Authors' answers to Reviewer's comments

We would like to thank the Reviewer for the comments and suggestions that helped improving the quality of the manuscript. Below is an item-by-item reply to the Reviewer's comments and suggestions. The Reviewer's comments are reported in italic (1), the answers to the Reviewer are reported in blue (2), and the modifications to the manuscript are reported in green (3). The corrections will be inserted in the revised manuscript as indicated.

The authors evaluate relatively long (2002-2012) records of precipitation estimates from different sensors and devices (satellite, radar, rain gages). The focus is on the continental United States and analyses are performed at different temporal scales. The work and results seem solid and fit nicely within the line of research that has been pursued by the two authors. With that said, I have some major and minor issues with this study, as detailed below.

Major comments:

1. A big problem I have with this study has to do with its motivations. I found asking myself "why this study" quite a few times. If the focus is on a climatological analysis, then the satellite is not your best choice given that you have rain gages for a much longer period of time. The authors wrote, "this paper proposes to evaluate satellite precipitation estimates in the perspective of climate applications." Why? I would use 30+ years of rain gage measurements rather than 13 years of satellite estimates to accomplish that.

As indicated in the title, the primary goal of this paper is to evaluate different precipitations remotely sensed products like radar and satellite. This evaluation was performed over the period of record over which the different products are available. Stage IV is available from 2002 to present, while the Tropical Rainfall Measuring Mission (TRMM) suite of product is available starting in 1998.

If the longer-term perspective of this work is to use those remotely sensed estimates in the perspective of climatic application, a thorough evaluation of these products needs yet to be performed. The present study aims to serve as a benchmark in determining the difference and uncertainties between the different sensors suitability over CONUS for which we have extensive means of observation (in-situ, radar, satellites). According to our knowledge, there are no studies providing a comparison for in-situ, radar, and satellite measurements over an 11-year period.

To account for the Reviewer's comment, we modified the Introduction section and removed all source of confusion.

We will remove (p. 11490, ll. 22-23) *"Remotely sensed precipitation products are now coming of age where they can be considered for climatological applications.*"

We will replace (p. 11491, l. 23) *"Over the last 30 years …*" By: "*Over the last decades …*"

We will replace (p. 11491, l. 21) *"… and capture precipitation extremes in a climatological perspective.*" By: "*…and capture precipitation extremes over a multi-annual time frame.*"

We will remove (p. 11491, ll. 28-29) *"… and provide a long-term picture of the evolution of precipitation over time.*"

We will remove (p. 11492, ll. 11-12) *"Therefore, this paper proposes to evaluate satellite precipitation estimates in the perspective of climate applications.*"

We will modify (p. 11491, ll. 19-21) *"Although 11 years is not a long climatology, the duration of the study will be informative enough to derive long-term trends, assess systematic biases, and capture year-to-year and seasonal variability."* By: "*The duration of the study will allow to assess systematic biases and capture year-to-year and seasonal variability.*"

2. I don't think that the United States is not the right place to perform a study of this kind because of the large rain gage network. I would also argue that these results cannot be generalized to other regions of the world for two main reasons: 1) different climatology, synoptic conditions and types of event; 2) lack of a dense rain gage network (compared to the United States) to bias-correct 3B42. A study of this kind would have made more sense to me if the focus had been global and/or if analyses had been performed at the sub-daily scale (even though I would still argue that in the United States there are rain gages providing data at the hourly/sub-hourly scale for a period of time longer than the satellite).

Indeed, the generalization of this study cannot be generalized to a global scale. A similar study would be necessary to assess those biases over different climatological zones. However, the goal of this study was to evaluate radar and satellite products with respect to in-situ observations over CONUS. We agree that a global evaluation of satellite products would be pertinent but this will be the object of a future communication. A long-term global analysis of the most common satellite QPE products including adjusted datasets (3B42, CMORPH-ADJ, PERSIANN-CDR) and their unadjusted counterparts (3B42, CMORPH, PERSIANN) is currently being conducted.

