We firstly would like to acknowledge the editor comments. Based on the comments from referees and editor, we have revised the manuscript and uploaded in the separated file. In the revised manuscript, the modified contents are in blue while the original contents remain in black.

In addition to the responses to the referees that have already been uploaded to the discussion page, we provide the updated line numbers in the revised manuscript regarding to the referees comments. The response to editor is provided in the last section of this document.

<u>Response to Dr. Sylvain Ferrant</u>:

1. Comment1: There is a gap between the results analysis and the conclusions: Page 11857, line 3: "Though it is encouraging that GRACE assimilation improved the estimated streamflow, these results demonstrate that it clearly cannot replace high quality forcing data or good model calibration" P11859, line 3: in conclusion "GRACE assimilation is clearly beneficial ...From my point of view, there is no clear evidence of improvement between ENOL and ENKF. The small to really small differences between both cases shown in Figure 11 should be used to demonstrate that there is not much improvement in this specific study

The modified contents are located at <u>lines 632 – 636</u> in the revised manuscript.

2. Comment2: The abstract does not reflect the results presented in this study. The analysis showed a noticeable improvement in groundwater estimates when GRACE data were assimilated, with an overall improvement of up to 71% in correlation coefficient (from 0.31 to 0.53) and 35% in RMS error (from 8.4 to 5.4 cm) compared to the reference (ensemble open-loop) case.

The modified contents are located at <u>lines 536 – 537</u> and <u>lines 33 – 38</u> in the revised manuscript.

3. Comment3: Groundwater results are presented in abstract but Figure 7 does not give a clear idea of the stream flow improvements with GRACE assimilation. On the contrary, ENKF simulation of the TWS is really close to the GRACE derived TWS. This indicates that the assimilation process reach good results but the model is not able to take advantage of this to simulate better the water cycle. "Only a slight overall improvement was observed in streamflow estimates when GRACE data were assimilated. Even not any improvement. I doubt this could be explained only by the forcing data errors.

The explanation is provided in the response to S. Ferrant. There is no modification in the revised manuscript regarding to comment 3.

4. Comment4: One major water flux that is not taken into account is the water withdrawal for human and agriculture consumption. A recent study has used GRACEderived TWS to validate the calibration of an agro-hydrological model by taking irrigated water withdrawal into account (Ferrant et al., 2014, in Nature

ScientificReport; http://www.nature.com/srep/2014/140115/srep03697/full/srep03697.html). This part of the water consumption has a huge impact on the TWS anomaly derived from GRACE, and is not taken into account in this study. This should be discussed as the Rhine river basin is highly inhabited and include high industrial and agricultural activities

The modified contents are located at lines 175 – 180 in the revised manuscript.

5. Comment5: Page 11850, line 2. The calibrated model is calibrated on spatial soil moisture whereas averaged soil moisture is used for the non calibrated model. Please detail. This is not obvious for the reader. What kind of soil moisture data is used? Is it remote sensing soil moisture products? In that case, it is difficult to get an idea of the soil water storage from a surface soil moisture estimate.

The explanation is provided in the response to S. Ferrant. There is no modification in the revised manuscript regarding to comment 5.

6. Comment6: Section 5.2 Here the improvements of the TWS assimilation on groundwater are not obvious and are discussed in details. It seems that calibrated soil moisture does not lead to appropriate groundwater during the assimilation process. Groundwater data should be discussed regarding the accuracy or representativeness of piezometric data. Local fluctuations of the water table cannot often be considered as representative of the basin average.

The explanation is provided in the response to S. Ferrant. There is no modification in the revised manuscript regarding to comment 6.

7. Comment7: Page 11858 line 19, "GRACE could be combined with a hydrological model in a data-sparse region to yield additional insight into the variations in terrestrial water storage." I doubt this study demonstrates this. GRACE could be used as an extra observation to validate model, especially in a data-sparse region where any additional observations are welcome. Furthermore, TWS from GRACE is highly correlated to climate variables that are not always representative of a region in the case of global meteorological forcing data. The assimilation process will lead to redirect water fluxes between soil, groundwater and river to compensate the lack or the excess of water.

The modified contents are located at lines 618 – 623 in the revised manuscript.

Response to Referee 1:

1. Q1: Page 11841: You cite Güntner et al. (2008) for satellite altimetry. However, this paper is not really about altimetry and quite significant progress has been made in recent years regarding the accuracy of radar altimetry and its applicability to inland surface water bodies. I therefore suggest to reference a more recent state of the art publication on this topic.

