Reply to anonymous referee #2, Maxwell et al GMD- 2014-198 "Simulation of groundwater and surface water…". Original reviewer comments in **bold** *replies in italics.*

The paper is very well written and I appreciate the immense effort to set up and run the model. However, for a number of reasons the paper is not publishable. There are number of issues, both obvious and hidden ones. The general claims based on model outcome are neither substantiated nor discussed in detail and/or are mainly trivial.
A critical discussion on how the extreme uncertainties (e.g. the underlying geology) might challenge the bold claims on scalability is conspicuously absent. What is most troubling though is the presentation of the results. The paper suggests a reasonable to good model performance. Unfortunately, critical data that would allow the reader to evaluate how well the model performs in reproducing hydraulic heads are missing. We see scatterplots over scales that simply cannot be wrong. Over large spatial scales groundwater always roughly follows topography, and a plot with altitudes ranging from 0-4000m will always look good, no matter how bad the model is. What would be really interesting instead is a scatter plot showing simulated and observed depth to ground- water. This would allow more transparent insights into model performance, but we are only shown histograms. Why? The flow model could be completely wrong in predicting spatial patterns yet the match of the histograms might still be excellent. I find this presentation very misleading and not an appropriate way to demonstrate model performance. Also, the comparison between simulated and observed river discharge looks terrible, even on a log-scale. The model should simply not be used. There is a lack of novelty. Instead, there are several claims on novelty that in fact are not at all new at all (see below).
The conceptual setup of the model is somewhat incomplete and under- mines the philosophy of a fully integrated approach by externally calculating infiltration rates (P-ET). Fractured and karstic systems are not modelled, important processes are missing yet little is said about this. For all these reasons, the paper should be released. Below I provide some discussion concerning the points I made above.

While we respect the Reviewer's opinions and thank them for acknowledging the effort that goes into this type of simulation, we feel that there is a fundamental difference of opinion between the viewpoint of the Reviewer and what is put forward in this work.

The model is not meant as a watershed model that has been calibrated to e.g. discharge data and/or groundwater level data. This is not the objective of this study. The study is designed as a numerical experiment*, simulating integrated surface-subsurface flow at continental scales to interrogate spatial scaling behavior across 4 orders of magnitude spatially. The numerical experiment is* <u>*directed at*</u> *the CONUS domain and hydrology, however it does not aspire to include all aspects, and may as a matter of fact miss important aspects of the hydrogeologic systems, such as karst as* *mentioned by the Reviewer. We do not intend to include this type of complexity at this point.*

We do not intend to calibrate either, but compare in an ad-hoc fashion to observations as a sanity check. This check exhibits surprisingly good agreement even without calibration despite the simplifications and shortcomings, which lends confidence in the experiment in representing key process of surface-subsurface flow at the continental scale using a physics-based approach. To our knowledge no other numerical experiment exists with this type of model at this resolution at continental scales, thus the presented study is indeed novel and results are of interested to the hydrologic community. We are resolving and analyzing hydrologic variability over 4 orders of magnitude in space, which provides new insights in e.g., potential scaling laws and has not been done before in coupled surface-subsurface hydrogeology honoring 3D variably saturated flow.

We are very much aware of all aspects of uncertainty and have dealt extensively before with uncertainty quantification and propagation (Maxwell and Kollet 2008; Kollet 2009; Meyerhoff and Maxwell 2011; Meyerhoff et al. 2013). However, in the current numerical experiment, the impact of uncertainty is secondary, since we do not intend to fully reproduce and predict the real-world system i.e. the CONUS domain. Future studies will also include a detailed uncertainty analysis.

While the immense challenges and benefits associated with the presented numerical experiments may not be apparent to everyone, those who are active in the integrated modeling community know that spatially resolving hydrologic variability across multiple orders of magnitude using mechanistic models is an ambitious goal that cannot be achieved if we do not support efforts in model development and numerical experiments as they are presented in this study.

