This is my second review of the paper entitled "ERA-Interim/Land: a global land water resources dataset" by Balsamo et al.
My previous review is archived as HESSD, 10, C7409–C7413, 2014.
The reply by the authors is archived HESSD 10, C8316–C8324, 2014.

The summary of the paper given in the first paragraph of my previous review still applies.

Unfortunately, the paper is still not ready for publication in HESS. While the authors have improved the manuscript, and while I still think that the paper is of great interest to HESS readers, I still have concerns about technical issues and the presentation of the material. I also feel that the authors did not satisfactorily address most of the major concerns that I raised in my last review.

I therefore recommend that the paper be returned to the authors again for another round of MAJOR revisions.

This review first answers to the authors' response in regards to my previous review. Thereafter, new comments are listed. Most of the new comments are minor, but the fact that there are quite a few of them, the fact that the text is still sloppy in many places, and the fact that most of my major comments from the previous review still require discussion leaves me disappointed with the manuscript.

Previous major comments and answers to the authors' responses:

1) Throughout the paper, the authors only discuss two datasets, ERA-Interim/Land and ERA-Interim. By comparing only those two datasets, it is never really clear whether the differences between ERA-Interim/Land and ERA-Interim are due to the changes in the land surface model or whether they are due to changes in the precipitation forcing (GPCPv2.1 corrections). […]

R: Yes the Reviewer observation is correct, we concentrate on the ERA-Interim/Land to ERA-Interim comparison. There are several years of research and several publications in support of the ERA-Interim/Land and it is not possible to report results in those publications in the current one. The results in Albergel et al. (2013) are duly cited for this purpose.

Answer:
If results have been published in Albergel et al. (2013) they do not need to be replicated here verbatim, but this reader still would like to know whether any improvements in ERA-Interim/Land over ERA-Interim are due to new model physics or due to the use of observations-based precipitation forcing. The authors' response is not detailed enough to advise me of where changes might have been made in this regard.
Albergel et al. (2013) is cited only once (Line 473) where no information is provided on the relative contributions of the model physics changes and the observations-based precipitation forcing. Any sense of what the reason might be for the improvement is still

missing from sections 3.1.1 (fluxes) and 3.1.2 (discharge). In section 3.2.1 (soil moisture), the text implicitly suggests that improvements in the time series variability are due to changes in the model physics (greater dynamical range), but this is not stated explicitly, and Albergel et al. (2013) is *not* cited. Moreover, the lack of improvement in the time series correlation is indicative of the fact that the *monthly* precipitation observations are not helping improve that metric, but again this fact is not discussed in the text. Section 3.2.2 does suggest that the GPCP precipitation corrections are detrimental to the skill of snow estimates.

2) For a user in search of a global land reanalysis dataset suitable for their application, it would be tremendously helpful to include other available datasets in the comparison. Primarily, I would think that this should include MERRA-Land.

R: The article is meant to provide a reference for ERA-Interim/Land with description of its components and a selection of results. The comparison with MERRA-Land is very informative as well as other global datasets, but this has to necessarily realized in other studies.

Answer:
I understand the authors' reluctance to delve into the additional work required to add another reference dataset. However, it does leave the reader wondering whether ERA-Interim/Land is an improvement over MERRA-Land or not, and where ERA-Interim/Land fits into the bigger picture. Perhaps the results from Albergel (which include MERRA-Land) could be referenced in this context?

3) The title and the last sentence of the abstract mention "water resources". While soil moisture and SWE can be considered water resources, the perhaps more common understanding would also include lakes and reservoirs as well as groundwater. The latter components are not part of ERA-Interim/Land, however, and I think the term "water resources" is not appropriate here. How about "a global land surface reanalysis dataset"?

R: We believe the proposed title would be also appropriate but we would prefer to keep the current version that put the accent on water. The word "reanalysis" can however be added and we leave the decision with the Editor recommendations. We clarify in the abstract that this intent comes with the limitations of the modeling system utilized, and such limitations can be found in all global datasets. We also stress this point in the conclusions and mention ongoing research at ECMWF to embrace a more holistic representation of the land surface via a modeling cascade and an improved representation of bio-geo-physical processes.

Answer:
I note that the title has not been changed and still believe that a different title would be more appropriate.

4) Section 2.1.4, Fig 9, Table 3: Why is the validation vs. in situ soil moisture observations limited to a single year (2010)? Many of the in situ datasets cover much longer periods that should be used for validation.

