We have approved after the consideration of both reviewers and the editors comments. The paper more or less went through a large 're-vamp', where we mostly have addressed the main objections for this paper, which were:

- The pitch and presentation of the paper should be clearer.
    - We have re-written the introduction so that it should be clear now, that the presentation is: (1) global-scale models are important for global-scale research; (2) advanced local model are important for local studies; (3) but that national-scale modelling is also important and cannot currently be addressed by either global-scale or local-scale models' (4) that we have developed a possible solution that bridges that gap.
- The method description should contain no confusion on steady-state or transient.
    - We addressed this. See also replies to referee comments below.
- The paper should have more testing results, e.g., sensitivity analyses.
    - We have put in testing on calibration and different spatial resolutions, but also performed a sensitivity analyses that looks at the sensitivity of the water table to recharge and conductivity.
- The paper should contain more information on why the improved national scale (NWT) model is better than the original global-scale (EWT) model;
    - We have put in more and clearer comparisons on how the NWT model improves compared to the EWT model in two case studies.
- There should be less confusion about 'scales'.
    - In the abstract, introduction, discussion and conclusion we are as much as possible sticking to the same presentation as described in the introduction.

We will go over all smaller comments of the reviewers below. We have addressed all comments below (*referees comments in Italic* and our reply in red):

A1) There are linguistic weakness

(e.g. in the abstract: inconsistent terms and meanings with respect to smaller scale/

small-scale/local models and larger-scale/global models, l9: "because the quality

of their, coarse and global-scale, input data is large", l1: "larger, i.e. global", while larger

here should also refer to national).

Because of the different structure of the introduction, these terms and sentences have been removed altogether. We now tend to use more consistent descriptions of scale.

A2) The scientific terminology is not always used correctly, and some statements seem to be wrong (e.g. p6, l4 "ground-based, satellite-observed and modelled parameters"; p8 l16: g is not the gravitational constant but gravitational acceleration, and it is not 9.90 m2/s but 9.81 m/s2. Why "rainfall recharge" instead of "diffuse groundwater recharge"?

We've changed this to "ground-based, satellite-observed and modelled data".

p8 l16: “gravitational acceleration”. We’ve corrected this typo to 9.8 (the approximate value in NZ is 9.80).

A3) The reason for setting parameter values are not always clearly given, e.g. p8 l27: “As cell resolution of the NWT model is 200 m, the values of a, b and fmin of Eq. A2 were changed accordingly, to 75, 150 and 4”. To what extent does the cell size leads to is e.g. setting the value to 75 instead of 120 in the global-scale model?

We assume that the referee requires explanation as to why these values were used, which is given in the next sentence, that says: “These values were also used by a 200 m resolution EWT model in the Amazon basin (Fan and Miguez-Macho, 2010a) .”.

A4) The manuscript does not explain, except in the Appendix, one major simplification of both the global (EWT) and the New Zealand (NWT) model: There is no hydraulic gradient dependent interaction between groundwater and rivers; where the water table reaches the land surface, the groundwater is assumed to flow out. This, however, may be the main reason for the dominant overestimation of groundwater table elevation as compared to observations as river levels may be below the land surface elevation. Also, losing rivers cannot be simulated.

According to the earlier discussions on this reviewer comments, we have dedicated a paragraph (p14, l3-18) to this.

B Methodological weaknesses, combined with confusing presentation: In the manuscript, it is stated at various locations that the model was run in daily time steps for 100 years. However, a steady-state model cannot be run at daily time steps, because per definition there is no time variable in a steady state model, and the change in hydraulic head over time is zero. In addition, to do transient runs, one would need to set a storage coefficient, which is not mentioned in the manuscript.

We were referring with this comment to the strategy used in achieving convergence in the model result, yielding the sought for equilibrium water table. In the original global calculations with the model, iterations where performed with annual recharge values (i.e. yearly time steps). These speeded up convergence in most parts, however causing fluctuations in high slope terrain with deep water table and substantial recharge, as was the case of the mountains of New Zealand. For this reason, we tried iterations representing smaller time-steps, aiming at limiting the aforementioned fluctuations, and run for a number of steps that we quantified in number of years. Perhaps mentioning years was not a very appropriate choice, since as the reviewer points out, it can cause confusion. We have now better clarified in the revised manuscript that the model result is indeed steady-state and not transient, and refer to the number of iterations needed to achieve convergence with the raw number, instead of using a time-measurement equivalent, such as years.

