Authors' response to reviewers comments for:

GRACE storage-runoff hystereses reveal the dynamics of regional watersheds

The authors would like to thank both Referees for their comments, which we incorporated into our revised manuscript. This response section provides a detailed response to the comments, or provides the line numbers in the revised manuscript that address the comments. We have also numbered are responses for easier reference throughout the document.

Comments from the Referees are posted in italics.

And our responses are all in a normal font.

Anonymous Referee #1
Received and published: 5 December 2014

This paper addresses:

- The possibility to forecast runoff at certain times during summer from terrestrial water storage in spring measured by GRACE.

- The behavior of TWS and groundwater storage during the seasonal cycle

This study uses GRACE observations of terrestrial water storage observations to expand upon a fundamental concept in watershed hydrology – that the temporal relationship between storage and runoff can be used to quantify complex watershed behavior at broad scales, including groundwater recharge amounts and timing, baseflow recession characteristics, and long lead-time streamflow prediction. The methods is implemented over three catchments of minimum size (for the application of GRACE data) and similar climatic conditions

Interesting highlights in this study:
- Clock-wise behavior of groundwater hysteresis in the presence of a counter clock-wise behavior of the hysteresis for the GRACE signal.

- Prediction of seasonal runoff using GRACE data

1. *However, the focus of the presented results is not reflected in the title of the paper nor in the abstract. Both, title and abstract claim a much more general result with respect to the dynamic behavior of systems. This is not covered in the presented investigations.*

Thank you for your feedback. We have amended the abstract to incorporate your perspective, however we feel that the title is direct and reflects our actual results, and so have opted not to change it.

2. *The claimed methodology cannot be applied to other regional scale studies.*

We respectfully disagree with this statement, GRACE is a global dataset. The Referee is correct in stating that applying the methods we present to other regional watersheds would require some adaptation. But this is the norm for hydrological studies and research in general. Concerns with our methods provided by Referee #1 are included in the revised manuscript.

3. *As the authors mentioned in the text, the paper is based to a large extend on the work by Riegger & Tourian 2014, who have investigated and modeled the runoff–storage(R-S) relationship for large scale catchments in different climatic regions in detail.*

Here in this paper there is a lot of text claiming an explanation for system dynamics by hysteresis only in a qualitative way, mainly repeating the results of Riegger & Tourian yet not being supported by own investigations. There is also a lengthy text trying to explain the deviations from the expected results by anthropogenic management, pumping and local groundwater level measurements.

The work by Riegger and Tourian provided new insights into regional watershed dynamics, which we acknowledge several times throughout our manuscript. To further highlight the contributions of Riegger and Tourian we have added additional commentary in lines 91-95. We have also revised the methods, results, discussion, and conclusion section to address the Referee's suggestions.

We respectfully disagree with the Referee's opinion that we have repeated the results of Riegger and Tourian. Their paper included evapotranspiration as a focus of their research, where we focused on climate, topography, and geology. Topography and geology were not included in the Riegger and Tourian analysis.

The work of Riegger and Tourian (2014) was published prior to this manuscript, and we fully acknowledge their contributions throughout the paper. However, the methods that we applied in this paper were developed independently from the work of Riegger and Tourian. The conceptual idea for this research was presented at a GRACE workshop in 2013 and at the Fall meeting of the AGU later that year.

Sproles, E.A., Leibowitz, S.G., Wigington Jr., P.J., Patil, S., Reager, J.T., and Famiglietti, J.S.: Multi-scale analysis of terrestrial water storage and stream discharge in the Columbia River Basin, Using GRACE Data for Water Cycle Analysis and Climate Modeling Workshop Abstracts, 2013

Sproles, E.A., Leibowitz, S.G., Wigington Jr., P.J., Reager, J.T., Famiglietti, J.S., and Patil, S.D.: GRACE storage streamflow hystereses reveal the dynamics of regional watersheds, AGU Fall Meeting Abstracts, 2013

4. *The investigations presented here are not performed in a sound scientific way and are insufficiently documented, because for an assessment of the forecast potential a comparison of predicted and measured discharge is needed. This comparison should be supported by reporting appropriate metrics like RMSE, Nash-Sutcliffe and correlation (The claimed correlation coefficients seem to correspond to the fitting curves and not to the observed data).*

The Referee provides an excellent point, and we have augmented our analysis in this revised manuscript. We divided the nine years into two sets of data and used a double pass approach to empirically fit the TWSA-Runoff power function. The first pass used one set of data to fit the model, which was then evaluated against the second set. The data sets were then switched for empirical fit and evaluation.

