Response to comments from Zachary Zubin

We wish to thank the reviewer for his positive review and the useful comments and suggestions on our manuscript. We have modified the manuscript accordingly and detailed corrections are listed below. In the following the reviewer's comments are given in italics and our replies in regular font.

Short comment by Zachary Zubin

Based on an initial reading, this manuscript seems thorough and represents an important contribution to the literature. I glossed over some areas beyond my expertise (e.g., the river parameterization). Below are some minor comments / suggestions:

The methods should explain briefly (in addition to the references) how the surface fluxes / vegetation / atmospheric boundary conditions in the PCR-GLOBWB are determined. Is this a full land-surface model like that found in a global climate model?

We explain the used land-surface model into more detail in the revised manuscript and added the paragraph below in section 2.1.1 p. 5, l. 15 until p. 6 l. 11.

"For a detailed description of the model PCR-GLOBWB we refer to van Beek et al 2011, and a summarized model description is given here. PCR-GLOBWB was run at 6' resolution using a daily time step. Monthly climate data were taken from the CRU TS2.1 (Mitchell and Jones, 2005) with a spatial resolution of 0.5^{o} and downscaled using the ERA-Interim reanalysis (Dee et al., 2011) to obtain a daily climatic forcing (see de Graaf et al., 2014 for a more detailed description of this forcing dataset). Each grid-cell contains a land surface that is represented by a vertical structured soil column comprising two soil layers (maximum depth 0.3 m and 1.2 m respectively), an underlying groundwater reservoir, and the overlying canopy. Sub-grid variability is included with regards to land cover (in this case using fractions of short and tall vegetation), soil conditions, and topography. The model employs the improved Arno Scheme (Todini, 1996; Hageman and Gates, 2003) to simulate variations in the fraction of saturated soil in order to quantify direct surface runoff. Each time step, for every grid cell the water balance of the soil column is calculated on the basis of the climatic forcing that imposes precipitation, potential reference evaporation, and temperature. Actual evapotranspiration is calculated from potential evaporation and soil moisture conditions. Vertical exchange between the soil and groundwater occurs through percolation and capillary rise. Specific runoff from the soil column, comprising direct surface runoff, interflow and baseflow, is accumulated along the drainage network that consists of laterally connected surface water elements representing river channels, lakes or reservoirs. The accumulated runoff is routed to obtain discharge using the kinematic wave approximation of the Saint-Venant equations at a sub-daily time step."

Eq. 3-5 are a bit hard to follow and may need some additional explanation.

According to the reviewer's suggestion, we rewrote point 3 (p. 9) and point 4 (p. 9 – p. 10) in the manuscript to explain the used equations more extensively.

Eq. 6 seems to be mixing the concept of near-surface permeability with the deep-groundwater permeability as determined by Gleeson et al 2011. Is it realistic to decay to zero below the depth alpha, or should there be a minimum bedrock permeability for the thickness of the aquifer? Give some idea of the range of alpha. Is alpha the soil depth, the regolith depth, or the depth to impermeable bedrock?

To estimate aquifer permeability at greater depth we combined the concept of exponentially decreasing permeability of the continental crust with depth (e.g. Ingebritsen and Manning (1999)) with data on near surface permeability from Miguez-Macho et al. (2008).

The near surface permeability is prescribed by the sediment-bedrock profile at a location, which depends strongly on terrain slope; the steeper the land, the thinner the regolith and the sharper the decrease in permeability with depth. Two figures are presented in the supplementary material of this paper; presenting the range of e-folding depth (Fig. 2) and its spatial distribution (Fig. 3). Also, we extended the description of the usage of the e-folding depth to explain this better, p. 10, l. 10-16.

The permeability diminishes exponentially with depth from a known value of near surface permeability (k_0). The transmissivity can then be calculated with the presented integral of Eq 6 (p. 10). As Eq 6. is an exponential function, permeabilities will approximate zero.