R: In the soil moisture verification section, 2010 was chosen as the most recent year in this verification dataset coinciding with ERA-Interim/Land and also as the year with the largest amount of available observations. The same criteria was chosen also for identify good years for observation availability of other networks. Extended soil moisture verification is certainly possible and a very important component. This is realized in other studies (e.g. Albergel et al. 2013) but if redone here would change the balance of the topics. In this paper we aim at providing verification material for the different water cycle components within the land surface, both water fluxes and water reservoirs (those represented) to provide an overview of the surface water cycle performance in ERA-Interim/Land.

Answer:
I do not see why a longer (and therefore statistically more robust) soil moisture validation cannot be done here if it was done in Albergel et al. (2013). Also, I don't see how using a longer soil moisture validation period would change the balance of the paper in any way. This non-response to my comment suggests that the authors shied away from the additional effort required to extend the validation time series for this paper, even though this effort should be small given that the longer period was already used in Albergel et al. (2013).

5) Figure 2: If the underlying distribution is based on the 32 values for January 1 (or July 1) 1979-2010, then the 95th percentile is essentially the same as the second-largest value (because of the granularity of the distribution). It does not make sense to me to derive the "95th percentile" from just 32 values.

R: We tend to disagree, as Figure 2 is a meaningful illustration. For instance it provides visual information on which portion of land would be snow-covered in an exceptional year. Or similarly which area of the globe might be subject to extreme (note that also the median is provided in Figure 1, therefore 2 points in the distribution). Note that the 95th percentile is calculated for each grid-point independently and the resulting global map is not a predictable or trivial pattern. We agree however that 32-years is probably not an ideal length for reaching statistical robustness in extremes characterization and we clarify this caveat in the text. The length of the considered period is related to ERA-Interim availability and one should recognize that a 32-year reanalysis is a non-negligible effort requiring sizeable computational and research resources. Seasonal forecasting systems (e.g. EUROSIP participating systems) normally consider shorter hindcasts of the order of 16 to 20 years. To further address the Reviewer concern, in the conclusion a sentence is added to point to statistical robustness introduced by new and ongoing reanalysis in the ERA-CLIM project (such as ERA-20C) and illustrated in Dee et al. (2013)

Answer:
I agree that something like the figure in question is meaningful and offers valuable insights.
My point here was that the granularity of the percentiles is misleading, and had nothing to do with spatial patterns.
From just 32 values the 95-th percentile simply cannot be computed robustly. I could not find the added sentence on statistical robustness in the Conclusions, neither was I able to find a caveat

on statistical robustness in Lines 283-313. The authors' response generally lacks line numbers and quotes from the revised manuscript, which makes re-reviewing the paper much more difficult than need be.

6) Eg., page 14717, lines 1-2; page 14720, lines 18-20; caption of Fig 11: It is not always clear whether the "ERA-Interim" data that are analyzed here are from the original ERA-Interim dataset (derived with the coupled atmosphere-land modeling and assimilation system) or from offline simulations of TESSEL with ERA-Interim forcing. The latter should be very similar to the original ERA-Interim dataset, but it cannot be identical. At the very least, the difference needs to be mentioned clearly. The authors should also state how different the two "ERA-Interim" dataset are.

R: The differences are shown in the Figure 4 that allow the appreciation of the magnitude of the land surface revision introduced in ERA-Interim/Land. The text above is now included.

Answer:
The authors' response does not address my comment, which is still valid. I encourage the authors to re-read my comment carefully. My comment is *not* about the differences between ERA-Interim and ERA-Interim/Land. The comment is about differences between the original ERA-Interim dataset and an off-line run that essentially recreates ERA-Interim by using the same land model as ERA-Interim and ERA-Interim forcing (without any observations-based corrections). Was such an off-line integration used in any of the Albergel references? This matters because differences between such an off-line integration and the original ERA-Interim data would impact any analysis of the cause of the improvements in ERA-Interim/Land.

7) page 14718, line 21 - page 14719, line 2 and Figure 7: While the improvement in soil moisture […]

R: The misrepresentation of soil moisture at particular sites […]

Answer:
Thank you for the clarification.

8) Figure 9 includes confidence intervals and nicely demonstrates that for soil moisture, ERA-Interim/Land and ERA-Interim have essentially the same skill in terms of R. The same information is not available for other comparisons, e.g., Table 2, Fig 5, Fig 6, where it is not clear whether the skill differences between ERA-Interim/Land and ERA-Interim are significant.