This work is a <u>step</u> towards hyper resolution, physics-based, continental scale models of the coupled water energy budget. We are confident that this work is an advancement that moves us closer to the types of modeling and analysis advocated by Wood et al (2011) and Bierkens et al (2015) and that this contribution will be well received by others who are working in this area. We feel that it is important to move beyond calibrated watershed models and apply realistic integrated physics-based models at the continental scale to study scaling behavior and perform e.g. water resources assessments and scales that are relevant for societies. For example, physics and atmospheric sciences are excellent examples of stimulating and testing scientific hypotheses based on numerical experiments using physics based models.

Therefore we have made a strong effort to respond to all comments provided by the Reviewer. In addition to our response to specific comments we provide the following responses to the general concerns noted by the Reviewer:

1. ***This work is not novel***
   The Reviewer makes several comments referring to the fact that this type of simulation has already been achieved and that this work is not novel. We strongly disagree with this assertion and would like to point out that the Reviewer has failed to provide any concrete examples of others who have accomplished what we present here. The only example provided is the work by Kollet (we assume they are referring to Kollet et al., 2010). Kollet et al., 2010 simulated a simple domain using a synthetic dataset without topography. This study simulates a real domain with topography, heterogeneous subsurface and forcing. Furthermore, our model is not comparable to other studies (not listed by the reviewer), which have simulated the entire US with separate groundwater or surface water models. Our model provides fully integrated physically based flow through a 3D variably saturated subsurface and surface. This is a critical difference that allows exchanges between the surface and subsurface to evolve dynamically. We challenge the Reviewer to produce references of studies of the same scope. Interestingly, Reviewer 1 states this is a novel advancement, while this Reviewer acknowledges that while we do get the rivers in the correct place that this is a detriment. The development of integrated models and their differences with respect to coupled and single component models are well documented (Maxwell et al. 2014). While, it seems apparent from their comments that the reviewer does not value these differences, we assert that there is significant interest of the hydrologic community, published articles of different groups and number of citations supporing the presented approach.

2. ***The results of this modeling effort are trivial***
   The paper focuses on detailing the development of the model, demonstrating that the model is simulating realistic hydrologic response with publically available datasets at the continental scale, and is useful in studying hydrologic scaling behavior across a number of spatial scales. All three aspects are far from being trivial. This includes the finding of generality of the broad range of scaling relationships of flow with basin area, which has not been established before with physics-based models across continents. As a matter of fact, one of these references claims there is no generality in this relationship.

3. ***Inputs are too uncertain***
   In hydrogeology, this is a trivial comment, since data scarcity is ubiquitous. In the presented numerical experiment, uncertainty is a factor of course, but of secondary importance at this stage of the work, because we are presenting first results e.g. the general scaling behavior. The potential impact of uncertainty will be explicitly discussed in the revised manuscript and will be subject of future numerical experiments.

4. ***The model does not perform well and we should not use it***
   There are differences between model results and observations and we are transparent about this. Therefore, the comment of the Reviewer of a misleading

presentation is not acceptable. We are presenting a numerical experiment that is directed at reality without calibration, thus we fully expect that there will be differences between outputs and observations, which are reported. As a matter of fact, the agreement with observations is surprisingly good including stream discharge and hydraulic head considering an uncalibrated numerical experiments based on uncertain input data sets. In case of hydraulic head the comparison is presented in a way that is consistent with other studies of this scale (Fan et al. 2007; Fan et al. 2013).

Replies to the specific comments of the Reviewer are provided below.

- "Over large spatial scales groundwater always roughly follows topography, and a plot with altitudes ranging from 0-4000m will always look good, no matter how bad the model is."
  We agree that hydraulic head (which is the plot we assume you are referring to since you note the axis range of 0-4000m) will very closely follow topography (we make this point on lines 284-286). This is why we have included a histogram of water table depth as well (note that the range is 0-100m for this plot) and we discuss on lines 286-289 that this plot reveals that we have a wet bias. Nevertheless, it is obvious from first principles that hydraulic head is the driving force for lateral groundwater flow and thus an important metric of model performance.

