<u>Responses to Dr. C. H. David Referee #2</u>

The following is a review of manuscript HESSD-10-2117-2013 entitled "Detection of global runoff changes: results from observations and CMIP5 experiments" by R. Alkama and collaborators. This study investigates the trends and statistical signify cance of potential changes in stream flow for 8 regions of the globe. Three types of stream flow data are used: 1) stream flow estimations from in-situ observations where no data gaps exist in the study period, 2) reconstructed stream flow estimations from in-situ observations with gaps filled, and 3) modeled stream flow from Atmosphere-Ocean General Circulation Models. 1) and 2) are based on a dataset collected and gap-filled by Dai et al. (2009), and 3) uses modeled runoff from CMIP5 runs. The method used to evaluate the statistical significance of potential changes is the Temporal Optimal Detection (TOD) of Ribes et al (2010).

The text and the figures of the manuscript are generally clear; with the few exceptions mentioned below. In my opinion, the work presented is important because it sheds further light on the unresolved issue of detecting potential global stream flow changes and because the conclusions seem well supported by the study. The authors do a commendable job at presenting the strengths and limitations of previously-published approaches and also those of their own work. I was particularly seduced by the thorough checking of results every step of the way from raw observations to gap-filled observations to model results. In doing so, the authors really give a good picture of how much is known, and how much isn't. Also, I really liked the authors' ability to present separate regional trends while keeping a single value of the statistical significance test globally.

However, I find that the article would likely benefit from a more rigorous use of some terms, clearer explanation of the statistical method used, and stronger justifications of why the assumptions made in the statistics test are valid. Also, I did my best to document a few typographic errors as I know (first-hand!) that technical writing in English as second language is a challenge, but I would suggest the authors to further check their manuscript prior to their future submissions. I hope that my comments below given in decreasing order of importance can provide further information on potential ways to strengthen the manuscript.

Granted that these modifications are addressed, I would enthusiastically recommend this manuscript be accepted as a full HESS paper.

We thanks the reviewer for his positive comments

Specific Comments.

It seems to me that the word "significant" is used interchangeably to qualify whether a given change is of large magnitude (usually referred to as "significant"), or whether the change is shown to be not attributable to chance ("statistically significant"). These are two different things as change can be very small but highly statistically significant, and vice versa. I would encourage the authors to go over the manuscript and check that every instance of "significant" is further qualified in these terms. The choice of words is critical here, particularly in light of the magnitude of the spread in multi-model stream flow computations compared to the magnitude of the modelled change.
Ok, done. All of "significant" meaning large magnitude are replaced by "important or considerably" in the manuscript.

I can appreciate that the authors do not want to burden the reader with excessive information on the TOD test since it is available in published literature. However, I found myself wondering "what exactly does the TOD test do?" a few times and I suggest that a little more information would be beneficial to clarity since the TOD test is a central part of the study. Maybe splitting the presentation of TOD into two distinct paragraphs may help: 1) a general paragraph on what the test does, its inputs, its outputs, and what the assumptions are; and 2) a specific paragraph on how it is applied to stream flow and why the assumptions of the test hold in this study. My understanding is that the test assumes the spatial-temporal behavior of equation (1) for stream flow and in which for each given location, the variations of stream flow are a simple first order linear system and to which is added variability. Further assumptions are also made on the shapes of functions x(t) and epsilon(s,t). With these assumptions, the TOD test seems to provide (for any given temporal range) one value of magnitude of the trend for each region of the world, and one unique p-value for the statistical significance of all global trends combined. This understanding is based on multiple readings of the manuscript, but I admittedly am unsure that I understood fully. I still have a hard time grasping the full meaning of Figure 2 which apparently justifies the applicability of the approach and the values used.
This section has been widely re-written, and split into two subsections, as suggested by the referee. The understanding of the referee was right, and the new version of the manuscript explains it more clearly.

The choice of the function x(t)=t is fine since it allows to find the slope of a linear trend and because such linear trend values are easily understood by all and as such are very powerful scientific information. However, the authors justify this choice because "the non-linearity of the change is probably not the dominant feature" without providing further explanation or reference to published literature. Maybe would an argument related to the simplicity of the approach be more appropriate unless published literature can support this.
We do agree with the referee, and justify our choice as follows: "In order to base our study on a very simple temporal pattern that is not model-dependent, we used only linear trends *(ie x(t)=t)*." We then further comment on previous published literature on this specific choice: "In addition to be very simple, this choice is consistent with several previous studies dealing with potential changes in the globally hydrology *(e.g. Labat at al. 2004; Gedney et al. 2006; Dai et al. 2009, Alkama et al. 2010, 2011)*." The comment regarding potential non-linearity is not removed, however, as, in our view, it provides a useful additional information.

