Abstract. Line 11 (and throughout the paper), none of the two products validated in this work are strictly “forecasts”: I suggest to refer to the validated fields as “estimates” or “products” instead.

We changed the text accordingly.

Introduction. This section has to be improved. First, no mention of the most popular satellite derived precipitation products is given. Global products such as TMPA (Huffman et al., 2007. J. Hydrometeor.,8, 38–55), C-MORPH (Joyce et al. 2004, J. Hydrometeor., 5, 487–503) PERSIANN (Behrangi et al. 2009, J. Hydrometeor., 10, 1414–1429), H-SAF (Mugnai et al., 2013, NHESS, 13, 1959–1981) and the newest IMERG (Huffmann et al. 2015, http://pmm.nasa.gov/sites/default/files/document_files/IMERG_doc.pdf, should be at least mentioned. Moreover, the introductory discussion of literature about evaporation (now on lines 23-30 on page 13) and HOAPS/ERA-I validation (lines 31-page 12 to line 14-page 13) should be moved here.

We followed the suggestions of the reviewer and revised the introduction as well as page 12 and page 13. The revised introduction is given below:

Precipitation is one of the key parameters of the global water cycle. Therein the precipitation over the ocean is especially important since it contributes more than 75 % of the global annual total precipitation (Schmitt, 2008). Due to climate change it is reasonable to expect changes in both the pattern and amount of precipitation (Liu and Allan, 2013, O'Gorman et al., 2012 and Trenberth, 2011), as indicated by changes in the horizontal surface salinity distribution over the ocean (Durack et al., 2012). However, Valdivieso and Haines (2011) figured out that near the surface, much of the Atlantic is generally saltier compared to the climatology, although not uniformly so. They suspect that precipitation from the ERA-Interim (Dee et al., 2011) product might have errors. On the other hand Grist et al. (2014) found that a large part of the spread in the estimates of the mean surface-forced circulation of the sub-tropics is associated with biases in the global ocean heat budgets implied by the atmospheric reanalyses. Indeed for saltiness an unbiased estimate of freshwater fluxes is needed, good knowledge about evaporation is also essential. Regarding precipitation robust trends in regional precipitation over the ocean have not been detected to date because in-situ precipitation measurements are sparse and uncertain (Rhein et al., 2013). However, the progress in satellite technology has provided the possibility to retrieve global data sets from space, including precipitation. In addition to the Hamburg Ocean Atmosphere Parameters and Fluxes from Satellite data (HOAPS, Andersson et al., 2010 and 2011)), remote sensing based precipitation data sets are alternatively available e.g. from the GPCP (Global Precipitation Climatology Project, e.g. Huffman et al., 1997), CMorph (Climate Prediction Center MORPHing technique, Joyce et al., 2004), TMPA (TRMM Multisatellite Precipitation Analysis, Huffman et al., 2007), Persiann (Precipitation Estimation from Remotely Sensed Information using Neural Networks, Behrangi et al., 2009), the HSAF (Hydrology Satellite Application Facility) precipitation products, Mugnai et al., 2013) or IMERG (Integrated Multi-satellitE Retrievals for Global precipitation measurement, Huffman et al., (2015). However, concerns have been expressed over the need for further work to evaluate these data sets (Rhein et al., 2013).
Recent progress in reanalyses shows improved accuracy in precipitation (Dee et al., 2011), although tests for internal consistency among different components of the hydrological cycle in reanalysis data still reveal some issues (Trenberth et al., 2011). This is supported by a study of Andersson et al. (2011), who pointed out that state-of-the-art satellite retrievals and reanalysis data sets still disagree on global precipitation amounts, patterns, variability and temporal behaviour, with the relative differences increasing in the poleward direction. Pfeifroth et al. (2012) showed that HOAPS tends to underestimate precipitation by 6 to 9% in the tropical Pacific, where ERA-Interim (Dee et al., 2011) gives an overestimation of 8 to 10%. Note that these comparisons were done against measurements on atolls, although the representativeness of the atoll gauges of open-ocean rainfall is still an open question (Wang et al., 2014). Andersson et al. (2011) have shown that HOAPS gives lower precipitation rates than ERA-Interim, except for small areas in the northern sub-tropics, the north-western and south-western Atlantic Ocean. They also figured out that over the mid–high latitudes between 40° and 70° the precipitation (only liquid precipitation) in GPCP (Adler et al., 2003) is also systematically 10%–30% higher relative to HOAPS. Locally the values exceed 50%. North of 60°N Lindsay et al. (2013) detected an overestimation in