The results of all 9 years were then used to calculate RMSE, Nash-Sutcliffe, and R2.

We provide a more complete methodology (lines 237-248) and results (lines 325-357) in the updated manuscript (Tables 1 and A3).

5. *No reasons for the selection of points in time are given, neither for the use of TWSA (March) nor for the predictions of Qseason and QAug. No alternatives months are investigated or discussed.*

We have described the selection of months in greater detail in the manuscript (lines 120-123, 231-233, 481-482). In additional to testing for TWSA March and April we have also included February. We have also included runoff for the months of July and September.

6. *If there are anthropogenic impacts in the chosen catchments, why isn't other catchment or time periods used for which no anthropogenic effects occur?*

We appreciate the Referee's comment and agree that ideally regional watersheds that are not managed would benefit this research. However such regional watersheds do not exist in the western portion of the conterminous United States, where this study was focused, or in many areas of the world. The reality is that major dams on the Columbia River date back decades and the GRACE data record dates back to only 2002. Future research could include catchments in Alaska, but were beyond the scope of this study.

7. *Details in the calculation of GWSA are not given nor are the corresponding data sets (time series of mean GW-level, recharge, soil moisture, snow water equivalent, reservoir volumes, TWSA, discharge) displayed. Thus the calculation steps and conclusions with respect to the GWSA hysteresis cannot be retraced. An appropriate visualization of time series of different compartments is needed.*

The description of the methods used to isolate GWSA are provided in 176-180. Additionally, Figure 2 now includes a time series of the different components and the data as a supplemental file.

8. *In the comparison of TWSA vrs point specific well data (Fig.7) no explanation for the inconsistency of GWSA and mean groundwater level is given. Instead further detailed studies are proposed.*

We would like to thank the Referee for identifying this source of ambiguity. We have added additional discussion on lines 462-471.

9. *The authors should clearly describe the behavior of the groundwater system GWSA compared to TWSA and provide possible physical reason for the behavior. For a better understanding, one might think of showing the hysteresis of slow and fast discharge vrs precipitation and different storage compartments like soil moisture, snow water equivalent etc.*

We apologize if the original manuscript was not fully developed. We have added additional methods (GW + Soil Moisture), and the results are provided in lines 283-297). Additional discussion is provided in lines 429-461.

The separation of slow and fast discharge was not a focus of this study, and therefore were not included.

10. Results and conclusions need to be related to the presented investigations.

We apologize if the results and conclusions appeared not to be related to investigation. We have updated the manuscript with this goal in mind.

11. Conclusions have to be explained in detail.

The conclusions section was written as a brief summary of what was included in the research and their implications. The discussion section was written in a more thorough format, which has been added onto in this revised manuscript.

12. Title and abstract have to reflect the actual investigations

Please refer to comment and response #1.

Based on the aforementioned points, the paper can only be accepted after a major revision, in which the aforementioned points are taken into account. Specific Comments:

13. 12029 L14

The correlation values seem to correspond to data and fitting curves rather than to predicted and measured runoff.

Please refer to comment and response #4.

12029 L15

This is very general sentence. In fact to apply the same methodology, one should characterize each basin individually. Indeed, this characterization sometimes becomes cumbersome due to the heterogeneous behavior of many large scale basins. In fact, the prerequisite of applying this method is not only availability of GRACE data. So please be precise in your statement.

Please refer to comment and response #2.

14. 12032 L8

There might be confusion in the citation of the respective publication of Reager et al. 2014: In the cited paper the relationship between regional water storage and specific streamflow is not addressed neither the corresponding hysteresis. In the cited paper GRACE TWSA is just used as additional forcing term in an auto-regression approach. Possibly it is from another publication, please cite the correct paper.

Thank you for introducing this topic. We have also included an additional citation (Reager et al., 2009). We still include the Reager et al (2014) citation as this paper uses GRACE as a major component of an analysis focused on specific streamflow event (Missouri River in 2011).

15. 12032 L10

In the paper of Riegger and Tourian 2014 the hysteresis between runoff and storage is described in detail for different climatic zones. For boreal regions they show that runoff is linear to coupled liquid storage. They claim that beside the time lag for runoff the uncoupled solid components of storage are responsible for the hysteresis. However their calculation of runoff from GRACE mass is based on a homogeneous distribution of aggregated snow mass and is not directly applicable to mountainous areas as in-vestigated here. In mountainous areas the snow mass distribution very much depends on local conditions like topography, elevation etc.

