Anonymous Referee #2
The original comments of Referee #2 are in black color and indicated by "R2:". Replies by the authors ("A") are colored in green. Actions are introduced by "Action:", changes in the manuscript are in italics.

This paper explores the sensitivity of a number of hydrological indicators to five perturbations of the WaterGAP GHM; changes in 1) climate (using different climate forcing datasets for 1971-2000), 2) land-cover (using different homogenised land-cover datasets), 3) structure (not different model parameters but an exploration of the effect of implementing new parameterisation schemes), 4) human water use, and 5) whether the model is calibrated or not.
It's nice to see a paper that explores uncertainties within a macro-scale hydrological model. Papers like this are quite common within the catchment-scale hydrological modelling community so it's good to see the practice transfer to larger-scale models because there is a need to understand the uncertainties that are inherent within individual global hydrological models (GHMs) and not only across different GHMs. To this end, I consider the article to be a valuable contribution but I would like the authors to address my comments below.

A: We thank the reviewer for the detailed and constructive comments. We will answer the comments below.

General comments ————————-
R2: 1) It could be argued that there is little value in the NoCal-STANDARD comparisons that appear throughout the paper. This is because all GHMs to some extent will be calibrated, whether it is through a comprehensive basin-level calibration with a runoff correction factor (e.g. WaterGAP), or simply by tuning model parameters until there is a good fit (e.g. MacPDM). No model will ever be used without some level of tuning having taken place during the model development stage. Thus the NoCAL-STANDARD maps that show large differences (e.g. Fig 7 and Fig 9) is hardly surprising and offers little value to the paper. It could also be argued that the inclusion of this comparison skews the conclusions of the paper, because it would be considered infeasible and inappropriate to run a model that had not been tuned or calibrated at any point during its construction. To put it another way, it's a bit like comparing the STANDARD model with a model where the LAI parameter for each vegetation type is increased by 200%, to explore the effect of vegetation parameterisation. While this would tell us something about how hydrological response is sensitive to LAI, the leaf sizes would be unrealistic in some cases, so a comparison of the two simulations would be nonsensical. In the same way, here the authors are indirectly suggesting that it would be feasible to run an un-calibrated model, when of course it is not. I suggest that the authors either miss out the NoCal comparisons from a revised version of the manuscript or that they provide a rigorous justification for why they make the NoCal comparisons. If the authors decide to keep the NoCal comparisons, then can they also please explain clearly and concisely, which parameters are calibrated in the STANDARD version – i.e. is it only γ, CFA and CFS that are perturbed during calibration, or are other parameters related to, for instance, interception capacity, rooting depth, and field capacity etc. perturbed too?
A: We thank the reviewer for raising this important point and the reviewer's concerns. We have done the NoCal – analyses to analyze, how a basin-specific objective calibration of one to three parameters to long term averaged discharge impacts the model results. The

motivation of WaterGAP is to assess water relevant problems, therefore a good representation of observed long term averaged discharge is a prerequisite. During development of WaterGAP this could not be achieved by tuning model parameters globally (i.e. adjusting certain parameters like vegetation-specific LAI) in the way the reviewer described it in the first sentences without leaving plausible ranges or modify parameters e.g. for specific regions. Looking at results from global models without basin-specific calibration, they often do result in quite biased mean river discharge for invidual basins.

During the applied calibration, only γ and up to two correction factors are modified by using an objective calibration approach (all other parameters are not perturbed and therefore equal throughout all model variants). This is state-of-the-art in (catchment) hydrology, even if normally more parameters are calibrated. We feel that the effect of such basin-specific calibration is of interest. In the revised manuscript we still include the NoCal comparisons but explain the motivation better by adding a paragraph in the introduction and the new section 2.2.4 where the reason for investigating the NoCal variant is described.

R2: 2) It's good to see that the authors acknowledge on line 11 of page 1589 that they neglect parameter uncertainty. It is not unreasonable to assume that the uncertainty that might have arisen from running a perturbed parameter ensemble with the GHM would have been as large as the greatest source of uncertainty identified by the authors and this should be mentioned in the Discussion. This is often the case with climate modelling experiments that have run PPEs (e.g. see Rowlands et al. 2012 and Collins et al. 2006). To some extent, recent multi-GHM climate change impact assessments (Prudhomme et al. 2013; Davie et al. 2013; Hagemann et al. 2013) provide a glimpse of the range of uncertainty that can arise from different GHM structures, although strictly, GHM structure is not explored systematically in those cases.
A: We thank the reviewer for the explanation on parameter uncertainty and the additional references. In the revised manuscript, we added to the introduction "*The impact of parameter uncertainty is not investigated in this study as previous studies have already identified its strong impact on WaterGAP output (e.g. Kaspar 2003, Werth and Güntner, 2010). In addition, we added as the last sentence of the discussion: "Please note that this discussion on the dominant type of uncertainty does not take into account parameter uncertainty which is expected to be a major additional source of uncertainty.*"

R2: 3) This is related to 2) above. It is worth discussing that one of the reasons the simulations appeared less sensitive to land-cover than the other modifications, is because the parameters associated with each land-cover class were kept constant. The differences would probably have been much larger if land-cover-associated parameters were modified (e.g. LAI, interception capacity, surface roughness etc.), especially those related to PET.
A: As we stated already in the objective of the presented study, we here wanted to concentrate on impacts of model structure and spatially distributed input, and explained there that parameter modification can lead to large changes in the model output. Also we hope to deal with parameter uncertainty by basin specific calibration, at least to a certain extent. The reason for the relatively small differences of LANDCOVER compared to STANDARD is that there are not too many regions where land cover class attributes are changing abrupt even if the type is differing, except some regions described at Sect. 3.2 and

3.3. Compared to that, changes of climate forcing or model structure are of higher importance. Action: we feel that no action is required.