In the determination of the 6' gridcell properties, I'm not sure if it is done at 30" and aggregated up, or if the 30" data is only used to calculate the floodplain depth for the 6' cell and the average depth is used in determining that 6' cell's properties.

We used the 30" data to determine the floodplain elevation at 6'as follows: Within each 6'cell and using the HydroSHEDS dataset, we identified the lowest elevation at 30'' (maximum 144 values for a cell comprising only land area), and assigned this as the floodplain elevation for the entire cell. We clarified this into more detail in the revised manuscript in section 2.2, point 1, p. 7-8.

What is the difference between the "true" and "apparent" MODFLOW grid cell area?

We rewrote this part in the revised manuscript, p.11, l. 13, and clarified our explanation. By naming the cells 'true' and 'apparent' we wanted to clarify the difference between the projected cell area used by PCR-GLOBWB and the cell area used by MODLFOW, as the latter assumes rectangular grid cells. This means there is a difference in area, for which we should correct (as done in Eq 7).

The Figure 6 caption needs correction.

As suggested, we have corrected the caption to read: "Scatter plot of observed heads against simulated heads for sediment basins (red) and mountain ranges (blue)."

Figure 8 caption: where are the white areas referred to?

The white areas are no-data values. This is added to the figure caption now.

Good luck on a productive review and publication.

Thank you for your kind wishes.

Response to comments from Nir Krakauer

We thank the reviewer for the positive response to our paper and thoughtful comments that allowed us to significantly improve the paper. We have modified the manuscript accordingly, and address the reviewer's comments in detail below. In the following the reviewer's comments are given in italics and our replies in regular font.

Reviewer 1: Nir Krakauer

The work presented here involves a global simulation of steady-state groundwater movement that employs the popular groundwater model MODFLOW, which is usually used in basin-scale and regional studies. Global datasets of lithology and permeability are combined with some assumptions and calibration to observed groundwater levels. While this work is interesting and valuable in that it moves in the direction of expanding the capabilities of groundwater models and integrating them with land surface models over large domains, there are some changes that could be made to render it of wider applicability.

The authors should mention new work by Krakauer et al. (ERL 9:034003 2014) on the relationship between model spatial resolution and lateral flow volume and by Gleeson et al. (GRL, in press) on higher-resolution global permeability and porosity maps.

We took notice of the suggested literature and added this to the introduction of our manuscript. (p. 3, l. 9 and p.3 l. 20.)

The water level observation database needs a citation (probably Fan et al. 2013), not just a URL.

We included the reference to the paper of Fan et al. 2013, where the data is presented (p. 14, l. 22).

It will be very difficult for anyone else to reproduce the global configuration of MODFLOW developed, particularly given the reliance on stochastic methods for parameter estimation. Therefore, I strongly recommend that the authors post their MODFLOW input files and control scripts under an open-source license in a suitable repository such as GLOWASIS, so that they can be evaluated and improved on by the hydrology community.

In our study we aim to explore the possibility to develop a high resolution groundwater model using existing globally consistent datasets. As such, our parameterization is subservient to the sensitivity analysis that we have performed here and samples the uncertainty of existing datasets, most notably that on permeability of Gleeson et al. (2011) and the data we have compiled here. The resulting products are not meant to have any definite status as a unique global geo-hydrological parameterization and for that reason, we will not publish the data online in their present form. However, we are willing to provide the input data to researchers who are interested in exploring the parameter space covered by our realizations .

Figure 3: It is not clear what "cumulative probability" means.

With the 'cumulative probability' we mean the frequency distribution of the aquifer thickness, as described in the text. We changed Figure 3 (according to the suggestion of reviewer 3, Mary Hill) now presenting calculated aquifer thickness and transmissivities. The likelihood of finding a thick aquifer (F'(x) in eq. 1), used to estimate aquifer thicknesses is placed in the extra material now (Figure 1).

