Hydrol. Earth Syst. Sci. Discuss., 8, C2776–C2786, 2011
www.hydrol-earth-syst-sci-discuss.net/8/C2776/2011/

Hydrology and Earth System Sciences Discussions

Interactive comment on "Skill assessment of a global hydrological model in reproducing flow extremes" *by* N. Candogan Yossef et al.

N. Candogan Yossef et al.

ncandogan@hotmail.com

Received and published: 14 July 2011

Comment: This is a well-written paper which addresses an important research need within the hydrological modelling community to corroborate hydrological extremes simulated by global hydrological models. The paper is clearly written and utilises some effective means of data analysis, which are to be commended. However, the use of bias correction on hydrological model output may prevent the publication of this manuscript. This approach is highly questionable in the circumstances it is used, and without further elaboration on the reasoning for this or details of it, it is difficult to approve this paper for publication. The content of the paper is also weighted too heavily towards data and methodological considerations, leaving insufficient time for description and explanation

of results. In general, there is a lack of rigour in the discussion, with results presented often without providing any explanation. Should these issues be resolved, there is enough promise demonstrated by this paper that it could proceed towards publication (following major revisions).

Reply: We thank Anonymous Referee 2 for appreciating the clarity of our paper, the efficiency of the analysis and the importance of the research need we address. We also thank him for his very useful comments which helped us improve our paper. In this reply we would like to present our reasoning on the issues that he raises and hopefully resolve them. Before addressing the technical comments, we would like to clarify one point. The main purpose of this exercise is to test the maximum skill that can be achieved (when the systematic bias is eliminated), so that we have an indication of the potential skill which we may expect in actual forecasting, after a simple post-processing.

Comment: The main issue with this paper is the use of bias correction of global hydrological model (GHM) output. The presentation of raw output from the GHM (i.e. not bias corrected) in this paper does not give any consideration to the possibility that there may be issues with the GHM (structure, parameterisation, etc.), and I am not sure I agree with the assertion that non-bias corrected data look 'reasonable'. I find it difficult to believe that the often vast differences between simulated and observed data are entirely determined by the climatic input.

Reply: We mention in Section 3.1 that "Model bias due to errors in the input data, model parameters, or simplifying assumptions, can highly degrade the quality of the output of a hydrological model." This is true for our model as well. We discuss possible reasons for the errors in the simulations in Section 4.1, where we mention some of the known problems in the climatic input. We indeed do not discuss what the possible issues with the GHM may be. This point is discussed in other papers by our research group (e.g. Van Beek, Wada and Bierkens, 2011), but it is not the subject matter of this paper. However, we certainly acknowledge that the bias is partly due to model

errors and we have modified the text to make this clearer. We have also removed the assertion that the non bias-corrected simulations are in reasonable agreement with the streamflow records for most river basins, because this is somehow subjective.

Comment: If bias correction is justifiable, the reasoning given for its use (Pg 3476, Ln 5-6) is too brief and not sufficient, and there is insufficient information provided on the correction procedure. I am not surprised that bias correction improves the simulated data (Fig. 3), given that the procedure is simply based on comparing to the mean monthly flows of other years. Some of the non-bias corrected hydrographs suggest that it may not be necessary to bias correct. Given that this paper is supposed to be a 'skill assessment of a global hydrological model', it appears that any analysis of data after bias correction is more of an investigation of how successful bias correction has been. To truly explore the ability of the model in reproducing flow extremes, it would be more pertinent to analyse the non-bias corrected data, which should give a better reflection of potential shortcomings in the skill of the GHM. See Haddeland et al. (in references) for an example of a study which assesses GHM outputs (without bias correction) against observed data, attributing differences between models and observations to varying model structure. See also Doll et al. (2003) for an example of a study of a GHM that analyses model skill against observations using non-bias corrected data. The WATCH Special Issue of the Journal of Hydrometeorology also features a number of papers within which GHM outputs have been used without bias correction. In this instance, it seems as though bias correction may have been adopted in order to improve agreement between simulated and observed data presented in the hydrographs, whilst turning a blind eye to the issues with the raw simulated data and therefore the GHM.

Reply: We agree with the referee that the reasoning for the use of bias correction is too brief and insufficient in the text, therefore we elaborate this further. The explanation of the correction procedure is also brief but we believe is sufficient, since it describes the method fully.