We have included a more detailed discussion of the Riegger and Tourian (2014) paper in the updated introduction (lines 95-103).

16. Please report the scale factors of studied basins.

These have been included in the appendix (Table A5).

17. 2035 L2

How the leakage is quantified. The scale factor would only deal with signal attenuation.

Leakage is much more complicated to be quantified by a simple scale factor. Please also describe the measurement error as well. Does the measurement error come from propagating the calibration error of spherical harmonic coefficients?

The scale factors that are applied at 1-degree resolution following the methodology outlined in Landerer and Swenson [2012] are designed to estimate and reduce the "leakage" error in the GRACE solutions. The following assumptions are made:

1) That GRACE estimates of water storage variation contain signal degradation due to the orbital configuration of the satellites and the nature of the observed quantity (i.e. the decorrelation length scale of regional water storage variability) [Wahr et al., 2006]. These "measurement" errors manifest as noise (i.e. random errors that increase in amplitude with increasing spherical harmonic degree) and as systematic errors that are correlated within a particular spectral order [Swenson and Wahr, 2006; Landerer and Swenson, 2012]. These errors are generally improved by truncation and filtering, though this compromises high-resolution signal. These errors should be maintained and propagated through a scaling approach.

2) That the signal attenuation effects of GRACE solution processing on an underlying hydrological signal result from the truncation of spatial harmonics at a degree and order selected to minimize noise and maintain spatially-variable signal for each monthly solution, and from the algorithmic filtering of the resulting harmonics to remove systematic correlated errors ("stripes" described in Swenson and Wahr [2006]) and also random errors (300km gaussian). This cumulative error is referred to as "leakage" and depends on the filtering process as well as the characteristics of the original signal. Here, a truncation to degree and order 60 was performed before application of the destriping and gaussian filters mentioned.

3) That there is some hypothetical "true" global hydrological state that can be estimated by a land surface model (LSM: a numerical simulation) that is forced by precipitation and solar radiation, and modulated by various and land surface characteristics (topographic variability, vegetation type, soil type, etc.). Here, this simulation is performed at 1-degree resolution by the NCAR CLM model as described on the TELLUS website (grace.jpl.nasa.gov).

4) That the errors resulting from these "leakage" effects can be modeled to 1-degree resolution by placing the monthly 1-degree LSM simulation outputs through processing equal to that performed on the GRACE spherical harmonic solutions (i.e. truncation and filtering). These errors can then be quantified by measuring the signal loss between the original simulations and those processed with "leakage" effects.

Scale-factors can be calculated based on a least-squares minimization of the residual between the unprocessed and processed time series.

In summary, because the spatially-distributed scale factors have a sub-GRACE-resolution structure, and because scale factors can have a value of less than 1, they attempt to recover the best-case underlying hydrological signal according to the assumptions above. This is how the spatial scale factors estimate and correct signal attenuation caused by "leakage" errors. There is however some residual uncertainty associated with this process, and this process does not remove all of the "leakage" error in the GRACE observations — it only acts to reduce it. Conversely, the measurement error is actually amplified through the scaling process, along with the TWSA estimates. Of course these error estimates are both included in the regional application of GRACE TWSA data.

It is necessary to use some sort of regional scaling in the application of GRACE observations for hydrology [e.g., Famiglietti et al., 2011; Landerer et al., 2010; Swenson and Wahr,2007; Klees et al.,2007; Chen et al.,2007]. That is because GRACE observations are intrinsically more coarse than other hydrologic observations (e.g. stream gage), and a failure to account for and represent signal loss in TWSA would result in a mismatch of data [Swenson et al., 2003], leading to results that are based only on proportionality and are consequently of limited utility/accuracy [e.g. Riegger and Tourian, 2012; Tang et al., 2010].

All GRACE results, as the Referee has identified, can also possess complex errors, such as the existence of neighboring regions that are out-of-phase (e.g. the Orinoco and the Amazon) whose signals can interact. For this reason, any scaling approach needs to include some representation of measurement and leakage errors.

The 1-degree scale factors represent the opportunity to apply GRACE observations at a scale that is consistent with the hydrology simulations in global climate models (i.e. LSM's) and with many other global gridded hydrology data sets, and the error estimates that accompany them represent a more robust approach to quantifying potential limitations in hydrologic applications of GRACE.