Figure 5: Most of the land area has a beige tone in all 4 panels. What does this indicate?

The beige tone indicated the value (around) 0 (indicated in the colour bar).
What can be seen from Figure 5a is that overall the coefficient of variation of groundwater depth is small. Higher values are found for low recharge areas (e.g. Sahara , Australian dessert) and for areas with shallow groundwater depths with higher transmissivities and recharge rates (e.g. Amazon basins, Indus basin). Figure 5b shows saturated conductivity is the main driver of changes in groundwater depths.

Accordingly to the comment of reviewer 3 (Mary Hill) we improved this figure by showing the coefficient of variation of the overall analysis (with changing parameters for aquifer thickness, groundwater recharge, and saturated conductivity) and for saturated conductivity. The other two parameters are of minor importance for the variance in groundwater table depths.

Figure 6: What is the criterion for considering a water table to be "local and perched"?

The spatial difference between observed and simulated groundwater depths (Fig. 7) show that in general groundwater depths are overestimated compared with observations for steeper and higher elevated terrains. This is explained by the grid resolution (p.17, l. 1-6), which is to coarse to capture small local valleys. As a result, smaller local aquifers are left out and the simulated groundwater heads present the regional scale continuous and deep groundwater tables, rather than local and perched water tables. We assume that for steeper and higher elevated terrains local and perched water tables are likely to occur, but stay uncaptured by the model. The shallow groundwater depths observed for these areas confirm this assumption (shown in Fig. 6).

The criterion to consider groundwater tables local or perched is now better explained: p. 16. l. 22-27, and p. 18 l. 6-8.

Additionally we changed Figure 6 (following the suggestion of Mary Hill). This figure now presents the scatter plot of observed vs simulated heads of the best performing overall run for sediment basins and mountain ranges.

Figure 7: I am not sure if displaying "relative residuals" makes sense here. Would this quantity be infinite when the observed water level is zero?

Indeed the relative residuals will be infinite if observed groundwater depths are zero. But in the used dataset of Fan et al. (2013) no groundwater depth of zero occur.

The relative residuals showed here to make clear that, although the absolute error for higher elevated and steeper terrains is large, the relative error is small. In other words, this means that there were deep groundwater is simulated observed water can be more shallow, however it still is deep. Nevertheless, we decided to leave out this figure, as it causes confusion. Furthermore, the same information is given with absolute errors and the histogram.

Figure 8: The long groundwater flow paths, for example around the Gulf of Bothnia and Gulf of Finland, are remarkable. I am not very familiar with these areas, but maps such as http://www.ymparisto.fi/en-US/Waters_and_sea/Hydrological_ situation_and_forecasts/Hydrological_forecasts_and_maps/Hydrological_forecasts_ and_maps(26174) for Finland show them to be fairly well drained. Is the simulation resolving the river network in such regions as a groundwater sink?

With our groundwater model we simulate the regional scale deeper groundwater flow. Due to the grid resolution, which is too coarse to capture small local valleys (see also p. 18. l. 16-19), smaller local aquifers area left out. For the steeper and higher elevated terrains local and perched water tables are likely to occur, but stay uncaptured by the model.

In the case of Finland deep regional scale groundwater depths for the basement rocks are simulated, as these rocks are defined as acid plutonic and metamorphic rocks in the used lithological map (GLiM Hartmann and Moorsdorf 2012). Consequently long flow paths are simulated. Local streams draining saturated cover materials (e.g. glacial till, blanket peats etc.) are not captured by the groundwater model, but are part of the land-surafce model PCR-GLOBWB.

Response to comments from Mary Hill

We thank the reviewer for the positive response to our paper and thoughtful comments that significantly improved the paper. We have modified the manuscript accordingly, and detailed replies to her comments are listed below. In the following the reviewer's comments are given in italics and our reply in regular font.

