Summary: This Discussion Paper presents on model development in the H08 Global Hydrological Model intended to expand the model's representation of human water appropriation. The reported model enhancements are important in an age of large and increasing human modification of the water cycle, and the authors' approach to model structures and data will be of interest to HESS readership. The Discussion Paper presents both the implementation of model enhancements and the results of model simulations globally and for major river basins. As a presentation of model methods I find the paper to be a strong and useful contribution to the literatures. it is clearly written, appropriately detailed, and reports an impressive set of model development and data wrangling efforts. GHM of this type are not my primary research focus, so I cannot comment on the completeness of referencing and the place of this paper in the broader GHM literature, but it appears that the authors have taken care to put their work in the context of related efforts with different models.

We are grateful that you have evaluated our paper so highly.

As a presentation of simulation results I believe the paper succeeds to some extent. The authors present reasonable estimates of water sources at basin and global scale, and they discuss potential sources of error and uncertainty. Perhaps inevitably, however, both model evaluation and quantification of uncertainty are quite thin. As a result one is left with point estimates of large quantities with significant uncertainties, where it would be considerably more informative to have ranges reported on the basis of some kind of quantitative error estimate.

Thank you for this comment. For technical reasons, the addition of systematic/formal error bars to our estimates was quite challenging. We have now added standard deviations to the mean annual estimates, which were derived from 30 years of simulation, for the key simulation results (Tables 2 and 3, Abstract). This addition presents basic information about one aspect of the uncertainty.

That said, I believe that the Discussion Paper is important in that it presents on a large investment in model development for H08, and my comments below are suggestions for improvements rather than requirements for final publication.

Thank you for your valuable comments. All of your points are well taken and we have tried our best to incorporate them in the revised manuscript.

Suggestions:

[R3-M1] P. 4, line 31-2: This is the first instance I noted in the paper of parameters presented as single point estimates with no sensitivity test and little in the way of justification. The same happens at several other points in the text, either explicitly or implicitly via citations. This left me wondering how sensitive model results are to the choice of parameters that are, at best, only roughly constrained by data. Later in the paper the authors identify a number of major sources of uncertainty, but parameters within the hydrological model or management modules are not specifically discussed. Would it be possible to perform targeted sensitivity tests of some of these parameters? Not only would this enhance confidence in model results, but it would provide useful guidance for later model development regarding the relative importance of various parameters to model results.

> We have enhanced the discussion of the hydrological parameter uncertainties in Section 3.4.3 (Potential sources of uncertainty): "Moreover, the hydrological parameters were not tuned to individual basins, yielding a generally lower reproducibility of historical river flow observations (e.g., Hattermann et al. 2017). In cases in which the H08 model was applied to specific basins, sensitivity testing and hydrological parameter calibration were conducted systematically using reliable long-term observations (e.g., Hanasaki et al. 2014; Masood et al. 2014). Conversely, when H08 is applied globally, as in this study, these procedures are difficult to perform because observations are not available for vast areas and simulation periods. Without ground truthing, the sensitivity test cannot be interpreted, and parameter calibration cannot be performed. This is particularly true for groundwater parameters because very few reliable observations representing the grid-cell size (0.5°) are available."

[R3-M2]. p.5, line 27-28: The assumption regarding the division between surface water and groundwater is quite a large assumption, and it doesn't account for regions in which farmers use groundwater to make up for surface water shortfalls. I don't have any better idea about how to deal with these complications in a global simulation, but it would be useful if the authors could spend a sentence or two justifying the assumption and explaining potential limitations.

> We have taken your advice and added a new simulation option that extracts additional renewable groundwater when surface water is depleted. The relevant methods and results are shown in Supplemental Text S5, as follows:

“As described in Sections 2.1.2. and 2.1.7., the water source at an individual grid cell is assigned to the surface water and groundwater parts using the fixed local parameter, termed the fraction of the water requirement assigned to groundwater (f_{gw} in Eqs. 4 and 5). We added a simulation option (hereafter SWT) to abstract additional renewable groundwater when surface water is depleted. This option reflects the ability of some water users to switch water sources by taking availability into account.

The results are shown in Table S4. Compared with the ALL simulation, SWT uses as much as 213 km^3 yr^{-1}, or approximately 30%, less unspecified surface water. Groundwater abstraction increased by 345 km^3 yr^{-1}. We used this gap to compensate for the reduction in river water abstraction (123 km^3 yr^{-1}). Additional groundwater abstraction depressed the storage of renewable groundwater and consequently the baseflow, which eventually reduced the availability of river water. Comparing the total groundwater use of ALL and SWT, the estimation of ALL is closer to the range of statistics-based literature (639–765 km^3 yr^{-1}, according to FAO 2016 and IGRAC 2004). This result implies that although water users may switch water sources flexibly from surface water to groundwater in some regions, this appears not to be the case in many parts of the world.”