The modified contents are located at line 72 in the revised manuscript.

2. Q2: Page 11847, lines 17ff: It is not entirely clear to me what the impact of GLDAS is here. If you use soil moisture from GLDAS to determine groundwater variations from the measurements, is this really an independent observation of groundwater? How meaningful is this observation after mixing it with GLDAS? Or do you rather validate against the soil moisture compartment of GLDAS? Please discuss this issue with a bit more detail.

The modified contents are located at line 265 – 291 in the revised manuscript.

3. Q3: Chapter 4.1: How did you set up your ensemble Kalman filter procedure? Did you use an available software package (such as, for example, DART)? Or did you implement the procedure individually?

The explanation is provided in the response to Referee 1 (A3). The modified contents are located in Sect. 4.1 (from line 298) and Sect. 4.2 (from line 333) in the revised manuscript.

4. Q4: Chapter 4.2: It does not become clear to me how you use the GRACE observations. Do you use them as a basin mean averaged over the Rhine catchment? Or did you calculate the GRACE TWS values on some grid?

The modified contents are located at lines 305 – 306, lines 357 – 358, and lines 369 – 370 in the revised manuscript.

5. Q5: Page 11849, lines 14ff: I do not really understand how the vertical distribution of the GRACE information into the soil moisture (SM) and the groundwater compartments (LZ and UZ) works. If I understand correctly, SM is adjusted first and if this storage reaches its upper or lower limit, the rest of the increment is applied to LZ and UZ (?). However, I would assume that the information about the distribution of the increment among the different model compartments can be obtained directly from the Kalman filter itself? Should this information (given a reasonable ensemble model covariance matrix, see also my question regarding Chapter 4.3 below) not be provided by the Kalman gain matrix? Please give some more details on this and why you chose to carry out the vertical distribution the way you do.

The modified contents are located at lines 357 – 368 in the revised manuscript

6. Q6: Page 11849 lines 21/22: You apply an observation error of 2cm for your GRACE TWS observations. This appears to be a rather simplistic assumption. First of all this number disregards the recent improvements of GRACE accuracy (Klees et al. 2008 used RL04 data). Furthermore, the Klees et al. refer to this accuracy for river basins above 1 million km2, which is significantly larger than the Rhine. Have you performed any kind of error propagation to test whether this assumption is valid for your test area? Or did you carry out any tests on how different GRACE error estimates would affect your assimilation results?

The modified contents are located at lines 369 – 375 in the revised manuscript.

7. Q7: Page 11850: lines 5ff: In your "non-calibrated" case, you set each parameter value to its mean value over the whole basin. This way, on average the non-calibrated and the calibrated cases agree. Is this really the case in data sparse regions? I would assume that even the mean value of the non-calibrated parameters might differ quite significantly from the mean value of the calibrated parameters. Therefore, I believe that setting the mean values equal is over-optimistic and not a necessary assumption. I would expect that your results might show the positive impact of GRACE in data sparse regions even better, if you would not assume a "correct" mean value for the parameters.

The modified contents are located at lines 386 – 399 in the revised manuscript.

8. Q8: Chapter 4.3: How do the model uncertainties enter the Kalman filter algorithm? Do you determine a full empirical ensemble covariance matrix with the dimensions of all of the model grid cells and the three model compartments? Or do you use only the variances? Please give a few more technical details on your approach.

The explanation is provided in the response to Referee 1 (A8). The modified contents are located in Sect. 4.1 (from line 298) and Sect. 4.2 (from line 333) in the revised manuscript.

9. Q9: Additional references: Quite recently, there have been additional studies on assimilating GRACE data into hydrological models (see full references below). First of all, Forman and Reichle (2013) discuss the effect of spatial aggregation of GRACE TWS estimates before assimilating them into a hydrological model. And Eicker et al. (2014) discuss the introduction of the full GRACE error structure into the assimilation procedure. (A more detailed treatment of the GRACE error from the product itself is an issue you also mention as topic for future research in your conclusions). I would suggest that you include references to those new studies in your manuscript.

The modified contents are located at lines 641 – 648 in the revised manuscript.

Response to Referee 2:

1. *Q1: Are GRACE observations averaged across the basin, or are they assimilated as gridded data? If the latter then how were horizontal error correlations taken into account?*

The modified contents are located at lines 305 – 306, lines 357 – 358, and lines 369 – 370 in the revised manuscript.

