Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2019-50-RC3, 2019

Hydrology and Earth System Sciences
Open Access
Discussions

Interactive comment on "Using R in hydrology: a review of recent developments and future directions" *by* Louise J. Slater et al.

Michael Stoelzle (Referee)

michael.stoelzle@hydro.uni-freiburg.de

Received and published: 2 April 2019

First of all, I really support the idea of the paper and I think the presented review is fully in the scope of HESS. I like how the relationship between hydrology and R is presented and the authors gave a broad overview of historical and recent developments in the R-Hydro community and sketch a interesting roadmap for the future of R in hydrology. The paper is from my perspective also a good example for a joint effort from young scientists in our hydrological community. However, since the first two reviews are full of praise I try to be a little bit more critical putting together some thoughts that hopefully improve the paper. My major "overall" suggestion is to revise the paper going a little bit away from just having a list of available packages or presenting potential possibilities in the R-Hydro universe towards a) more "best practices" for certain challenges and b)

more comparisons of similar packages. However, I absolutely recommend to publish this paper in HESS. See more specific comments below.

Best regards, Michael

Major comments

#### The authors give a lot of technical information about R in different section in the paper (e.g. GNU S background in sect 2.3, R script workflow in Sect. 2.1, R output format in the Introduction). As all the parts are really important to learn about the fundamental principles of R for me it would be more helpful to have the technical and historical information about R in one separate section. This might be also helpful for readers who are new to R coming from other programming languages.

#### I miss more discussion about different philosophies how to program in R. From my perspective there is, for example, often the question whether you are doing your visualizations with ggplot2 or in base-R or doing data wrangling with dplyr or in a complete different "style" with data.table. I don't know if there are comparable examples for hydrological packages (doing more or less the same thing but with different techniques / packages). The paper could, however, be improved by presenting opposed approaches to solve problems in R. For example, data.table is known to be very fast in data wrangling but on the other side the dplyr package (and piping %>%) is known to support highly readable code. Here the trade-off between performance and communication could be discussed (i.e. Sect 4.3 and Sect. 2.1/2.2). My experience from teaching R in hydrology is that students really like the concept of piping (%>% read as "and then do that") and can do the step from base-R to tidyverse in a couple of sessions. Tidyverse and Pipes could also be mentioned in Sect. 5.3 as a good possibility to teach "highly readable code" in the classroom.

#### Regarding Sect 2.2 it would be helpful to draft a way forward how developers of R-Hydro packages can develop better R packages with a clear description of input data types, a comprehensive documentation on package usage. For example, tidyverse is

set of packages that work well together as all packages use the same data representations and API design. Is this also a valuable approach for, let's say, developer of R packages that include hydrological models? For example, is it possible to use the data retrieval packages from Table 1 in a coherent way or is the syntax between all packages different? Is there an effort to combine data retrieval packages from different regions with a consistent R syntax/usage? Using different hydrological R packages with the same syntax might be of great value for the community. A short discussion about this issue can draft a way forward in the R Hydrology community (as mentioned in Sect 4.1 where the need of good vignettes in R is highlighted).

#### The introduction of section 3 and also Fig. 3 are really nice and well-conceived. I like the idea to split the data analysis into different parts and subsections in the paper (Sect. 3.x).

#### In general, the paper is full of valuable package, links, and corresponding citations. However, from my perspective often only the package names are given and more detailed information how the package is working or why it is especially useful/important for the hydrological community is missing. Of course, it is not possible to describe all the functionality of all the packages in detail but the authors could pick 2-3 cases where they really compare the advantages and disadvantages of using different packages for the same purpose / challenge. To be honest (and I also think that the author team is aware of that), you can relatively easy check on the internet if there is a package for a certain task and what are the technical details or data requirements of a single package. So, the great list of available packages cannot be the main message of this compelling paper! Comparing different packages among each other could really be of great value for the hydrological community and this paper is a good place to do that. A good example is Sect 3.4 describing spatial data analysis and visualization in R. In this section a lot of important R packages are mentioned but the reader gets less information which package is better or has the same functionality as other packages etc. In Sect 3.5 the information about snow functionality in the hydrological models is a good

example how this can be done.