I wish I was more familiar with "auto-regressive process of order 1", "red noise", "non-white". These terms may be well known in the fields of frequency analysis and climate detection but I feel many hydrologists and earth system scientists (including myself) would benefit from a thorough definition of these terms. Including equations would likely help.
The AR1 processes, which are equivalent to red noise, are now more carefully introduced and defined, in Section II.2.

The "red noise" structure of epsilon(s,t) is central to the argument made. Please provide further definition of the Hurst phenomenon (and a reference) and how it leads to "red noise" internal variability of stream flow.
This section has been revised, with further explanation on the "red noise" structure. The « Hurst phenomenon » was inappropriate here and has been removed. Hurst phenomenon is related to long-range memory effect, while red noise assumes short range memory effect. In

our view, there is no evidence of long term memory effect from pre-industrial control simulations.

Please clarify lag-one autocorrelation. Is this a lagged autocorrelation for each gauge (basin) using a one-month lag? How is this computed? The value of alpha is used in the arguments made and I don't understand how it is obtained. Please provide an equation relating modeled stream flow of each basin (all basins?) and alpha. My understanding from the text is that the lagged-autocorrelation of stream flow at a gauge decreases exponentially with the increasing value of the lag ("red noise"?) and that this exponential decrease is characterized by a factor alpha.
Alpha represent the autocorrelation of one-year lag. This is clearly stated in the new manuscript. For each Picontrol simulation, alpha is estimated as the correlation between y(t) and y(t-1), on the global mean time-series (Fig 2). Allpha has also been estimated, for each model, over each of the 8 studied zones (not shown). These values are generally greater than 0.1 and did not exceed 0.3. Our choice of alpha =0.2 is coming from the analysis of the global mean time-series in Figure 2a, and from Fig 2c, which shows that the distribution of the P-value is uniform only in the case of alpha =0.2.
These explanation are added in the text (see section II).

Why is alpha estimated from CMIP5 model runs instead of from available observations?
It would help further justify the validity of the structure of epsilon and of the valid values for alpha if it did not come from model runs. If observations don't allow to obtain alpha or give unexpected values, please mention this.
In each detection study it is crucial to separate the natural internal variability from total signal. The goal is to be sure that the detected change cannot be explained by the natural internal variability alone. Alpha is a memory factor which should be estimated from internal variability only (in particular, without anthropogenic forcings). Due to the difficulties to separate the natural and anthropogenic signal in the observations, one alternative commonly used in D&A studies is to use Picontrol simulations. Indeed, in Picontrol simulations all of the boundary conditions (green house gases, land use, aerosols, …) are constants.
These information's are added in the text (see second paragraph pf the section II.2.b).

The magnitude of the relative bias in modeled stream flow is estimated at 25% (P2128, L12) and at 50% (P2131, L14). Looking at Figure 4 it seems to me that some basins easily reach 100 to 200% relative bias. I am not shocked by such large spreads as stream flow computations can still be improved in climate models, but I would highly welcome the addition of a table summarizing the average value for all models and for observations for all regions, and the bias for all models and for all regions.
We think that figures are more speaking than large tables (14 models + observations over 8 regions). Two figures showing the error in percent of both mean and standard deviation of river discharge are included in the current version of the manuscript (see section III.2 and figures 6 and 7).

Technical corrections.

P2118, L10. Please provide a reference for CMIP5. P2118, L10. Please define the acronym RCP and provide a reference for it.
Ok, done.

P2126 L14-18, is there a reference for these values or were they computed by the authors? I can’t tell from the text.
The global land area is known, where the area of the river basins used in this study are coming from Dai et al. (2009). The river discharge of the individual rivers and the estimation of the global land discharge are also coming from Dai et al. (2009).

Please clarify. P2118 - L18. “rivers discharges” should be replaced by “river discharge”.
Ok, done.

P2118- L19. Do the authors mean “sought” or “thought”? P2118 - L26. “River gauged stations” should probably be replaced by “gauging stations”. This differs from a "gauged river" which is a river that has a gauge (or gauging station). I’ve never seen these three words used together.
P2119 – L6. Labat et al. (2005) is listed in references as 2004. Which is it?
P2119 – L20. “Hight latitude” should be “high latitude”.
Ok, done.

