Reviewer # 2

The paper describes a method for estimating hourly rainfall over Southern Africa, based on a neural network approach, using MSG SEVIRI observations for the estimation and rain gauges data as ground truth. The results are compared to those obtained from IMERG of the Global Precipitation Measurement mission. The paper is interesting because it addresses an important and complex issue, as is the estimate of the surface precipitation in a region (the African continent) with sparse rain gauge and radar networks. I would like to recommend that this paper could be published after major/minor revisions to address the following comments.

Response

Thank you very much for taking your time on our manuscript and your helpful comments! In the following we would like to outline our response (green color) to your concerns (red color) as well as the subsequent changes that we made for the final version of the manuscript (green, italic).

Major revisions:

1 – The description of some important aspects of the study is often done in a concise, not sufficiently complete and precise way to allow a direct and complete understanding. This fact is partly due to the use of some too general references (e.g. a conference (P2 L6 : IPWG, 2016) or books (P6, L13 : Venables and Ripley, 2002) or (P6, L17 : Kuhn and Johnson, 2013)), where more precise/accurate references (the paper in the conference or the section/pages in the books) would facilitate the understanding of the specific topics. In part it is due to the use of references that seem irrelevant/inconsistent with the text (P5, L8-9 : xxl technology .... OpenCL acceleration (see https://github.com/umr-dbs/xxl)). In part it is due to the use of specialized terms generally difficult to understand/interpret (P6, L15 : stratified 10-fold cross-validation). More attention to the aspects mentioned and a clearer description of the different topics would make it easier to read the text and would better highlight the most innovative aspects of the study.

Response

We changed/added the following information towards a comprehensive description of the applied methods:

1) References: IPWG refers to a website which is THE reference to compare different rainfall retrievals for South Africa (see also comment from Reviewer 1) and must be mentioned here. Venables and Ripley, 2002 refers to the software implementation being used. The nnet package is supposed to be cited in this way (see https://cran.r-project.org/web/packages/nnet/citation.html), though we agree it's a very general reference. We therefore added the direct reference to the software package:
Ripley, B. and Venables, W.: nnet: Feed-Forward Neural Networks and Multinomial Log-Linear Models, http://CRAN.R-project.org/package=nnet, r package version 7.3-12, 2016.

For Kuhn and Johnson we adapted the chapter and pages.

2) According to your suggestions we expanded the description of the input variables being used (see comment 2), the preprocessing of the satellite data, the architecture of the neural network (see comment 2ii) and the cross validation.

About the preprocessing of the data: *“MSG SEVIRI Level 1.5 data (EUMETSAT 2010) were preprocessed to radiance values according to EUMETSAT (2012a) and BBT values according to EUMETSAT (2012b) using a processing scheme based on a custom raster processing extension of the eXtensible and fleXible Java library (see https://github.com/umr-dbs/xxl) which enables parallel raster processing on CPUs and GPUs using OpenCL.”*

About stratified 10-fold cross validation: *“Thus, the training samples were randomly partitioned into 10 equally sized folds with respect to the distribution of the response variable (i.e., raining cloud pixels, rainfall rate). Thus, every fold is a subset (1/10) of the training samples and has the same distribution of the response variable as the total set of training samples. Models were then fitted by repeatedly leaving out one of the folds. The performance of a model was then determined by predicting on the held-back fold. The performance metrics from the hold-out iterations were averaged to the overall model performance for the respective set of tuning values. For the rainfall areas classification models, the distance to a "perfect model", based on Receiver Operating Characteristics (ROC) analysis (see cite{Meyer2016} for its application in rainfall retrievals) was used as decisive performance metric. For the rainfall quantities regression models, the Root Mean Square Error (RMSE) was used.”*

2- Since the neural network is a key point in the study, more clarification on its design and its architecture would be appropriate. The references to texts (e.g. P6, L17 : Kuhn and Johnson, 2013) or packages (P6, L13 : “nnet” package (Venables and Ripley, 2002); P6, L14 : “caret” package Wing et al (2016)) do not lead to a direct understanding of the actual network used. The following points should be clarified:

i) How the network input variables were selected (P5, L30 and P6, L1-2). The reference P6, L1 : Meyer et al. (submitted) is not available.