monthly precipitation in ERA-Interim ranging from 10 to 25%. Kidd et al. (2013) found that precipitation in the tropical Atlantic is skewed to the east towards the African coast by ERA-Interim. On the other hand, several studies have shown that in the area of the Iberian peninsula (Belo-Pereira et al., 2011) or in the area of four African river basins (Thiemig et al., 2012) ERA-Interim overestimates the frequency of rain events, as well. Furthermore, with respect to solid precipitation, Klepp et al. (2010) demonstrated the ability of HOAPS to detect even light amounts of cold season snowfall; with a high accuracy (96%) between point-to-area collocations of ship-based optical disdrometer data (ODM 470) and HOAPS data.
In the present study we use in-situ ship rain gauge (Hasse et al., 1998) and optical disdrometer (ODM 470) data (Großklaus et al., 1998), gained on board of research vessels, to validate two data sets, the HOAPS and the ECMWF (European Centre for Medium-Range Weather Forecasts) ERA-Interim reanalysis data set. For the present study HOAPS has been chosen as an example of a satellite derived precipitation estimate, because it is only derived from SSM/I (Special Sensor Microwave Imager) radiometers in their native sensor resolution without any further interpolation.
Section 2 gives an overview over the data and instruments used. The collocation of the data is described in section 3, followed by an overview of used validation methods in section 4. In section 5 we present our results, followed by a discussion of the results and the summary and outlook in sections 6 and 7.

Section 2.
Page 2. Here there is no real reason to mention the exact position of the data and I suggest not mentioning Figure 6 (it is also repeated in similar sentences at lines 24 and 28).

We deleted mentioning Fig. 6

Page 3, lines 9-10. Later in the paper, it is stated that that the ships' speed is 20 Kn. This information should be anticipated here, indicating if this value is a kind of average of the cruise speed or it is just an order of magnitude.

Cruise speed is that of the merchant ships over the Baltic Sea to relate temporal and spatial scales, we changed the sentence to:
A decorrelation length of 30 km, which corresponds to a temporal correlation length of 45 min assuming a merchant ship's speed to be of the order of 20 Kn.

Figure 2 is confusing, and does not help in understanding section 2.5. I suggest to better explaining all the features in the figure (e.g. what is the meaning of the colors?), to improve the caption and to improve the clarity of section 2.5.

In fact the colors indicate only same elements of the time series used to derive simulated fields, which is now mentioned in the capture to Fig.2:

Figure 2. Scheme for estimates of simulated precipitation fields from precipitation time series gained on the main building of the GEOMAR in Kiel (54.3° N, 10.2° E) at a time interval of 8 min. The numbers, additionally highlighted by colours, indicate the nth 8 min interval of the time series. Assuming a speed of motion of precipitating clouds of 5 m s-1, each 8 min interval is equivalent to a displacement of 2.4km. Thus, a spatial field of 75 km x 75 km is simulated by 31 elements x 31 elements of the 8 minutes time series. Since ERA-Interim precipitation is accumulated over 3 h, simulated data have to be averaged over 23 consecutive fields (23·8 min=184 min). Left: Scheme for a spatial field representing ERA-Interim resolution at the beginning of a 184 min period, right at the end of this period. Start elements are indicated by the red colour. The X indicates an observation, which is taken randomly from elements n = 1 to n = 68 of the time series for ERA-Interim simulations and from elements n = -5 to n = 37 for HOAPS simulations due to collocation criteria (chapter 3.1 and 3.2).

Section 2.5 has been changed to improve the clarity:

To get an idea of reasonable numbers for the statistical analysis of collocated along-track measurements with areal/temporal averaged estimates, simulations of in-situ observation data sets and corresponding areal averages have been estimated from 8 min time series of precipitation measurements performed on the main building of the GEOMAR in Kiel, Germany. To derive simulated areal averages from a time series the scheme given in Fig. 2 has been used based on the assumption that Taylor's principle of frozen turbulence can be applied to

precipitation fields in a similar manner, assuming a speed of motion of 5 m s^{-1} of precipitating clouds. Thus, an 8 min interval is equivalent to a displacement of precipitating clouds by 2400 m. Areal averages R_{field} have been computed according to

