Atmos. Meas. Tech. Discuss., 8, C2678–C2685, 2015
www.atmos-meas-tech-discuss.net/8/C2678/2015/

Atmospheric Measurement Techniques Discussions

Interactive comment on "Retrieval algorithm for rainfall mapping from microwave links in a cellular communication network" *by* A. Overeem et al.

M. Heistermann (Referee)

heisterm@uni-potsdam.de

Received and published: 25 August 2015

Subject and scope

In their manuscript "Retrieval algorithm for rainfall mapping from microwave links in a cellular communication network", A. Overeem and colleagues present details of a retrieval algorithm that has been originally published by Overeem et al. (2013) in PNAS. Making the idea of rainfall retrieval operationally viable is certainly subject to intense research and the subject fits the scope of AMT well.

Innovation

The authors do not hide the fact that the actual retrieval algorithm had already benn published in a different paper. The intention of the present manuscript is rather to present details about the algorithm and, alongside, an implementation of the algorithm in the programming language R. As a result, the manuscript does not present any original research. The content would be ideally suited for a format generally known as "technical note" which is, however, not available at many journals, including AMT. So regarding the lack of other appropriate formats at AMT, I think it is justified to still submit this as a research article.

The innovation of the present submission can be seen in the attempt to provide an open source implementation of the algorithm in order to stimulate further development and applications in this field of research. Or as the authors put it, "the purpose of this paper is to provide a detailed description of a slightly modified version of the algorithm of Overeem et al. (2013)" (p. 8193, ll. 18-19), and "to promote the application of rainfall monitoring using microwave links in poorly gauged regions around the world" (p. 8211, ll. 8-10).

I certainly endorse this motivation, and I would like to thank the authors for making that effort. Let us be frank: In the scientific sector, publishing a paper is still more rewarding than publishing a software code, although publishing a software code might advance scientific progress just as much. I also think that it makes sense to accompany the publication of an open source code by a related article in a scientific journal. Admittedly, the article of Overeem et al. (2013) would have been a good opportunity to publish the code alongside. Still, I think of it as appropriate to publish a paper in AMT as both entry and reference point of an open algorithm, particularly since no other group has made their algorithms openly available, yet (at least to my knowledge).

Apart from that, I have to say that I have some major concerns about the manuscript and the code which I would like to point out in the following.

Major concerns

My main concerns are about the design of the software code, the presentation of the software code and the balance between technical details and scientific discourse in the manuscript itself. The algorithm is implemented in R which is good since R is also an open source environment. The code runs fine at least on my machine which is also good. So the authors achieved their main objective to allow the community to run the code together with the set of sample data. However, I think the presentation and the design of the code are not up to the standards of scientific software development and this will hamper making progress as a community:

1. I think it is not appropriate to publish the code as a supplement. This way, it is difficult for the authors or anyone else to improve the code and make these improvements available to the community. It is state-of-the-art (and also easy) to use public file hosting services such as GitHub or Bitbucket which also offer version control and ample tools for collaborative development and public code review.

2. The software code is basically a collection of scripts without any modularity and a lot of stuff hard coded. Just as an example, the script "WetDryClassification LinkApproach.R" contains a lot of stuff apart from the actual wet-dry classification such as the determination of reference signal levels and corrected received powers as well as outlier filtering. It would be much more suitable to design a library (or a package) with different modules/functions, a clear application programming interface (API) and examples on how to combine the API functions in processing workflows such as the one presented in the manuscript. The code should be designed in a way that modules are reusable (so it is up to the user how to combine modules), and in a way that the various parameters of the approach are clearly assigned as function arguments (which then might have default values using the authors suggestions). This way, it is easier to add new functions e.g. for wet-dry

classification or spatial interpolation which could then be combined with already existing functions. Functions should be able to exchange data in memory instead of using file I/O as a detour that compromises computational performance. Overall, I suggest redesigning the code as an R-package which would also address issues of documentation, distribution, and dependency management.

3. The readability of the manuscript suffers very much from the technical details (and also from a lack of conciseness). I think most of these details should be part of the software documentation (the API reference is an intrinsic part of an R package and can be enhanced by external sources such as web pages containing e.g. theoretical underpinning, technical guidance for system setup, and tutorials etc. – such pages can also be easily developed and hosted via e.g. GitHub).

Removing many of the technical details from the manuscript will increase readability and should make it much shorter. In particular, the authors should find a smart and elegant way to condense and reorganise chapter 3. I think of it as an opportunity to make the paper brief, crisp and informative, but still, the question remains what should actually remain in the manuscript. Surely, it is the authors to answer this, but I think such an article could really be an opportunity to familiarize non-experts (such as myself) with this methodology and at the same time provide an in-depth discussion of the limitations and challenges that should be addressed in future (community) efforts in order to make this approach more widely applicable and more compatible with other approaches (other sampling strategies, other interpolation methods, other rainfall observation methods etc.). The authors could briefly present a typical workflow just as they did, but with graphical material more to the point, showing ideal behaviour but also highlighting typical cases of failure. Finally, they should provide a more concise discussion of how the specific aspects of the implementation limit transferability (such as the data situation in the Netherlands, the daily accumulation interval, the interpolation approach etc.).