Response

Meyer et al. (submitted) was in review when this manuscript was submitted. It is now published in Remote Sensing Letters so we could include the correct reference:

Meyer, H.; Kühnlein, M.; Reudenbach, C. & Nauss, T.: Revealing the potential of spectral and textural predictor variables in a neural network-based rainfall retrieval technique. *Remote Sensing Letters*, **2017**, *8*, 647-656.

Concerning the choice of predictor variables: We agree that a paragraph describing the general idea of the relation between MSG channels and cloud properties in section 2.2.2 is missing. We therefore added the following information:

“The rainfall retrieval technique presented here works under the assumption that VIS, NIR and IR channels of MSG SEVIRI provide proxies for microphysical cloud properties, which are, in turn, related to rainfall. VIS and NIR channels have been shown to be related to cloud optical depth (Roebeling et al., 2006; Benas et al., 2017) and cloud water path (Kühnlein et al., 2014b) where the NIR channel is further related to cloud particle size (Roebeling et al., 2006). The IR channels have been shown to provide information about the cloud top temperature which was used as a proxy for cloud height (Hamann et al., 2014). The cloud droplet effective radius as well as liquid water path during night was approximated using IR channel differences (Merk et al., 2011; Kühnlein et al., 2014b).”

From a technical perspective, it is no problem to insert all available information, even though individual channels might only have minor relations with rainfall. We added a note on that issue in the method section:

“The function of the neural network is then to learn the relations between the spectral information and rainfall areas or rainfall quantities, respectively. In this context, a sophisticated pre-selection of input variables is not required, as the network is able to deal with correlated and even uninformative predictors unless their number is very high (Meyer et al., 2017), which was not the case in this study.”

---

ii) What is the network architecture (number of hidden levels and perceptrons) and how it has been designed. The text P6, l6-17 : The number of hidden units were tuned for each value ...., is not clear in this regard.

Response

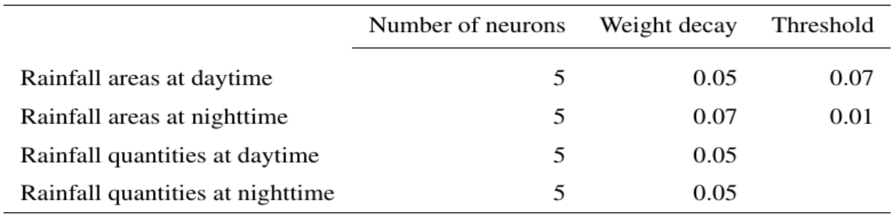
We made the architecture clear and improved the description of the hyperparameters that required tuning.

“A single-hidden-layer feed-forward neural network was applied as machine learning algorithm. The spectral channels of MSG SEVIRI as well as the channel differences served as input nodes (predictor variables). The neural network was then applied to learn the relations between these spectral information and rainfall areas or rainfall quantities, respectively. In this context, a sophisticated pre-selection of input variables is not required, as the network is able to deal with correlated and even uninformative predictors unless their number is very high (Meyer et al., 2017), which was not the case in this study. For the technical realisation, all steps of model training were performed using the R environment for statistical computing (R Core Team, 2016). The neural network implementation from the "nnet" package (Venables and Ripley, 2002; Ripley and Venables, 2016) in R was used in conjunction with the "caret" package (Kuhn, 2016) that provides enhanced functionalities for model training, estimation and validation.“

“Neural networks require two hyperparameters to be tuned to avoid under- or overfitting of the data: the number of neurons in the hidden layer, as well as the weight decay. The neurons in the hidden layer represent nonlinear combinations of the input data and their number influences the performance of the model (Panchal et al., 2011). Weight decay penalizes large weights and controls the generalisation of the outcome (Krogh and Hertz, 1992). “

The actual number of neurons in the hidden layer results from the model tuning. We added a table showing the final model settings.

