Response to Reviewers

Reviewer #1

RC1

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.

Change made: a cut-down and tidied version of the code and data used in the paper has been made available at https://github.com/LeahJB/gbn-vansjo. I will archive this (Zenodo) and add a reference and doi to the paper once the revision process is finalized.

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.

   Change made: As described in our response to reviewer #1, this wasn't a problem with the discretization of the data, but the weighting of the prior. We looked into the imaginary sample size (iss) parameter in more detail, and experimented with different values. As described in our original response, this parameter effectively controls the prior's weight compared to the data counts. In our original submission we just used bnlearn's default value of iss = 1. However, we found that using a larger value of 15 smoothed the CPTs enough that at least some of the unexpected behaviour was removed. i.e. our new network suffers much less from over-fitting due to a stronger weighting of the prior.

   We have provided a better description of our method for fitting the discrete network's CPTs, including a mention of the iss parameter, in the last paragraph of Section 2.6.2. As this involved updating the fitting of the discrete BN, we have also updated all the stats and parameters associated with the discrete network throughout the text (including the cross validation results).

(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.

Changes made: we have clarified in section 2.6.2 (first paragraph) that three classes were used for colour_prevSummer, with justification. In this same paragraph, we have improved our description of the method for discretizing using regression trees, and added text to emphasize that chla and chla_prevSummer (and all the other current vs previous summer variables) can be different for what should be the same year.

(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?

Changes made:
- As we said in our original response, we agree that using expert opinion to decide on the priors in the discrete network would likely have given better results. However, as it wouldn't have been a fair test compared to the GBN, we didn't explore this when revising.
- We have revised the Discussion (Section 4.2, end of 3rd para), to mention that, rather than using a GBN, a discrete network could have been used with specified probability distributions, which certain software then discretizes. This should give near identical results to our GBN (in the case where normal distributions were assumed), and 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.

Changes made: We have updated the first paragraph of Section 2.7.1 to include a more detailed explanation of the cross validation procedure. We also changed the cross validation to be leave-one-out (rather than nearly leave-one-out, as used in our original submission).

Minor comments

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

   Change made: Added flow directions to Fig. 1, changed Mosselva to Moss River in the text, and mentioned where the outlet from Vanemfjorden is 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?

   Changed: Short addition made to text (penultimate para of Section 2.1).

6. L176: Should it be 1998-2013? At least in L179 you seem to suggest NIVA for 2013 as well. Changed: made it clearer why we used MOVAR data for the period until 2012, despite having decent NIVA data from 2008 (second para of Section 2.3)
7. L188: specify that it is River Hobol. Done
8. L192: Change “As the aim” to “The aim”. Alternatively combine the two sentences in L192-195 and remove “therefore” on L194. Done, thanks for spotting.
9. Figure 2: You could consider plotting error bars to give an idea of the variation in the different parameters.

   Response: we looked into doing this, but the plot looked strange with error bars for some variables (those that had been averaged), and not for others (those which had been summed or were maxima). We would rather leave them out, as the focus wasn’t really on this in this model.

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?

    Changes made: updated first para of Section 2.4 to better explain the choice of features, given the aim of producing a model for operational seasonal water quality forecasting.

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).

Changes made: We have redone Table 1 to include all features explicitly and deleted Table 2 (which was redundant).

12. Table 2: Are the features chl-a_prev, cyano chl-a and cyano_prevSummer for the lake? Yes (see the 'Description' column in Table 1)
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). Change made: removed this information into new Table 2
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?

    Changed: Expanded Section 2.6.2 first paragraph to provide a more complete description of the discretization method used.

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.

    Fixed typo, thanks.

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

    Yes, but unlikely to be a causative relationship.

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

    Changed: description added to figure captions

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? Added, thanks for spotting
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). Changed.
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?

    Changed: Originally when we back-transformed the cyanobacteria predictions, we did not adjust for bias introduced by the transformation. So the back-transformed value was the median, rather than the mean. We have now replaced this with a bias-adjusted back transformation (equation is here: https://otexts.com/fpp2/transformations.html). This results in much reduced bias in the cyanobacteria predictions, i.e. a much more realistic-looking forecast.

Because of this change in the way we calculate cyanobacteria, we have updated all the Tables and Figures in the text that relied on cyanobacteria forecasts. We have also updated the text (results and discussion) to reflect the improved cyanobacteria forecast performance statistics.

21. L656: change wasn't to was not. Done

Reviewer #2

RC2

General comments:

This study presents an application of a Continuous Bayesian Network (CBN) to seasonal (6-month average) algal forecasting in a northern lake. This is likely the first use of CBN for this purpose. In general, the model performs similarly to a traditional (discretized) BN and a naïve model (using the mean from the previous year). It could be a good fit for this special issue, but I do have several concerns, as outlined below.

I'm not really sure that there is a strong contribution, as the CBN does not perform particularly well.