We would like to try to convince the referee that in this case it is justifiable to bias correct the simulation results. First of all, we would like to emphasize that we use uncorrected data for the skill assessment in reproducing monthly anomalies and extreme events. For the categorical and binary contingency tables, thresholds for observations and simulations are calculated separately so systematic errors are eliminated and we can use the simulation results without bias correction.

For the skill assessment in reproducing hydrographs, we use uncorrected results as well as bias-corrected results, and we apply the verification methods on both sets of results. This serves two purposes. Verification with uncorrected data presents a better reflection of potential shortcomings in the skill of the GHM as the referee correctly points out. It also gives the reader the opportunity to compare our simulations with the results of other studies which use uncorrected data, such as the ones mentioned by the referee. Verification with bias-corrected data, on the other hand is relevant for the assessment of forecasting skill, which is the ultimate purpose of this study. It provides an indication of the maximum skill that can be achieved when the systematic bias is eliminated (whether it is due to model errors or forcing). It is also necessary for the sake of consistency, since the two following types of skill assessment eliminate systematic errors inherently.

Comment: The selection of some of the observed data from the GRDC can also be questioned. Some of the observed river flow timeseries are very short (e.g. the Ganges, Indus, Brahmaputra and Zambezi). For the Zambezi in particular, it is difficult to validate model output for 1958-2001 on four years of data, and removing this would still leave sufficient geographical coverage of four basins in Africa. If two basins are sufficient for North America and South America, then four should certainly suffice for Africa. It seems strange to have five of the twenty basins, a quarter of the total, in Africa, with far fewer per continent elsewhere. It is also unclear why the Murray, Zambezi and Parana have been selected when the GHM cannot simulate agricultural impacts on the hydrological regime. If the model cannot hope to produce reasonable

results for these three basins, this does not appear to add much to the paper. Furthermore, I would have expected more than three of the selected basins to be affected by artificial influences. Perhaps consider including 16 basins, and use a minimum criterion for data availability of at least half of the 1958-2001 study period.

Reply: We explain our criteria for the selection of basins in section 2.3. We indicate that we want to represent all the continents, a wide range of climate zones and latitudes as well as a variety of precipitation regimes. It is true that Africa is over-represented with five basins whereas only two basins are selected from North America for instance. Our actual selection of basins can be explained again by concerns about consistency with our forthcoming studies. For our ultimate purpose of operational seasonal forecasting, we are more interested in the developing regions of the world which lack efficient forecasting systems. Our preliminary research into the potential value of forecasts show that some of the basins with short observation records, and high artificial impacts are those for which forecasts could be most valuable. For this reason we prefer not to remove these basins from our current analysis.

Comment: A further issue is the use of monthly data itself in a study on the ability of a GHM to reproduce hydrological extremes, for which daily data are most appropriate. Twelve of the twenty basins selected have daily data on the GRDC for at least half of the study period, including some of those basins that have very short monthly records (e.g. Zambezi). It may also be possible to find medium-sized basins which do have daily data for more than half of the timeseries, with the gain of validating using daily data outweighing the loss of areal coverage. Given that the GHM runs on a daily timestep, and that daily discharge data appear to be available, and that the study concerns hydrological extremes, it may be prudent to analyse the model on a daily timestep; currently, aggregation to the monthly timestep represents a loss of information on model performance and does not capitalise on the full extent of the opportunities in this study.

Reply: To our knowledge, no GHM has been tested on daily data, without basin-specific

calibration. For this reason, although our GHM runs on a daily time-step, and even for the basins for which daily records are available, we restrict our analysis to monthly data.

Concerning hydrological extremes, for forecasting high or low flows, we believe a monthly time-step is appropriate since these extremes are determined to a large extent by persistent characteristics. A daily time-step seems to be more suitable for floods, whereas a monthly time-step is more appropriate for droughts. It can be argued that a monthly time step is too coarse to correctly predict flood sizes. However we demonstrate in Appendix 1 that monthly high flows will certainly be indicative for increased probability of floods for large rivers. The skill we want to assess in this section is the skill in forecasting increased probabilities of flow extremes rather than exact discharges, and we want to make these forecasts on monthly/seasonal lead times. For this reason we believe a monthly time-step is more appropriate and more promising.