2. Q2: GRACE observations are assimilated once every 5 days and, if I understand correctly, no smoothing is applied. Are any temporal discontinuities seen in model state variables or related fluxes (e.g., ET, runoff) due to this episodic application of increments? None are obvious in the time series presented in the paper, but it *would be useful for the authors to comment on any artefacts that do exist or to discuss how this was avoided.*

The modified contents are located at <u>lines 339 – 345</u> in the revised manuscript.

3. Q3: GRACE products are now distributed with gridded error estimates, and a method for estimating basin-wide error using these estimates is provided on the GRACE Tellus website. How does the error calculated from these estimates compare to the 20 mm estimate used in this study?

The explanation is provided in the response to Referee 2 (A3). The modified contents are located at <u>lines 369 – 375</u> in the revised manuscript.

4. Q4: Also related to the question of GRACE errors: did the authors perform any sensitivity study by varying the GRACE error estimate? Figures 5,7, and 8 indicate that the DA run copy falls very close to GRACE, suggesting that the observations were weighted very heavily in the EnKF update. Is this optimal? A higher GRACE error estimate would relax the DA simulations back towards OL, and it would be interesting to see how this affects metrics of simulation performance.

The explanation is provided in the response to Referee 2 (A4). The modified contents are located at <u>lines 369 – 375</u> in the revised manuscript.

5. Q5. I am confused by the authors' comments regarding adjustment for "dry snow." Why, exactly, does this need to be corrected for in the GRACE observations?

The modified contents are located at <u>lines 350 – 356</u> in the revised manuscript.

<u>Response to Referee 3</u>:

1. Q1: Comment on the calibration vs. non calibration experiment: My first guess when I read the experiment setup was that results will not change much if the parameters were not calibrated but assumed to be the average over the basin. Even if you are using gridded (1deg) GRACE products, the spatial representation of GRACE is much courser than that so I would have guessed that the impact of a detailed (high spatial resolution) calibration of the model parameters does not have a major impact on your results if the spatial average of the parameters are used instead. In my opinion choosing an average of the calibrated parameters as the "non calibrated" case may be too optimistic and not representative of a region with limited observations. I would suggest to add/substitute this case with one where the parameters are not known (e.g. for example maybe just derived from a global land classifications such for example: http://webarchive.iiasa.ac.at/Research/LUC/External-World-soil database/HTML/ or other globally available database)

The explanation is provided in the response to Referee 3 (A1). The modified contents are located at <u>lines 386 – 399</u> in the revised manuscript.

2. Q2: Comment on the verification methods: The whole section about how/why you choose to scale groundwater in situ observations from piezometric to storage units needs some work. It is not clear to me why if you remove the soil moisture temporal mean from GLDAS you can get \Delta_GW_{in-situ}? Where does the \DeltaSM_{GLDAS} come in the context of equation (1)? If you remove a constant (average SM) from the GRACE aren't you effectively obtaining the same time series just shifted by a constant value?

The explanation is provided in the response to Referee 3 (A2). The modified contents are located at <u>lines 372 – 382</u> in the revised manuscript.

3. Q3: Treatment of snow: It is unclear to me what is the need to remove snow from the GRACE observations prior assimilation? Why don't just include it in the assimilation scheme? And include a snow term in the calculation of the modeled TWS?

The explanation is provided in the response to Referee 3 (A3). The modified contents are located at <u>lines 350 – 356</u> in the revised manuscript.

4. Q4: Actual EnKF scheme: It is assumed that a single observation is acquired in the middle of the month, however GRACE TWS is assumed to "represents the surface mass deviation for that month relative to the baseline average over Jan 2004 to Dec 2009." therefore this has to be considered as an average TWS variation for the entire month. This is effectively the reason why existing GRACE-EnKF techniques used a "two-step" approach (Zaitchik et al., 2008, Forman et al., 2012) where a single month was modelled twice: one time to obtain a "monthly average" observation prediction (from an open-loop simulation of the entire month, and not simply from the TWS modelled at a single day; and a second time to apply the increments computed from the EnKF. Are you also using a two-step approach or a straightforward application of the EnKF (as a real time assimilation scheme)? How would results change if instead the observation was assumed to be taken of the end of the month?

The explanation is provided in the response to Referee 3 (A4). The modified contents are located at <u>lines 305 – 306</u> and <u>lines 339 – 345</u> in the revised manuscript.