Reviewer 2 Mary Hill

1. Scientific Significance: Does the manuscript represent a substantial contribution to scientific progress within the scope of Hydrology and Earth System Sciences (substantial new concepts, ideas, methods, or data)?

The study is good to excellent. The global simulation of groundwater flows is very interesting. Unfortunately, this article is Fair to Poor. The article is too confusing and important contributions are not presented in a very informative way. The article lacks a depth of consideration needed to translate model development and results to useful take away lessons. This is addressed further under Presentation Quality using some specific examples. However, only examples are presented. The same kind of analysis presented for those examples needs to be considered for every aspect of the paper.

2. Scientific Quality: Are the scientific approach and applied methods valid? Are the results discussed in an appropriate and balanced way (consideration of related work, including appropriate references)?

I see no real problems with the scientific method to the extent I understand it. The problem is that it is poorly explained.

3. Presentation Quality: Are the scientific results and conclusions presented in a clear, concise, and well-structured way (number and quality of figures/tables, appropriate use of English language)?

The article is very confusing as presently written. There are problems with both the description of methods and the presentation of results. These are solvable problems within the present length of the article.

We are thankful for the rigorous assessment of our paper and glad that the scientific quality of our work has been recognized despite the issues with the presentation and the formulation that the reviewer identifies. We have taken her concerns at heart and improved the internal logic and reformulation of our manuscript and included a clear description of the general implication of our findings at the end of the Conclusion. Overall, we tried to improve the explanation of the methods used (see also the comments by Reviewer Zachary Subin and Nir Krakauer) and presentation of results.

Two examples of difficulties with the description of methods:

a. There are three comments about the linear store in the PCR-GLOBWB model that are confusing. I think probably the method used is ok, but the description is too confusing to tell. The three comments are as follows.
i. Page 5222, around line 20, says that the linear reservoir of the PCR-GLOBWB model

is replaced by MODFLOW.
ii. P. 5332 around line 21 says that "Because of the offline coupling and the lack of
topographical detail in a 60 cell, the linear groundwater store is maintained, specifically
for calculating baseflows above the drainage level to the surface water network using
a cell-specific recession constant which accounts for aquifer properties and drainage
density."
iii. P. 5228 describes how the MODFLOW head is used to calculate streamflow gains
and losses. The role of the linear reservoir in calaculating baseflow is not described.

We understand the raised confusion, and explained this concept clearer and in more detail to ensure any confusion is removed: p. 6, l. 17, and p.13 l. 11- 16.

In the original version of PCR-GLOBWB the groundwater reservoir, underlying the two soil layers, is described as a linear reservoir. We replaced this with a one-layered MODFLOW model, which makes simulated of lateral groundwater flow possible. In the MODFLOW model groundwater-surface water interactions are considered via river drainage along the main stream as represented in PCR-GLOBWB. Drainage via groundwater-surface water interactions as simulated in MODFLOW forms the main contribution to the baseflow. However, at 6' resolution, the main stream is insufficient to represent truthfully where groundwater levels intersect the terrain and additional drainage is needed to represent local sags, springs and streams higher up in valleys in mountainous areas. To resolve this issue, groundwater above the drainage level (taken equivalent to the river floodplain elevation, explained in section 2.2), can be tapped by local springs which is represented by means of a linear reservoir.

b. The method behind characterizing ranges and sediment basins is really confusing.
I broke the confusion down into the following pieces.
i. A small thing is that the term ranges is vague – perhaps use mountain ranges or high
elevation, steep terrain?

We agree with this suggestion and changed it to read mountain ranges accordingly.

ii. On p. 5224, line 9, is it correct that it could be "Next, all 6' cells with floodplain: : :"? If
this is correct, then note that the text is clear on how the floodplain level of each 6' cell
is determined, but is not clear on how the surface level of each 6' cell is determined.