The goal of the paper was not to provide more research on leakage and scale factors, and thus will include the same brief description with references as found in the initial manuscript.

Refs:

Chen, J. L., C. R. Wilson, J. S. Famiglietti, an d M. Rodell (2 007), Attenuation effect on seasonal basin-scale water storage changes from GRACE time-variable gravity, J. Geodesy, 81, 237–245, doi:1 0.1007 /s00190-006-0104-2.

Famiglietti, J. S., M. Lo, S. L. Ho, J. Bethune, K. J. Anderson, T. H. Syed,S. C. Swenson, C. R. de Linage, and M. Rodell (2011), Satellites measurere rates of groundwater depletion in California's central valley, Geo-phys. Res. Lett., 38(3), L03403, doi:10.1029/2010GL046442.

Klees, R., E. A. Zapreeva, H. C. Winsemius, and H. H. G. Savenije (2007),The bias in GRACE estimates of continental water storage variations, Hydrol. Earth Syst. Sci., 11, 1227–1241.

Landerer, F. W., J. O. Dickey, and A. Guntner (2010), Terrestrial wate rbudget of the Eurasian pan-Arctic from GRACE satellite measurements during 2003–2009, J. Geophys. Res., 115(D23), D23115, doi:10.1029/2010JD014584.

Tang, Q., H. Gao, P. Yeh, T. Oki, F. Su, and D. P. Lettenmaier (2010),Dynamics of terrestrial water storage change from satellite and surface observations and modeling, J. Hydrometeorol., 11(1), 156–170,doi:10.1175/2009JHM1152.1.

Swenson, S., and J. Wahr (2002), Methods for inferring regional surface-mass anomalies from gravity recovery and climate experiment (GRACE)measurements of time-variable gravity, J. Geophys. Res., 107(B9), 2193,doi:10.1029/2001JB000576.

Swenson, S., and J. Wahr (2006), Post-processing removal of correlated errors in GRACE data, Geophys. Res. Lett., 33(8), L08402, doi:10.1029/2005GL025285.

Swenson, S., J. Wahr, and P. C. D. Milly (2003), Estimated accuracies of regional water storage variations inferred from the gravity recovery and climate experiment (GRACE), Water Resour. Res., 39(8), 1223,doi:10.1029/2002WR001808.

Wahr, J., S. Swenson, and I. Velicogna (2006), Accuracy of GRACE mass estimates, Geophys. Res. Lett., 33, L06401, doi:10.1029/2005GL025305.

18. 12036 L9-17

The simultaneous display of TWSA and GWSA in one figure for each catchment might help to highlight the different dynamical behavior.

Thank you for the suggestion. We tried having the TWSA and GWSA in the same graphic, but found it visually confusing. We have also added subfigures for subsurface water, which highlight the combined signal of GW and soil moisture.

19. 12036 L22 Similar confusion in the citation of the respective publication of Reager et al. 2014 is seen as above (12032 L8).

We would like to thank the Referee for their perspective regarding this article. Figure 2 in Reager et al., 2014 represents the relationship between TWSA and Q, and

highlights the 2011 flood events in the Missouri River. Thus we have included it in the manuscript.

20. 12036 L22
In order to describe the systems independently from catchment area I propose to use runoff rather than discharge for a comparison

We would like to thank the Referee for this comment. We did use runoff throughout the study, and have clarified the labeling in the updated manuscript.

21. 12037 L25 – 12038 L8 To support this paragraph I suggest to show time series of soil moisture or distribution of snow coverage in time.

This suggestion has been incorporated into Figure 2 of the updated manuscript.

22. 12039 L14 Fig.3b shows the total runoff i.e. surface runoff and baseflow vrs. GWSA. Thus total runoff should be separated into its fast and slow components and each of them being displayed vrs. GWSA. The Clock-wise behavior of groundwater hysteresis is one of interesting finding of this study. Therefore, the authors should provide a physical explanation for that. Explanation of Fig 3b is not clear: Vertical branch Jun-Oct: How can runoff decrease with a nearly constant GW storage? Left branch Oct – Mar: how can runoff increase with a decreasing GW storage? Is it matter of surface runoff? For an understanding of the behavour it is essential to display time series of SM, SWE, RES, TWSA and RGW and RSW. A display of precipitation and evapotranspiration would be very helpful.

