Geoscientific
Model Development
Discussions

Open Access

EGU

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

Anonymous Referee #1

Received and published: 14 June 2020

General Comments:

Overall, I would say that this paper has an interesting concept but is hard to interpret the main contributions and is overall hard to read. For example, in Section 2 there needs to be more information about Figure 1 in this section (Material and Methods), the authors go straight into the steps (1-6) without a high-overview first. I was just confused of the paper structure without reading Fig. 1 first, which is hard to interpret without a better explanation. There are many arrows and connections that are not explained well, because of this, I don't really have a good understanding the main concepts and think the authors need to make things clearer and easier to interpret. I just would like to have a better understanding of how decision trees can be used for DBS and other methods (e.g. k-means) and believe if the authors can make this clearer then it would

improve the overall paper. For example, can you add a Figure/Algorithm that explains Step 3 to Steps 5/6 better?

Other Comments:

This lacks recent explainable frameworks for clustering, for example the paper "An Explainable Artificial Intelligence Model for Clustering Numerical Databases" is very similar and does a nice job at comparing other clustering approaches (k-means) too.

I also think the authors should either explain better why this is considered an AI system. I don't really understand this until Section 2.3 (cluster analysis) where they mention that DBS uses a mix of game-theory and neural networks. This information should made more clear in the Introduction.

Section 2.1: Where are this values collected at (Germany?)? There should be a figure of the location and points on a map, this would help with the interpretation of the topographic results. Do you need to state the variables that were removed from the analysis since they have high-overlap of similarity? Shouldn't the AI system do this for you?

Section 2.2: Why is a distance metric important for the algorithm, this isn't explained well. For example, I think you need to explain the DBS system (Section 3) before this section stating why this important for the algorithm.

Figure 1: This needs to be made clearer. This is hard to interpret and not sure what's going on.

Figure 2: This needs a legend.

Figure 4: How many clusters are here? There needs to be a legend.

Figure 5: This is hard to interpret. Where are the four clusters? Can you add these to the heat map?

Figure 6: Do you need this? Seems not really relevant to the paper or combine this

with Figure 4.

Figures 8/9: Seem like these can be merged (Fig 8 a/b).

Figure 10: Is this important? Seems like this can be excluded.

Labeling of the Figures are off, Figure 6 is mentioned in the Discussion after Figure 9.

Discussion: Is it possible to include other methods (k-means, DBSCAN) in this framework?