We have provided point by point responses below with author response is in bold.

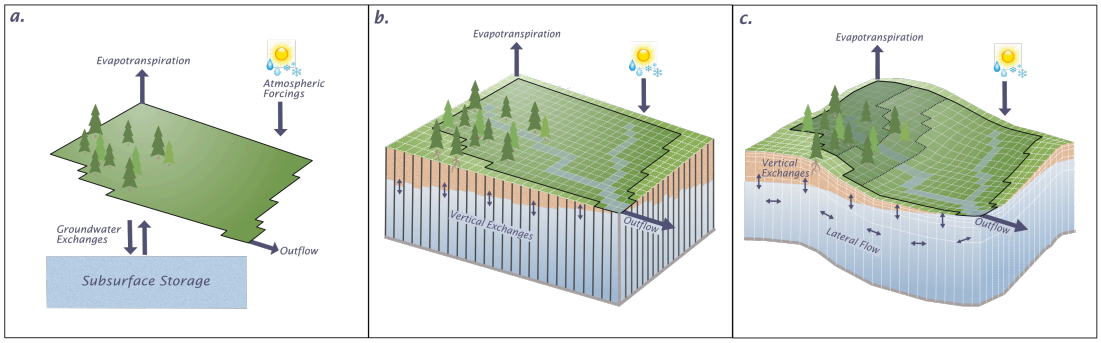
Overall, I find this manuscript to be interesting and in general, well written. It is also impressive to see these types of large, physically vigorous integrated GW-SW models applied towards problems that have typically been addressed with predominantly empirical methodologies. That being said, there are a number of issues with this work that need to be addressed prior to publication. Especially, I believe the authors need to address the major limitations in their approach.

We thank the referee for their thoughtful review of our manuscript. We completely agree that, as with other approaches, there is no perfect solution and there are still many limitations to this type of modeling. We hope that the reviewer will see that we have kept these limitations in mind when designing our experiment and in determining what type of questions these simulations are well suited to answer. We have clarified these points in our responses below and have also made significant revisions to manuscript to more clearly explain the purpose of our comparative analysis as well as its limitations.

Issue 1: Representation of the groundwater flow system.

As I understand from the description of the model domain, groundwater flow is primarily constrained within a single layer that extends from 2 m to 102 m below ground surface. Conceptually, this must mean that groundwater movement is restricted to a 2- dimensional plane parallel to surface. With such simplification, can the model really be used to assess the groundwater component of the water balance within nested watersheds that range from 10's to 100,000's of sq-km? If one can assume the presence of Tothian flow systems across much of the domain, how can local groundwater flow conditions develop when in essence the local systems will be overprinted by large regional flow? If the results of the work did not have such strong dependence on the model's ability to simulate the water balance in small watersheds this would not be such an issue; however, given the large number of small watersheds that are considered in the analysis, I believe it is crucial to have a more highly resolved subsurface domain. Even one or two additional layers within the groundwater flow zone would allow 3 dimensional groundwater flow systems to develop, and hence facilitate the model's ability to capture local flow systems that are a key part of the water balance, especially in humid areas with notable topographic variability such as the Eastern and Northwestern regions in the domain. Furthermore, a graphic that depicts how the watershed nesting has been conceptualized within the model domain framework would be of value to readers.

The referee is correct that we have a simplified representation of the subsurface and we agree that additional layering and increased resolution would improve the simulation. However, we would like to clarify that the current approach does not mean that all groundwater flow is occurring parallel to the surface. Previous work has evaluated groundwater flow systems within this model and showed power law residence time distributions as well as nested systems of local and regional groundwater convergence [*Condon and Maxwell*, 2015; *Condon et al.*, 2015; *Maxwell et al.*, 2016]. We do observe systems of local convergence and our ability to balance water does not depend on the watershed size. ParFlow is simulating a gridded domain and it balances water for every grid cell and time step. The nested watersheds are only used to group results for analysis as a post processing step and are not units of simulation. To make this point more clearly we have added the following conceptual figure to the methods section which compares our integrated hydrologic model (subplot c) to lumped parameter models (subplot a) and gridded models which incorporate only vertical interactions with the subsurface (subplot b). Subplot c of this figure also illustrates the nested watershed approach used here.



Also, we would like to stress that the purpose of our analysis is not to predict the magnitude of storage changes across the US for water year 1985. Rather, we are using this simulation as a way to sample a broad range of groundwater storage changes in order to illustrate the way these changes would impact various Budyko approaches. We agree that improving the subsurface characterization could change the level of storage changes that we simulate, however these changes would not impact our findings. We do not think that this point was made clearly in the original manuscript and we have significantly revised the text to better clarify our goals.