These suggestions have been incorporated into the updated manuscript. We have added additional methods (incorporating GW + Soil Moisture), and the results are provided in lines 283-297. Additional discussion describing these processes is provided in lines 429-461.

The separation of runoff into fast and slow components is an interesting idea, but fall outside the scope of this paper and therefore was not included.

23. 12039 L19
Confusion about Fig. numbering: possibly Fig 7 is meant

Thank you for identifying the mislabel. This has been corrected in the revised manuscript.

24. 12039 L21–23
To me taking TWSA_march for prediction looks like cherry picking. Why not March and not February for TWSA or June, July for discharge, for instance? Please provide reason for taking this month!

Please refer to comment and response #5.

25. Also, confusion about Fig. numbering: possibly Fig 6 is meant It is not clear from the text how stream flow is predicted and how this is related to Fig6: Qseason in Fig6: is it the mean observed discharge or predicted discharge? If it is the later, how is it calculated?

We apologize for any confusion. This is the observed discharge and text describing this has been added in the revised text caption.

26. Are the correlations displayed in Fig6 and Table 1 the correlations between the measured Qseason and the fitted curves (power functions with which parameters?) or between the measured Qseason and TWSA (the high correlation value rather represents the curve fit than Qseason vrs TWSA).

The correlations displayed in Figure 7 of the original manuscript represent the fitted curves, as do the metrics in the figure. Tables 1 and A3 provide a more detailed look at the results from testing the empirical fit using a double-pass calibration and validation described in the updated methods (lines 226-335). Table 2 and A4 provide the metrics for the complete data set.

27. The fitting curves in Fig6 do not already represent predictions, yet are the basis for predictions using measured TWSA to determine forecast discharge (via the curve fits) (as Fig6 shows, that Qseason and QAug are not very much depending on TWSA for smaller values of TWSA, yet only for bigger values. The calculation scheme is essential for an assessment of the method). For an evaluation of the predictive potential of the investigated methods the parameters of the fitting curves determined on a training period should be used to calculate predic tions of discharge in an independent prediction period. The predicted discharge values should then be displayed versus the measured for the prediction period in a scatter plot and correlations should be calculated for forecasts vrs measured.

As a predicted value should always be better than a simple use of mean monthly values from the training period of forecasts, on top of conventional Nash Sutcliffe (NS) coefficient, in which the values are assessed w.r.t long term mean, the NS coefficient w.r.t. the seasonal signal (using the monthly residuals of the training period) should also be presented in table1 (NS_cycle: in the denominator instead of \bar{Qo} you should use monthly mean).

The Referee provides an excellent point, and we have augmented our analysis in this revised manuscript. We divided the nine years into two sets of data and used a double pass approach to empirically fit the TWSA-Runoff power function. The first pass used one set of data to fit the model, which was then evaluated against the second set. The data sets were then switched for empirical fit and evaluation.

The results of all 9 years were then used to calculate the standard hydrological metrics RMSE, Nash-Sutcliffe, and R^2.

We provide a more complete methodology (lines 224-248) and results (lines 325-357) in the updated manuscript (Tables 1, 2, 3, A3, A4).

12040 L1
Again why August and no other month? How a reader should follow the story here? What does a seasonal average/aggregation of discharge mean for possible applications?

Please refer to comment and response #5.

28. 12041 L5–29
Over boreal regions the R-S hysteresis is determined by (Riegger and Tourian 2014):
-Climatic impacts i.e. the relative importance of aggregated solid precipitation (repre-sented on the lower branch)
-The runoff time constant determining the slope of the linear part (upper branch)
-The time lag between mass and runoff being responsible for (a smaller) part of the hysteresis
The different forms of the hysteresis thus can be explained by the corresponding hydraulic time constants, which is shortest for steep slopes and fractured systems. This explains that the upper branch is steeper for the Upper Columbia than for the Snake River. This should be considered and discussed in this section.

We thank the Referee for this suggestion. We have addressed this concern using an alternative approach. We removed only the Snow and Reservoir signal to examine watershed hysteresis, and we tested the ability of SWE to predict Runoff.

The findings are included throughout the results and discussion (for example lines 375-399).

29. 12042 L10–17
See comment above on prediction accuracy in table 1 There are 9 years of measurements available. This period could be split into a training and a prediction period for a better estimation of prediction accuracy

We have augmented our analysis in this revised manuscript. We divided the nine years into two sets of data and used a double pass approach to empirically fit the TWSA-Runoff power function. The first pass used one set of data to fit the model, which was then evaluated against the second set. The data sets were then switched for empirical fit and evaluation.

