Reviewer #1

General Comments:

`This paper describes the eartH2Observe Tier-1 dataset and presents a benchmark for the key water cycle variables simulated by a suit of land surface models (LSMs) and global hydrological models (GHMs). The authors provide an overview of current state-of-the-art models and analysis framework along with tools that enables benchmarking repeatable as new improvements are made to the models and forcing datasets. I'm impressed and excited about the open access and completeness of the datasets and results of the paper. I think that the authors did great job designing the framework for identifying the model consistency/inconsistency via the use of common forcing and the SNR analysis, including both GHMs and LSMs and investigating the uncertainty in the precipitation forcing, and verifying with the benchmark dataset. The paper contains valuable findings to the modeling community, in which the strength of ensemble mean over any single model is demonstrated in some variables and the areas of importance for further model dataset development are identified.

We would like to thank the reviewer for taking the time to review our paper and for the positive comments. We are glad that our open data policy is appreciated. We have numbered the specific comment for clarity below.

Specific Comments:

1) Page3, Line8: I was not sure if the second question "is the current modeling protocol with one forcing dataset and the selected output variables sufficient for evaluation of (global) water resources" was answered

We agree that the answer to this initial question may not have been provided explicitly but can be deduced from the paragraph on page 22 starting at line 23. To clarify this we will add the following to the conclusions and outlook on line 24 after '... to global discharge.': " Beck (2016) also shows that the ensemble mean comes close to the best (calibrated) models with respect to discharge." Next, further on at the end of the paragraph: "The above indicates that the ensemble mean of the present dataset could be used to evaluate water resources."

A caveat remains and that is that we use only one forcing dataset. This is especially important for precipitation as the SN ratio analysis indicated. In the current work we focus on the uncertainties between models and the added value of a set of models. But, we plan to include an ensemble of precipitation products in a future version (see also point 9 below).

However, we agree that the answer remains partial and this question should probably not have been a main research question. Therefore, we have reformulated the section as follows:

> *"In this paper we present the first version of the dataset, which is based on the current state-of-the-art of the contributing modelling systems and will provide a benchmark to evaluate improvements made to the models and forcing data in the coming years. The main goal of this paper is to provide a multi-decadal dataset of water balance components from an ensemble of models that is open and of use for further research and applications. Secondly, we investigate if the ensemble mean in*

this dataset is superior to the individual models given the diverse set of models, and if so, for which variables."

2)"Continental water budget"
is referring to water budget over land? I thought of individual continents (i.e. mean over North America, etc) but just global budget was presented.

The term may indeed be a bit misleading. With continental water budget we mean the terrestrial water budget. We will clarify this in section 3.3 and adjust the section header of section 3.3.

3) Page13, Line 25: "the
spread in ET is large" and that the model estimates are higher than the reference datasets are indeed concerning points. Only ORCHIDEE and WaterGAP3 include irrigation or water-use currently, but incorporating irrigation in other models will likely increase ET even more.

ET estimates from the land surface and hydrological models are very sensitive to the precipitation input, particularly where ET is limited by water availability. This also applies to the reference datasets that also rely on a significant amount of modelling and share similar uncertainties. Miralles et. al. (2016) demonstrated that reference datasets we used have similar problems. Miralles' results also indicate that the partitioning between the different ET fluxes vary widely between the different products. Precipitation remains one of the most uncertain inputs and specifically rainfall intensity is very important. The rainfall interception schemes incorporated in some of the models are very sensitive to changes in intensity and can thus influence the total ET significantly. Similar for the runoff generating mechanisms with the models. Although, adding irrigation to more models would yield an increase in overall water use if everything else remains the same a (small) change in precipitation input could have a larger effect either increasing or decreasing the total ET.

4) -Is there a reason why you didn't use the snow cover from GLOBSNOW-2 and used IMS instead?

For snow cover we only used IMS as it provides a consistent gap-free dataset. At first we only compared model output of snow cover as it is usually more reliable than SWE estimates. Because only a limited number of models supplies snow cover (all supply SWE) we opted to also include SWE which is not supplied by IMS and we used the GLOBSNOW product for SWE. However, all these datasets have their limitations, and more datasets could be included (e.g. GLOBSNOW for snow cover and MODIS), but such a detailed focus on cold processes is beyond the scope of this study.