Issue 2: Applicability of the Gleeson et al. (2011) dataset.

While it does need to be recognized that parameterizing the subsurface component of these large integrated models is challenging, and the use of large scale homogeneous datasets is particularly attractive, modellers must be cognizant of the limitations that these datasets impose on the model. This point is highlighted with the Gleeson et al. (2011) subsurface permeability map, which is an extreme simplification of subsurface hydrostratigraphy in order to facilitate global coverage. For an application such as the one of focus here, the Gleeson et al. dataset does not provide adequate spatial accuracy/resolution for credible model results to be generated for smaller-scale watersheds, and considering the large weighting that smaller-scale model results for the groundwater component of the water balance are given in the analysis, this is very problematic.

We agree with the referee that subsurface datasets remain a significant limitation for large scale groundwater simulations. The permeability map developed by Gleeson is not perfect, but it is the only consistent dataset available for the entire continental US (it should be noted also that the resolution of this dataset is higher for the US than it is globally). If the goal is to build a model to precisely predict local groundwater levels in small scale systems, we agree with the referee that this is not the ideal dataset to start from. However, this is not the intention of this work. We focus on using this model as a tool to characterize behaviors across a broad range of physical settings and spatial scales that incorporate realistic heterogeneity. We have intentionally designed our experiment so that the credibility of our results is dependent on the physical processes we are simulating and not any local calibration. We have added the following text to the methodology section to make this point more clearly to other readers (new text is underlined):

> ***"The 1-year simulation presented here intentionally violates the steady state assumption. The purpose of our analysis is to evaluate the impact of net storage changes on Budyko relationships, therefore a steady-state simulation is not the goal. It can also be argued that storage changes will vary from year to year or depending on the multi-year period analyzed.***

The 1985 simulation year is not presented as a prediction of long-term storage variability, it is simply used to sample a range of groundwater surface water exchange across variable climates and physical settings. We present a general framework for understanding the impacts of storage changes in various Budyko formulations using water year 1985 as a representative example.

Similarly, because we are focused on a comparative analysis within the Budyko framework the results are not dependent on local calibration between simulated results and observations. The discrepancies between approaches stem from differences in the variables used to create a water balance (refer to sections 2.3 and 2.4); these findings are not sensitive to parameter uncertainty in the model. Still, the transient simulation has been rigorously validated against all publically available observations for water year 1985. This includes transient observations at varying frequencies from 3,050 stream gauges, 29,385 groundwater wells and 378 snow stations for a total of roughly 1.2 million comparisons points. Flux tower observations were not available over this period, but latent heat fluxes were also compared to the Modern Era Retrospective-analysis for Research and Application (MERRA) dataset. Complete details of the model validation are provided in the supplemental information of Maxwell and Condon [2016].

Although there are of course limitations to the model and significant uncertainties in spatial model parameterization, especially for the subsurface, overall comparisons between simulated and observed values demonstrate that the modeling approach is robust. Stream-flow timing and magnitude are generally well matched in undeveloped basins, snowpack timing and melt is accurate and spatial patterns in latent heat flux are reasonable. Most importantly for this analysis, the model validation shows that ParFlow is accurately capturing the relevant physical processes. Uncertainty in subsurface parameterization, bias in atmospheric forcing data and lack of anthropogenic activities were identified as key areas that could improve the local predictions of the model. However, as discussed above, the purpose of this work is not to predict Budyko curve parameters for water year 1985. The uncertainties listed here are therefore important to note, but do not limit the utility of this tool as a test bed for evaluating interactions across spatial scales and complex physical settings."

Issue 3: The use of modelled ET.

Typically, and especially in large scale models, it very difficult to accurately simulate ET. In this work the authors use simulated ET as a surrogate for measured ET across much of continental USA. Considering the extreme importance of ET in the water balance analysis, I wonder if biases or errors in simulated ET may not be skewing the results. This point may be highlighted in Figure 2d, where it appears that groundwater recharge is unrealistically high across much of the Great Plains. Furthermore, as Figure 10 highlights, the analysis results that are dependent on simulated ET show a strong deviation from the results generated from the other two water balance calculation methodologies.

