Hydrology and Earth System Sciences
Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC1
https://doi.org/10.5194/hess-2021-621-RC1, 2022

Comment on hess-2021-621

Anonymous Referee #1

Referee 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-RC1, 2022

Overview

This paper presents a Gaussian Bayesian Network (GBN) for seasonal lake water quality (TP, chl-a, cyanobacteria and colour) forecasting. The GBN was developed and applied to Lake Vansjø in southeast Norway. The GBN was found well-suited for seasonal water quality forecasting and could be parameterised purely using observed data, despite this dataset being small. The forecasting performance of the GBN was assessed using a cross validation scheme, and the performance was also compared to that of a discrete BN (with the same structure) and a naïve forecast model; it was found that the 3 models performed similarly largely due to low interannual variability and high temporal autocorrelation in the study lake.

General comment

Overall, I think this is an interesting study that is very relevant to the HESS special issue. Although the forecasting results with the GBN were considered a mixed success at the study site, I do agree with the authors that the GBN seems to be a sensible and promising approach for water quality forecasting, and I think that by sharing all the code on GitHub the authors have provided a very useful tool for others to use and adapt. I very much enjoyed reading the paper which is both well-written and well-presented, and I believe it can be accepted after some minor revisions.

Below are my comments which I hope the authors will find useful. Lykke til!

Specific comments

- My main concern is related to the discrete BN and the comparison to the GBN. It is interesting that the discrete BN did a mixed job of representing the relationships, however, I don't understand why this happens and I think this could be elaborated on further. Specific comments in relation to this:
  (i) I'm not sure how the method you used to fit the CPTs works, but considering you have a small dataset and that you are using flat priors, I'm surprised that the fitted CPT in Table 6 seems to suggest that the evidence was strong (i.e., most of parent state combinations results in low-high probabilities of around 99%-1% and 95%-5% or vice versa). Intuitively, I would have thought that the probabilities would still be influenced by the flat prior given the small dataset, but the priors have been completely "outweighed" by the data. To me this suggest that there is something odd about the discretisation of the data and/or the target node states.
  (ii) I had a brief look at what I believe is your discretised input data files on Github (..\BayesianNetwork\Data\DataMatrices\Discretized\), and I think these look a bit strange (although I appreciate these may not be the final version). First of all, the 'colour_prevSummer' node seems to have been given 3 states (L, M, H) contrary to what is stated in the manuscript. It also looks like the value for 'chla' does not always match the value of 'chla_prevSummer' the previous year. The same is the case for 'colour' and 'colour_prevSummer'. I would urge the authors to double check these data files and see if this possibly explain (at least partly) the results of the discrete BBN.
  (iii) Finally, I wonder if it would not have been better to use expert opinion to reflect the priors in the discrete network before training, especially as you have a small dataset? To me this would seem sensible, and you already use expert opinion to inform the structure of the network. I also wonder whether you could just have discretised your GBN after it was created (in software like Netica and Genie you can specify continuous distributions and then subsequently discretise these distribution) and how the discretised model would then perform?
- I'm not sure I fully understand how the leave-one-out cross validation works and I think it would be great if the authors could make this a bit clearer in section 2.7.1. Do you leave one data point (i.e., a year?) out at the time and then fit the GBN to the remaining data and see how well the GBN predicts the target node time-series? Or how well the GBN predicts the data point that was left out? Or something else? I also don't really understand why the cross validation is stochastic and why it was run a default 20 times.

Minor comments

- Author name: I believe it should be James E. Sample. Alternatively, change JES to JS in author contributions (L670).
- L21: change "wasn't" to "was not"
- L63-64: maybe worth explaining what polymictic and dimictic lakes are; at least I'm not familiar with these terms.
- Figure 1: where is the outlet from Vanemfjorden? At Moss River?
- L127+: Can you explain briefly why Vanemfjorden with its short residence time is more susceptible to eutrophication and cyano blooms than Storefjorden, and why it does not seem to be related to the major input source from River Hobol?
- L176: Should it be 1998-2013? At least in L179 you seem to suggest NIVA for 2013 as well.
- L188: specify that it is River Hobol
- L192: Change "As the aim" to "The aim". Alternatively combine the two sentences in

L192-195 and remove “therefore” on L194.

- Figure 2: You could consider plotting error bars to give an idea of the variation in the different parameters.
- L227-229: I’m not sure I understand why these features would have to be included as latent variables. Because they are not measured? From Figure 1, it looks like there are monitoring stations in the eastern lake basin (the same as Storefjorden?), so would you not have water quality data from here?
- Table 1 and Table 2: I find it slightly confusing what features are included. Are all the features for the 6-month growing season as well as for the previous winter season (Nov-Apr), i.e., the number of features used for all variables are at least 2x13? Looking at Table 2, and if I understand the caption correctly, it looks like cyano has 8 additional features, so 34 in total (not 33).
- Table 2: Are the features chl-a_prev, cyano chl-a and cyano_prevSummer for the lake?
- L293-300: I think this would be better presented as a table, where you clearly state what is defined as Low and High in the model. The specific comments related to the water quality parameter in question could then be added in a separate column (e.g., that L and H for TP is in fact lower and upper moderate and so on).
- L304+: I don’t follow this part of the discretisation process and why you get unbalanced class sizes. Are the variables still transformed in the discrete version and fairly normal?
- L348+ and Figure 3: Is the relationship between number of calm days and TP negative? To me it looks like the two are positively correlated.
- L355: Are wind speed (winter_wind) and TP(PS) positively correlated?
- Figure 3-6: What are the bell-shaped curves and how were they derived?
- Figure 7: Is TP_prev supposed to be linked to chl-a_prev? If so, should chl-a_prev not have a beta1_TP_prev coefficient?
- L456: Should it not say: “For parentless nodes…”? Some of your nodes are both parent and child nodes (e.g. lake TP is the parent of lake chl-a but the child of TP_prev)
- L526: As you say, this bias in cyano is likely due to the box-cox transformation. Rather than the mean, would it not have been better to use the median (or mode)? Also, did you calculate the mean before or after back-transformation?
- L656: change wasn’t to was not