#### For me a major deficiency of the paper is that the topic “colours” is completely missing. Hydrology is a multi-facetted discipline and hydrologists doing a lot of visualizations for posters, presentations and papers. Hydrology is also often based on multi-dimensional analyses with different variables in space and time. To visualize hydrological processes and hydrological change colour is often the first choice to compare data sets and to investigate relationships. As we all know, a sophisticated colour choice is often a huge challenge. I encourage the authors to implement a “colour section” where a short discussion about, for example, the basics of appropriate colour gradients for graphs and maps is given (one-colour gradient, two-colours gradient, discrete vs. continuous variables). In large-scale hydrology “rainbow” colour gradients are still state-of-the-art (to generate fancy and colorful world maps) but a lot of papers and blogpost have taught us that the “rainbow” isn't perceptual uniform and not colourblind-safe.

Minor comments

P2L20 Give 1-2 examples of hydrological models coded in R.

P2L33-35 It might be helpful for the reader to compare the fundamental principles of R against other programming languages (e.g C, Java) to better describe the archíture of R (e.g. in terms of object-orientantion, complilation of code,

P3L4 Are there any reasons for the large number of updates in R-Hydro packages (related to Fig.1)?

P4L11 The platform Stackoverflow (from Sect. 2.5) can also be mentioned here as there often also information on relevant R packages are given.

P5L31 What is meant with the “standardized format”? Here it might be also valuable to add packages like roxygen2 that help to develop, create and document R packages.

P6L27 It will be easy to find some justification for this statement (e.g. number of R topics on stackoverflow.com compared to other programming languages).

P6L3-14 The paper would be improved by adding various possibilities to share and publish code during and after paper publication (e.g. github repository, gists or other tools and platforms). Is there a possibility to share R code with a DOI?

Sect 2.3 For me this section contributes less to the strength of the paper – what is the key message here? The specific statistical packages that are relevant for R-Hydro are given in Sect 3.6 (as mentioned). The link to CRAN is perhaps too unspecific here. Yes, R provides a vast number of statistical tools for hydrologists but then you have to explain in more detail why and what advantages R here has to offer (compared to other languages?).

Sect 2.5 What are the rotating topics and why have they been chosen? The picture might be nice to see but the space could instead better be used to highlight the EHGU SC topics from the last years with a short information why the topics were of great interest for the community. It might be also important to highlight the possibility to raise issues for certain R packages on github.com to foster an online-documented discussion with the package developers (i.e. interaction in the community). At least a cross reference to Sect. 5.4 could be made here.

Table 1 I am not sure on which basis you have choose the 11 packages? Are they the most important ones? Perhaps a thematic order would improve the table (sort by data type, spatial extend, multiple proposes or not,. . .)

P11L5-9 Consider to mention the rio package here as it is really powerful. The tidyverse package is cited as a package that reads multiple file formats – I think this is not the main purpose of the tidyverse package.

P11L10 Which netcdf package is the best? Or let’s say, are there more information which package to use for a specific application.

P12L15 For all ggplot2 users it is really important to mention that the ggplot2 3.0 version offers support to visualize sf objects directly with a specific geom (geom_sf). So, if one is familiar with ggplot2 and want to visualize spatial data (raster, shapefiles etc.) this is a really easy way to do that? As the ggplot2 community is growing and growing this might be a really important information for some readers of the paper.

Table 2 Should be included in Sect 3.5

Sect. 3.6 Package “skimr” might be valuable for this section.

P15L10 I agree, one strength of ggplot2 and faceting is the ability to generate multiple views. This is especially important to avoid overplotting or – as the authors mentioned – to split a data set in different plots using meta information of this data set (i.e. season). If you have a nice example for faceting here would be a good place to show this principle for the readers.

P19L17 Is it true that publishing a R package on CRAN is free of quality checks? However, a short comment on the differences during package publication (CRAN vs. GitHub) might be helpful here.

P19L21-26 I agree but the paragraph could be more specific on how we should develop R packages. Is there a good article or blogpost describing the progress of publishing a R package? What are the requirements to generate good tutorials or readme files? Might it be useful if journals like HESS recommend to publish paper code on github?

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2019-50, 2019.