All of our analysis is based on simulated results because our goal is to characterize the way that storage changes perturb points within the Budyko space. We have attempted to clarify this point more fully in the following text that was added to the start of the methods section:

"We use an integrated hydrologic model to simulate water and energy fluxes in both the surface and the subsurface. Here we apply a high resolution (1 km^2) simulation of the majority of the continental U.S. which covers more than 6 M km^2 and simulates hydrologic systems across a broad range of physical settings and storage change magnitudes. The model is driven using historical observed atmospheric forcings such as precipitation and temperature, and

provides gridded outputs of all water and energy fluxes throughout the system. We use simulated surface water flow, evapotranspiration and groundwater surface water exchanges to calculate Budyko relationships using three different approaches to estimate fluxes:

1. ***Calculating evapotranspiration from simulated runoff and precipitation***
2. ***Using simulated evapotranspiration values directly***
3. ***Using simulated evapotranspiration values directly and taking into account storage changes.***

Differences between the approaches are compared with storage changes in each basin to evaluate the systematic impacts of these changes on Budyko relationships.

The numerical modeling approach used here provides several important advantages for this type of analysis. Within the numerical framework, groundwater surface water exchanges for every watershed in the system are fully characterized. This guarantees perfect closure of the water balance and means that we can mimic all three approaches within a consistent numerical framework where storage changes are directly accounted for. Furthermore, because the goal is to understand differences between approaches, and not to predict local Budyko parameters, the key advantage here is the ability to evaluate physically realistic behavior across a variety of physical settings and spatial scales where groundwater can be fully accounted for. Within this context, it should also be noted that the focus is on how groundwater storage changes perturb relationships. Therefore, uncertainty in local model parameters is much less important than realistic simulation of physical interactions for a range of storage changes and aridity values within a controlled numerical framework."

Also, we have added the following note to the description of the direct evapotranspiration method to emphasize this point:

"*Note that in this case we are still using simulated E not observations. The intention is to treat the model as our simulated truth and compare variations within this framework*"

Figure 10 does show clear differences between the direct evapotranspiration approach and the other approaches, but this is not an indication of model bias because all of the results in this figure are based on the same simulated outputs. The point of figure 10 is to show that the way storage changes are accounted for in different methods will systematically impact results. In all cases, the data underlying the plots is exactly the same; we are just showing here that you get a different answer depending on what parts of the water balance you choose to analyze. Based on this comment and others we think that this point was not made clearly enough in the original manuscript. We have expanded the discussion of all of the results figures to try to emphasize these points better.

A few other minor comments are as follows:
L16: be careful with use of 'realistic' this work is more conceptual in nature
We agree that 'realistic' is a relative term and in the revised abstract we do not include this language.

L96: . . .a physically. . .
This is no longer included in the revised manuscript

L120: abcd?
'abcd' is the name of the hydrologic model applied by Du et al.

L135: expense,
This sentence has been revised.

L142: comma not needed
This has been corrected in the revised manuscript.

L150: technical feasible yes, but how realistic is it to extract local scale information from continental scale models?
This is of course an open point of debate. We would argue that even if there is local uncertainty there is still much that can be learned about spatial organization and process interactions across scales from these models. Please refer to our responses to the major comments above for a more detailed response on model limitations.

Eqn 1: Are the units expected to balance?
Yes, we have confirmed that units of this equation do balance. The units of every term are 1/time.

L167: Verify units for q
The units for q are correct as stated (i.e. 1/time). Fluxes are normalized by the cell thickness in ParFlow so the length dimension is canceled out in this formulation. We have added the following text to the manuscript to clarify this point:

> ***"Note that units of T^{-1} for the flux terms reflects the fact that they are scaled by the cell thickness."***

L201: Ev, and. . .
This has been corrected in the revised manuscript.

L263: Is a single year really ideal?
In this sentence we were intending to say that the pre-development approach was ideal. We have revised the sentence as follows:

> ***"Therefore, the simulation represents natural flows in a pre-development scenario, which is ideal for Budyko analysis."***

General comment: a histogram showing watershed size distribution would be valuable.
We appreciate the suggestion. In response to earlier comments we decided to add an additional conceptual figure explaining our modeling approach and the system of nested watersheds we are using. We have decided not to add additional histogram figure here; however, we hope that the reviewer will see from this response that our findings are not sensitive to the size distribution.

L298: balanced
This has been corrected.

L314: for
This has been corrected.

L331: opposed to
This has been corrected.

L340-343: This statement is not really supported by Figure 2, which is presumably (authors should state this in caption) depicting ratios in annual totals. This statement is also irrelevant to main objectives, suggest deletion.
The text on lines 340-343 explains that groundwater contributions fractions are significant throughout the domain and therefore the system is not in a steady state over the annual simulation period. We are a little confused by the comment that this statement is irrelevant to the main objectives of our work because the entire purpose of this analysis is to evaluate the impact of storage changes on Budyko relationships. It is in our opinion critical to show that we have selected a simulation period where storage changes are occurring. We have clarified in the caption of Fig. 2 that subplot d is plotting G/P and we have revised the sentences in question as follows to hopefully be more explicit about this point:

> **"*Within this annual simulation, subplot d shows that groundwater surface water exchanges (G/P) can be a substantial portion of the water balance in much of the domain. This indicates that the system in not in steady state over the simulation period. As discussed in Section 2.2 the one-year simulation time was intentionally selected for this reason. Here, we take advantage of the ability to directly calculate groundwater surface water exchanges within a controlled numerical simulation where such exchanges are prevalent in order to evaluate the impact of storage changes on Budyko relationships across a range of spatial scales and climates*."**

L396: one,
This coma has been added.