5) -Table4: It states that the difference in model mean ET (and products?) are due to different periods used for the comparison. Do they match over the common overlapping period, 2003-2011? Additional information on spread of the three ET products can be helpful as a first cut uncertainty estimate, given that quality of ET validation datasets is difficult to assess.

We may not have been completely clear in our description. We refer here to the different means of the model output which differ when calculated for the same reference period as the respective reference datasets because these each cover a different period. The mean for the different products used as reference (GLEAM, MODIS) is not the same, even for overlapping periods. We will revise the caption of the Table to explain this better.

6) -Page15, Line10: “Although this may seem to be a large mismatch: : :.” I don’t see how this makes it more comparable. Could you elaborate?

By using monthly averages in the comparison we smooth the higher frequency dynamics that are typical for the surface soil moisture signal. Thus, we argue that for monthly averages the temporal dynamics of the topsoil signal becomes very close to the root-zone signal. Various studies have successfully used satellite-based surface soil moisture at the monthly time scale to represent root zone soil moisture characteristics, e.g. as a driver of vegetation dynamics or agricultural drought (See Barichivich et. al. 2014, Dorigo et. al. 2012, Muñoz et. al. 2014 and van der Schrier et. al. 2013)

7)-Page19, Line 3 and top panel of Figure9: is the precipitation increase after 1997 evident in the reanalysis observation based datasets as well or has it been evaluated elsewhere?

The present paper does not aim to investigate the trends in precipitation over land and as such we have not compared the trends in different products. An increase in global average precipitation over land has been reported by Ren et. al. (2013) in the station based REC product. As can be seen in the figure below it is most probably related to the gauge corrections of WFDEI as the increase is present in the WFDEI precipitation we used as forcing but not in the raw ERA Interim precipitation. MSWEP 1.0 plotted in the figure using a red line also shows some trend but less so than WFDEI. Most datasets that try to use as many observations as possible include a gauge distribution that changes over time and with it the way precipitation is sampled. This may lead to a dataset that is not homogeneous in time and thus doesn’t allow for trend analysis. This is similar for the temporal varying inclusion of satellite precipitation data (MSWEP). However, as Beck et. al. 2017 showed, adding gauges does increase the usefulness for hydrological application significantly. Furthermore Loew et al. (2013) showed that trends in WFDEI and GPCP over parts of Africa were very similar (increasing in the recent periods) while EI showed a decline over the same period. We are not confident in stating that the trend is significant or not, it is also outside the scope of this paper but it does show that the choice of precipitation input in a study like ours may have a large influence on the outcome.

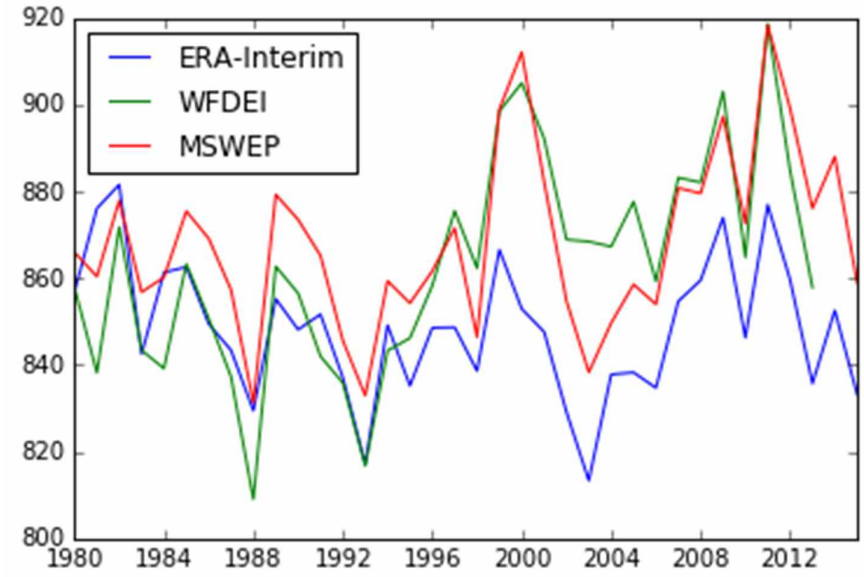

Figure: WFDEI: Global terrestrial precipitation (green line), ERA Interim Global terrestrial precipitation (blue line) and MSWEP 1.0 terrestrial merged precipitation (red)

8) Just curious, what does eartH2Observe stand for?