Comment: The use of terminology such as ‘forecasting’, ‘hindcasting’ and ‘reproducing’ needs to be rationalised throughout the paper. It is important to note that anything related to past hydrological extremes cannot be termed forecasting, as it is on some occasions throughout the paper. Since the GHM is being used to simulate past observed hydrological extremes, I would favour the term ‘reproduce’.

Reply: We have changed the text to use only the term ‘reproduce’ for simulation of past extremes, and removed the term ‘hindcasting’. The term ‘forecasting’ is used only to refer to future predictions.

Comment: Pg 3470, Ln 22-24: I would be careful with this statement (repeated towards the end of the conclusion; Pg 3485, Ln 17-18). This paper has shown that there is some potential for reproducing past hydrological extremes, but that is somewhat different to using the GHM to forecast monthly river flows into the future.

Reply: We agree with the referee that demonstrating potential skill in reproducing past hydrological extremes is different to using the GHM to forecast monthly river flows

into the future. The conclusion of this paper is that "the prospect for using PCR-GLOBWB for monthly and seasonal hydrological forecasting is positive". However we also add that "This assessment in hindcast is a preliminary one; and it shows a potential skill given the current GHM, with a meteorological forcing based on observations.". We also mention that the true skill should be assessed in forecasting mode using monthly/seasonal meteorological forecasts. This is what we intend to do in our forthcoming studies.

Demonstrating skill in reproducing past hydrological extremes does not guarantee a skill in actual forecasting. Nevertheless we believe that for a forecasting window limited to few months, the demonstrated skill indicates a positive prospect, which should be tested further.

Comment: Pg 3477: the use of 25th and 75th percentiles suggests that anomalous flows occur 50% of the time (25% highest flows, 25% lowest flows), which would make them just as likely as 'normal' flows, therefore not particularly anomalous. For an indication of high flow and low flow anomalies, I would suggest using 15th and 85th percentiles as a minimum (perhaps even 10th and 90th), which would be a better test of model skill in reproducing anomalous flows. The reproduction of the anomalies is much better than the reproduction of the hydrographs because the target window (top 25%, bottom 25%) is so large. It would be interesting to see the effect of varying percentiles in the second analysis and varying return periods for the third analysis.

Reply: We agree with the referee that it would be interesting to see the effect of varying percentiles in the second analysis and varying return periods for the third analysis. Therefore in the revised manuscript we have repeated the second analysis for the 10th and 90th percentiles for all basins. In the third analysis, our decision to use 5-years return period was a compromise between the rareness of the events on one hand and the limited record of discharge observations on the other. It is not possible to use significantly higher return periods for all basins because of short records. Therefore, we have included an appendix (Appendix B) where we present the results for 5 and 10-

year return periods, not for all basins but for the two basins with the longest records, the Danube and the Mississippi.

Comment: Pg 3482, Ln 1-16: the results presented in Fig. 2 have been caveated in this paragraph, although it is surprising that deficiencies in the GHM are mostly neglected as a potential explanation. I would not expect issues concerning climate input data to be entirely responsible for all of the differences between simulated and observed river flows presented.

Reply: In Section 4.1, we discuss possible reasons for the errors in the simulations. As we have stated earlier in our reply to the first Technical Comment, in this section we mention some of the known problems in the climatic input but we indeed do not discuss possible deficiencies in the GHM. These issues are discussed in other papers by our research group but it is beyond the scope of this paper. For an in depth discussion of this point, we refer the reader to Van Beek, Wada and Bierkens (2011). However, we certainly acknowledge that the bias in the simulations is due not only to errors in the climate input but also to model errors. We have modified the text in section 3.1 of the revised paper to make this clearer.

Comment: Pg 3483, Ln 4-6: all results prior to this were discussed in terms of MSESS, with R2 and NS neglected. This imbalance needs to be corrected, or alternatively, if R2 and NS do not lend anything to the discussion, perhaps they should be excluded.

Reply: MSESS is used as the main skill score for verification of the hydrographs, since it provides a relative skill measure, i.e., relative to the climatology. This is consistent with the measures used in the second and third analyses, which are also relative measures. Consequently, the results are discussed in terms of MSESS. The two other scores R2 and NS are presented additionally simply because they are the two most commonly used verification measures in hydrology. They could be important for some readers as an indication of model performance, and as a means of comparison with other models. For this reason we prefer not to exclude these measures from the anal-

yses. Besides, as the referee points out we refer to these scores at one point in our discussion, where they provide additional information.