Regarding the precipitation analysis at the hourly/sub-daily scale. The satellite products used in this study have a three-hourly temporal resolution. While the Stage IV is available at the hourly, 6-hourly, and daily scales, the hourly Stage IV precipitation estimates should be used with caution especially when using them to compare to other data sets due to the automated quality control at the RFC level that cannot identify bad rain gauge reports (See Fig. 3 in Nelson et al. 2015). For more details about the products used, see the revised version of the manuscript.

Reference:

Nelson, B. R., Prat, O. P., Seo, D. J., and Habib, E.: Assessment and implications of NCEP Stage IV quantitative precipitation estimates, *Weather Forecast*, in review, 2015.

3. There are a number of additional rain gage gridded rainfall products with high space-time resolution (e.g., see data by Ed Maurer at Santa Clara University).

As mentioned by the reviewer, there are other rain gauge gridded rainfall datasets available. In addition to the one named before, we could also mention the gridded estimates from GPCP (Global Precipitation Climatology Project), or GPCC (Global Precipitation Climatology Centre). In this work, we selected the GHCN-D dataset (more precisely the US-COOP subset of the GHCN-D dataset: See the answers to the second Reviewer for more details on the selection of the in-situ data). Regarding the dataset mentioned by the Reviewer, the dataset seems to be available for the period 1949-2010. Furthermore, it uses the same (or a portion) COOP stations used in this study (Maurer et al. 2002). Our goal in this study was to use directly the point stations for the direct comparison with the satellite and radar data. For consistency purpose, we used the gridded PRISM (Parameter-elevation Regressions on Independent Slopes) Model that incorporates GHCN-D in situ observations for annual and seasonal precipitation characteristics. By using point measurements (GHCN-D), (1) we avoid uncertainties and processing artifacts generated by the gridding procedure, and (2) we are able to select the rain gauges that were in operation over the entire period of record (we selected the rain gauges that reported at least 90% of the time over the period 2002-2012 (about 50% of the stations).

Reference:
Maurer, E. P., Wood, A. W., Adam, J. C., Lettenmaier D. P., and Nijssen, B.: A Long-Term Hydrologically-Based Data Set of Land Surface Fluxes and States for the Conterminous United States, *J. Climate* 15, 3237-3251, 2002.

4. I would assume that the vertical bars in Figures 4 and 6 are standard deviations. If that's the case, I doubt that the differences in mean are statistically significant. Please test this formally.

The vertical bars indicate indeed the standard deviations. We tested the statistical significance of the differences between the different datasets (PRISM, Stage IV, 3B42, 3B42RT) when compared to GHCN-D. The table below reports the results of the significance test for the annual (Fig. 4: Table 2) and seasonal (Fig. 6: Table 3) at the 5% significance level (Y=significance at the 5% level).

ID	YEAR				DJF				JJA			
	PRISM	St.IV	3B42	3B42RT	PRISM	St.IV	3B42	3B42RT	PRISM	St.IV	3B42	3B42RT
CONUS	Y	Y	Y	Y	-	Y	-	Y	Y	Y	Y	Y
ABRFC	-	-	Y	Y	-	-	-	-	-	Y	Y	Y
CBRFC	-	Y	-	Y	-	Y	Y	Y	-	Y	-	Y
CNRFC	-	-	Y	Y	-	-	Y	Y	-	Y	-	Y
LMRFC	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	-	Y
MARFC	Y	Y	-	Y	Y	-	-	Y	Y	Y	Y	-
MBRFC	-	-	Y	Y	-	-	Y	Y	-	Y	Y	Y
NCRFC	Y	Y	Y	Y	-	Y	Y	Y	Y	Y	Y	Y
NERFC	-	Y	-	Y	Y	Y	Y	Y	-	Y	-	Y
NWRFC	-	Y	Y	Y	-	-	Y	Y	-	Y	-	-
OHRFC	Y	Y	Y	-	Y	Y	Y	Y	Y	Y	Y	Y
SERFC	Y	Y	Y	-	Y	Y	-	-	-	-	-	Y
WGRFC	-	Y	-	Y	-	Y	-	-	-	-	-	Y

In the revised version, we will add a sign '' next to the values that are statistically different. The legend for Tables 2 and 4 will be completed by: "The asterisk indicates that the datasets are* *statistically different at the 5% significance level with respect to the surface observations." The revised sections 3 and 4, will include the result of the statistical tests.*

We will add (p. 11498, l. 10):

The differences are statistically significant at the 5% significance for about half of the RFCs (5 over 12).