The results of all 9 years were then used to calculate RMSE, Nash-Sutcliffe, and R2.

For more, please refer to comment and response # 28.

30. 12042 L18
Probably Fig 6, and Fig 7 is meant

We have addressed the mislabeling.

31. 12042 L21-26

The fact that Q is insensitive to TWSA < 100mm is reflected by the curves in Fig 6. This means that prediction could only be made for TWSA > 100mm. If these catchments are managed, a reliable prediction from TWSA cannot be made!! So either other, un-managed catchments have to be chosen or the authors should only consider the time periods with no management. The explanation of the Q – TWSA relationship for TWSA < 100mm by water resources management is not sufficient to explain quantitative effects.

We appreciate the Referee's comments and agree that ideally regional watersheds that are not managed would benefit this research. However such regional watersheds do not exist in the western portion of the conterminous United States, where this study was focused.

From a data perspective, we are also limited with regards to the length of the GRACE data record. However the available data demonstrates a threshold behavior for total seasonal runoff in each of the three regional watersheds when TWSA in March is less than 100 mm (lines 492-494). When TWSA in March is above 100 mm, GRACE measurements also provide a high degree of skill in predicting season total seasonal runoff.

From a scientific perspective these findings describe the measured relationship between storage and runoff at regional scales. These insights identify potential ways to apply in understanding watershed behavior and predicting streamflow.

From a management perspective, the skill of TWSA from a single month to predict seasonal streamflow provides value as well. It would allow managers to know how with a fairly high level of confidence how much water to expect in the system for the remainder of the water year. This is important in the Columbia River Basin where management balances protecting endangered species in a region with extensive industrial agriculture and hydropower generation. Since the stated goals of HESS are "the advancement of hydrologic science … to serve not only the community of hydrologists, but all earth and life scientists, water engineers and water managers," we think such an emphasis is appropriate.

Even in a managed basin such as the Columbia River you will have years of high and low flows, each of which can introduce management challenges (e.g., http://www.oregonlive.com/business/index.ssf/2011/05/bpa_curtails_wind_farm_generat.html; http://www.forbes.com/sites/jamesconca/2012/08/05/hydro-forced-to-take-a-dive-for-wind/, http://wdfw.wa.gov/news/aug2101a/).

The management of the Columbia River Basin is regulated by a treaty between the United States and Canada dating back to 1964, which is currently being renegotiated.

Our approach in this study was not to let what we cannot do limit what we can do.

32. 12042 L21-26 –12043 L11
Lengthy text, clear quantitative consequences are missing

Thank you for the suggestion, and we have shortened this text (lines 498-501).

33. 12043 L16-19
Explanation of GWSA-Q not understood (see also above)! During the winter period in snow covered areas discharge is released only from the groundwater system. The groundwater recharge from the surface in this case is zero, i.e. GWSA should decrease with Q. How is distinguished between surface runoff from snowmelt and runoff from groundwater. Possibly the behavior of the model-based calculation of soil water content helps.

Thank you for this comment, and we have included changes in subsurface water (soil moisture + groundwater) in our revised manuscript.

34. 12043 L25
GWSA is nearly constant from June to October. Why? This does not fit to the timing of pumping test! What is the purpose of the pumping tests mentioned in this paper?

We apologize if this was unclear. The purpose of including pumping tests is the timing of tests for management purposes corresponds to the highest groundwater levels. This could potentially lead to management strategies that do not capture periods of lower groundwater levels. We are publishing this paper from a government agency, and thus cannot include recommendations regarding changes to regulations. However we can highlight ways that understanding systems could potentially be improved.

35. 12044 L3
Probably Fig 7 instead of 6 Meaning and conclusion of Fig 7

Thank you, we have corrected the label.

36. 12044 L3–10
GWSA does not fit to the overall GW-levels from observations. So either the calculated GWSA is wrong or the selection of observation wells is not representative for the general GW level. GWSA correspond to the total volume of the groundwater system and not to groundwater levels!

The GW-storage coefficient determines the relationship between volume and level and is not mentioned here as it is probably not known on catchment scale. There is no scientific consequence from this observations mentioned in the paper. If this part is presented here, there should be a more detailed description of GW-level

measurements, storage coefficients from hydrogeology, selection of observation points and a detailed discussion of the results and consequences for the message of the paper.

