Hydrology and Earth System Sciences
Discussions

Hydrol. Earth Syst. Sci. Discuss., author comment AC1
https://doi.org/10.5194/hess-2021-621-AC1, 2022

Reply on RC1

Leah A. Jackson-Blake et al.

Author comment on "Seasonal forecasting of lake water quality and algal bloom risk using a continuous Gaussian Bayesian network" by Leah A. Jackson-Blake et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2021-621-AC1, 2022

Many thanks for a very useful review, we really appreciate your careful reading of the paper. You have made some very good comments, as well as pointing out some errors and areas where our descriptions were lacking.

In the attached Reviewer1_response.pdf file we provide comments in response to every point raised and suggestions for changes we would make to the text during revisions. Main points in this response include:

Response to major comments:

Firstly, it is great that you think that the data and code might be a useful tool for people. The active repository is a bit of a live (and therefore somewhat messy) workplace, and I will aim to produce a cut-down working version of the code and data used in the paper and archive it in e.g. Zenodo before final publication.

1.(i) We should have provided more detail in the text on the method used for fitting the discrete network. We used the default uniform prior in BNLearn, the very simple so-called 'K2' prior, which just adds 1 to the count of every state before counting how often each state of the variable occurs in the data (conditionally on the parent states, where relevant) to estimate the values of the CPDs. The prior doesn't therefore have much weight compared to the data. However, we could instead have specified an imaginary sample size (iss) >1 (where the iss, or pseudo-counts, are equivalent to having observed iss uniform samples of each variable). Use of a higher iss would have resulted in a smoother (and potentially more realistic) posterior, and I suggest we explore this when revising.

1. (ii) Thanks for digging into the data! This comment can be split into two:

- You're right, we used 3 states for colour_prevSummer, not the two that we said in the manuscript (and that were used for all other variables), and the text needs correcting. The hope was that in using three classes we would make the most of the extremely strong correlation between colour_prevSummer and colour.
- Yes, the class of e.g. chl-a in 2018 may not be the same as the class of chl-a_prevSummer in 2019 (and likewise for the other lake water quality variables), despite these summarising the same lake growing season. This is because we were tied to using WFD-relevant thresholds to discretize lake TP, chl-a and cyanobacteria for the current season which we aimed to forecast, to make them management-relevant.

However, for all other features, and including lake observations from the previous summer, we could instead use a discretization method that would give us better predictive power. In our case, we used regression trees between parent and child variables to pick the thresholds to use in discretization of the parents. We do say this very briefly in Section 2.6, last paragraph, but I suggest we add extra text to clarify and justify the method, including more text to describe the regression-tree based discretization (and its limitations).

1. (iii) Using expert opinion to inform the priors in the discrete network would indeed likely have given better results. However, it would not then have been a fair test compared with the GBN. Your second point here is a good one, and is something we would incorporate in a revised "Discussion": rather than using a GBN, a discrete network could have been used where continuous distributions were specified and then discretized. This would have resolved the small sample size issue, and I think it should give near identical results to the GBN (assuming normal distributions were assumed), and would not have the same parametric constraints as other continuous distributions could be specified instead. Although it is a slightly clunky solution compared to just developing a GBN, it could be a good alternative for people who use software that does not have GBN capabilities.

2. The description of the cross validation (CV) scheme used needs updating and we will attempt to improve it to make it clearer what was involved. In fact, the method description is currently slightly outdated, which we apologize for. We used k-fold CV, not leave-one-out CV, but with a high value of k (20) so that it approached leave-one-out CV for cyanobacteria (n=23). In k-fold CV the data are randomly assigned to the k subsets, hence the stochastic element.

Response to minor comments:

20. This comment made us reassess our approach to back-transforming the cyanobacteria predictions (which were Box-Cox transformed for fitting the GBN and producing predictions, and then the expected value was back-transformed to the original data scale). Extra reading has made us realise that straight back transforming from a Box-Cox results in a prediction of the median (e.g. Hyndman and Athanasopoulos, 2018, Chapter 3.2). We will explore using a bias-adjusted back transformation to instead calculate the forecasted mean on the original data scale instead. This should reduce the bias in the GBN predictions and may therefore make the GBN perform better, potentially altering some of our results and their interpretation.

References:

Hyndman, Rob J., and George Athanasopoulos. Forecasting: principles and practice. OTexts, 2018. https://otexts.com/fpp2/

Please also note the supplement to this comment:
https://hess.copernicus.org/preprints/hess-2021-621/hess-2021-621-AC1-supplement.pdf