Response to Reviewer #1

Many thanks to Reviewer #1 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. Below, we provide comments (in red text) in response to each point raised and suggestions for changes we would make during revisions.

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.

Response: Thanks for the positive feedback. 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.

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

Specific comments

1. 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.

Response: It’s not a problem with the discretization or target node states I don’t think, but rather we didn’t do a good enough job of explaining how the discrete network was fitted. Giving more weight to the prior would help smooth the fitted CPTs. For the benefit of those who are new to BNs: the simplest way of estimating the values of the CPDs at each node is just by counting how often each state of the variable occurs (conditionally on the parent states, if the variable is dependent on parents). Or you can start with a prior, which is then updated using the counts from the observed data. The priors are basically pseudo state counts added to the actual counts before normalization. Here, we used the default uniform prior in BNLearn, the very simple so-called ‘K2’ prior, which just adds 1 to the count of every single state. The prior doesn’t therefore have much weight compared to the data. However, we could instead have specified the equivalent or imaginary sample size (iss) to be >1, and therefore use a Bayesian Dirichlet equivalent uniform prior. Then the pseudo-counts are equivalent to having observed iss uniform samples of each variable. Use of a higher iss in our discrete network would have resulted in a smoother (and probably more realistic) posterior, and I suggest that we explore increasing the iss value in a revised version of the paper. We couldn’t find any rules of thumb for what value of iss might be most appropriate, so it will be a case of trying different things out and seeing whether the CPTs look more realistic. As well as exploring higher iss values, we would improve our description of the discrete BN fitting procedure.

(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.

Firstly, thanks for digging into the data!

- You’re absolutely right, we used 3 states for colour_prevSummer, not the two that we said in the manuscript. It was the only variable we used 3 for, and the hope was that in doing so we would make the most of the extremely strong correlation between colour_prevSummer and colour. We will update the text to correct this oversight.
- You’re right that chla and chla_prevSummer (and all the other current vs previous summer variables) can be different for what should be the same year, well spotted. We used WFD-relevant thresholds to discretize lake TP, chl-a and cyanobacteria for the current season. For all other features, and including lake observations from the previous summer, we used regression trees to pick the thresholds to use in discretization. We do say this briefly in Section 2.6, last paragraph, but I suggest we add extra text to emphasize that (1) classification boundaries were different for the two (previous season; current season) variables; (2) that this was done because we did not have to be constrained by the need to produce management-relevant predictions when it came to discretizing the previous summer’s values, and so we opted to choose the

discretization that we hoped would give us the strongest relationship between variables (as identified using regression trees), rather than WFD-relevant boundaries; and (3) that it is very possible this wasn't the best approach, and that better results might have been obtained if we had maintained the same classes through time. This relates to the more general point that discretization is subjective and time-consuming.

(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?

- Using expert opinion to decide on the priors in the discrete network would I'm sure have given better results. However, it would not then have been a "fair" test compared with the GBN.
- Your second point is a really 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 you first assumed and specified continuous distributions, and then discretized these. This would have resolved the small sample size issue, and I think it should give near identical results to our GBN (in the case where normal distributions were assumed). Although it is a slightly clunky solution compared to just developing a GBN, and is not something we have explored ourselves (so I'm not sure how well it would work in practice), I imagine it could be a good alternative for people who use software that does not have GBN capabilities built in yet.

2. 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.

   Response: We will certainly clarify this section, and in fact it is slightly outdated compared to the final method used, which we apologize for. We in fact used k-fold cross validation, but with a high value of k (20) so that it approached leave-one-out cross validation for cyanobacteria (n=23). In short, the cross validation was repeated for each node that we wanted to estimate predictive error for (chla, cyano, TP, colour). For each of these "target" nodes in turn, the algorithm randomly assigns the time series data (i.e data for all nodes for a given year) into k subsets (20 in our case, so for the cyanobacteria data many of these subsets will just have one year in them). Then one subset is left out at a time, the BN is fitted using all the remaining subsets, and then the fitted BN is used to predict what the target node would be for the left out subset (using only the nodes that we would have data available for at the time of issuing a seasonal forecast). As the data are randomly assigned to the 20 subsets, results differ between runs (less so for cyanobacteria than the other variables), and so we repeated the procedure 20 times. Each loop through all k subsets produces a single time series of predictions, and then there are 20 sets of

these due to repeating the procedure. Each of these was then compared to observations to generate model performance statistics, and we took the mean of the model performance statistics. I hope that makes more sense?

