Zhang et al. describe the development of a new climate data record that provides monthly values of precipitation, evapotranspiration, runoff and total water storage changes at 0.5 degree resolution globally from 1984-2010. Their approach combines a variety of remote sensing, reanalysis and land surface model products using a weighting scheme based on the variance of each data source from the ensemble mean. Water budget closure is enforced using a constrained Kalman filter to attribute the sources of budget imbalance to individual water budget terms. I think developing a complete climate data record that is internally consistent and ensures water budget closure is an important data need that would be useful for many other scientific applications, and the authors do a good job of pulling together all of the relevant global datasets. Unfortunately, as detailed below, I have significant concerns about the approach used to ensure closure and the assumption that variability between data sources is representative of uncertainty and error. While I acknowledge that the authors are doing the best they can with what is currently available, I am not convinced that the approach used here is sufficient to overcome these data limitations and achieve water balance closure in a meaningful way.

General Comments:

1. The biggest concern I have with this approach is the reliance on the assumption that variability between data sources is a proxy for error individual products. I understand that this assumption arises from a lack of data for direct error analysis, but I still have significant concerns about its validity. At a minimum, I think the authors need to include some analysis demonstrating that the variability between approaches is similar to this error in locations where there are observations to compare to.
2. I'm also concerned with the weightings that emerge from this assumption. On Page 8 line 22 the authors note that this is 'optimal merging weight,' but it's not specified what this is optimal with respect to. Given that many of the data sources are not actually independent and some approaches contribute more datasets than others, this will result in a mean that is skewed toward the approaches with the most datasets regardless of how much unique information is being provided. I think a much more thorough analysis of what is redundant in the datasets is needed to identify when 'agreement' is actually indicating certainty as opposed to repetition of inputs and assumptions that arise from data limitations (i.e. greater uncertainty).
3. The weighting is particularly problematic for the total water storage calculations which rely on VIC and GRACE. It is assumed that the uncertainty of VIC is 5% and GRACE is 10% (Page 10 lines 17-18) and therefore when both datasets are available VIC is weighted higher than GRACE. I have concerns about using VIC at all given that it is not actually simulating deeper groundwater storage and it does not make sense to me to weight VIC higher than GRACE when GRACE is much closer to an observation of TWS than VIC is.
4. I disagree with the de-trending adjustment to ensure zero water storage changes over the 1984-2010 period (Page 11 lines 6-15). It's not clear to me why this assumption is necessary and in many developed locations sustained groundwater depletions over this time period have been well documented.
5. I think that additional discussion and analysis of the impacts of human development on this approach is needed. The outputs are verified only against basins without significant human

development (e.g. excluding basins with large dams, urban or irrigated area >2% or >20% forest cover change); however, gridded values are being provided globally both in developed and undeveloped locations. The developed climate dataset does not reflect natural conditions because some of the input datasets used reflect human activities (e.g. remote sensing ET and storage losses from GRACE) while others (e.g. simulated runoff) do not. I am concerned that it's not clear in the manuscript (1) exactly what assumptions are being made about human impacts on the individual hydrologic budget terms in the calculation and (2) that the biases causes by human activities are not well understood in this approach and may be incorrectly adjusted for with the closure adjustments made with the Kalman filter.

6. The verification datasets used here are not necessarily independent of the input datasets themselves. I suspect that for example the flux towers used here are also used to validate (and/or calibrate) many of the remote sensing and land surface models used here. While this is probably unavoidable given the limited number of global observations networks I think this should be evaluated and discussed because it's if these aren't really independent points, it's likely that performance based on these points is a best-case scenario.

7. In my opinion, the scientific motivation and conclusions of this work do not come out clearly enough. I think the introduction should be refocused on the strengths and weaknesses of existing datasets and the motivation for this work rather than starting with an outline of government organizations. For example, the paragraph starting on page 2 line 22 covers all of the remote sensing products as well as bias in inferred runoff and precipitation and challenges with water budget closure. I think this discussion as well as the motivation provided in the paragraph starting on Page 3 Line 25 should be expanded and should appear sooner in the introduction.