P2120, L11-15. “First”, “Secondly”, and “Third” please use the same form for all three.
Ok, done.

P2122, L24. “late” should be “latter”.
Ok, done.

P2122, L28. Not sure “up stream rivers” is the right use of words. Maybe “head waters”?
“up stream rivers” is replaced by “upstream rivers”

P2123, L19. “rich” or “reach”?
P2123, L20. “actual” should be “current”.
P2123, L27“African’s” should be “Africa’s”.
P2124, L9, “coherent” should be “consistent”.
P2130, L22, “where” should be “were”.
Ok, done.

Fig 1: I can’t see circles.
The circles are clearly shown in the new figure 1.

Fig 3: left and right panels are inverted in the legend. Please include the time frame at which the trend is calculated also in the right-hand side map (not only the legend) to help clarify.
Ok, done (see new figure 3).

We thanks the reviewer for his careful reading which help us to improve our manuscript.

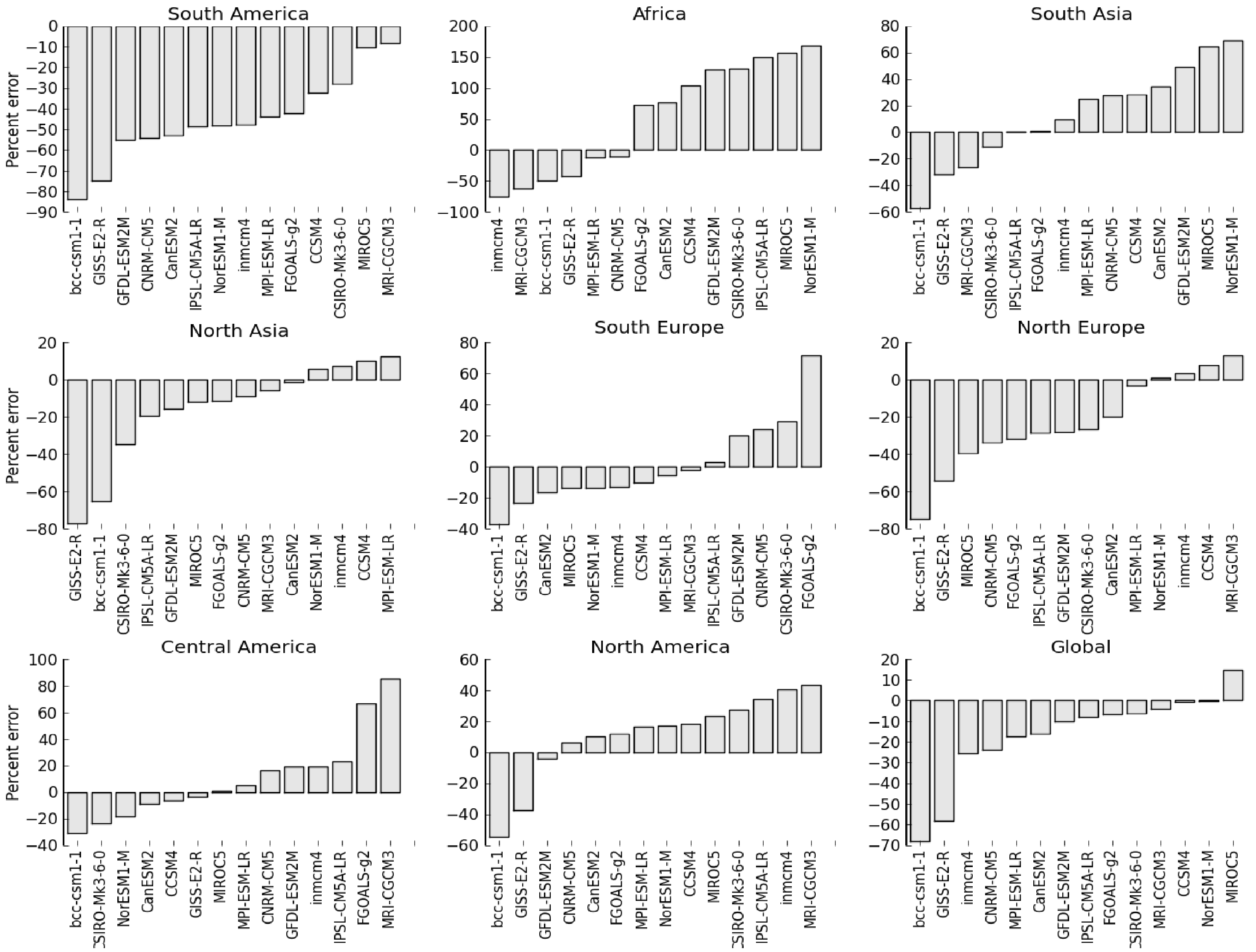

Fig III. Percent error (100(Sim-Obs)/Obs) of simulated runoff averaged over 1958-1992.