Thank you for introducing this discussion. We also understand that GWSA data (volume) do not equal well levels, and for this reason we normalized the standard deviation across the time series to compare the temporal signal. We concur with the Referee that the groundwater observation well levels provide no relationship with GWSA data. We debated whether or not to remove this component of the research. In the end we included it as the lack of correlation highlights the variability of site characteristics across a region, which is of scientific consequence.

The revised discussion (lines 454-471) addresses these points.

37. 12045 Conclusions

The conclusion should represent the conclusion of this work. In case the authors would bring in arguments from other published works the bridges between studies should be clear. The last paragraph of conclusion is too general and does not reflect the results of the study.

We have included updated results in the conclusion.

38. 12045 L10–13

Please report the prediction results here including RMSE, NS, NS_cycle, correlation

We have included model performance metrics in the conclusion.

39. 12045 L23 Please provide a citation for the background research in the text, otherwise please rephrase the sentence as this statement contradicts with your earlier statement in 12040 L22.

Thank you for pointing out any potential ambiguity. The statement in acknowledgements is factually accurate. The approach we present was developed concurrently, but independently from the research provided in Riegger and Tourian (2014). Because their work was published prior to this paper, we have cited their efforts in the research portion of the paper. Again our approach was developed independently, and implying that our approach was developed from Riegger and Tourian is not factually accurate.

Anonymous Referee #2
Received and published: 11 December 2014
General comments:
In its exploration of the use of GRACE gravity data for improved understanding of the hydrology of regional watersheds this paper presents intriguing results and points a way forward to further applications of this approach. To that extent it appears to merit publication.

This paper can be viewed within the general context of an increasing attention for changes of water storage in a watershed as the driver of streamflow. The classical rainfall-runoff models of hydrology avoid the obvious fact that streamflow is driven by storage in the watershed and not by precipitation as such, but the intermediate storage change step was skipped because there were no adequate means to observe storage changes in most regional watersheds. New observation techniques such as GRACE allow closer consideration of storage.

The prognostic ability demonstrated in this paper of the GRACE signal to predict seasonal runoff is impressive and indicates a potential for the use of GRACE results to enhance the reliability of seasonal water supply predictions.

Specific comments

40. *The distinction between soil moisture and groundwater appears rather arbitrary and re-quires more scrutiny. It might be preferable to combine the two as subsurface moisture storage because for much of these watersheds soil moisture changes on a monthly time step are likely to be closely linked to groundwater storage changes. The groundwater storage includes water storage in the capillary fringe above the water table and the top of the capillary fringe is likely to be above the 2000 mm below ground level over much of the time and space of the analysis. Thus it is possible that the soil moisture changes as estimated in this paper include much of the groundwater storage changes. That may be one of the reasons why the modeled groundwater storage changes appear to be small and have almost no correlation with the observation well records.*

Thank you for this suggestion. We have added a separate subsurface water signal (TWSAsub = TWSA – SWE – RES) into our revised manuscript.

41. *Fig 7. The almost total lack of correlation between the groundwater levels and TWSA serve to underline the questionable assumption that GWSA can be estimated from TWSA – SWE – SM. Likely the problem lies both in the uncertainty of the SM estimates and the high variability of groundwater dynamics across the whole basin from low to very high elevations. It is also likely that the groundwater observations are practically all for valley bottoms and do not represent the GW storage at higher elevations and on steep slopes. By contrast the TWSA is dominated by high-elevation snow.*

The paper should include plots, analysis and discussion of the changes of SWE and SM (or of SWE and SM+GW). These are just as critical an aspect of the components of the TWSA as the estimated groundwater changes. One would expect that the SWE can be validated fairly well on the basis of various point observations, at least much better than SM.

Thank you for this suggestion. We have included it into our revised manuscript as $TWSA_{sub}$.

42. *It would be intriguing to attempt a water balance for the watersheds by including pre-cipitation estimates. Since evaporation is relatively minor during the winter months P*
TWSA ~Q for the winter and this would provide a test of the consistency of these components with the conservation of mass. However, this perhaps such analysis lies outside the scope of the present paper.

We agree on both parts of this comment. It would be interesting, but it does in fact lie outside the scope of this paper.

Technical comments:

43. *P. 12029 L 22. Topography is clearly a major watershed descriptor apart from climate and geology, as also implied in this paper by the contrast between the steep slopes of the Upper Columbia basin and the relative flatness of the Snake River watershed.*

We have added the role of topography as a watershed descriptor throughout the manuscript. Thank you for the suggestion, it helps frame our discussion all the better.