“Optimal hyperparameters for the individual models revealed during the tuning study and applied in the final model fitting.”



	Number of neurons	Weight decay	Threshold
Rainfall areas at daytime	5	0.05	0.07
Rainfall areas at nighttime	5	0.07	0.01
Rainfall quantities at daytime	5	0.05	
Rainfall quantities at nighttime	5	0.05	

---

iii) What is the training procedure used in the study. Section 2.3.3 does not appear clear on this subject both for the language and the references provided (see point 1 above) and because the cited paper Meyer et al. 2016 does not provide more details about this procedure (apart from the threshold tuning methodology).

Response

Meyer et al. 2016 was included as a reference for the threshold tuning. We now made clear that the final step is fitting the model to all training data using the optimal set of hyperparameters:

“The optimal values for the hyperparameters that were revealed in the tuning study (Tab. 1) were adopted for the final model fitting. In this step, the model is fit to all training data using the optimal hyperparameters.”

We also included a short paragraph on the spatial estimations (see also minor comment 4):

“2.3.4 Spatial estimations of rainfall

Final models were applied to all hourly MSG SEVIRI scenes from 2010-2014 for the Southern Africa extent *to obtain spatio-temporal estimates of rainfall. Therefore, the clouded areas of a scene were first classified into rainy or not rainy using the respective model. The rainfall quantities were then estimated for the estimated rainfall areas. To ensure consistency within one scene, the choice of the model being applied (either the daytime or nighttime model) was made according to the mean solar zenith angle of the respective scene. If the mean solar zenith angle was <70°, rainfall for the entire scene was estimated using the daytime model. For scenes with a mean solar zenith angle > 70°, the nighttime model was applied."*

Thus, our description of the training procedure now contains the selection of predictor and response variables as well as their preprocessing, the network architecture, the model tuning and cross validation approach, the final fit of the models and the validation as well as spatial model estimations Please let us know if you still miss information.

---

3 - The use of rain gauges as ground truth requires checks on the data quality. In the paper some aspects of this issue should be developed, e.g check on no-data or no-rain, consistency between data from different networks. Is the retrieval quality depending on the rain gauges density?

Response

We totally agree that the quality of the ground truth data is an important issue! Yes, the data distinguish between zero and "no data" otherwise it would not be possible to train a model for rainy and non rainy clouds. We now included information about the pre-processing of the data:

"The data passed general provider-dependent quality checks before it was used in this study. This includes for example filtering of data beyond common data ranges, or situational checks for consistency with related parameters (e.g. air humidity) by SASSCAL. The data was then included in an on-demand processing database system (Wöllauer et al. 2015) where it was automatically cross-checked for reliability by filtering values < 0 and > 500 mm of rainfall per hour. All station data that provided sub-hourly information was aggregated to a temporal resolution of 1 hour within the database."

Unfortunately, we don't see a way to ensure a consistency between data from different networks: First of all, we can't analyze weather inconsistencies exist or not because this would require having stations from the different networks at exactly the same location which we don't have. Second, even if we had, how to correct for inconsistencies? There would probably be ways in areas where a high number of stations are available so that systematic inconsistencies in the data could be studied in a robust way. But this is not possible for our study. Our study relies on (comparably) sparse data from different sensors and provider-specific ways to process data. We don't see a way to check and/or correct for inconsistencies. However, we now accounted for this point in the discussion:

"Therefore, different sensor and data provider dependent calibration techniques, gaps in the time series of the data as well as the general problems associated with rain gauge measurements might lead to inconsistencies and uncertainties. However, no reliable alternatives are available and rain gauge measurements are still considered as most reliable source of rainfall data. "

It's further difficult to test if the retrieval quality depends on the gauge density. An obvious test would be to correlate the station density with the performances. However, that wouldn't lead to meaningful results since areas with many stations allow for a robust signal while areas with low density of stations don't because there are simply too few data.

However, we assume that different densities don't affect the model performance for the following reasons: We assume that the density would certainly have an effect if we trained on few scenes only, because then the current conditions of the areas with high station densities are highlighted. However since we trained on a long period, the model saw different conditions from dry to wet to learn from. Therefore it should be able to learn the relations between cloud properties and rainfall without the risk of over-fitting towards areas with high station densities.