We will add (p. 11498, l. 11):

… and statistically significant differences for 9 of the RFCs.

5. Another element that is not discussed and that could affect the evaluation of the products is related to the fact part of the study region is outside of the orbit of TRMM (35N/S).

The products 3B42RT and 3B42 are available from 50S-50N and are here used over CONUS for the domain 24N-40N. The TRMM Multisatellite Precipitation Analysis (TMPA) uses input datasets from low earth orbit satellites (TMI: The Microwave Imager; SSM/I: Special Sensor Microwave Imager; AMSR-E: Advanced Micro-wave Scanning Radiometer-Earth Observing System; AMSU-B: Advanced Microwave Sounding Unit-B). Apart from TMI, which has a footprint covering the band 38S-38N, all the other sensors (SSM/I, AMSR-E, AMSU-B) have coverage beyond the band 50S-50N of TMPA. While the sensor differing coverage might have an impact of the quality of the precipitation estimates for latitude above 38N (for the band 38N-40N), this point is beyond the scope of the study. More detail on the algorithm 3B42RT and 3B42 can be found in Huffman et al. (2007) and Huffman and Bolvin (2013).

Reference:

Huffman, G. J., and Bolvin, D. T.: TRMM and Other Data Precipitation Data Set Documentation, Lab. for Atmos., NASA Goddard Space Flight Cent. and Sci. Syst. and Appl. Inc., 2013. [Available at: ftp://precip.gsfc.nasa.gov/pub/trmmdocs/3B42_3B43_doc.pdf, accessed 28 Jan. 2015].

We synthetized this point in the section 2.4 related to 3B42 and 3B42RT product description.

6. Section 5 requires some additional work. I don't think it makes sense to compare "extreme" rainfall at fine resolution (Stage IV or even worse rain gages) with respect to a _625 km^2 pixel. Analyses of this kind should have been performed by both regridding Stage IV and interpolating the rain gages, or by using thresholds associated with the rainfall distribution for each pixel/location and products (e.g., 95th or 99th percentile).

The goal of this study being to use the different products at their native resolution, we do not see the necessity of regridding the different datasets from point measurement (in-situ) or 4-km (Stage IV) to the 25-km resolution of TMPA 3B42. However, we agree that this point, also mentioned by the other Reviewer requires clarification. The pertinent question is whether the correlation distance (defined as the distance at which the pixel-to-pixel correlation decreases below 1/e≈0.37) is greater than the native resolution of the sensors that is 4-km for Stage IV and 25-km for TMPA.