Comment: Pg 3484: it is correctly stated twice on this page that the GHM shows more skill in reproducing floods than droughts, although no potential explanations are highlighted. It would be useful to propose some ideas on the possible reasons for this.

Reply: We have added a paragraph to Section 4.3, in which we discuss model structure and process descriptions that explain the difference in skill in reproducing floods and droughts.

Comment: Pg 3485, Ln 21: I would be interested to know which GHMs are analysed in Sperna Weiland et al. 2010b. In general, I would be wary of talking about other GHMs that have not been analysed within the same framework as is used in this paper, particularly when it is unclear whether the other GHMs have been bias corrected.

Reply: This paragraph has been modified to include LSMs as well as GHMs. The references to the papers of Sperna Weiland et al. have been updated. Table 3 on their paper which is now referred to as Sperna Weiland et al. 2010, presents a comparison of discharges from several GHMs including WBM, VIC and WaterGAP. It can be seen here that the results using PCR-GLOBWB are comparable to the results with these other GHMs. The paper under revision by the same authors, now referred to as Sperna Weiland et al. 2011, demonstrates that runoff fields produced by the LSMs of the two general circulation models ECHAM5 and HadGEM2 are also comparable to those by PCR-GLOBWB. While we suggest that our conclusion can be extended to similar GHMs and LSMs, we acknowledge that they have not been analyzed within the same framework. Therefore we base this argument purely on their similarity in model structure, parameterization and forcing data set, as well as their comparable performance.

Specific Comments:

Comment:- global hydrological model (GHM) is preferred over macro-scale hydrological model (MHM)

Reply: The term macro-scale hydrological model (MHM) has been replaced by global hydrological model (GHM).

Comment: - Pg 3471, Ln 4: Nijssen et al., 2001a is used to support model development over the past two decades, yet it does not cover the last ten years; more up-to-date reference required

Reply: The reference to Nijssen et al., 2001a has been removed.

Comment: - Pg 3472, Ln 11: CRU should be written in full, as it has not been used yet

Reply: CRU has been written in full.

Comment: - Pg 3472, Ln 25: GRDC should be written in full, as it has not been used yet

Reply: GRDC has been written in full.

Comment: - Pg 3474, Ln 5: ‘in surface runoff’ should read ‘into surface runoff’

Reply: The expression ‘in surface runoff’ has been replaced by ‘into surface runoff’

Comment: - Pg 3475, Ln 7: ‘non-regulated’, ‘unmodified’, ‘natural’ have same meanings; one will suffice

Reply: The terms ‘natural’ and ‘unmodified’ are deleted and ‘non-regulated’ has been used.

Comment: - Pg 3477, Ln 20: need references for Heidke skill score and the Peirce skill score here

Reply: References for Heidke skill score and the Peirce skill score have been added.

Comment: - Pg 3477, Ln 21: Gandin and Murphy already abbreviated to GM earlier;

can just use GM

Reply: GM has been used for Gandin and Murphy.

Comment: - Pg 3484, Lns 8-18: all numbers written as digits (e.g. 0) should be written as words (e.g. zero)

Reply: Numbers written as digits have been written as words.

Comment: - Pg 3496: tables are the same as those on the previous page; second ten should be shown

Reply: We will make sure the tables and figures are correct when the revised version is online.

Comment: - Fig. 3: blue colour on graphs is not as distinct from black as red on previous graphs

Reply: The two superimposed simulations and observations in the discharge time series figures (red and black as well as blue and black) can be distinguished sufficiently only when they are viewed in a larger size.

Comment: - Fig. 4: unclear what the different colours represent in the reliability diagrams

Reply: It has been clarified in the figure caption that the different colours in the reliability diagrams represent different months of the year.

Comment: - Fig. A1: might be useful to have index of correlation on this graph

Reply: Actually Figure A2 conveys sufficient proof for our argument but we also present Figure A1 because it shows the relation between daily and monthly extremes in a more visually straightforward way.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., 8, 3469, 2011.