$$R_{field} = \sum_{n=1}^{n\max} w(n)\, R_{timeseries}(n) \ , \qquad (1)$$

where n indicates the n^{th} element of the time series, $w(n)$ is a weighting function according to the number of same elements of the time series $R_{timeseries}(n)$ at a certain time n used for averaging, normalized by the total number of values used for averaging. Simulated in-situ measurements were taken randomly from the same time series. With respect to the different resolutions of HOAPS and ERA-Interim and the assumed displacement of 2.4 km within any 8 min interval, we used a 21 x 21 field (21·8·60 s·5 m s^{-1}=50.4 km) for HOAPS averaging and a 31 x 31 field for ERA-Interim averaging. (31·8·60 s·5ms^{-1}=74.4km). While for HOAPS a simulated field was estimated at a certain starting time, simulated fields for ERA-Interim were estimated as an average over 23 consecutive fields in time (23·8 min=184 min), each constructed according to Fig. 2, to simulate the time increment of 3 hours. Starting with element n=1 this gives nmax=31 for simulated fields of HOAPS and nmax=68 for simulated ERA-Interim fields. For HOAPS-simulations, simulated point measurements were taken from those elements of the time series, which are within the temporal threshold used for collocation (see chapter 3.1), for ERA-Interim simulated measurements were taken from all members of the time series used for calculating the simulated fields. In case of HOAPS, simulated fields having a precipitation rate below 0.3 mm h^{-1} are set to zero, according to the lower threshold in the HOAPS data.

Section 3. Page 7, lines 6-9. The discussion on the use of an interpolation system for rainrate data is probably out of the scope of the paper (many interpolators do not affect rainrate maxima and minima). I suggest to simply say that nearest neighbor is used following Bumke et al., 2012.

We deleted the first paragraph and started with:
As in a similar validation study over the Baltic Sea (Bumke et al., 2012), the nearest neighbour approach was chosen for collocation. Therefore, it must be …

Page 7, line 14. It is 30 km a weighted mean of the decorrelation distance? How it is computed?

It is simply an average value from different estimates of the decorrelation lengths based on 8 min timeseries, where we assumed the lower decorrelation length for stratiform/frontal precipitation;
(17km + 25km + 46km) : 3 = 88km : 3 ≈ 30km
We changed the text and added Puca et al. an additional reference:

Using 46 km as decorrelation length for stratiform/frontal precipitation these three numbers give an average decorrelation length of about 30 km which agrees well with a study of Puca et al., 2014. A decorrelation length of 30 km corresponds to a temporal correlation length of 45 min assuming a merchant ship's speed to be of the order of 20 Kn.

Since the data domain ranges over all latitudes and seasons, it would be possible (and useful) to apply different correlation lengths, according to the precipitation type (convective/stratiform)?

Unfortunately the data do not include any information about the precipitation type. So we used an average decorrelation length for all collocations, derived for 8 min time series. On the other hand we merged the data to events, thus, each event represents, also with respect to measurements, a temporal/spatial average. In fact measurements of collocated HOAPS events represent an average over about 1 hour. Indeed this means that decorrelation lengths also should increase, see for example (http://postel.obs-mip.fr/IMG/pdf/CSP-0350-ATBD_Precipitation-I2.00.pdf). That the chosen decorrelation lengths are uncritical is supported by test runs, where we changed decorrelation lengths for collocation purposes by +-10km. This had no significant influence on the statistics.

Section 4. Page 8, line 11. About the considered indices, I suggest the following. 1)

avoid to use the “accuracy”: it depends on the number of correct negatives, and this number varies with the data acquisition strategy, affecting the results; 2) replace accuracy with some equitable skill score, such as the ETS or HSS (mentioned below in the manuscript) to assess the skill of the product with respect to random rain assignments; 3) rename hit rate with Probability of Detection (POD), and write explicitly that the success ratio can be considered as 1-FAR (False Alarm Ratio): POD and FAR are more self-explaining (and popular) names for these indicators (see, among others, Puca etal., 2014, NHESS, 14, 871-889).