We used the 30" data to determine the floodplain elevation at 6'as follows: Within each 6'cell and using the HydroSHEDS dataset, we identified the lowest elevation at 30'' (maximum 144 values for a cell comprising only land area), and assigned this as the floodplain elevation for the entire cell. We clarified this into more detail on p. 7, l.22 - p8, l. 5 in the revised manuscript (see also the raised question on this point by Zachary Zubin).

iii. The z-score of eq. 2 either needs motivation and clearer description, or to be
removed from the paper, described more clearly in another report or paper, and referenced
from here. One example of confusion is that equation is presented in the context
of the surface level and, while the figure being referred to (fig 3a) is described as "Cumulative
probability of aquifer thickness". Associated with this, figure 2 lower panel is
not referred to in the paper.

We concur with the reviewer that the underlying motivation for our choice may have been somewhat opaque. We have clarified this in the revised manuscript: p. 8 l. 10-15.

By definition basins are linked to sedimentary environments in fluvial systems and deltas. Sediments are deposited perpendicular to the main gradient (constituting the transversal axis of the basin), with grain size and volumes decreasing at greater distance away from the transversal axis. Grain size also decreases along the transversal axis, distinguishing proximal (near the source of sediment) and distal parts. We assume gradation in grain size are somehow captured in the GLiM but differentiation in depth is not. However, since we can locate the transversal axis along the main stream, we can use relative elevation as a measure of proximity to the transversal axis and as an indicator of the associated depth. We standardize the relative elevation to obtain the standard normal ordinate $Z(x)$ and use this to define the distribution of aquifer depths using a log-normal distribution, assuming depth is non-negative and positively skewed.

I think part of the confusion comes from it not being clearly stated that the PCRGLOBWB model is transient while the MODFLOW model is steady-state with forcings equivalent to long-term averages from the PCR-GLOBWB model.

We clarify this more on p. 5 l. 8- 9, p. 11 l. 5-6, and p. 1 l. 21, where we describe the land-surface model PCR-GLOBWB and inputs for recharge and river discharge. Long term averages for recharge and river discharges (outcomes of PCR-GLOBWB run for 1960-2010 (Wada et al. 2014)) are used to force the steady state groundwater model with.

The presentation of results is difficult because the global scale figures are hard to evaluate and there are too many of them. Some ideas are as follows.

Following the suggestion of the reviewer, we have rigorously evaluated the figures in our manuscript and reduced and focused the presented material and improved the overall visual presentation.

However, in this study we want to give global pictures and overviews of the different parameters and simulated groundwater heads, so that larger scale patterns are shown and can be understood. To focus the discussion at some points on specific processes we used insets, as the reviewer suggests.

We will shortly discuss what we changed in the figures, following the reviewer's comments.

a. Figure 3. put the top panel in auxiliary material or another pub with a better description of how it is obtained. This allows the thickness and T maps to be larger and easier to read.

As answered above (answer to iii), we clarify the use of the "frequency distribution of aquifer thickness" better to resolve the confusion. We agree with this suggestion to put this figure to the auxiliary material.

b. Figure 4. Choose a location and show detailed inset for each figure. This might require another figure. Choose the location to that a them can be followed and a transferable lesson told.

For the top figure "groundwater recharge", we want to show the spatial differences over the globe. At the regional scale groundwater recharge is one of the main controls for groundwater levels, we want to show the differences at the global scale.

For the bottom figure "rivers and drains" we agree with the reviewer that insets for specific regions will be useful. In the revised manuscript we use Europe and Africa, as this are also the regions of Fig.

9. The main message of this figure is to show that for 'big rivers' we consider the main drainage only, smaller drainage occurs are a 'drain'. This is explained more clearly in now on p. 11 l. 17-19.

c. Figure 5. Just show panel B and in the caption say that this is the dominant parameter. Say the other two parameters had very little sensitivities and the figure with all parameters varied looks a lot like the one shown. In the new figure, include an inset for the same area in the new figure.