4 – Figures 3 and 4 show the box plots concerning the POD, FAR, PDF, HSS, RMSE and rho evaluated considering the whole set of data; It would be more effective to evaluate these indexes considering different ranges of precipitation values (e.g. 0-25 mm, 25-50 mm etc).

Response

Thank you for this comment! We now compared POD for different measured rainfall quantities and we compared FAR for different predicted rainfall quantities. The results give valuable information about the performance: high rainfall rates could be very well recognized as rainy clouds by the model. Though the model overestimated rainfall (high FAR), the predicted rainfall quantities for these false alarms were comparably low. We accounted for this in the results:

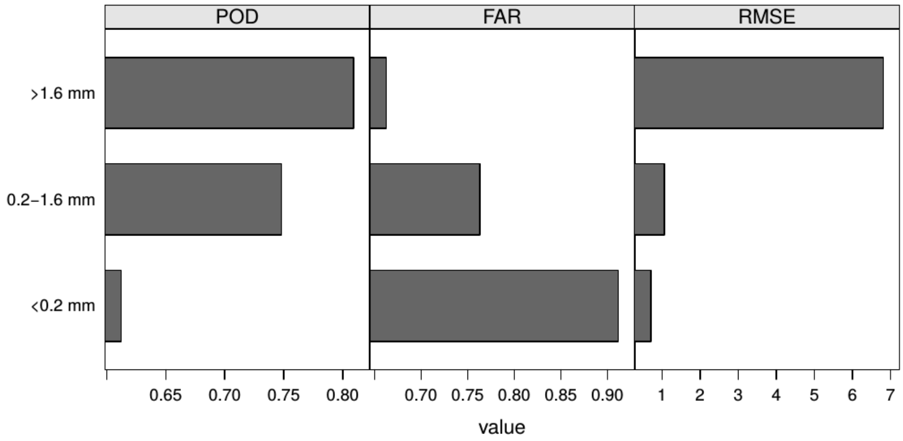
"The POD was highest for high measured rainfall quantities and decreased for lower rainfall quantities (Fig. 4). FAR was highest for low predicted rainfall quantities and decreased for higher predicted quantities. [...] Especially data points with low or medium measured rainfall could be estimated with low RMSE (Fig. 4).

And in the discussion:

"The strength of the retrieval in terms of rainfall areas classification was a high POD for heavy rainfall events. The rainfall quantities for the heavy rainfall events were, however, underestimated in most cases. The major problem of the model was the overestimation of rainfall events leading to an overestimation of rainfall quantities. However, false alarms in the retrieval were generally predicted with low rainfall quantities."

We presented the figure as barplot since a similar boxplot representation is not possible in this context: The boxplots base on the POD/FAR/etc on a scene basis, thus the data points of the boxplot could not be assigned to a unique rainfall quantity class.

POD can't be calculated when no rainfall is measured and FAR/POFD can't be calculated when rainfall is measured/no rainfall is predicted. Therefore, we can't make sense of a HSS for different rainfall classes since it bases on POD and FAR. For that reason we only compared POD and FAR. We didn't follow your suggestion of a class-based comparison for the correlation coefficient. For the correlation we want to know the model's ability to distinguish between low and high rainfall. When we only consider small parts of the gradient we would lose too much information.