- "What would be really interesting instead is a scatter plot showing simulated and observed depth to ground- water. This would allow more transparent insights into model performance, but we are only shown histograms. Why?"
  We show the observations prior to our model results in the manuscript and we use a number of approaches to compare to these observations that have been established elsewhere and we subdivide these comparisons regionally. We highlight areas where input datasets have a strong influence on results and discuss how this will impact prediction. We do this in an open-access journal. We feel this is a pretty transparent way to present the model and comparisons that allows the reader to understand the full nuance of performance.

- "The flow model could be completely wrong in predicting spatial patterns yet the match of the histograms might still be excellent. I find this presentation very misleading and not an appropriate way to demonstrate model performance."
  The suggestion of misleading presentation is not acceptable.
  We checked the spatial performance and analyzed the outputs consistently. All results can not be presented in the manuscript. While we appreciate that the reviewer might prefer some different figures, we cannot respond to this comment unless specific suggestions are provided.

- "Also, the comparison between simulated and observed river discharge looks terrible, even on a log-scale."
  What does the Reviewer mean by terrible? In case of the presented uncalibrated numerical experiment, the comparison to discharge shows surprisingly good agreement. Again, we feel this is a misleading comment by the Reviewer.

5. ***This is not a fully integrated model because you are specifying P-E***
   This is correct, the model is not closing the full hydrologic cycle. Fully integrated refers to surface-subsurface flow. This will be clarified in the revisions.

6. ***Important processes and Karst systems are not included***
   This is a numerical experiment directed at the CONUS domain, not a representation of reality. The model will be improved successively in future. We have made every effort in this manuscript to be clear and transparent about the capabilities, and limitations, of our model and we present this work as a building block for future advances.

Novelty:

The numerical model is not new. A proof of concept on high resolution modelling has already been published by Kollet. That the spatial scale can be increased with increasing computational power is in itself not a novelty. Going to much higher resolutions across this spatial scale (e.g. 20*20m horizontal and centimeter scale vertically) would however merit a publication.
The comment on novelty has been discussed extensively above.

.

Page 7326, line 27: The authors claim here that it is a novelty that streams form without predefining their presence in advance. This is simply not true. The models discussed in the introduction are all capable of this. Just by increasing the spatial scale of a feature does not make it new in any way. Also, the authors claim that their approach will capture the complete steam network. This is in theory true, but to what extent the model captures the network is dependent of the spatial resolution and quality of the DEM. On a 1 km scale as used here one is very far from capturing the complete network.
R1 mentions this and considers it a novelty. We have already responded to this comment for R1 and added some clarification to this statement.

There have already been previous attempts to model the entire US. That a fully integrated model was used here is not the great step ahead, especially given that integrated refers only to the interaction between surface water and

groundwater as mentioned above, not to the calculation of the important dynamics between infiltration, ET and recharge.
This comment reflects only the opinion of the Reviewer. We consider the application of a 3D variably saturated integrated groundwater-surface water flow model in a numerical experiment at continental scale a great step forward. No experiment of this order has been published so far. It is up to the Associate Editor and Editor to make a final discussion on this disagreement. Integrated models, as defined in Maxwell et al WRR 2014 include 3D Richards and some form of the shallow water equations. We apply potential recharge on the top of the domain from a publically available product (Maurer et al. 2002) that has been widely used in the literature (cited at least 479 times according to Web of Science).

The paper claims to provide one means to bridge point measurements of hydro- logic states and fluxes to continental scales (page 7319, line 5). This is an extreme overselling of the model capabilities, especially given its poor performance even in steady state, undiscussed uncertainties and its incomplete process conceptualization.
The performance of the numerical experiment is not poor as discussed above. In the revision, language will be revised and uncertainty will be discussed.

The general hydrologic behavior in steady state is well established for the US. There is nothing added with this model.
Again, we respectfully disagree with the reviewer, steady-state or not, as mentioned earlier this is the first numerical experiment with a 3D variably saturated integrated groundwater-surface water flow model at continental scales. Please also refer to our responses above.