We agree with the reviewer at this point. The improved figure shows now: "A) coefficient of variation in groundwater depth of 1000 runs with different parameter settings for aquifer thickness, groundwater recharge, and saturated conductivity. B) coefficient of variation of 100 runs with different parameter settings for saturated conductivity."

d. Figure 6. Make the print on the axes much bigger. The caption says A and B are pretty similar, so it is not clear that the distinction shown is very important. Omit one panel? I think the red points are in the sedimentary basins and the blue are in the ranges, but this is not in the caption. I may have missed item but I do not recall that the obs consistently being larger than the simulated values is discussed. Any thoughts on why this is so?

The spatial difference between observed and simulated groundwater depths (Fig. 7) show that in general groundwater depths are underestimated compared with observations for steeper and higher elevated terrains due to the grid resolution, which is to coarse to capture small local valleys. As a result, local drainage is underestimated for these terrains, and groundwater levels are lower. This is explained better now on p. 16, l. 22- l. 27. In fact, the simulated result shows the regional scale groundwater depth pattern, where the local groundwater is (most often) sampled. We assume that for steeper and higher elevated terrains local and perched water tables are likely to occur. The shallow groundwater depths observed for these areas confirm this assumption (see also reaction on comment Nir Krakauer).

The difference between the two scatter plots is the parameter set that is used (explained at p. 5233 l. 10-15 old manuscript). But we agree with the reviewer that the figure can be reduced to one scatter only. The new Figure 6 presents the scatter plot of observed heads against simulated heads of the best performing run. The distinction is made between mountain ranges and sediment basins, explained p. 17 l. 8-9.

e. Figure 7 is very hard to read. Decide what message is most important and focus the figure accordingly.

We rearranged this figure, now showing absolute residuals and histograms. With the maps of spatial distribution of residuals we want to show were the absolute differences are. The histogram shows the distribution of the absolute errors linked to the observed groundwater depths and show that larger residuals are mainly found for area with larger groundwater heads and vice versa. We removed the relative residuals, as this shows the same.

f. Figure 8 is not presented at a scale that allows the dotted areas to be distinguished, as far as I can tell. I am not sure what "lighter/washed-out colors" refer to – the lighter shades? Perhaps take the same area for which detailed insets are provided and do the same for this figure?

In this figure we want to show the global scale patterns or groundwater depths. We improved the figure captain: "…" washed-out colours indicate deep groundwater regions, where most likely shallow perched and local water tables occur, which are not captured by the model. The hyper arid regions area distinguished at the grid resolution.

g. Figure 09. This is the inset idea, but I think at too large a scale to make the figure very meaningful. Use the same location used for the other figures?

In these figures we want to show long inter-basin flow paths as well shorter ones that exist. We chose Europe and Africa as there is a clear difference in groundwater recharge en discharge for these regions.

Again, I commend the authors for their fine study of an important topic.

We thank the reviewer for her kind words and invaluable comments.

Response to comments from reviewer #3

We thank the reviewer for this positive response to our paper and thoughtful comments that significantly improved the paper. We have modified the manuscript accordingly, and detailed corrections are listed below. In the following the reviewer's comments are given in italics and our reply in normal font.

Reviewer 3

Review of "A high resolution global scale groundwater model" By I. De Graaf et al.

Scientific Significance:
Does the manuscript represent a substantial contribution to scientific progress within the scope of Hydrology and Earth System Sciences (substantial new concepts, ideas, methods, or data)?

The paper addresses a very interesting topic which is definitely suitable for the journal. I believe it is important to have a better consideration of groundwater at the global scale: this will help understanding and refining the results of the global hydrological model and it is definitely justified by the increase amount of data (also geological data) that are becoming available at the global scale. We are all aware of the limitations of the global set of data, but this should not prevent us from using these data for global scale models just clearly showing all the potential drawbacks.
The paper is definitely within the scope of HESS.