In the submitted version, although not written explicitly, we stated that changes in the accuracy between both periods, was caused by differences in the data acquisition scheme, i.e. the coupling with a rain sensor. Instead of deleting the results of the accuracy we mention now clearly that it depends also on correct negatives and that it can be considered as 1-FAR as suggested. We added now figures for the HSS and changed everything accordingly:

Starting with the last paragraph of chapter 4:
This allows us to derive the accuracy, the bias score, the probability of detection (POD), the success ratio and the Heidke skill score. The accuracy is the fraction correct, 1 indicates perfect accuracy. It also depends on the number of correct negatives. The bias score answers the question: How does the HOAPS/ERA-Interim frequency of "yes" events compare to the observed frequency of "yes" events? A score of 1 means that both data sets include the same number of precipitation events; a larger bias score indicates that there are more precipitation events in the HOAPS or ERA-Interim reanalysis data. The POD gives the portion of measured precipitation events, which can also be found in the HOAPS or ERA-Interim reanalysis data, with 1 indicating perfect agreement and 0 no agreement, while the success ratio gives the fraction of precipitation events in HOAPS or ERA-Interim reanalysis data that are also seen in measurements. Again, a score of 1 indicates perfect agreement while 0 indicates maximum disagreement. It is equal to 1 minus the false alarm ratio, where the false alarm ratio gives the fraction of the observed no-events which were incorrectly forecasted as yes. The Heidke skill score measures the fraction of correct estimates after eliminating those which would solely be correct by random chance. 0 indicates no skill, perfect score equals to 1.

5 Results

The results are given for the earlier period (1995-1997) in terms of accuracy, bias score, POD, success ratio and Heidke skill score as a function of an assumed lower threshold of measured precipitation rate (Fig. 4), below which the measured precipitation rates were set to zero. The statistical parameters are derived separately for rain events, snow events and all events; for comparison the results of the simulations (chapter 2.5) are shown. The results for the 2006 to 2008 period are depicted in Fig. 5. The results derived from simulations (chapter 2.5) and the rain only events from the earlier period (1995-1997) are shown for comparison. The numbers of events with measured precipitation are given in Tab. 1 as a function of the lower threshold for both periods and both collocated data sets. As expected numbers decrease rapidly with increasing threshold, since rain rate is distributed as a power law.
Taking into account that ERA-Interim data has no precipitation threshold, only rain rates below 0.01 mm h^{-1} are set to zero for the statistical analysis. Best agreement is expected if no threshold is applied to the measurements. For HOAPS, having a lower threshold of 0.3 mm h^{-1}, best results can be expected for a lower threshold in measurements similar to that of HOAPS. This is seen in the estimated accuracy, where HOAPS shows the best performance for measurement thresholds ranging between 0.1 and 0.5 mm h^{-1} , while ERA-Interim reaches highest values in accuracy when applying no threshold to observational data. Obviously the performance of HOAPS to detect snow is lower compared to ERA-Interim (0.6 to 0.8), but it is still close to the results for simulated data, deviations are less than 0.04 in terms of accuracy for thresholds of 0.1 to 0.5 mm h^{-1} in measurements. On the other hand, estimated snow accuracies for ERA-Interim decrease rapidly for higher thresholds applied to the measurements. The rain performance of ERA-Interim is weaker than that of HOAPS in terms of accuracy, and for a wide range of thresholds applied to the measurements it is even lower than estimated for simulated fields. Note that the different behaviour of simulations with an increasing lower threshold in the measurements is due to the assumption of a lower threshold of 0.3 mm h^{-1} in the simulated fields for HOAPS
The bias score, for which perfect agreement is given at value 1, indicates that ERA-Interim’s frequency of precipitation exceeds that of the measurements, strongly increasing with increasing threshold applied to measurements. HOAPS’ bias score is close to 1 for a lower threshold in the observation data ranging from 0.1 to 0.2 mm h^{-1}, the increase with increasing lower threshold applied to the measurements is less prominent. The

results for the simulations indicate that this is mainly a result of averaging over larger areas due to the resolution of the ERA-Interim analysis, although the model description states that ERA-Interim grid point values do not represent an areal average (ECMWF, 2015).

It is obvious that the POD (Fig. 4) should increase with the lower threshold in measurements, because intense measured precipitation events should be easier to detect than low intensity precipitation events. PODs for HOAPS are of the order of 0.5 to 0.7; the smallest values are estimated for solid precipitation. Nevertheless, the values of simulations indicate a good agreement between HOAPS and the measurements. The ERA- Interim precipitation estimates show PODs close to 1, not depending on a lower threshold applied to measured data. These high PODs mainly originate from the high values of the bias score, i.e. when ERA-Interim frequently gives precipitation it also shows precipitation in more or less all cases where precipitation was measured. The POD numbers derived from simulations also indicate that this might be partially a result of the coarser spatial resolution.