R: We have now included confidence intervals in Table 2 consistently with the statistical significance method used for soil moisture results and reported in Table 3. We have not included confidence interval when single station Root-Mean-Squared-Differences were displayed (Fig.5) or when Cumulative-Distribution-Functions are used (Fig. 6, 8, 11).

Answer:
I appreciate the addition of confidence intervals in Table 2 (surface heat fluxes). However, the confidence intervals suggest that the improvements (eg., from R=0.81 to R=0.84) are well within the estimated 95% confidence intervals (+/-0.10), which is *not* clear from the text. Rather, the text (Lines 323-329) simply omits the obvious caveat that the improvements are not statistically significant. Please add an explicit caveat.

f) I do not understand Figure 6 and page 14717, lines 16-17): If the y-axis shows the cdf, then a horizontal line through y=0.5 should intersect the dotted blue line where the correlation is at the median value (x-axis). Similarly for the red line. But then the fact that the dotted blue line is "above" the red line would suggest that ERA-Interim has higher skill. Put differently, a cdf that rises more slowly has fewer low (correlation) values. I am getting this wrong? Or is the figure mislabeled? Also, why do some of the cdfs not end up at y=1 (for x=1)??

R: We believe the doubts on the interpretation are due to missing information on our side and this is now added in the text. Note that the CDF frequency curves do not start necessarily at 0 and end up at 1 in all cases (this is very evident for Asia) because the range of correlations obtained by the river discharge comparison with measurements can obtain also negative values that are excluded from the graph and the statistics as not meaningful. Large-scale models face enormous challenges when compared to point observations at river outlet (here the size of river catchment is not a criteria for exclusion from the comparison). The plot (now better explained in the text) is however very informative, in our view, of the general impact that ERA-Interim/Land revisions bring to water cycle at rivers level. The area comprised between the blue and red curve thus defines the figure of merit for ERA-Interim/Land. If the blue curve is above and the area is large this is a genuine improvement of skill at continental scale

Answer:
I appreciate the clarification about possibly negative correlations and agree that the cdf need not start at 0. However, correlations cannot exceed 1, so it is still not clear to me why the cdf does not approach 1 (y-axis) as the correlation approaches 1 (x-axis).
Moreover, I am still confused about the fact that the curve that is on top (greater y-value for the same x-value) would represent *better* skill. I still think it should be the other way around.
Turning my original argument around, why would a greater frequency (cdf-value) for a given correlation be better? For example, the Asia subplot suggests that for x=0.4, ERA-Interim has a cdf value of 0.3 whereas ERA-Interim/Land has a cdf value of 0.4.
For ERA-Interim, a cdf value of 0.3 at a correlation of 0.4 suggests that 30% of the correlation values are below 0.4 (that is, 70% of the correlation values are above 0.4).
For ERA-Interim/Land, a cdf value of 0.4 at a correlation value of 0.4 suggests that 40% of the correlation values are below 0.4 (that is, 60% of the correlation values are above 0.4).
Put differently, more correlation values are above 0.4 for ERA-Interim than for ERA-Interim/Land.

Again, what am I missing? Isn't the perfect cdf staying close to y=0 for as long as possible as x increases?

New comments:

The comments are in no particular order. Note that comments N-22, N-24, N-27, and N-28 are perhaps the most important and go somewhat beyond minor clarifications/edits.

N-1) Lines 129-148, (new) Figure 1: The new text and figure are very helpful and make it possible to understand how the surface meteorological forcing dataset was constructed from ERA-Interim data. However, now that I understand what was done, I am wondering about the potential inconsistencies between drawing the instantaneous fields (air temperature, humidity, wind and surface pressure) from different forecasts (03-12 h lead) than precipitation and radiation (09-21 h lead). Why are the instantaneous fields not also taken from the 09-21 h forecasts? Presumably, the air temperature (etc) forecasts that are valid at the same time but are from different forecast lead times are very similar. However, the text is silent about why this approach is preferable to using the same forecast lead time for all surface meteorological forcing fields. At the very least, the authors should add a statement to that effect.

N-2) Line 53: suggest replacing "modern era" with "satellite era"

N-3) Line 144: replace "instantaneous and accumulated fluxes" with "instantaneous fields and instantaneous fluxes"

N-4) Line 150: replace "The GPCP dataset" with "The monthly GPCP dataset" (it should be made clear in this subsection that the precipitation observations used here are monthly totals).

