). In their perceptive discussion of trend detection in hydroclimatological time series, Cohn and Lins (2005) also wrote "From a practical standpoint . . .. it may be preferable to acknowledge that the concept of statistical significance is meaningless when discussing poorly understood systems." When the "data" to be analysed are not measurements recorded by human observer or instrument, but are simulations produced by GCMs and/or hydrological models, the opinion of this Reviewer is that the use of classical statistical procedures is even more open to question. The "data" then being analyzed may be deterministic, in the sense that a model will reproduce the same sequence, given the same initial values, without any random component: and even if a random component has been added post hoc to a simulated output, this random component may have little to do with the randomness found in natural processes. Much more work is required to assess the uncertainties – that is, to calculate standard errors - in measures of future hydrological change, when such measures are derived from model simulations.

In the revised manuscript we only apply the t-statistics for the calculation of the significance of change obtained from a single GCM relative to its own inter-annual variability to identify notable changes. For this t-statistics the equal variance is not required ($\mu past = \mu fut$). In equation 8 we calculate the combined standard-deviation and use this as input to the t-test in equation 7. Within this t-test we test whether there is a difference in the twenty-year average discharge for the current and future climate.

References

Cohn, T. A., and H. F. Lins (2005), Nature's style: Naturally trendy, Geophys. Res. Lett., 32(23), L23402, doi:10.1029/2005GL024476.

Koutsoyiannis, D., A. Efstratiadis, and K. Georgakakos, Uncertainty assessment of future hydroclimatic predictions: A comparison of probabilistic and scenario-based approaches, Journal of Hydrometeorology, 8 (3), 261–281, 2007. Koutsoyiannis, D., and A. Montanari, Statistical analysis of hydroclimatic time series: Uncertainty and insights, Water Resources Research, 43 (5), W05429, doi:10.1029/2006WR005592, 2007.

Von Storch, H. (1995) Misuses of statistical analysis in climate research. In: Analysis of Climate Variability: Applications of Statistical Techniques, edited by von Storch H. And Navarra, A., pp. 11–26. Springer-Verlag, Berlin.

Von Storch, H., and F. Zwiers (1999), Statistical analysis in climate research, Cambridge University Press, Cambridge.