C Lack of new information/innovation that is of general scientific interest The analysis is lacking components that would lead to improved scientific understanding. I suggest to analyze the specific reasons for the better fit of the NWT model results to observations. In a type of sensitivity analysis, variants of the NWT model could be run, in which only one “improved” data set is included while the other data sets remain those of the EWT model. Or alternatively where all but one data set is improved. For example, to understand the impact of the new hydraulic conductivity approach, do one NWT variant in which the approach used in the EWT model is applied. This would be a useful analysis to support the suggestion in the last paragraph of the discussion to use the hydraulic conductivity approach used for NWT (Gleeson data) also for the global-scale EWT, to improve it. But it was not

shown in the study whether with the EWT hydraulic conductivity approach the fit to observations in New Zealand would have been better. Similarly, the groundwater recharge estimate used for EWT could be used as input of another NWT variant, and the resulting water table elevations could be compared to the standard NWT results to understand the importance of improved/national groundwater recharge estimates. Then, the presumably large role of the DEM in improving results would be clearer, and your concluding statements would be more firmly based.

As we submitted to the Special Issue "Integration of Earth observations and models for global water resource assessment" forthcoming from the EartH2Observe programme, we addressed the issue that was also addressed in the EGU 2017 EartH2Observe session called HS1.12 (full session description at http://meetingorganizer.copernicus.org/EGU2017/session/23938), i.e., the collaboration gap between global modellers and local, catchment-scale, modellers.

The improved scientific understanding in our manuscript is that there is a possible role for global-scale models for smaller-scale studies (e.g., national or catchment-scale). After the reviewer's comments we chose to show improvement of NWT to EWT results (Section 4.2 , Table 2, and Section 4.3).

I would also suggest adding to Fig. 8 the simulation results of EWT to directly visualize the improvement of NWT over EWT, and adding to Fig. 11 also the results of the standard NWT with a spatial resolution of 200 m, not only the LiDAR-based 100 m variant that is shown (but not indicated in caption).

We have added this, as well as a substantial addition where model resolution is analysed.

Reviewer 2: specific comments:

The writing, and therewith the presentation and discussion of the research, should be significantly be improved. In addition to the points razed by R1, I suggest to rewrite the abstract and introduction and specifically focus on logic of the reasoning (meaning is a statement followed by the right argument and is the argument clear) and being as clear as possible. For example, abstract L2-3 reads: Large-scale models are simplified and not used at smaller-scales, because hydrology and water policy are constrained at the catchment scale. This does not make sense. What the author meant to say is that large scale models, are not useful for smaller scale groundwater assessments yet, because of the simplifications (and the coarse resolutions), therewith are not useful for e.g water policy. The next line reads: However, ... However, the statement in this line cannot be linked to the previous statement. Something like "for water policy smaller-scale models are more useful. However, ..." should be included. This are just two examples within the first three lines. Also, be careful using "this" "that" "it" without a summary word.

Overall from the abstract and introduction it was not clear for me what the main motivation and goal of this research were and how it will help us to improve current modelling efforts; to improve the EWT model but also be more useful for water managers? The lack of a logical structure and the bad writing are not beneficial for a clear understanding.

We have restructured the abstract and introduction so that the message comes across better now.

- I found the manuscript very limited in discussion of previous work, methodologies, results, and relevance of the work done. For example, on discussion of previous work: P4 L4 "many studies . . .." And then only one reference is a bit limited, as it is not a review paper you refer to.

We have improved it, so all of the mentioned topics now have two references. Also, we use the word 'existing' instead of 'many'.

P4 L7 "De Graaf apply a model. . .. Global-scale input data" This is too generalized, it should be a bit more specific what is meant with "a model" and "input data". Especially as you give some details for the Fan et al 2013 model. 1 to 2 Lines extra focusing on the differences between the two models referred to is needed. I know the models are quit difference. I little review here will also connect to the discussion, and will help you getting your point across why your model is better than the large-scale models available currently (see also my points later on)

We have incorporated this into our introduction.