N-5) Line 200: It looks like Table 3 is discussed before Table 2. Please rename or reorder.

N-6) Line 238: typo "span-up" should be "spin-up"

N-7) Line 283: "ERA-Land" should be "ERA-Interim/Land"

N-8) Lines 284-286: "showing the added value…" This sentence is not in the right place because the results discussed in this portion of the text do *not* show "added value". They are just illustrative. The sentence is more appropriate after the section 3.1 heading. In fact, it is not clear why the discussion in Lines 293-313 is not in a separate subsection.

N-9) Lines 303, 304, 307: The figure numbers here seem to be following the original manuscript.

Line 303: replace “Fig 3a and Fig 4a” with “Fig 4a and Fig 5a”?
Line 304: replace “Fig 4a” with “Fig 5a”?
Line 307: replace “Fig 4” with “Fig 5”?

N-10) Line 310: replace “applied by data assimilation” with “applied by the screen-level data assimilation”? I presume that it is the screen-level DA that is relevant here.

N-11) Lines 326-329: This text should quote numbers from Table 2. See also comment above about lack of statistical significance in the improvements.

N-12) Line 336: “(blue dashed line)”? There is only a “blue solid line”

N-13) Lines 339-341: This sentence is misleading. As written, I would expect to find quantitative results indicated within each subpanel. In any case, such quantitative *should* be provided and discussed.

N-14) Line 343: The term “modelling cascades” should be avoided or explained.

N-15) Lines 366-367: Are the RMSD numbers quoted here from Albergel et al (2012a) or are they from the ERA-Interim/Land results discussed in the present paper? This is not clear from the text. Please clarify. Maybe “is shown” in Line 365 can be changed to “Albergel et al 2012a show that …” (adjust grammar accordingly)

N-16) Line 392, Figure 10, 11: Previous figures use red for ERA-Interim and blue for ERA-Interim/Land. Figures 10 and 11 switch the colors around, and Figure 12 uses red and green instead. Please use colors (or dashed lines) consistently throughout all figures.

N-17) Line 414: “ERA-Interim/GPCP-rescaled”? Should this read “ERA-Interim/Land”??

N-18) Line 427: “FCA=1 being the best value” There is still a mismatch in units. In Line 438, FCA values are in percent, but the percent sign is missing. I already commented on this in the previous review but the changes were not made consistently.

N-19) Lines 430-432: This text appears to be taken verbatim from a different tech memo or paper. What are the “two offline simulations” discussed here? Based on the authors’ responses above, I thought that the present paper only discusses ERA-Interim (which is not an off-line simulation) and ERA-Interim/Land (which is). Again, the distinction between the various runs discussed here and in the various Albergel references requires much clarification.

N-20) Lines 432-433: “Fig 12 (left)” should read “Fig 12a”, and “Fig 12 (right)” should read “Fig 12b”

N-21) Line 444: “Fig 12c” should read “Fig 13c”

N-22) Lines 441-446: The 3-panel Fig 13 is discussed in just 5 lines! At about 20 double-spaced pages, the paper is still relatively short. There is no excuse for such a marginal effort in putting text to a figure. Also, area-averages of the differences would be helpful.

N-23) Lines 471-475: This paragraph appears to address one of my previous major comments. However, it does not appear to be in the best place. This paragraph would much better fit where the objectives and scope of the paper are discussed.

N-24) Line 491: The earlier discussion clearly states that the GPCP-based precipitation forcing is detrimental to the snow estimates of ERA-Interim/Land, but the summary statements here do not reflect this finding.

N-25) Table 2: The "model" description (first column) explicitly states "HTESSEL" vs. "TESSEL", but the fact that ERA-Interim/Land and ERA-Interim are also different in terms of precipitation forcing is not reflected. This is misleading.

N-26) Figure 7: legend in top-left panel: "ERA-Interim-GPCP-offline-HTESSEL" should read "ERA-Interim/Land"???

N-27) Figure 9: The y-axis label is
"RMSD(ERA-Interim/Land) minus RMSD(ERA-Interim)"
The fact that numbers along the y-axis are positive suggests that ERA-Interim/Land has the higher RMSD. Is this mislabeled?

N-28) Figure 9: The caption speaks of "black dots", a "black solid curve", a "continues [sic] line", and a "dashed line". The graphic shows a red solid line with circles as markers and a blue solid line. It is impossible to interpret this figure given the obviously mismatched caption.