[R3-M3]. Table 4 presents some model evaluation, but no significance tests are presented to show whether ALL is significantly different from NAT for each basin, or whether either simulation is significantly different from observation. Please provide tests of significance for these differences, accounting for temporal autocorrelation as appropriate.

We have added statistical testing of the bias, correlation coefficient, and slope results presented in Tables 4 and S3. Please note that we excluded the Nash–Sutcliffe Efficiency because the authors do not feel that there is an established method to conduct statistical significance test of it. In short, we first added statistical significance information to the correlation coefficient and slope (i.e., annual trend) for TWSA. We then added statistical significance testing to the difference between the NAT and ALL simulations for bias in river discharge, correlation coefficient, and slope in TWSA.

[R3-M4]. Irrigated area: Perhaps this is covered in an earlier H08 publication, but how does the model decide on what fraction of area equipped for irrigation is active in any given year?

In my own work I've found this to be a challenge, particularly when it comes to interannual variability in irrigation demand under extended droughts. e.g., when farmers fallow irrigation fields due to water shortage. Is this addressed in the model, particularly when it comes to trends in water stressed regions?

> As described in Section 2.2.1, all land use was fixed throughout the simulation period for this study and the irrigation water requirement was estimated based on this fixed land use. Although interesting to analyze, little geographical information is available on the variability and change in irrigated area (and crop type). We rephrased the text in Section 3.4.3 (Key uncertainties) as: "We note that Siebert et al. (2015) developed the global distribution of irrigated areas from 1900 to 2005, which would be an important contribution to simulations incorporating inter-annual variation in the irrigation water requirement. We fixed the irrigated area throughout the simulation period, however, because little information is available on annual variation in crop practices (e.g., crop type, crop intensity, fractions of surface water and groundwater dependence)."

[R3-M5]. Comparisons with GRACE: the authors have compared to a single GRACE product. While I expect that different flavors of the spherical harmonics GRACE simulations will be similar in most basins, the more recent mascon solutions have emerged as likely more reliable for terrestrial applications (http://onlinelibrary.wiley.com/doi/10.1002/2016WR019494/full). The authors should consider adding a mascon analysis to their evaluation, both to quantify observation based uncertainty and because the mascons might indicate that the TWS trends are actually larger than the spherical harmonics solutions indicate, and are in better agreement with ALL simulation results.

> We obtained the Mascon data (Scanlon et al. 2016) for 12 basins and drew the same figures as Figures 10 and S4. We found only marginal differences between them (Figures R1 and R2 show TWS anomaly in the Mississippi River that adopt CSR and Mascon as observation respectively). The most notable difference was observed in the Ganges River Basin; Mascon showed twice as large a decreasing trend in TWS (-19.59 mm/yr) than in CSR (-10.54 mm/yr), which is consistent with the H08 simulation (-21.16 mm/yr). As no significant difference was observed, we decided to continue using the CSR product.