I am very much aware that all these requirements imply a lot of additional efforts. However, I am convinced that Open Source Software can only effectively and efficiently support scientific progress if the developers are willing to take the extra effort.

Overall evaluation

I think that this submission has the potential to actually add value to the original publication of Overeem et al. (2013) and to provide a valuable service to the atmospheric sciences community. In order to make this happen, though, both the manuscript and the code would require a major revision.

Specific comments

p. 8193, ll. 11 ff.: I think that "countrywide“ is not a helpful scale concept. Basically, the scale of such rainfall maps in terms of extent is not limited to the upper end. The question would rather be about the effective resolution that can typically be achieved.

p. 8193, ll. 13 ff.: “Rainfall estimation employing the cellular communication network was shown to work successfully for a relatively large data set in space and time.” What do you actually mean? I’d rather drop the sentence...

p. 8194, ll. 9-10: “The transmit power may be assumed constant.” – May or is?

p. 8194, l. 12: “This may lead to quantisation errors.” – May or will? Could you please elaborate on the extent of these errors?

p. 8194, ll. 21-22: “on average 2309 links and 1440 link paths over all 96 time intervals of 15 min” – the difference between link and link path should be explained to the reader

p. 8195, ll. 23 ff.: “The code will also...” – shouldn’t this be separated by another paragraph?

p. 8196, ll. 2-3: “Though many missing data may lead to rainfall intensities not being calculated.” – Why? Under which conditions?

Fig. 2: I don’t think that a figure is needed here...at least it does not really add much to the main text.

p. 8197, ll. 3-6: “It is important that the value of the exponent b in Fig. 5 (right) is close to 1, which is the case for a range of frequencies. Here, only links with microwave frequency from 12.5–40.5GHz are selected. The chosen frequencies can be altered in the script.” – wouldn’t it make sense that the user can pass adequate exponents instead of limiting to frequencies with an exponent of approx. 1?

p. 8197, step 8: Why is it mandatorily required to produce an intermediate file output instead of passing the data in memory? Usually, this will speed up performance. Or is this only for demonstration?

Section 3.2 appears not well structured and organized.

p. 8199, step 2: What is the implication of that 15 km radius in case of small scale convective events? There is a follow up on p. 8200, ll. 10 ff., but it does not discuss all the implications. It think it is hard to justify a statement such as “The 15 km radius is of the order of the decorrelation distance of rainfall in the Netherlands” (p. 8200, ll. 12-13) since the decorrelation distance depends on the type of rainfall event.

p. 8203, ll. 10-12: “The values of a and b used in this study are derived from measured drop size distributions (Leijnse et al., 2008), and are shown as functions of link frequency in Fig. 5.” – why not add this as a function or look-up table to the code?

p. 8205, ll. 9-15: The present discussion of implications for using horizontally polarized signals is not concise: It sounds like “a and b are valid only for vertically polarized signal, but they can also be used for horizontally polarized signals, but not really.” I would replace that statement rather by something like “Note that the coefficients a and b as shown in Fig 5. are only valid for vertically polarised signals.“

p. 8207, ll. 4-6: “Representativeness errors may play a large role given the local convective character of this rainfall event [...]” – Could this be tested by revising the

assumptions made on the distribution of R(s) based on the radar observations? Could you show a figure of the radar rainfall along the link?

p. 8207, ll. 19-20: “Kriging is well-suited for dealing with heterogeneously distributed data locations.” In comparison to what? The advantage of Kriging is the explicit consideration of spatial autocorrelation of a random variable as exhibited by observations. I am aware that the approach to quantify the variogram parameters as a function of season and accumulation interval has been successfully published by Van de Beek et al. (2012), however, I would like to see evidence that this approach actually allows for a better prediction of rainfall at unsampled locations (i.e. interpolation) in comparison to simpler benchmark methods (e.g. constant variogram, IDW, ..). Van de Beek et al. do not provide that evidence. I really have doubts as to the efficiency of this approach (both in terms of interpolation quality and computational performance). I am fully aware that this is beyond the scope of this study. However, this is another reason for breaking down the procedure into a more modular approach in which users might decide to use a different interpolation method. In fact, other interpolation methods have been suggested that explicitly take into account the path-averaged nature of the measurement instead of simply assuming the measurement to be representative for a point in space (which it is not!).

p. 8210, ll. 19-20: “Note that some large areas do not have link data, as shown in Fig. 7 (left panel).” I think I do not yet understand how missing data is handled in the interpolation. Are the “white” links and points in Fig. 7 considered missing data or zero rainfall?

Fig. 8: Colorbar is inconsistent in case value < 0.1 mm are not to be shown.

The “Conclusions” are not really conclusions, but rather a brief summary and a (quite unstructured) collection of challenges that do mostly not relate well to the previous sections. In case the authors intend the initiation of a collaborative effort, I would suggest that this section should provide a well structured discussion of current limitations and

challenges that need to be addressed in order to achieve a broader application of the approach in other environments.

Interactive comment on Atmos. Meas. Tech. Discuss., 8, 8191, 2015.