The success ratios show reasonable results for both data sets. Success ratios for HOAPS compared to measurements reach values of 0.7 to 0.9 for the lowest threshold applied to the measurements and decrease with increasing lower threshold, as expected. These numbers are comparable to those of the PODs and indicate again a good agreement between HOAPS and measurements. Values are a little higher than those estimated from simulations, and differences between different kinds of precipitation are small again. The success ratios of ERA-Interim reanalysis data compared to measurements are much smaller, which is a consequence of the high frequency of precipitation events in the ERA-Interim reanalysis data as also reflected in the bias score. I.e. when in ERA-Interim estimates the frequency of precipitation is too high, precipitation is also not measured in all cases. The Heidke skill score is in general higher for HOAPS than ERA-Interim indicating a better performance of HOAPS in detecting precipitation with slight weaknesses in the case of snow. As expected the Heidke skill score decreases with increasing lower threshold applied to the measurements. Nevertheless, a comparison to the results for the simulations also depict a good performance of the ERA-Interim reanalysis data in correctly predicting the occurrence of precipitation events, taking into account the coarser resolution of the ERA-Interim reanalysis data. Overall, validation results for solid and liquid precipitation show a similar level of detection for rain and snow.

Figure 5 gives the results for the second period, 2005 to 2008. For comparison the results for the period 1995 to 1997, as well as for the simulations are also provided. Since the frequency of events with no precipitation has increased compared to the earlier period, because instruments were running continuously (chapter 2), accuracies have enhanced considerably, but ERA-Interim still gives a less accurate estimate than HOAPS. This is mainly due to an increasing frequency of precipitation events in ERA-Interim, as indicated also by the bias score, which indicates that the frequency of precipitation in ERA-Interim data is about three times higher than in the measurements and also considerably exceeds the simulated bias score. This is in line with the estimated PODs, which are again close to 1. The values of the success ratios for HOAPS are close to the estimates of the first period, while ERA-Interim give success ratio numbers that are only about half the values of the earlier periods. The Heidke skill scores are in general slightly higher, but still remain on a low level especially for ERA-Interim. In summary, HOAPS performs similarly during both periods while ERA-Interim performs better in the earlier period.

Lower threshold (mm/h)	Number of events with measured precipitation			
	1995-1997 HOAPS	1995-1997 ERA-Interim	2005-2008 HOAPS	2005-2008 ERA-Interim
none	354	353	364	1365
0.1	243	242	286	916
0.2	392	191	235	721
0.3	358	158	195	572
0.4	327	127	146	460
0.5	315	115	123	368
0.6	301	101	111	318
0.7	92	92	94	276
0.8	86	86	77	250
0.9	79	79	68	222
1.0	72	72	64	191
1.1	67	67	58	172
1.2	63	60	52	153
1.3	58	57	49	143
1.4	53	53	43	133
1.5	48	48	40	123
1.6	45	45	34	111
1.7	44	44	31	105
1.8	44	44	30	101
1.9	41	41	25	94

Table 1. Number of events with measured precipitation, where the precipitation rates exceeds the given threshold.

Figure 4. Accuracy, bias score, hit POD, success ratio and Heidke skill score (from top to bottom) against precipitation measurements for the period 1995-1997 (left HOAPS, right ERA-Interim) as a function of a lower threshold applied to the measurements.

Figure 5. Accuracy, bias score, POD, success ratio and Heidke skill score (from top to bottom) against precipitation measurements for the period 2005-2008 (left HOAPS, right ERA-Interim) as a function of a lower threshold applied to the measurements.

Section 5. The discussion on the thresholds (Figs 4, 5) should include the analysis of the number of "wet" events in the database, that are expected to decrease rapidly with increasing thresholds, since rainrate should be distributed as a power-low. This is the reason I suggest using equitable skill scores instead of accuracy.

Table 1 has been added giving the numbers of wet events, see above.

Section 6. The second part of this section (below line 31 on page 12) is a review of literature results and should go in the Introduction.

Went into the introduction, please see above.

Section “Data availability”. I suggest to cancel this unnumbered section and to include data providers in the Acknowledgement section.

Is now included in the acknowledgements