Also, the model appears to be based on existing software (an R package), so there isn't new methods development. If the objective of the study is to provide a thorough demonstration of CBNs for algal bloom modeling, that could potentially be an important contribution. In this case, I'd like to see more demonstrations of how the CBN approach (e.g., Figure 7) can be advantageous for studying a system or supporting management. In my opinion, the current discussion is too focused on skill assessment (e.g., R2), which probably doesn't do justice to the CBN approach. Also, probabilistic predictions using various linear covariates can also be obtained through multiple linear regression (frequentist or Bayesian), so why use a CBN? I think there are potentially good reasons for using a CBN, but they aren't compellingly demonstrated in the current manuscript.

Also, I'd like to see more discussion of how this effort compares to other CBN (or BN) applications for water quality or environmental sciences, more broadly.

Changes made:

- Made the aims of the paper clearer in the last para of the introduction
- Tried to make the novelty of the seasonal forecasting aspect more prominent (rephrasing parts of the introduction to describe the setting of the WATExR project, better explanation for the choice of variables to include in the analysis and how these would need to be replaced with seasonal climate forecasts in any operational model, more mention of this in the discussion)
- Added a section demonstrating the use of the GBN for supporting management (Section 3.4)
- Added a discussion comparing GBN and multiple linear regression in the context of the case study (Section 4.2, last para).

Major comments:

The paper includes a tangential analysis on making predictions at smaller time scales (e.g., Lines 208-215). I recommend removing this material, as it doesn't seem relevant to the main focus of this paper (no CBN was used). Furthermore, this additional analysis doesn't provide

new insights (that aren't available through existing phytoplankton literature). It seems a bit "tacked on". If you do keep this analysis, the data should be presented (as in Figure 2 for the six-month model).
Changed: moved to Appendix A

The variable selection process seems ad hoc (Section 3.1.1), making it somewhat hard to follow and likely difficult to reproduce. Some of the explanations seem questionable. For example, the article cites previous literature showing that "windier summers" are relevant, but the CBN uses winds from the previous 6 months (prior to summer), right? I have two general suggestions. First use clear and consistent terminology that clarifies which time periods you are talking about (also use consistent notation across the different figures and tables). Second, drop wind from the 6-month analysis altogether. Much of the text is a somewhat tortuous explanation (at least for this reader) of reasons to include/exclude wind speed, while in reality, the authors readily acknowledge that wind speed is only relevant at smaller time scales (e.g., Lines 443-445: "wind would likely only have an immediate and relatively short-lived effect…"), not ~6 months in advance.

Changed:
- We have re-done Table 1 and added a new column to clarify the aggregation period used for each of the variables.
- We have gone through the paper and made sure the notation is now consistent for each variable across the different tables and figures
- We have added a clearer justification at the start of Section 2.4 (Feature generation) for the choice of variables to include in the analysis
- Re-written Section 3.1 (Results of feature selection) in an attempt to make it less long-winded, and remove the emphasis on the wind discussions.

Detail-oriented comments:

Line 11: Clarify in the abstract that you are predicting a May-October average (rather than daily predictions).
Done

Line 20: The term "purely parameterized" is used multiple times throughout this manuscript, but I don't understand what it means or how it is justified. As noted above, the parameterization process seems somewhat ad hoc to me.
Changed: mentioned that expert knowledge is often used to parametrize CPTs when sample sizes are small (Introduction, discussion),

Line 23: Suggest clarifying what is meant by a "naïve forecast" here.
Done

Line 44: Models for Lake Erie cyanobacteria blooms (including Bayesian models) predict the maximum bloom size months in advance.
Altered the text

Line 56: Could you explain why "colour" is particularly relevant to water treatment or provide a reference?
Reference added

Figure 1: Suggest including arrows to show dominant flow directions.
Done

Table 2: Clarify what averaging periods were used.
Added new column to the Table (and merged old Tables 1 and 2 into new Table 1)

Line 273: Clarify what normality test was used.
Done

Figure 3, 4, 5: Clarify why only certain features are shown in each figure.
Added text to figure captions

Table 4: The "Feature subset" column is confusing. Use consistent terminology and explain in the caption.
Amended this table slightly and the figure caption

Line 370-371: Revise for clarity.
Done

Line 422: Suggest "wind-related" instead of "related" for clarity.
Done (now in Appendix A)

Line 458: The term "credible" usually refers to the uncertainty in a parameter. It could be good to present actual parameter uncertainties (e.g., credible intervals). Also, I don't think relationships matching the simple bivariate correlations necessarily makes them "credible" in any sense. For example, see literature on Simpson's Paradox.
Changed:

- Replaced "credible" with "plausible", and tried to make the sentence more nuanced: "Fitted coefficients for the Gaussian BN were all plausible, and matched the simple bivariate relationships between variables seen in the exploratory data analysis".
- Added 95% confidence intervals to a new Table B2 in Appendix B.