The project title tried to encompass earth observation (earth and observe) and includes the subject of our research, water (H2O). It is pronounced "earth to observe". The project focusses on the evaluation of the applicability of EO datasets for (local) water resources assessments.

9) Is there Tier-2 dataset (is it going to be the next round with error estimation and higher resolution etc.)?

This is certainly intended. We plan to achieve a higher common resolution (0.25 degree) and to include runs based on an ensemble of precipitation products accompanied with error estimation.

10) Technical Corrections: -Page1, Line11: "at" -> remove -Page3, Line8: "modelling" -> typo -Page22, Line 25: "bets" -> typo -Figure 3 needs description on the line, box, and error bars. -Table 9 and Table 10 seem to be identical. I doubt that it is true since the global summary Table 6 shows different values for snow cover.

We fixed the typos at Page 1, page 3 and page 22.

The caption for Figure 3 has been extended to read: Distribution of the signal-to-noise ratio of monthly anomalies over different biomes (horizontal axis, see Figure 4) for Evapotranspiration (a), Runoff (b), Root Zone Soil Moisture (c) and Precipitation (d). The boxplots represent the spatial variability of the individual pixels of SNR in each biome

extending from percentile 5 to 95 (wisher), percentiles 25 to 75 (box) and median (horizontal line).

We apologise for the missing Table 10. The proper data for Table 10 can be found at: http://earth2observe.github.io/water-resource-reanalysis-v1/results/table_snowc.html . We will update the table with the correct information in the manuscript.

References

Barichivich, J., Briffa, K.R., Myneni, R., Schrier, G., Dorigo, W., Tucker, C.J., Osborn, T.J., Melvin, T.M. (2014). Temperature and Snow-Mediated Moisture Controls of Summer Photosynthetic Activity in Northern Terrestrial Ecosystems between 1982 and 2011. Remote Sensing, 2014, 6(4), 1390-1431.

Beck, H.E., van Dijk, A.I.J.M., Levizzani, V., Schellekens, J., Miralles, D.G., Martens, B., de Roo, A., 2017. MSWEP: 3-hourly 0.25° global gridded precipitation (1979–2015) by merging gauge, satellite, and reanalysis data. Hydrol. Earth Syst. Sci. 21, 589–615. doi:10.5194/hess-21-589-2017

Dorigo, W.A., De Jeu, R.A.M., Chung, D., Parinussa, R.M., Liu, Y.Y., Wagner, W., Fernández-Prieto, D. (2012). Evaluating global trends (1988-2010) in harmonized multi-satellite surface soil moisture. Geophysical Research Letters, 39, L18405. doi:10.1029/2012GL052988

Loew, A., Stacke, T., Dorigo, W., de Jeu, R., Hagemann, S., 2013. Potential and limitations of multidecadal satellite soil moisture observations for selected climate model evaluation studies. Hydrol. Earth Syst. Sci. 17, 3523–3542. doi:10.5194/hess-17-3523-2013

Miralles, D., Jiménez, C., Jung, M., Michel, D., Ershadi, A., McCabe, M.F., Hirschi, M., Martens, B., Dolman, A.J., Fisher, J.B., others, 2016. The WACMOS-ET project, part 2: evaluation of global terrestrial evaporation data sets. Hydrology and Earth System Sciences 20, 823–842.

Muñoz, A. A., Barichivich, J., Christie, D. A., Dorigo, W., González-Reyes, A., González, M. E., Lara, A., Sauchyn, D., Villalba, R. (2014). Patterns and drivers of Araucaria araucana forest growth along a biophysical gradient in the northern Patagonian Andes: linking tree rings with satellite observations of soil moisture. Austral Ecology, 39 (2), 158-169, doi: 10.1111/aec.12054

Ren, L., Arkin, P., Smith, T.M., Shen, S.S.P., 2013. Global precipitation trends in 1900–2005 from a reconstruction and coupled model simulations. J. Geophys. Res. Atmos. 118, 1679–1689. doi:10.1002/jgrd.50212

van der Schrier, G., Barichivich, J., Briffa, K.R., & Jones, P.D. (2013). A scPDSI-based global data set of dry and wet spells for 1901–2009. Journal of Geophysical Research: Atmospheres, 118, 4025-4048