Why was only a subsection of the US modelled? The authors had to make an award choice of a no-flow boundary condition.
The study is a numerical experiment directed at the CONUS domain. Boundary conditions are locations where all models suffer inaccuracies. As the goal was to model large-scale watersheds (Mississippi, Colorado) and clearly the subsurface flow conditions do not recognize surface topographical divides, we felt the most appropriate choice was to move the boundaries as far from the regions of interest as possible. ParFlow has the capability to resolve irregular boundaries and provide for complex and time varying boundary conditions, all of which could be added in future iterations of this model, however, we feel confident that the choice of any of these boundary conditions would not alter the simulation results.

It is a pitty that ET is not simulated, in my opinion the greatest advantage of fully integrated models. I am aware that to adequately simulate ET much finer vertical resolutions are required. But how good are the P-ET maps used? By not including these important dynamics the authors give away one of the greatest advantages of fully integrated models. The only advantage they retain is that the location of the rivers must not be predefined. However, I see no disadvantage to predefine the location of all medium to major rivers on this

spatial scale, given that fact that most of the rivers are too small to nicely flow downhill in a 1km DEM anyway.

The P-E dataset used is based upon a product derived from the VIC land surface model and has been widely applied in large-scale hydrologic studies (Maurer et al. 2002). This datatset is extensively evaluated in that paper and we refer the Reviewer (and the reader) to that analysis. It is planned to include the land surface moisture and energy balances in future. The comments implying the lack novelty have been discussed above.

Nothing is said about the correction of the DEM. In such fully integrated models, water will follow topography. Using a DEM with a 1 km resolution without any correction is not advised, as topography along rivers and drainage networks will go up and in the model, preventing the generation of small to medium streams.

In the original submission, lines 3-6, we state that the GRASS R.Watershed package was used to process the DEM. This is a standard technique commonly used to correct DEMs. Thus, the Reviewer's comment is misleading the reader.

The applied permeability map by Gleeson is not fit for purpose. In the Gleeson paper a number of warnings and constraints to use the map are highlighted that are relevant to this paper. The fact the permeability changes significantly across state borders hints to major unresolved issues on the subsurface conceptualization.

We use Gleeson exactly as described (large-scale, 100m depth) and it is incorrect that Gleeson states this dataset is inappropriate for this use, this is the exact reason he published the paper.

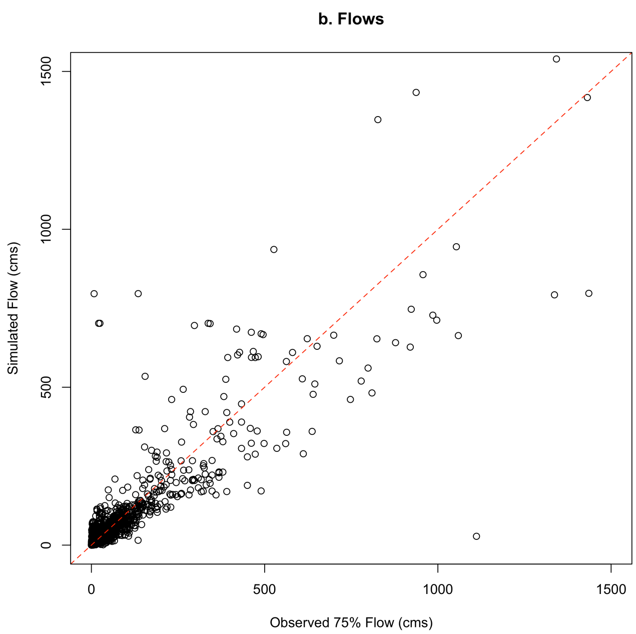
The vertical resolution is very rough to adequately model Richards equation.

We state that the vertical resolution is variable (p 7323 lines 26-27) with a fine resolution near the land surface (.1m) and coarser with increasing depth. We feel this choice represents a balance between adequately providing fine scale resolution near the ground surface and representing the resolution of the input data (~2m for soil, 100m for Ksat). Over the soil layers, this is exactly the resolution used by many land surface models, including Noah, and is reasonable to resolve near-surface processes.