L410: and often
This text has been removed from the revised manuscript.

L437 onwards: these results and discussion should be supported by at least some basic curve fit statistics.
It is not possible to fit the results to a single curve because the premise here is that different points are falling on curves with different shape parameters. We agree that some quantitative metrics are needed though. In response to this comment we have evaluated the number of points falling within the bounds of the curves defined with shape parameters of 1.6 and 3.6 and revised the associated text as follows:

> ***"Fig. 4 plots every watershed in the domain shown in Fig. 1 using the three approaches to estimate the evapotranspiration fraction. In all three figures, the watershed points follow the overlaid Budyko curves; 77% of the watersheds fall within the 1.6 to 3.6 shape parameter lines for the inferred evapotranspiration approach, 51 % for the direct approach and 72% for the effective precipitation approach."***

L514: correlation discussion should be supported by some r^2 values.
In this case it is not possible to calculate an r^2 value because we are just talking about the relationships at the lower limit of the plot. However, we have revised the text in question as follows:

"Finally, a scatter plot of shape parameters versus groundwater contribution fraction for the effective precipitation case (Fig 6c) shows similar patterns with aridity but no clear correlation between storage changes and shape parameters for the lowest shape parameter values."

Additionally, we have added r^2 values to the discussion of Figure 7 as follows:

"Also, there is a much stronger correlation between the inferred evapotranspiration and effective precipitation approaches (Fig. 7b) than between direct evapotranspiration and effective evapotranspiration approaches (Fig. 7c) (r2 value of 0.96 comparing inferred vs. effective as opposed to 0.32 for inferred vs. direct)."

L576: higher curve numbers? Visually the results look the same, could statistics be used here as well?
In response to this comment we have expanded this discussion as follows:

"Fig. 7b which showed strong correlations between the shape parameters of these two approaches (r2=0.96) but a slight positive bias with positive groundwater contributions for the inferred evapotranspiration approach; 62% of watersheds overall and 86% of watersheds with a positive groundwater contribution have a higher shape parameter using the inferred evapotranspiration approach."

L578: do you mean Fig. 6?
This should have been referring to figure 7b. We have corrected this in the revised manuscript.

L603: 100? Do you not mean 1000?
Yes, this should be 1,000 not 100. This has been corrected.

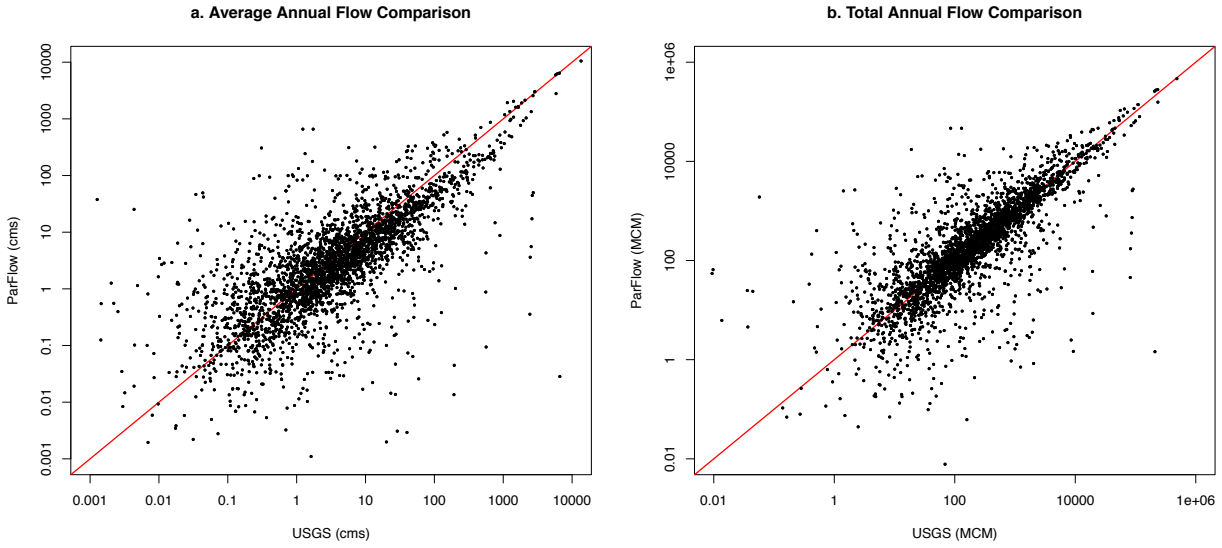
L607: References needed here.
We have added two references here: [*Budyko*, 1974] and [*Donohue et al.*, 2007].