8. Section 2 should be expanded to provide a better summary of the strengths and weaknesses of the different datasets without relying so heavily on the supplemental material (e.g. page 5 line 14 and section 2.1.2 paragraph 1). I think it's fine to refer to the supplement for the details of these datasets but additional discussion is needed in the main text to explain to the reader the strengths and weaknesses of these approaches and why they were chosen. For example, it is important to clearly explain here the difference between satellite data, reanalysis products and land surface models including what goes into each and what assumptions they rely on before comparisons are made. Some of this information comes up in the discussion of differences but it would be helpful to outline approaches upfront first.

9. The figures could be improved to provide more quantitative metrics of performance especially with respect to spatial and temporal variability. For example, Figure 11 maps all of the water balance components globally in a single figure for multiple time periods but each subplot is so small it's very difficult to note the connections the authors are discussing. Some cutouts or regional assessments would be useful. Also, Figures 2-9 are repetitive and I think some of these could be moved to the supplemental material or different plotting approaches could be tested to summarize this information with less figures.

Specific Comments:

1. The list of satellite products page 2 line 25 would be easier to follow in table form.
2. Page 4 lines 3: I think before the paragraph laying out the advantages of this approach a more thorough explanation of the weaknesses of previous approaches would be helpful. For

example, the first reason given here is the expanded use of the Constrained Kalman filter; however, the current limitations of the Kalman filter have not been explained.

3. Table 1 should clearly differentiate land surface models from remote sensing products.

4. Page 5 lines 2-7: This is very detailed for this intro to this section. I think it would be better to keep this high level, and provide an overview of the general approach and the organization of section 2 for the reader here.

5. Figure 2: A more detailed caption explaining the acronyms and the difference between the grey line and the colored lines is needed. Some of this is included in the * points. You should rewrite these to incorporate all of this into a single caption. This is also true of the subsequent figures, which should be adjusted accordingly.

6. For figures 2- 9: I think it would make more sense to plot the standard deviation rather than the coefficient of variation. The CV values clearly display a seasonal pattern caused by dividing by the mean. Since this information is already provided in the colored lines in my opinion it would be easier to understand if the grey line just showed standard deviation. This would also address the 'abnormal high spread' noted on page 5 line 25.

7. Section 2.1.1: Some aggregated statistics of differences in total precipitation for the major basin would be helpful to quantify the overall differences between approaches.

8. Page 6 line 12: The derivation of the other four satellite products is described but not the GLEAM dataset.

9. Section 2.1.3: I think this section should include a description of how runoff is calculated in each model and the strengths and weaknesses of each approach and their systematic biases.

10. Page 7 line 6: Can you be more specific about what type of discrepancy you are referring to (i.e. a low bias)?

11. Page 7 line 7: Can you be more specific about the type of 'disagreement' you are referring to?

12. Figure 13 should be figure 8 since it gets referred to after Figure 7

13. Page 7 line 14: Should be 'capture'

14. Page 7 line 14-15: This is unclear, can you expand on the uncertainty estimates you are referring to here?

15. Page 7 line 19: It would be helpful to define 'total water storage change' and 'total water storage anomaly' explicitly here before getting into this discussion.

16. Page 7: Equation 2 is not necessary in my opinion since this approach wasn't used.

17. Page 7 line 20: It would be helpful to explain what the significant differences in these three processing centers are.

18. Page 8 Line 10: It sounds like you are using the ensemble mean of GRACE here for future TWSC analysis and not using VIC at all but I don't think this is the case.

19. Page 9 lines 3-10: Some demonstration of the impact of this adjustment on the time series would be helpful here given that the authors argue it is a 'key step' for temporal consistency.

20. Page 10 lines 22-23: Globally mean TWSC may be small but this does not mean local changes are small and if the point is 0.5degree resolution I think this could be a limitation. Some discussion of spatial variability would be helpful here.

21. Page 13 Lines 6-7: What does it mean to be ‘filtering out those basins with non-significant correlations’? This sounds like an additional step beyond the filtering for different anthropogenic impacts. What was the threshold for this filtering and how many points were filtered because of it?

22. Page 14 lines 9-10: Even though ET is most dominant during the summer I think that the verification should not be limited to the warm season without further justification.