A range of processes are missing that could be captured with easier models. For example interception. How is this considered with the P-ET maps? Moreover, snow processes are not implemented, and as mentioned above the model cannot simulate Karstic or fractured systems. This is important, as Karst aquifers are frequently found in the US. I suggest consulting the nation-wide karst-maps provided by the USGS. Nothing is mentioned on these missing aspects of the model.

As stated earlier, the P-E dataset used is based upon a product derived from the VIC land surface model and has been widely applied in large-scale hydrologic studies (Maurer et al. 2002). As such, the product does capture interception, and calculates ET based on a full simulation of the land energy budget including snow. This dataset has been used in prior large-scale analyses (Fan et al. 2007; Gleeson et al. 2011), which *is precisely why it was chosen for the current study. As discussed above, karst and fractured systems are a complication in any hydrologic modeling endeavor and karst aquifers are commonly found within North America. However, we are presenting a numerical experiment at continental scales directed at the CONUS domain. Additional, improvements will certainly be made in the future.*

It is a major disappointment to see how bad the comparison between simulated and calculated stream discharge is, even though the model is in steady state and the results are presented on a log-scale scatter plot. The presentation of the somewhat arbitrarily chosen histograms (e.g. median and 75 percentile vs. a steady state simulation) cannot coat the fact that the model is far from capable to adequately simulate hydrological processes on this scale.
We include a comparison of observed and simulated flows in a linear plot shown below, we chose the log plot to better represent the true model comparison to observations as it emphasizes the misfit at low flows. The R^2 value actually increases to 0.8 over the log fit, which is a surprisingly good value for an uncalibrated numerical experiment. As stated in the manuscript, we used transient *gage data from a number of locations within the domain to compare to a steady-state simulation. Additionally, there are influences from water management on most of the observations used. This makes the comparison challenging, yet we still present it in an open and honest comparison with model results.*



Important information is missing, e.g. roughness in the overland flow domain or details on numerical performance and convergence.
This has been resolved in the revisions.

Given that I have very little confidence in the model I do not think it should be used to develop any general conclusions. Nevertheless, some conclusions have been made, most of them trivial. For example, the equal importance of hydraulic conductivity and recharge (page 7330 line 14) in reproducing hydraulic heads is a lesson learnt in the early if not first modelling lectures. That there is a relation between drainage area, aridity and surface flow (page 7331, line 14) is well known and not very surprising.
We respectfully disagree with the reviewer on these points. We feel that we have addressed this in prior responses.

The upper limits shown in the graphs are not further discussed. Given the major yet undiscussed uncertainties in the model there is no basis to make general conclusions.
We will add detailed discussion of model and input data uncertainty to the revised manuscript.

<u>Other comments</u>
The title is misleading. 1km of spatial resolution is far from "Hyper“ as suggested by the title, nor is it "high“ as suggested in the abstract. The authors might refer to their own publication (http://onlinelibrary.wiley.com/doi/10.1002/hyp.10391/pdf, section WG3) where hyperresolution is claimed to be <1km. Clearly, the term “Hyper” is always relative to the scale considered. However, the comparison should be oriented towards the scale of the relevant processes, not the scale of the model domain. With the exception of the largest streams, the majority of runoff-processes will be dominated by terrain features and processes that have to be resolved on spatial scales significantly below 1 km2. Finally, not the entire continental US is modeled as suggested by the title, it is a rectangular sub-area with no-flow boundary conditions.
We disagree with the reviewer's assessment. As stated, hyper is relative to the scale and extent *considered. While the community may be striving towards even higher resolution than 1km, we contend that at this large extent for this simulation, the use of the term is appropriate. We have clearly stated the limitations of a 1km lateral cell size. We have modified the title slightly to reflect that we simulate most of the continental US, though we do in fact capture the major basins within.*

The fonts on some graphs are very hard to read.
The font sizes for the legends in Figures 9 and 10 have been increased.