Specific comments ————————-
R2: 1) The abstract includes a lot of information and it is rather long. I suggest reducing the amount of text included in the abstract. Moreover, lines 15-25 of the abstract could be shortened and it could be stated more concisely the order (increasing or decreasing) in which each uncertainty impacts on one chosen hydrological indicator (e.g. Q, or AET).
A: We thank the reviewer for highlighting the necessity to write a concise abstract. Our abstract is long, but we feel that it contains the essence of the manuscript (which is also long). We could not find a way to me more concise without losing information we want to convey.

R2: 2) Lines 6-23, page 1592. Can the authors please clarify whether for both CLIMATE datasets they forced WaterGAP with monthly data, or daily data (disaggregated from monthly means).
A: We thank the reviewer for pointing this misunderstanding. For the CLIMATE variant we used monthly CRU/GPCC input which are disaggregated to daily values (described in Döll et al., 2003). For all other variants, we used WFD/WFDEI which are available in a daily resolution (as stated at P1591,L19-22). To make it more clearer, we added at P1592 L9 the sentence: *"Monthly means are disaggregated to daily values within WaterGAP (Döll et al., 2003)."*

R2: 3) Line 10, page 1592, should "e.g. wind undercatch" actually be "e.g. precipitation undercatch"?
A: We thank the reviewer. We modify the phrase to "*wind induced precipitation undercatch*".

R2: 4) Line 6, page 1593, please write fully what IGBP stands for.
A: IGBP stands for International-Geosphere-Biosphere Programme, we added this.

R2: 5) Figure 2 – it would be useful to include a third map that shows areas where there is different land-use according to the two sources of information. Grid cells where there is a difference could be shaded in a single colour. While this would be quite a simple map, it would make it easier to observe where the differences are.
A: We thank the reviewer for this good idea, and included such a map as a third map in the new Fig. 2.

References

Döll, P., Kaspar, F. and Lehner, B.: A global hydrological model for deriving water availability indicators: model tuning and validation, J. Hydrol., 270(1-2), 105–134, doi:10.1016/S0022-1694(02)00283-4, 2003.

Kaspar, F.: Entwicklung und Unsicherheitsanalyse eines globalen hydrologischen Modells, University of Kassel., 2003.

Werth, S. and Güntner, A.: Calibration analysis for water storage variability of the global hydrological model WGHM, Hydrol. Earth Syst. Sci., 14(1), 59–78, doi:10.5194/hess-14-59-2010, 2010.

New Figure 2, including a third map showing grid cells where land cover class has changed due to different landcover input.

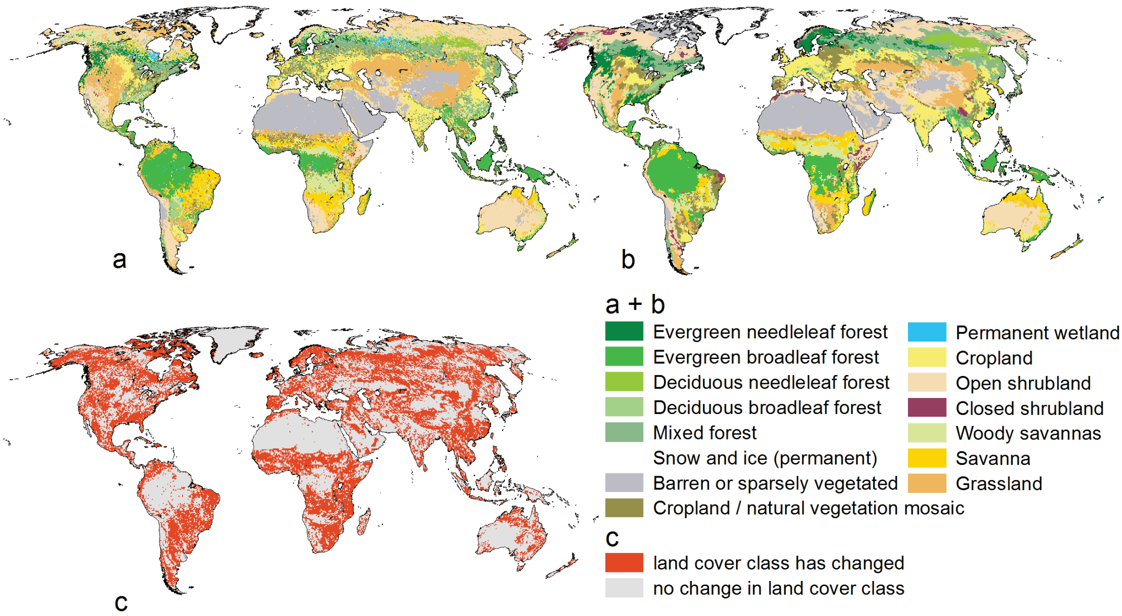

Figure 1. Land cover maps with a spatial resolution of 0.5° used as WaterGAP input based on MODIS observations for the year 2004 (variant STANDARD) (a), land cover derived from USGS GLCC but CORINE for Europe reflecting land cover distribution around the year 2000 (variant LANDCOVER) (b), and identification of grid cells where land cover class has changed due to different input data (c).