Line 470: Again, I'm not sure using simple bivariate correlations to evaluate a more sophisticated model makes sense.
Changed: re-written this paragraph (Section 3.2.2)

Table 6: To me, making some numbers bold isn't effective for highlighting unexpected results. It really depends on which particular pair of numbers is being compared. Also, I wouldn't describe some of these relationships as a "physical" response.
Changed: replaced Table 6 with a new Figure 8, which includes the fitted CPTs for the whole discrete network.

Line 569: This statement seems too strong and/or requires clarification.
Changed: added a qualification to the end of the line ("…a number of studies will have over-estimated its importance, by assuming that the within-year relationship between temperature and algal dynamics can be used to infer future algal responses to increases in summer temperature under climate change")

Line 644: This is clearly true (based on the general nature of a GBN), but it wasn't really explored in this study. I'm not sure why it is a conclusion.
Changed: deleted this sentence.

Line 659: This seems like a bit of a stretch. I'm not sure that any "expert" can predict an extreme event ~6 months in advance. Maybe the authors mean something else, but I can't imagine what.
Changed: added text to last para of conclusions to explain ourselves more clearly.

RC3

Quick question: How can this be considered a forecast model if it requires you to input wind speeds 6 months in advance? At best, we can only forecast wind speeds a week or two in advance.

RC4

I think the forecast development narrative is a bit muddled. To say that six-month-ahead wind speed forecasting "isn't quite there yet" is an understatement. If you are willing to consider 6-month-ahead wind forecasts, then why not also consider 6-month-ahead phosphorus forecasts? The latter is likely much more realistic.

Also, if the GBN can't provide any measure of credibility of the relationships (e.g., credible intervals for parameters), this is an important limitation that should be noted. I am not a GBN expert, so I can't provide guidance on how to do this. But it can obviously be done in most linear models (Bayesian or frequentist). Also, probabilistic predictions are easy to obtain from MLR models, so I'd be cautious about over-emphasizing this as an advantage of GBNs.

Overall, I'm not sure if I'll be able to recommend publication based on the proposed revisions. Of course, I defer to the editor.

RC5

Thank you for the additional notes. I think my last response was a bit hasty. At the same time, it's somewhat unclear why certain features (covariates) are based on observed data for the period of prediction (which can't be known at the time of forecast) and other features are based on forecasts of those features. However, I don't think this is a major sticking point, as the authors can further clarify these issues and their motivation in the manuscript.

I appreciate the authors exploring the credible intervals issue, and I think the proposed demonstration of an example forecast may be helpful. Given that that the GBN shares many of the same features as an MLR (linear relationships, Gaussian error distribution (usually), probabilistic predictions of continuous variables), it would be nice to clarify the potential advantages and disadvantages of the GBN approach. The authors provide some comparison with the discrete BN (Section 4.2), so perhaps something along these lines and with connections to your particular case study.

RC7

I agree it isn't necessary to create an MLR. For one thing, I suspect the predictive skill would be pretty similar to the GBN.

At the same time, it would be nice to elucidate how the GBN could be advantageous relative to more conventional linear statistical models for algal bloom forecasting (while also acknowledging GBN limitations). It's true that BNs have particular features documented in previous literature (e.g., Lines 76-83) but not all of these are unique to BN models, many weren't demonstrated in this case study, and some might be debatable for a GBN (given the linearity and distributional constraints). Perhaps one important distinction of the GBN is the multivariate structure illustrated in Figure

7. Perhaps the authors could explore and discuss this a bit further in the context of their case study.

Here are a couple of the relevant papers for Lake Erie. One uses an MLR and the other is similar to MLR in some ways (I think). So they might provide good context for this discussion, as well as the discussion of seasonal bloom forecasting, in general.

Obenour, D. R., Gronewold, A. D., Stow, C. A., & Scavia, D. (2014). Using a Bayesian hierarchical model to improve Lake Erie cyanobacteria bloom forecasts. Water Resources Research, 50(10), 7847-7860.

Ho, J. C., & Michalak, A. M. (2017). Phytoplankton blooms in Lake Erie impacted by both long-term and springtime phosphorus loading. Journal of Great Lakes Research, 43(3), 221-228.

Thanks for the interesting discussion. Good luck with your revisions.

Changes made in response to discussions during RC3 – RC7 and related ACs:

- We have re-written the text (Introduction, Section 2.4, Discussion) to provide a better background to our motivations for developing the seasonal forecasting tool, and to explain our choice of variables to include in the exploratory feature analysis.
- We have made some alterations to Figure 7 to make it clearer what the data sources are for the model development in this paper, and what they would be for an operational forecasting tool.
- We have added a new section to the discussion (Section 4.1.2) to clarify that seasonal climate data would be needed for met variables to be included in an operational model.
- Confidence intervals have been provided in new Table B2.
- We have added a new section (Section 3.4) which links to a prototype forecast developed for the case study site, with some discussion of how this might be useful to managers.
- Added a paragraph to the discussion (last para of Section 4.2) comparing the pros/cons of GBN vs MLR, in the context of our case study.
- Added the Ho & Michalak reference in relation to MLR vs GBN.