Minor comments

1. Author name: I believe it should be James E. Sample. Alternatively, change JES to JS in author contributions (L670). Yes, thanks
2. L21: change “wasn’t” to “was not” Ok
3. L63-64: maybe worth explaining what polymictic and dimictic lakes are; at least I’m not familiar with these terms. Will do
4. Figure 1: where is the outlet from Vanemfjorden? At Moss River? Yes (will change Mosselva to Moss River in the text, and mention this in the fig. caption).
5. 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?

   Yes, we can add this to the text. In brief, shallow lakes tend to have stronger interaction between the water column and the (P-rich) sediments, and so P concentrations are higher in the water. In addition, the local catchment surrounding Vanemfjorden is more agricultural than the larger Storefjorden catchment.

6. L176: Should it be 1998-2013? At least in L179 you seem to suggest NIVA for 2013 as well. No, NIVA data was ok after 2007, but lower frequency (and not in winter) compared to MOVAR data. We therefore used MOVAR data rather than NIVA data where possible (can add a sentence on this).
7. L188: specify that it is River Hobol. Yes, thanks.
8. L192: Change “As the aim” to “The aim”. Alternatively combine the two sentences in L192-195 and remove “therefore” on L194. Ok
9. Figure 2: You could consider plotting error bars to give an idea of the variation in the different parameters.

   We will look into this, at least for variables where growing season means are plotted (e.g. error bars of one standard deviation). For variables which are growing season maxima we could perhaps plot a single lower error bar down to the mean of the 3 or 5 highest values. I’m not sure what a meaningful estimate of variation would be for parameters which are summed over the growing season per year, so probably would not add anything for these.

10. 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?

    This data could be included in the forecasting model if they were measurements from before the time the forecast was issued in spring (e.g. from winter). But we couldn’t use Storefjorden water chemistry to forecast Vanemfjorden water chemistry in summer 2022 (for example), as we wouldn’t have that data available

yet and would then also need to forecast Storefjorden water quality. Hence the need for latent variables – probably important, but have to be predicted from nodes that you would have data for *at the time the forecast was issued*. We can emphasize this a little more in the text, as a forecasting model is somewhat different to a model that is simply designed to explain/expose interrelationships in the data.

11. 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).

    Yes, agree it is a little confusing. We will improve this part – perhaps easiest to just write all the features out in full.

12. Table 2: Are the features chl-a_prev, cyano chl-a and cyano_prevSummer for the lake? Yes (see the 'Description' column)
13. 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). Good idea.
14. 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?

    We will expand on this section. The regression trees, regressing a parent (independent variable) against a child (dependent variable) node, were used to pick out class boundary divisions for the parents. However, sometimes the regression tree division resulted in a split of the data where most were in one class, and just a couple of points were in the other class. No, the variables weren't transformed in the discrete network (that being one of the benefits of discrete vs GBN).

15. 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.

    Thanks, typo. There was a negative correlation with wind speed, not number of calm days.

16. L355: Are wind speed (winter_wind) and TP(PS) positively correlated?

    Yes, but given there is no way that lake TP concentration in e.g. summer 2020 can affect wind in winter 2020/21, we suspect this is correlation but not causation. I hope that came out ok in the text?

17. Figure 3-6: What are the bell-shaped curves and how were they derived?

    Thanks, will add description to fig. caption (they are probability density functions approximated using kernel density estimation).

18. 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? Very well spotted, thanks!
19. 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). Yes, thanks.
20. 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?

    The GBN prediction (mean of the cyanobacteria CPD on the Box-Cox transformed scale) was back-transformed to produce the forecasts on the original cyanobacteria data scale. However, extra reading (https://otexts.com/fpp2/transformations.html) has made us realise that the back-transformed numbers are in fact medians on the original data scale, not necessarily means, and in cases where the median and mean are different, a straight back-transformation introduces bias. We will explore using a bias-adjusted back transformation (using the formula given in the above link) to calculate forecasted mean cyanobacteria instead. This should reduce the bias, and it will be interesting to see how much by and whether it changes any of the conclusions.

21. L656: change wasn’t to was not. Ok