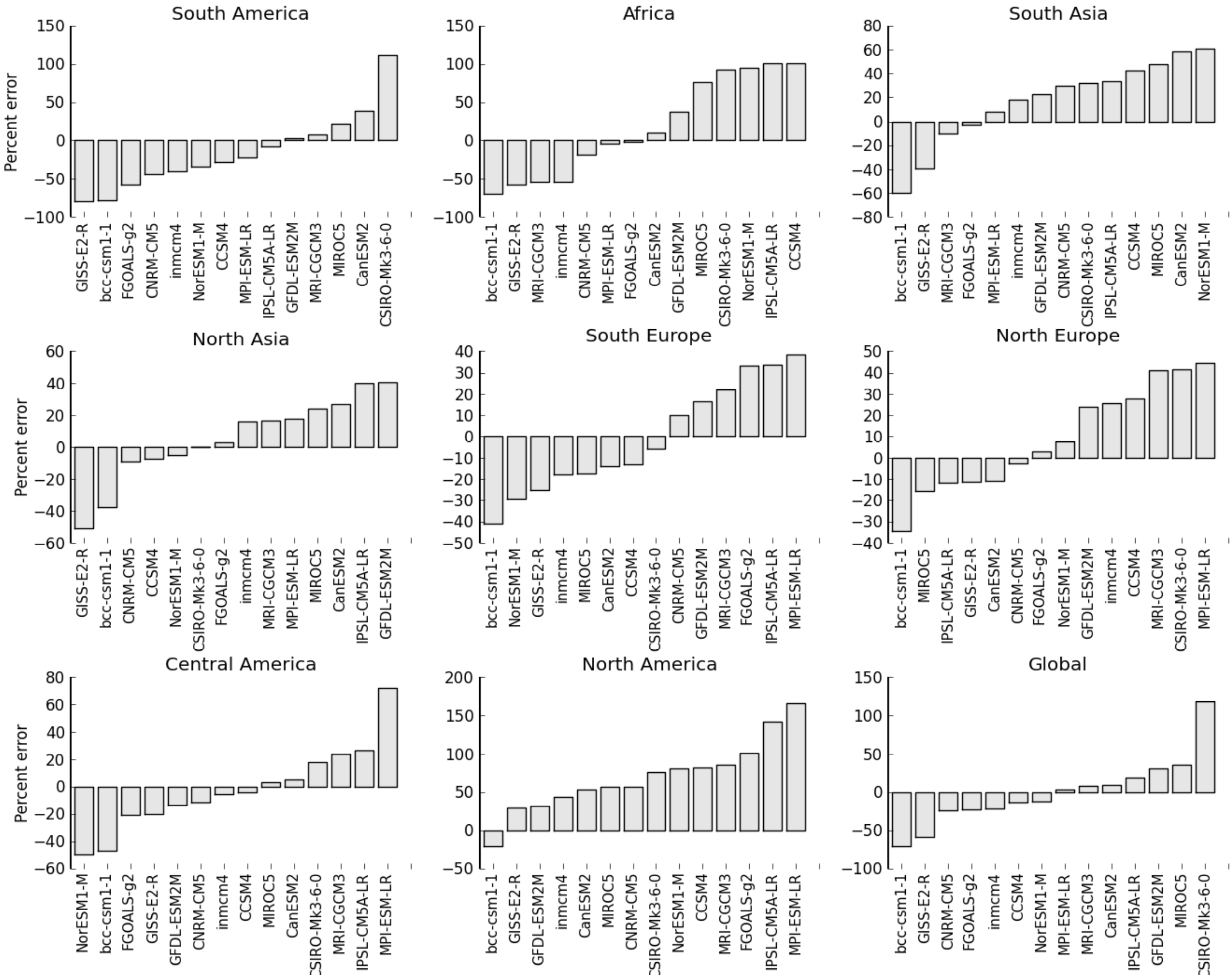

Fig IV. Percent error (100(Sim-Obs)/Obs) of the standard deviation runoff simulated by CMIP5 models over 1958-1992.