Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC1
https://doi.org/10.5194/gmd-2021-139-RC1, 2021

Comment on gmd-2021-139

Anonymous Referee #1

Referee comment on "AI4Water v1.0: An open source python package for modeling hydrological time series using data-driven methods" by Ather Abbas et al., Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2021-139-RC1, 2021

The manuscript describes the new python package AI4Water, intended as a modelling tool for hydrological predictions. It incorporates the basic steps of data-driven analysis and modelling - preprocessing, choosing one or several modelling approaches, post-processing including error analysis and visualization - and makes extensive use of existing python librariies. The focus is strongly on machine learning approaches and the package embraces many of the currently discussed and newly developed algorithms.

The paper is easy to read and, while surely leaving many details out and thus not a manual for the ambitious user, summarizes the fundamental steps in the modelling process in a concise manner. The authors do not develop their own routines or approaches, but the collection of state-of-the-art modelling utilities and approaches is impressive.

it would be nice if the authors could address three major issues relevant for anybody intending to analyze their own data:

1. It is not obvious how users might get their own data (time series) into the package. The access to two existing databases is implemented, CAMELS and LamaH, and the authors rightly remark that the different data formats, conventions etc. are an obstacle slowing down the analysis process. How generic are the input options to accomodate own data in different formats (text of Excel files, spatially extended time series in netcdf files, and the like)?

2. Expandibility: it might well be that for the specific data at hand or the particular user, other methods then the ones already provided might be desirable. An example would be gap-filling, but also others. The part of the manuscript describing that (chapter 3) is very vague and general, please be more specific.

3. Interpretation: it would be wonderful if the package could produce a comprehensive interpretation of the results achieved with the chosen model approaches. Interpretation also implies making connections to existing hydrologcal knowledge (process understanding) as well as local conditions (metadata) available for the site, its pecularities. However, in this context, interpretation is merely a visualization of the model architecture (e.g. the weights in the case of NNs). A more modest phrasing, e.g. "Model

Visualization" instead of "Interpret" as the class name, seems to be more appropriate.

The language quality is good to very good with very few typos etc. Some specific comments and corrections:

l. 78: "time series errors": do you rather mean performance measures rather than errors?

l. 112: "Fig. 3 shows examples of the three configuration files" -> "Fig. 3 shows three examples for configuration files"

l. 118: "obtain large and diverse data" - no, this cannot be guaranteed, and the hope is that modelling is also possible when there is only a limited amount of data from a given catchment, as is often the case!

l. 139: what is the difference between "scaling" and "transforming the data onto a different scale" ?

l. 142: EMD is a decomposition, not a transformation method, much like PCA. Of course, using IMFs as input rather than the original variables does change the model setup and has an impact on performance etc. as is correctly stated further down.

l. 146 "were" -> "are"

l. 153 "(McKinney, 2011) scikit" -> "(McKinney, 2011), 2) scikit"

ch. 2.4 Missing labels: it should be mentioned that this refers to a classification task only, not to regression. Also, what is the difference between "exclude examples" (l. 170) and "skip these examples" (l. 173)?

l. 179 "later" -> "latter"

l. 198 "time series weather data" -> "time series of weather data"

l. 205 how does the user provide HRUs / soil types, land use classes etc. ? Through shapefiles if available?

l. 212 "large" -> "many"

l. 272 "all possible results" -> "many different results"

l. 294 "cannot be defined" - why not?

l. 335 delete the first occurence of "training" in this line

l. 346 Christine, 2014 does not seem to be in the reference list

If these comments are taken into account by the authors, the paper should be published by GMD.