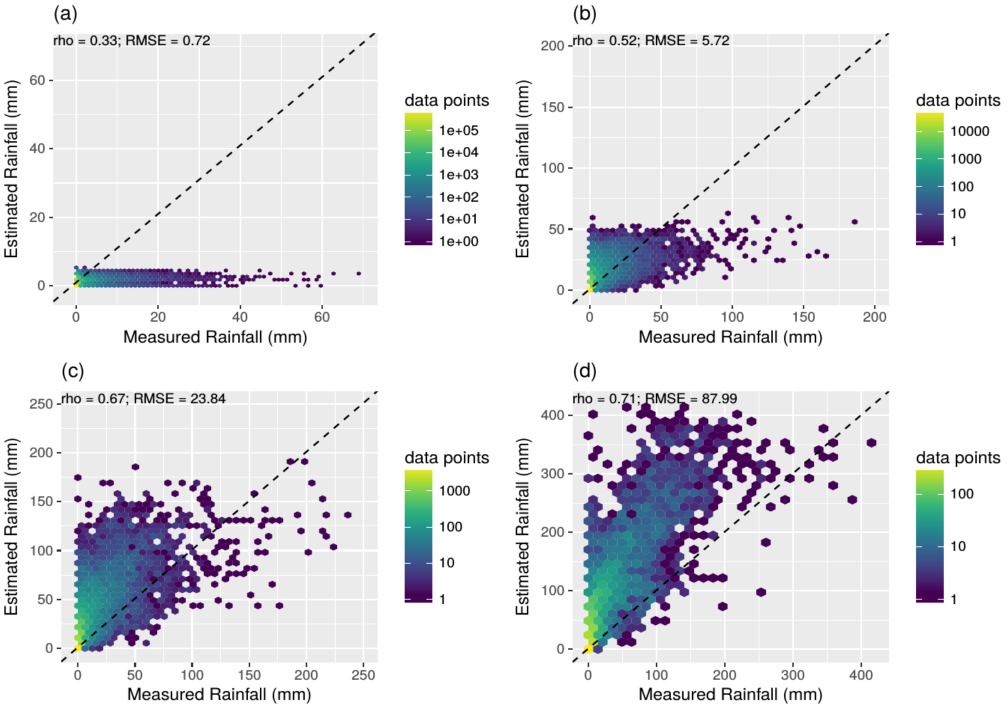
Minor revisions:
1 – The section 2.2.2 should be modified by introducing a short description of the ability of the Seviri channels to provide information on the state of the atmosphere and the ground. This is important to clarify the choices that led to the selection of the neural network inputs.

Response
See major comment 2i

---

2 – The performance of the retrieval technique (P8, L5-6) shown in fig. 5 (P11) could be presented in a more complete way by inserting in the four panels the corresponding RMSE and mean bias values. In the figure the colour bar (data point density) should be added.

Response
We improved the figure by providing a clearer binning of the values and a comprehensive color scheme with a legend showing the amount of data points. We also added the RMSE in addition to rho.



---

3 – The reference to Smith et al. 2007 (P7, L9) can be updated with: Hou, A. Y., Kakar, R. K., Neeck, S., Azarbarzin, A. A., Kummerow, C. D., Kojima, M., Oki, R., Nakamura, K., and Iguchi, T.: The global precipitation measurement mission, B. Am. Meteorol. Soc., 95, 701-722, doi:10.1175/BAMS-D-13-00164.1, 2014.

Response
done

---

4 – P6, L3 Please explain the criteria that has allowed to split the database into day and night.

Response

We now added the information about how the data were split into day and night:

“Since the VIS and NIR channels of MSG are not available during the nighttime, the dataset was split into a daytime dataset (data points with a solar zenith angle < 70°) and a nighttime dataset (data points with a solar zenith angle > 70°)”

We also added a section of how the spatial estimation of rainfall were created because in this case, the mean solar zenith angle of the entire scene was decisive for the choice of the model:

“Final models were applied to all hourly MSG SEVIRI scenes from 2010-2014 for the Southern Africa extent to obtain spatio-temporal estimates of rainfall. Therefore, the clouded areas of a scene were first classified into rainy or not rainy using the respective model. The rainfall quantities were then estimated for the estimated rainfall areas. To ensure consistency within one scene, the choice of the model being applied (either the daytime or nighttime model) was made according to the mean solar zenith angle of the respective scene. If the mean solar zenith angle was <70°, rainfall for the entire scene was estimated using the daytime model. For scenes with a mean solar zenith angle > 70°, the nighttime model was applied.”

---

5 – The paper contains a few typos that need to be corrected.

Response

We checked the manuscript again for typos.