5. Q5: Temporal correlations: Observations are assimilated every 5-days. This is done after the temporally interpolating observations. Isn't this interpolation introducing an implicit temporal correlation across the assimilated observations? The EnKF assumes that each observation is independent from each other but the 5-days temporal interpolation includes temporal correlation. Did the authors consider the effects of their 5-days interpolations in the assimilation scheme? For example, how would results change if instead a different temporal window (lets say daily or every 15 days) is chosen for interpolation? Or how would results change if none interpolation was done after all and perhaps observations were assimilated only at the end of a month?

The explanation is provided in the response to Referee 3 (A5). The modified contents are located at lines 339 – 345 in the revised manuscript.

6. Q6: Spatial correlations of the GRACE observations: I read from http://grace.jpl.nasa.gov/data/gracemonthlymassgridsland/ that "The spatial sampling of all grids is 1 degree in both latitude and longitude (approx. 111 km at the Equator). However, this does not mean that two neighboring grid cells are 'independent' because spatial smoothing has been applied" this means that spatial correlations between neighboring GRACE-TWS pixels should be applied. It seems that the authors did not consider observations spatial correlations in their EnKF, is it correct? If so what is the rationale for not including it?

The explanation is provided in the response to Referee 3 (A6). The modified contents are located at lines 305 – 306 and lines 641 – 648 in the revised manuscript.

7. Q7: Figure 2/or add to the text: : : can the authors add a schematic representation of the model? E.g. it would be useful to understand what exactly upper/lower (UZ/LZ) mean in terms of the actual model physics. In the same figure, of text can the authors described how is soil moisture (SM) defined (e.g. depth? rootzone only? surface+rootzone? etc)

The modified contents are located at lines 151 – 169 in the revised manuscript

8. Q8: Please avoid the usage of "later" e.g. in section 2 toward the end of the first paragraph

The usage of "later" has been removed from the revised manuscript.

9. Q9: Can the authors add orographic contours on the Figure 1. Also the text oftentimes refers to the "Alps" region, could you please add this label in Figure 1.

Figure 1 is modified.

10. Table 4-5 are very hard to read, maybe can group these by regions identified in Figure 1. Or perhaps help the reader by highlighting which stations improved or not upon the open loop case?

In the revised manuscript, the values are grouped based on the groundwater network (as in Table A1).

Response to editor:

1. Conclusions and recommendations should be more clearly presented and more consistent with findings presented in results (see "spontaneous" review for further details). It should be discussed whether the presented results really imply that GRACE was able to achieve an improvement.

According to the responses to S. Ferrant, additional statements are added into the revised conclusion, to read:

Lines 618 – 623: Given that the most significant improvements were observed in the NCG case, this suggests that GRACE observations are most valuable in data sparse regions. In these regions any additional observations, even those at coarse spatial and/or temporal resolution, are welcome. GRACE can provide essential independent observations for validation, and serves as a constraint for TWS within the assimilation process.

Lines 632 – 636: In conclusion, GRACE assimilation is beneficial, and the largest improvements are generally observed in the NCG (i.e. "data-sparse") cases. In addition to providing a modest improvement to the estimated streamflow, it may result in a noticeable improvement in TWS estimates, yielding an extra insight into the behaviour of the hydrological model, its forcing data and parameters.

According to the responses to Referee 1 (A9) and 3 (A6), the additional recommendations are also added into the revised conclusion, to read:

Lines 641 – 648: In addition, recent studies have explored the effect of spatial aggregation of GRACE TWS prior to assimilation (Forman and Reichle, 2013) as well as inclusion of the full GRACE error structure (Eicker et al., 2014). Combining the advances made in those studies with our assimilation framework is expected to yield even more realistic estimates. As shown by De Lannoy et al. (2009), working with a spatially distributed state vector (3D-EnKF) can lead to an improved estimate. Given the coarse resolution of GRACE, we expect that implementing our framework with a 3D-EnKF would lead to an improved performance.

2. Several aspects of the methodology have to be clarified. More details on the implementation of EnKF have to be provided. How were different compartments jointly treated/updated? Details on the following more specific points are also needed:

The details on the implementation of EnKF are added into Sect. 4.1, Lines 298 – 331 in the revised manuscript.

a. How were GRACE TWS-values assimilated, grid-based or domain averaged?