P4 L17: How do you know groundwater models are less reliable in data-sparse regions as there is no data to validate the results. In the case of a model calibration, like done in this study, you can say your model performs best for the regions where you do have data to calibrate on (the whole meaning of a calibration).

This text, with its paragraph, has been removed to have an altogether clearer introduction.

Methodology and results: In section 2 it is not explained what happens when water tables hit the surface, nor is it explained that this is not simulated as a head dependent flux and river infiltration (water entering your aquifer) is not included. How realistic is this in the real world? (this should come back in conclusion/discussion as well) Also, your model result look very biased toward shallow water tables, (however not discussed in the manuscript). I think this positive bias can be explained by the way drainage is estimated (see also comment R1). Another aspect I do not understand is the storage and the convergence criterium that is left out. I agree with R1 that 'steady-state' in combination with a timestep is a bit confusing. How I understand it, is that you run the model over 100 years forced with the same climate data until an equilibrium is reached (i.e. a steady-state). I think for this kind of procedures the term 'dynamic steady state' is used often. (I certainly would not call it transient). What I do not understand, for such a dynamic steady state you still need a storage coefficient, so how does that work? Also, it is not yet clear what you used as a criterion to stop your run. It is written that the convergence criterium is not used, as running the model beyond 100 years did not improve model performance. But how did you decide that 100years were enough; did you check your model outcomes, estimate R for water tables and when that looked good you stopped it. Or was it wallclock time driven, or CPU time driven? I think whatever criterium you used is fine, but now it raises questions.

I fully agree with R1 on point C and more extensive sensitivity analysis should be done. From the results it cannot be concluded which model change has the largest impact on the results.

We have addressed these topics:

In the model review. We have added extra text on the bias to shallow water tables, caused by the fact that the model does not incorporate pumping or drainage. Still, we try to keep the model review short in the main text (as the EWT has already been described in full detail in earlier studies) with further references to an appendix.

Furthermore, we have added text that explains that the NWT model still has a bias towards shallow water tables, although less than the EWT model because of the finer model resolution. However, this bias of shallow water table is also a correct indicator of the fact that most of the indicated

shallow water table areas used to be wetlands: an approximate 90\% of wetlands have been lost since European settlement in New Zealand, mostly to develop agriculture.

We have also addressed this topic in the discussion, where we explain that the model resolves for river better, but still does not incorporate draining features.

We now have a clearer description where we leave out the '100 years' if needed, so it is extremely clear we have a steady-state model.

In my opinion, a relevant aspect of the discussion that is not/not enough elaborated is where we stand now and how it will help is further. How useful is your model in reality, as it is a steady state model approach, not simulating groundwater gradients, calibrated for New-Zealand, under natural conditions only, only unconfined aquifer systems? Are there now model that can do this maybe better, and under real world circumstances (i.e. current climate conditions and human impacts). In other words, if you need to advice the New-Zealand water managers, how should they use the model and what do they need to know about the model structure and uncertainties to interpret the results correctly and use the model to its full potential? It for which purposes can the model not be used, and what should be improved to make the model useful for the more real world simulation (varying climate and human interactions).

We have now hopefully clarified better, through a better and clearer introduction, on the potential applications where the EWT could be used to solve issues relating to: data-sparsity; national guidelines that cut across regions.

Reading the authors comments on R1 point C I think the authors should be careful in saying that regions where not modelled before (is New Zealand not included in the large-scale models, I think so); stressing the computational efficiency (how efficient is the model, and how does this compare to other large-scale model efforts?).
We have rephrased, also saying that this is the first 'dedicated' national groundwater model for New Zealand.

Minor comments: In the introduction a bit more details on the modelling should be given: 1 to 2 lines saying it is a flux-based approach, simulating steady-state water table heads, using averaged climate conditions, run for 100 years etc.

We have added slightly more elaboration in the introduction, but also to the model review, where we think it is more appropriate.

P7-L16: "drained by humans"; artificial drainage? P8-L6-7 "who . . .. ." Leave this out, it is not relevant as you do not use the parameters of Gleeson. P6 L5 "the improved NWT"; is this the same at L4 "the NWT" or is there also an improved version (leave out improved). F8: it would be more logic to switch those scatters, so that wte, discussed first, becomes (a) and wtd (b) (same for the other scatters).

Thanks, we've improved those.