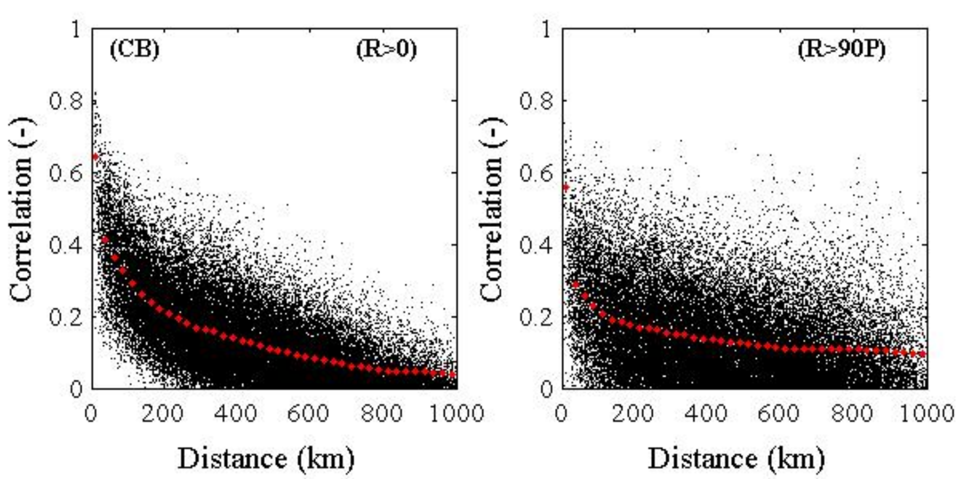
For unconditional rain rates, the correlation distance depends on the time scale considered. When it comes to extreme precipitation events (i.e. conditional analysis), this distance depends on the threshold considered (value or percentiles). The Table below shows the 90th percentiles of daily precipitation for the different RFCs at the seasonal (DJF, MAM, JJA, SON) and annual (YEA) scale derived from Stage IV (Nelson et al. 2015). As can be seen the there is an important variation between the seasons and the different RFCs. Roughly, the 90th percentile corresponds to the Wet Millimeter Day (WMMD: 18.8 mm/day).

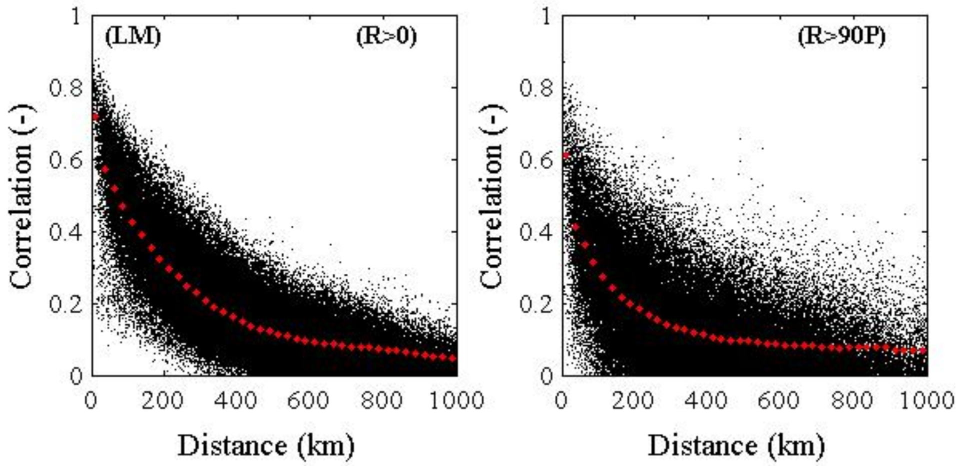
Percentile	Season	Rain rate (mm/day)											
		RFCs											
		NW	CN	CB	MB	AB	WG	NC	OH	LM	NE	MA	SE
90th	DJF	15.1	21.9	9.4	*6.4*	15.1	16.5	9.7	16.5	**28.6**	16.5	19.5	26.2
	MAM	11.3	13.5	*7.4*	14.1	23.0	23.4	16.5	21.4	**30.7**	17.1	18.8	28.5
	JJA	9.3	*6.3*	7.3	17.9	24.0	21.6	20.9	22.3	24.1	19.4	20.9	**26.0**
	SON	13.0	13.8	*10.1*	12.9	22.8	24.3	17.1	22.1	**30.5**	23.8	26.1	28.8
	YEA	12.5	15.9	***8.3***	13.9	22.0	21.8	16.8	**30.4**	**28.4**	19.1	21.2	27.0

**For each threshold and season, the max value is reported in bold and the min value is reported in italic. On average, the CBRFC (Colorado Basin) reports the lower values, while the LMRFC (Lower Mississippi) displays the higher values. The values reported in red are the values tested for the determination of the correlation distances for the 90th percentile.*

For each station with precipitation corresponding to different daily conditions (R>0, R>90th percentile), we computed the spatial correlation of daily precipitation events regardless of the daily values of the other stations for the CB and LM RFCs. The red dotted lines represent the average values obtained for 25-km intervals.