Full details of the assimilation scheme are given in Sect. 4.2, Lines 333 - 375 in the revised manuscript. GRACE-TWS values are assimilated at grid based (Line 357, "the GRACE TWS are calculated and assimilated at each 1-km model grid cell every five days")

b. Why was an observation error of 20mm chosen and what is the impact of this decision? How were off-diagonal elements of the covariance matrix of EnKF treated?

The role of 20 mm is given in Lines 369 – 375 in the revised manuscript:

"The 20 mm value is considered realistic as it was suggested by several independent assessments e.g., Klees et al. (2008), Wahr et al. (2006), Schmidt et al. (2008) and it also had been applied in previous GRACE assimilation studies (Zaitchik et al., 2008; Houborg et al., 2012). Our philosophy was to set the GRACE errors to realistic values determined from independent studies, so that the solutions were not guided towards any particular outcome."

As stated in Referee 3's response A6, the off-diagonal elements are not considered in our 1D-EnKF scheme. We included the additional statements and recommendations into the revised manuscript:

Lines 305 – 306: In this study, we implement a so-called 1D-EnKF (De Lannoy et al. (2009) in which each grid cell is updated individually.
Lines 369 – 370: The GRACE observation error is assumed to be 20 mm and horizontal observation error correlations are not considered.
Lines 641 – 648: In addition, recent studies have explored the effect of spatial aggregation of GRACE TWS prior to assimilation (Forman and Reichle, 2013) as well as inclusion of the full GRACE error structure (Eicker et al., 2014). Combining the advances made in those studies with our assimilation framework is expected to yield even more realistic estimates. As shown by De Lannoy et al. (2009), working with a spatially distributed state vector (3D-EnKF) can lead to an improved estimate. Given the coarse resolution of GRACE, we expect that implementing our framework with a 3D-EnKF would lead to an improved performance.

c. Why was snow removed prior to from TWS-signal prior to assimilation?

The treatment of snow component is explained in more detail in the revised manuscript:

Lines 350 – 356: Dry snow is also small averaged over the study area (approximately 2% to the estimated TWS in winter). Only over the Alp (see Fig. 1), the snow contribution is greater (approximately 7%). Therefore, we decided to exclude the dry snow from the state vector. To reconcile GRACE to OpenStreams wflow_hbv TWS, we then removed the dry snow component estimated from the nominal run from the GRACE prior to assimilation. Note that in catchments where the dry snow component is more significant, it should not be excluded from the state vector.

d. How were soil moisture data used?

Clearer explanation of the use of soil moisture data is given in the revised manuscript:

Lines 272 – 285: We adopt a similar idea by using the relationship between ΔTWS-ΔSM (TWS variation from GRACE minus SM variation) and the observed head to scale the observed head. Ideally, we would prefer to use in-situ soil moisture data to represent the SM term, but they are not available at the well locations, and the nearest station from the International Soil Moisture Network (ISMN: Dorigo et al., 2011) does not have data covering the GRACE observation period. The soil moisture estimated from remote sensing was also not appropriate because the penetration depth depends on frequency and would not be the same as that in OpenStreams wflow_hbv. Therefore, we decided to use GLDAS derived SM in this study. The SM variation from GLDAS (ΔSM_{GLDAS}) was computed by removing its long-term mean value. The long-term mean value was produced from all GLDAS SM data over the same period as the GRACE observations (see Sect. 5). The groundwater variations from GRACE ($\Delta GW_{\mathrm{GRACE}}$) were obtained by removing ΔSM_{GLDAS} from the GRACE observations every month. $\Delta GW_{\mathrm{GRACE}}$ was interpolated to daily values in order to compare it to the daily head variations Δh.

e. Was water extraction considered and how could it have affected results?

As discussed in the responses to S. Ferrant, the water extraction is not affected our result. The support statements are added into the revise manuscript:

Lines 175 – 180: Extraction of groundwater for irrigation is considered to be small over our study region. It accounts for less than 1 km^3/year. Industry is the largest user (Wada et al. (2014). However, The net removal is small as only 10% of the total water withdrawal over the Rhine is from groundwater and the water is re-introduced to the system after being used for industry. This is markedly different to the extraction of groundwater for irrigated agriculture observed in India (Ferrant et al. (2014)).

f. What was the role of GLDAS in the assimilation?

As stated in the Referee 1 and 3's responses A2, we used the relationship between GRACE minus GLDAS SM and the observed head to scale the observed head. Clearer explanation of the use of GLDAS is included in the revised manuscript Lines 265 – 291.