Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2020-87-RC2, 2020

Geoscientific Model Development Discussions

Interactive comment on "Explainable AI for Knowledge Acquisition in Hydrochemical Time Series V1.0.0" *by* Michael C. Thrun et al.

Anonymous Referee #2

Received and published: 1 July 2020

General Comments: This manuscript takes watershed time series data for nitrate concentration (NO3), electrical conductance (EC), and 12 other typical hydrology-related parameters from a previously published data set (reported on by Aubert et al., 2016) and attempts to understand conditions which lead to high and low nitrate and EC levels through the application of an AI procedure. The concept is straightforward and the value of deriving insight from large data sets is critical. The current manuscript suffers from poor description of methods and processes and could be improved through reorganization and removal of notable errors in grammar and writing. In order to follow the paper, I frequently had to go back and forth between sections and had to read the entire previous publication about the data themselves (Aubrey et al., 2016) in order to evaluate the current manuscript. What is perhaps what I find most troubling with the

current work is that the key findings of the study are not new and are concepts can be found in undergraduate level courses covering on water quality in natural systems.

Major/Global Comments:

1) The description of the data used in the study is poor. From the introduction, it is stated that the data come from the Swingbach catchment in Germany, and from the methods section that there are 32,196 data points for 14 variables. No discussion about how many sites there, where the data were collected, or if the data even overlap temporally. Instead of using standard terms for things like groundwater level or stream temperature, they are replaced with cryptic codes (e.g., groundwater level = GW13, GW125, or GW 132) that are make it difficult to follow the results. It took reading the previous paper to produce even a moderate understanding of what the variables are and why there appear to be duplicates. In terms of understanding underlying hydrological process, being able to discern that, for instance, that the three groundwater level measurements are for a lowland, hill slope, and riparian zone better tremendously.

2) Despite having a reasonable background, in terms of both the data and the methods applied in this paper, the current text provides a poor explanation of work that is spread out throughout the text. Take for instance a very simple examination of Figure 2, which plots probability distribution estimators of "distance" for the data, as well as Gaussian end-member populations which fit the data. But what distance? There are 14 variables in the study, so is this a multivariate distance of some sort? The corresponding text simply states that, "The Hellinger distance measure is selected. . . " (line 93) but Hellinger distance measures distance between distributions. So which distributions is it measuring distances between? Going back to the paper by Auger et al. (2106), they observed a tri-modal distribution of NO3 concentrations. . .is that what this is?

In another example, we can examine how the cluster analysis is explained (note that the section title is misspelled as "Clsuter Analysis). It begins by focusing on the Pswarm method, where DataBots move data that are similar towards one another on a grid or

map, but this is described using only qualitative means, such as searching for the “most potent scent” or moving towards DataBots with the most similar features. What type of similarity metric(s) is employed here? Later in the same discussion, it is explained that clustering can either be focused on compact clusters or connected clusters and that the authors have decided to emphasize the former (no justification given). The following text (line 14) then goes on to describe how “the choice of this parameter can be evaluated. . .”. What parameter? Is this related to compact clusters or connected clusters?

The final issue about the methods is that the reader basically must go back and forth through the paper to follow what was done and the arguments behind it. In the methods section on data pre-processing, mirrored density plots are described as being employed but no details about what they are or why they are used is provided (they look similar to violin plots but turn out to be somewhat different). However, a fuller explanation is provided much later, in the final paragraph of the methods, instead of when they are first introduced.

3) This work relies on and cites a large number of packages in R. While there’s nothing fundamentally wrong with that, citing a package without describing the methods it relies on is not beneficial. My recommendation is that a new table be created which lists all the packages used and their citations. Then in text, simply state the package title and what principles or techniques the package uses.

4) The fundamental processes identified by this investigation are largely rote conclusions for scientists who study water quality in paired stream-aquifer systems. For instance, input of groundwater into a stream corresponds generally with higher temperature (due to geothermal contribution) and EC (due to longer residence time to allow for water-rock interaction). Similarly, high nitrate levels during dry days and lower stream temperature is due to a lack of dilution effect during rainfall, bur still primarily a surface water vs. groundwater contribution to streams (i.e., lower temperature). What is the value added by this analysis?

I have numerous specific comments and noted many typographical errors as well, but the feel that the items above need to be rectified before providing further feedback.

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2020-87, 2020.