Scientific Quality:
Are the scientific approach and applied methods valid? Are the results discussed in an appropriate and balanced way (consideration of related work, including appropriate references)?

The applied methods that are applied are valid, but the general description of the approach is a little confused and mainly the discussion is not complete. For example, regarding the description of the methods, I found confused the explanation of the aquifer properties presented in chapter 2.2: I would suggest some rewording there .Regarding the discussion, I think it needs to be more focused on the very important and critical point . Just as an example, many times it is noted that there are problems because of the perched water tables in the mountains but it is not clear what is the actual impact of these observations (there is a figure that should show that, Figure 6, but it is impossible to understand that). Instead of concentrating on that I would find very interesting a discussion regarding the general overestimation of the results. What the authors think as the more reasonable explanation for that? I find it more interesting potentially with a higher impact on the future work than the problem with the perched GW table.

We are thankful for the rigorous assessment of our paper and glad that the scientific quality of our work has been recognized despite the issues with the presentation and the formulation that the reviewer identifies. We have taken his/her concerns at heart and improved the description of our methods and discussion of our results (see also comments from Zachary Subin and Mary Hill).

Presentation Quality:
Are the scientific results and conclusions presented in a clear, concise, and well-structured way (number and quality of figures/tables, appropriate use of English language)?

I will provide here a more detailed list of suggestions:
p. 5220 l. 10: the sentence is not clear

Following the reviewer's suggestion, we have rewritten this to read: "Their method however does not include hydrogeological information such as aquifer depths and tranmissivities, but uses estimates from soil data. Also, the hydraulic connection between rivers and groundwater, which is the primary drainage for groundwater in humid regions, is ignored." (p. 3 l. 29 – p. 4 l. 4)

p. 5223 l. 25: transmissivity should be in m^2/d so I do not understand why in the parenthesis k is correctly multiply by the thickness D but the units are m/d.

Correct, this is a typo

p. 5224 l. 20: I do not understand the difference between surface and floodplain elevation. Aren't they both obtained with Hydrosheds? And what happens when the aquifer thickness is zero?

We used the 30" data to determine the floodplain elevation at 6' as follows: Within each 6' cell and using the HydroSHEDS dataset, we identified the lowest elevation at 30'' (maximum 144 values for a cell comprising only land area), and assigned this as the floodplain elevation for the entire cell. We clarified this into more detail in the improved manuscript in section 2.2 p. 7, point 1. (accordingly to other reviewers Mary Hill and Zachary Subin) The aquifer thickness cannot become 0, as we assume a log-normal distribution. However, it can be very small.

p. 5227 l. 10-15: this should be better explained

By naming the cells 'true' and 'apparent' we wanted to clarify the difference between the projected cell area used by PCR-GLOBWB and the cell area used by MODLFOW, as the latter assumes rectangular grid cells. This means there is a difference in area, for which we should correct (as done in Eq 7). We explained this into more detail in the manuscript, p. 11, l. 5-15.

p. 5229 l. 3-5: am I right that this implies that smaller rivers will not lose water?

Only rivers with a width larger than 10 m are expressed in the river-package of MODFLOW which allows groundwater to drain to or draw from the surface water, given the gradient and riverbed conductivity. Rivers with a width smaller than 10 m are included via the drain package , which will only allow groundwater to drain to the surface once intersected by a drain placed at a particular depth (in this case surface level). This is better explained now on p. 11, l. 17-19.

p. 5229 l. 10-20: it is not clear which is the use of this in the MODFLOW model.

We understand the raised confusion, and explained this concept clearer and in more detail to ensure any confusion is removed (see also comment of Mary Hill).