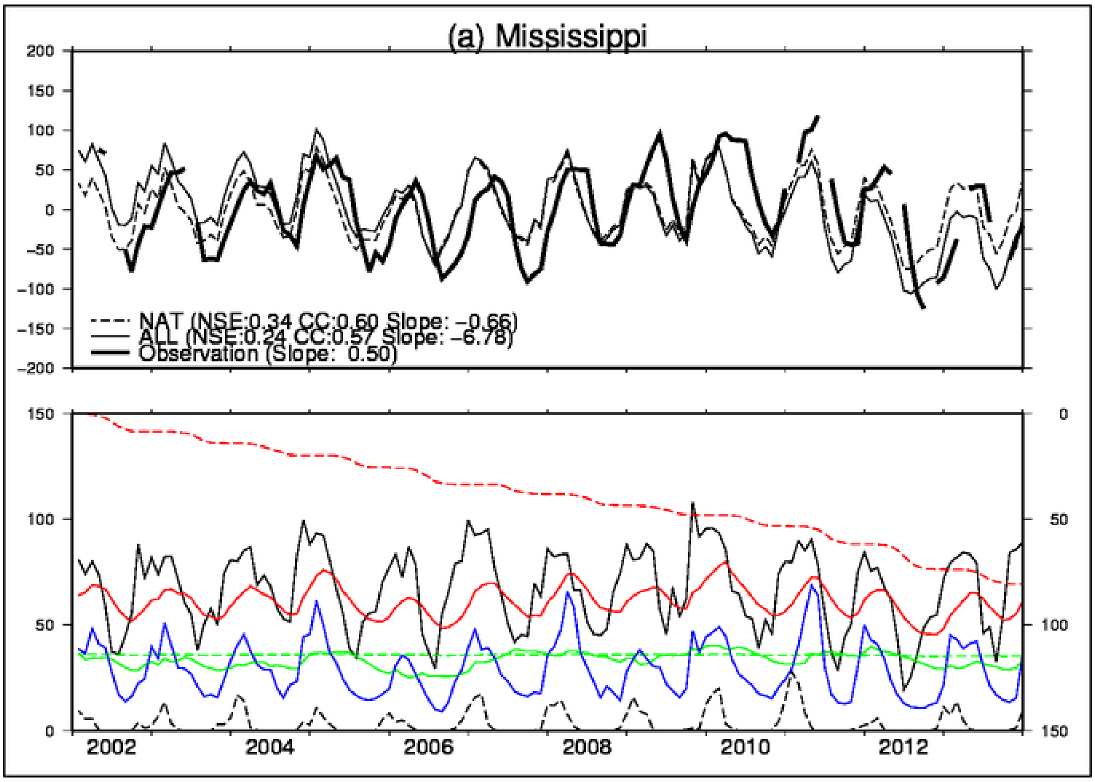

Figure R1. TWS anomaly in the Mississippi River. Observation adopts CSR.

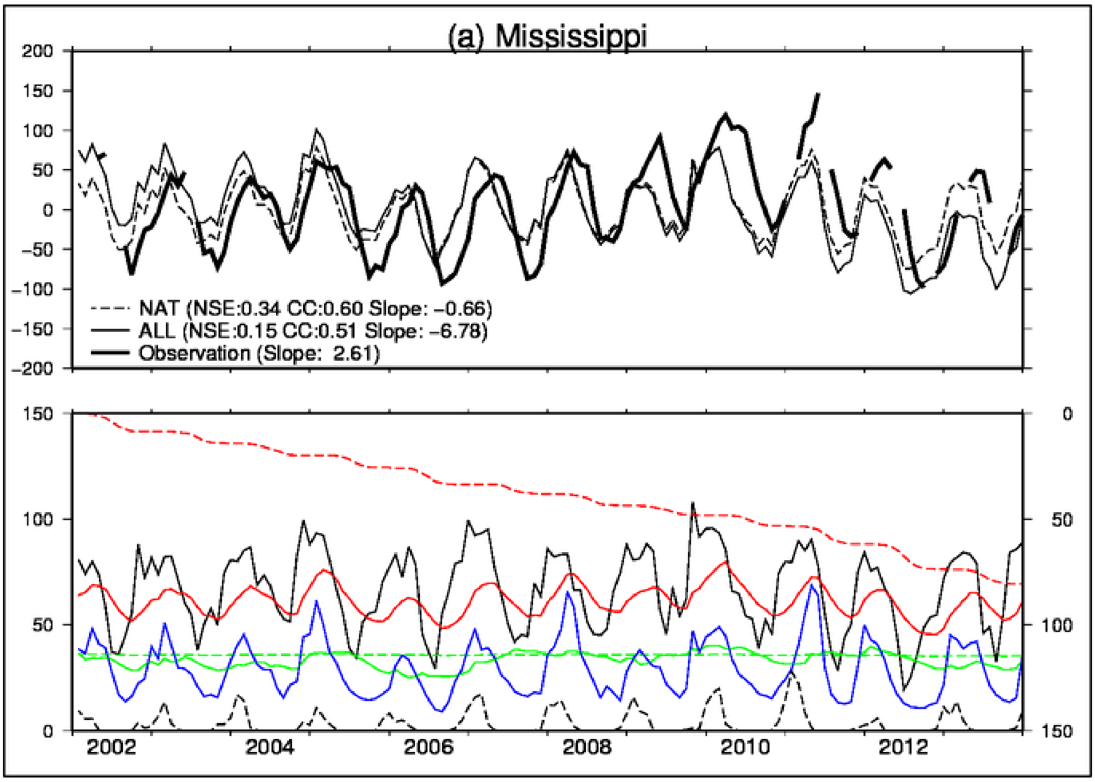

Figure R2 TWS anomaly in the Mississippi River. Observation adopts Mascon.

Minor comments:

[R3-S1] Abstract: The “R” in GRACE stands for Recovery, not Retrieval.

Thank you, this has been corrected.

[R3-S2] Section 2.2.2: A few words on the WATCH methodology would be helpful for those of us not familiar with it.

We have added further description of WFDEI, which reads, “The WATCH forcing methodology represents sub-daily reanalysis data scaled arithmetically to make the mean values and the range of variation consistent with spatio-temporal coarse-ground observation data.”

Section 3.2.2: Were scaling factors applied to the GRACE data?

We applied scaling factors. We now specify this in Section 3.2.2.