The correlations decrease rapidly with increasing distances between stations. The decrease is sharper for the CB RFC than the LM RFC and with increasing percentile of the daily value. For the 90th percentile, the average correlation distance (at 1/e) is about 30-80-km. Typically, the average correlation distance is greater than the representative footprint of each sensor: 4-km for Stage IV and 25-km for TMPA.





In the revised version, we synthetize those results. See also the answer to the other Reviewer. We added in the text related to Fig. 10 (p. 11507, l. 18)

However those results have to be interpreted with caution as they present a count of the daily events over the 11-year period. The number of events decreases with increasing rain rate and the WMMD correspond roughly to the 90th percentile precipitation events regardless of the RFC (Nelson et al. 2015). A test was performed to determine the interstation correlation of daily precipitation events corresponding to the 90th percentile (not shown). For each station, the correlation was computed using the daily events greater than the 90th percentile regardless of the values of the other stations. Results showed that for those high-intensity events, the average correlation distance was about 30-80-km which is comparable with the satellite footprint.

We also added (p. 11508, l. 13):

We will assume that the rain gauge is representative of the grid-averaged rainfall for Stage IV. The computation of the interstation correlation for daily events indicated that the correlation distance was greater than the 4-km spatial resolution of the radar (not shown).

Minor comments:

7. While the manuscript is generally well written, there are few typos here and there [e.g., pg. 11492, line 14 (remove parenthesis); pg. 11497, line 6 (stage IV); pg. 11499, line 15 (3B43RT)].

This will be corrected in the revised version.

8. Pg. 11505, line 9: rain gages have troubles measuring solid precipitation as well.

We added this point at the end of the revised section 4.

9. Pg. 11509, line 12: why "radar-only" if rain gages are used as well?

Here we refer to the radar pixel location that is defined as the closest radar pixel to a given rain gauge. This corresponds to a "false alarm" where precipitation is detected only at the radar pixel and nothing is detected by the rain gauge. For clarity purpose, we added the word "pixel" between "radar" and "only" as well as in the Figure 12 legend.

10. Please include confidence intervals in the qq-plots on Figures 3, 7, and 8.

The multi-paneled qq-plots and scatterplots are already delicate to decipher due to the number of points. Therefore it is doubtful that this would be of any interest to add the confidence interval on the figures. An option would be to replace the scatterplots by the confidence interval. However, we feel that the scatterplots provide a more pertinent illustration of the differences with respect to surface observations especially between the adjusted and un-adjusted versions of 3B42.

11. Figure 10: are the proportions different at the 5% level? Similar question for Figure 11.

We are not sure what the Reviewer means by that. The Figure 10 represents the number of stations or corresponding pixel that experience precipitation above a given condition (R>WMMD, EPD2, EPD4: 1 for yes; 0 for no) for each RFC.

12. Are the 3B42RT data rerun every time there was a change in the number of satellites providing data? If not, it is hard to make comparisons.

The 3B42RT product is the near real time satellite QPE and has a lag time of about 6-hrs. The product incorporates various satellite measurements. For a given time stamp (available at 00Z, 03Z, 06Z, 09Z, 12Z, 15Z, 18Z, 21Z), the precipitation estimate include the best available information that falls within the time window of +/- 1.5hr centered around the nominal time. The number of satellite providing data can be different for each time stamp and therefore differences in quality can be expected from one time stamp to another. However, this point is beyond the scope of this study. For more details we will refer to Huffman et al. (2007) and Huffman and Bolvin (2013).

The description of the satellite products 3B42 and 3B42RT that includes a discussion related to uncertainties was extended in the revised version. We added (p. 11495, l. 5):

The quality of the blended precipitation estimates depends on the number of satellite estimates available at a given time stamp and on the sensor characteristics.