Figure 9: General observation here: As the authors state, the analysis results that focus on the larger watersheds provides a better match to the idealized curves. Could this result be at least in part due to the major issues identified earlier in this review that highlight the weaknesses in the authors approach towards simulating the groundwater component of the water balance for small watersheds?
We refer the referee back to our earlier response with respect to groundwater simulation in small watersheds, which we do not think is a limiting factor here. In the detailed validation provided in *Maxwell and Condon* [2016] strong performance was observed across large and small watershed and there was no systematic bias with respect to drainage area. Also, we would like to clarify that what we are showing here is that the results converge more closely around a single curve for larger watersheds and that there is more variability for smaller watersheds. This finding is consistent with the original work by *Budyko* [1974] that showed a better convergence around a single curve for basins larger than 10,000 km^2 as well as subsequent research by *Donohue et al.* [2007] which explained the need to include more catchment specific effects as you move to smaller drainage areas. The reviewer is correct that uncertainty in local parameters can shift the points for individual watersheds; however, because we are looking at more than 30,000 watersheds here the overall result shown in Figure 9 is a function of the variability sampled at every scale. Furthermore, we are showing the effective precipitation approach here as an example, but as noted in the text, similar convergence behavior was observed for the other two approaches as well. Therefore, the convergence at larger scales is not a function of the approach used. This is also demonstrated in Figure 10 where you can see that the

three approaches have different central tendencies but all follow a pattern of decreased variance with increased drainage area.

L612: No doubt watershed characteristics are important, but what about limitations with the way groundwater flow in small watersheds is represented in the model, and the adequacy of the employed datasets for meeting such finely resolved objectives?

Again we refer the referee back to our earlier reply. We do not feel we have a specific limitation with respect to groundwater flow in small watersheds. While we agree that datasets are uncertain and could change individual point values, this would not change the overall point that variability between watersheds will be greater for smaller drainage areas than for larger areas. As noted in the manuscript, we have validated the model against more than one million observations points for the 1985 simulation period. The detailed analysis of these comparisons is presented in the supplemental information of *Maxwell and Condon* [2016]. While we refer the reader to the original publication for details on model validation, we would like to point out here that our results do not show a systematic bias with drainage area. For example, the comparison of simulated versus observed stream flows from *Maxwell and Condon* [2016] reproduced below does not show a systematic bias for smaller stream gauges.



Comparison of simulated vs observed (i.e. USGS gauged) streamflow [Maxwell and Condon, 2016].

Summary and Conclusions: General comment for this section is that it is too long and there is too much detail provided in the reiteration of methodology and results. Suggest shortening by at least 50% and focusing more on the significance of the results.

We appreciate the suggestion and have shortened the conclusions and refocused them on the significance of our findings.

Works Cited:

Budyko, M. I. (1974), Climate and LIfe, Academic Press, New York.

Condon, L. E., and R. M. Maxwell (2015), Evaluating the relationship between topography and groundwater using outputs from a continental-scale integrated hydrology model, Water Resources Research, n/a-n/a, doi: 10.1002/2014WR016774.

Condon, L. E., A. S. Hering, and R. M. Maxwell (2015), Quantitative assessment of groundwater controls across major US river basins using a multi-model regression algorithm, Advances in Water Resources, 82, 106-123, doi: http://dx.doi.org/10.1016/j.advwatres.2015.04.008.

Donohue, R. J., M. L. Roderick, and T. R. McVicar (2007), On the importance of including vegetation dynamics in Budyko's hydrological model, Hydrol. Earth Syst. Sci., 11(2), 983-995, doi: 10.5194/hess-11-983-2007.

Maxwell, R. M., and L. E. Condon (2016), Connections between groundwater flow and transpiration partitioning, Science, 353(6297), 377-380, doi: DOI: 10.1126/science.aaf7891.

Maxwell, R. M., L. E. Condon, S. J. Kollet, K. Maher, R. Haggerty, and M. M. Forrester (2016), The imprint of climate and geology on the residence times of groundwater, Geophysical Research Letters, 43(2), 701-708, doi: 10.1002/2015GL066916.