The authors confused all the first and second names in the reference to Nir.
This has been fixed in the references, interestingly this citation was based upon the .ris file provided by the publisher for this paper.

References

*Bierkens, M. F. P., V. A. Bell, P. Burek, N. Chaney, L. E. Condon, C. H. David, A. de Roo, P. Döll, N. Drost, J. S. Famiglietti, M. Flörke, D. J. Gochis, P. Houser, R. Hut, J. Keune, S. Kollet, R. M. Maxwell, J. T. Reager, L. Samaniego, E. Sudicky, E. H. Sutanudjaja, N. van de Giesen, H. Winsemius and E. F. Wood (2015). "Hyper-resolution global hydrological modelling: What is next?" Hydrological Processes **29**(2): 310-320, 10.1002/hyp.10391*

*Fan, Y., H. Li and G. Miguez-Macho (2013). "Global patterns of groundwater table depth." Science **339**(6122): 940-943, 10.1126/science.1229881*

*Fan, Y., G. Miguez-Macho, C. P. Weaver, R. Walko and A. Robock (2007). "Incorporating water table dynamics in climate modeling: 1. Water table observations and equilibrium water table simulations." Journal of Geophysical Research-Atmospheres **112**(D10): -,*

*Gleeson, T., L. Marklund, L. Smith and A. H. Manning (2011). "Classifying the water table at regional to continental scales." Geophysical Research Letters **38**(L05401): 6, 10.1029/2010GL046427*

*Kollet, S. J. (2009). "Influence of soil heterogeneity on evapotranspiration under shallow water table conditions: Transient, stochastic simulations." Environmental Research Letters **4**(3): 9,*

Maurer, E. P., A. W. Wood, J. C. Adam, D. P. Lettenmaier and B. Nijssen (2002). "A long-term hydrologically based dataset of land surface fluxes and states for the conterminous united states." Journal of Climate **15**(22): 3237-3251, 10.1175/1520-0442(2002)015<3237:althbd>2.0.co;2*

*Maxwell, R. M. and S. J. Kollet (2008). "Quantifying the effects of three-dimensional subsurface heterogeneity on hortonian runoff processes using a coupled numerical, stochastic approach." Advances in Water Resources **31**(5): 807-817,*

*Maxwell, R. M., M. Putti, S. Meyerhoff, J.-O. Delfs, I. M. Ferguson, V. Ivanov, J. Kim, O. Kolditz, S. J. Kollet, M. Kumar, S. Lopez, J. Niu, C. Paniconi, Y.-J. Park, M. S. Phanikumar, C. Shen, E. A. Sudicky and M. Sulis (2014). "Surface-subsurface model intercomparison: A first set of benchmark results to diagnose integrated hydrology and feedbacks." Water Resources Research **50**(2): 1531-1549, 10.1002/2013wr013725*

*Meyerhoff, S. and R. Maxwell (2011). "Quantifying the effects of subsurface heterogeneity on hillslope runoff using a stochastic approach." Hydrogeology Journal **19**(8): 1515-1530, 10.1007/s10040-011-0753-y*

*Meyerhoff, S. B., R. M. Maxwell, W. D. Graham and J. L. I. Williams (2013). "Improved hydrograph prediction through subsurface characterization: Conditional stochastic hillslope simulations." Hydrogeolgy Journal **in press**,*

*Wood, E. F., J. K. Roundy, T. J. Troy, L. P. H. van Beek, M. F. P. Bierkens, E. Blyth, A. de Roo, P. Döll, M. Ek, J. Famiglietti, D. Gochis, N. van de Giesen, P. Houser, P. R. Jaffè, S. Kollet, B. Lehner, D. P. Lettenmaier, C. Peters-Lidard, M. Sivapalan, J. Sheffield, A. Wade and P. Whitehead (2011). "Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring earth's terrestrial water." Water Resour. Res. **47**(5): W05301, doi: 10.1029/2010wr010090*