In PCR-GLOBWB the groundwater reservoir is described as a linear reservoir. This we replaced with one layer of MODFLOW, which makes simulated of lateral groundwater flow possible. In the MODFLOW model groundwater-surface water interactions are considered via river drainage along the main stream as represented in PCR-GLOBWB and recharge. This is the main component of the baseflow, especially for sediment areas where groundwater flow is relatively slow. However, at 6' resolution, the main stream is insufficient to represent how groundwater levels intersect the terrain and additional drainage is needed to represent local sags, springs and streams higher up in valleys in mountainous areas. To resolve this issue, groundwater above the drainage level (taken equivalent to the river plain elevation), can be tapped by local springs which is represented by means of a linear reservoir, explained on p. 13, l. 11- 16.

p. 5230 l. 3-10: I am not sure that changing together conductivity and recharge can provide useful results: usually they are strongly correlated

Groundwater recharge and saturated conductivity are independent parameters here. Groundwater recharge is the (potential) flux from the soil to the groundwater given the surplus in soil moisture content (calculated by PCR-GLOBWB for two soil layers). A high groundwater recharge flux from the soil and a low transmissivity leads to less infiltration along a steam. And thus less groundwater recharge.

p. 5230 paragraph 2.5: how have transient data been used

The dataset used only contains steady-state data.

p. 5230 l. 26: put the figure number where these results are presented
p. 5231 l. 14: figure number is missing
We refer to the Figures were results are discussed (p. 15 l. 24 and p. 18 l. 10)

p. 5232 l. 8-9: not clear

We rewrote this section to make it clearer: "Although different thicknesses do change transmissivities, impact on calculated groundwater depths is small.", p. 16 l. 7-8.

p. 5232 paragraph 3.2: did I understand correctly that the observations have been used as they are?

The observed data was used as they are, but only data was used where surface level was reported as well (added on p. 16 l. 17). The average of the reported data was used when more than one measurement was available in one 6' grid cell (discussed on p. 14 l. 22-23). The used groundwater head observations are presented in Extra Material Figure 4.

p. 5233 l. 13: I do not understand why the blue dots are still in the figure and also regarding the same figure (figure 6) the difference between the two versions is not clear and I would suggest to present only one of them

The blue dots present the groundwater depths for mountain regions. The difference between the two plots is the parameter set used. For A) the best performing run is selected. For B) the best performing run based on validation of groundwater heads in sediment areas is selected.

However, we decided to change this figure after the review of reviewer 2 to only the scatter of A. The main message of this figure should be to 1) show the correlation of the best performing run, 2) show the difference between sediment areas and mountain ranges.

p. 5233-5234: the last paragraph of 3.2 is very confusing

We understand the confusion and rewrote the paragraph. We also decided to rearrange this figure (accordingly to the suggestions of Mary Hill) to make the presented results easier to understand, see p. 17, l. 11-18.

p. 5235: the conclusion needs to be restructured: the problem with the perched groundwater tables is repeated again, but I think that a discussion on the potential reason for the residuals being always negative should be included

We discuss this issue in more detail in the discussion of the residuals (p. 16, l. 24-26).

Figure 5: it is really hard to distinguish and I suggest a better explanation in the caption

We rearranged this figure (according to the suggestions of Mary Hill).

The parameter that effects groundwater heads most is the saturated conductivity. The other two parameters, aquifer thickness and groundwater recharge, are of minor importance (as explained on p.16 l. 6-12). We focus in the new figure on the dominant parameter.
We rewrote the figure caption to: "Coefficient of variation of groundwater depth: (A) of 1000 runs with different parameter settings for aquifer thickness, saturated conductivity, and groundwater recharge; (B) 100 runs with different parameter settings for saturated conductivity, the other parameters remain fixed."

Figure 7: in this figure, what is presented in the three maps is hardly readable and I suggest to keep just the histograms

We rearranged this figure. We rearranged this figure, now showing absolute residuals and histograms. With the maps of spatial distribution of residuals we want to show were the absolute differences are. The histogram shows the distribution of the absolute errors linked to the observed groundwater depths and show that larger residuals are mainly found for area with larger groundwater heads and vice versa. We removed the relative residuals, as this shows the same.