44. *P 12033 LL 16-18. This characterization of aquifer storage capacity of the two watersheds is rather off-hand, and without any further explanation and references. The "well-developed soils" of the Snake River basin are perhaps relevant to soil moisture storage but not to aquifer storage which depends on the nature of the underlying subsoil and bedrock. Do the Snake River basalts in fact have much effective porosity at the water table (see also p. 12041, L 25)? The results shown later in the paper for groundwater storage changes in the Snake River watershed suggest very low groundwater storage capacity.*

Thank you for identifying any ambiguity. While the groundwater storage changes in the Snake River are more constrained, we interpreted this to represent that there is less change from max GWSA to min GWSA. This represents the fluxes of water moving through the system more slowly in the Snake. In part because 1) it is flat, 2) it is dry, 3) the aquifer can hold the water. This combination constrains the GWSA anomaly as compared to the Upper Columbia where 1) It is steep, 2) it is wet, 3) the aquifer

cannot hold the water. It moves large fluxes of water through it, which expands the range of GWSA.

We have also added additional text in the manuscript to describe these processes in section 5.1 of the manuscript.

45. P 12035 LL 20-25. In view of all the uncertainties in measuring or estimating regional soil moisture, as summarized in the introduction to this paper, these GLDAS-derived estimates of soil moisture are surely highly tentative at best. This would appear to be a very uncertain foundation for estimating changes of groundwater storage. The error estimate for SM is not adequately estimated on the basis of the "monthly standard deviation" (p. 12036, L 11) because there are likely large biases in the GLDAS model algorithms.

We understand the Referee's concern that there may be issues with large-scale modeled soil moisture, but this comment does not offer a direction forward. The Referee seems critical of the state-of-the-science of hydrology and of an already-tested approach, and not of this specific work. Of course, there have been numerous (~30) studies published using GRACE to detect groundwater variability globally, and all present a similar, if not less rigorous, methodology than the current work. Many of those studies have used well data to validate their results (e.g. Scanlon et al., 2011).

The NASA LDAS simulations used here were driven by observed precipitation and radiative forcing. Whether the model does a good job with soil physics and the parameterization of sub-grid scale processes is another issue: as explained in the text, we attempt to address model structural error by using an ensemble of model outputs, to give us an estimate of the bias contained in a given model solution relative to others, which we then propagate through our calculations. We also represent a range of possible scale-variance uncertainty by considering the GLDAS simulations (at 1-degree resolution) and the NLDAS simulations (at 1/4 degree resolution). For large areas (i.e., an area equivalent to that of the GRACE observations) this should be entirely appropriate., and it is the best methodology we know to be available.

A note on bias: because we are dealing with terrestrial water storage variability (anomaly), and specifically not a mean-value, the typical pathway for model bias to manifest (e.g. through differing numbers of soil layers) is eliminated by removal of the model soil moisture mean. We are really only considering variability in soil moisture. The use of multiple models in an ensemble provides a large range of uncertainty that can be propagated through groundwater calculations. Also, the removal of soil moisture tends to reduce the variability in GRACE time series, especially at the seasonal period. This is a likely outcome considering soil moisture memory with depth.

Finally, there is really no alternative approach, as soil moisture observations across the domain are not available. If they were, then scaling or interpolating individual point measurements would arguably contribute as much or more error than a

distributed model driven by observed forcing. While there is certainly space for continued work in hydrology on soil moisture variability and scale variance and invariance, it is really beyond the scope of the current manuscript to develop these analyses. Instead we rely on already proven, peer-reviewed methodologies to highlight new observations in the Columbia River basin. We humbly invite the Referee to collaborate on future refinement of these methods.

46. 12036, L 17 Insert "error" as in "individual ERROR components".

Thank you, and we have added this in the updated manuscript.

47. p. 12039 L 13. hardly "dramatic" since this is an obvious consequence of snow accumulation.

We agree and have changed the word to "distinct."

48. p. 12041, L 25. It is not obvious that the basalt provides excellent aquifer storage. The basalts provide excellent transmissivity for groundwater flow and discharge, but that is not the same as storage and in fact would go to counteract large changes of groundwater storage, as indeed is suggested by the analysis results of this paper (see Fig 2e).

Thank you for identifying this ambiguity and we have changed it in the updated manuscript (lines 400-403).