Guide to understanding and interpreting seasonal water quality forecasts for Lake Vansjø

Lake water quality forecasts are issued once a year in April for the period May-October, to coincide with the definition of the growing season used in the Water Framework Directive (WFD). Forecasts are for the western basin of Lake Vansjø (Vanemfjorden), and aim to predict WFD ecological status. Four variables are predicted: concentrations of total phosphorus (TP), chlorophyll-a (chl-a), lake colour and cyanobacteria. The forecast is for mean TP and chl-a and maximum cyanobacteria biovolume (consistent with WFD classification criteria). Mean colour is forecasted as it may affect cyanobacteria and also be of interest in terms of drinking water treatment.

The forecast tables for each lake variable contain the following information:

(WFD) class: classes considered when making predictions. The models used to make predictions are statistical, i.e. they rely on historic observations, and there was only enough historic data to allow two classes to be considered per variable. These two classes were chosen by simplifying the WFD classes, dropping classes which had few observations during the historic period. In the case of TP, almost all lake observations lay within the 'Moderate' class, so this was split into 'upper Moderate or better' and 'lower Moderate or worse'. The WFD does not specify lake colour limits that are relevant for Vansjø, so this was split into 'High' and 'Low' using the upper tercile (66th percentile) of the historic observed seasonal mean colour.

Probability of the class: The forecasted probability of the class occurring (note: not available for chl-a as a simpler model was used in forecasting; see below). Probabilities associated with each class were discretized into four categories, from very low to high, as follows:

Probability category	Probability of the class (%)
Very low	0 – 25
Low	26 – 50
Medium	51 – 75
High	76 - 100

Forecasted value: The most likely value for this variable, in terms of a single number rather than a class.

Historic skill statistics: These summarise how well forecasts performed over a historic period (1981-2019; 1996-2019 for cyanobacteria), compared to lake observations during that time. The following statistics are reported:

- **RMSE**: root mean square error, the likely difference between forecasted and observed values. Smaller RMSE is better, and the bigger the RMSE compared to the forecasted value, the less certain the predictions.
- **Classification error**: The percentage of time the model predicted the class incorrectly during the historic period. We might expect similar classification errors during the future, given similar conditions to the historic period.
- **MCC**: Matthews correlation coefficient, a measure of the overall accuracy of the classification model. Higher values are better: a value of +1 is a very good model, 0 means no correlation between predictions and observations, -1 is a very bad model. The MCC was used to summarise the historic skill of the models according to the following qualitative rules:

MCC	Historic skill
<0.2	None
0.2 – 0.39	Low
0.4 – 0.59	Medium
>0.6	High

Forecast summary: the most likely classification for the coming season, together with an indication of the overall confidence that can be placed in predictions. The overall confidence level is assessed as follows:

- If class probabilities are 'High' and 'Very low', the overall confidence is the same as the MCC historic skill.
- If class probabilities are 'Medium' and 'Low', then there is lower confidence in the forecast. If the MCC historic skill score is 'Medium' or 'High', the overall confidence is reduced by one (to 'Low' or 'Medium').
- For chl-a, the confidence level is always 'Medium'. Historic skill was 'High', but because of a lack of class probability information it is reduced a level.

Models used to generate the forecast

A number of forecasting models were built and their forecasting skill was compared during the historic period (1981-2019 for TP, colour and chl-a; 1996-2019 for cyanobacteria). The following models were considered, in order of increasing complexity:

1. Seasonal naïve forecaster: each forecast is simply the value observed during the previous summer.
2. Gaussian Bayesian Network (BN; Figure 1, and see Box 1 for an introduction to BNs), where meteorological conditions are set to the long-term average and kept constant when making predictions. This is equivalent to removing weather nodes from the BN structure shown in Figure 1.
3. Gaussian BN (Figure 1), where observed (and for future projections, forecasted) meteorological data are included when making predictions.

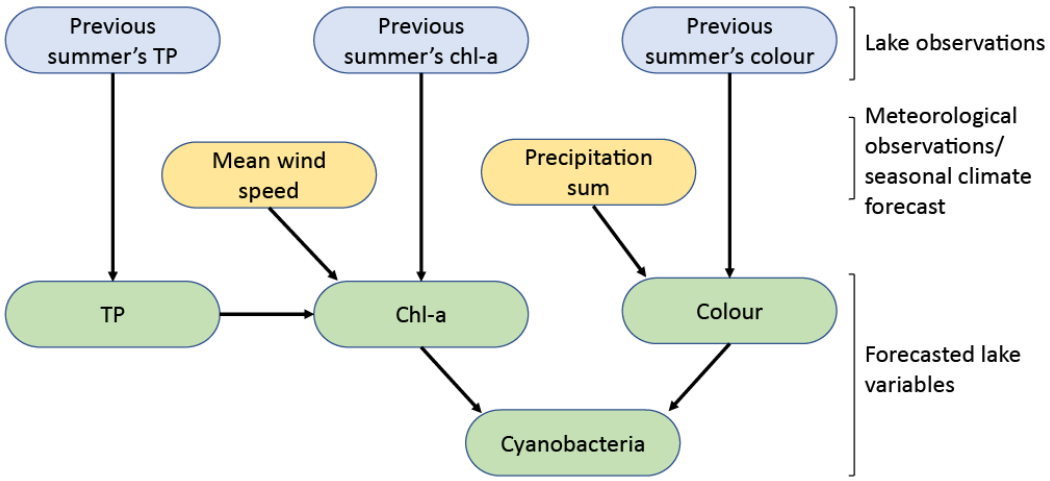

Figure 1: Bayesian Network (BN) structure used. See Box 1 for an introduction to BNs.

Historic skill was assessed through leave-one-out cross validation and a combination of skill metrics (predictive correlation, rmse, classification error, Matthew's correlation coefficient). The model with the highest skill was then chosen for operational forecasting, as follows:

- TP, cyanobacteria, colour: Gaussian BN (Figure 1), setting the meteorological nodes (mean wind speed and precipitation sum) to constant values, the ERA5-observed average for the historic period
- chl-a: seasonal naïve forecast

Note: the models chosen do not include any meteorological data information when making forecasts. The only data driving variation in forecasted values between years is the observed lake data from the previous year. At present, including meteorological information in the model reduced its predictive ability.

Box 1: Introduction to Bayesian Networks

A Bayesian Network (BN) is a probabilistic graphical model that represents a set of variables and their conditional probabilities in a causal network. In Vansjø, the BN represents the relationships between meteorological and historic lake variables, and the variables related to management targets (nutrient concentrations, algae, water colour). Given lake observations from the previous year and a seasonal climate forecast, the BN can be used to predict lake variables of interest in a probabilistic way. The underlying probabilities may be updated as more observed data becomes available, meaning the performance of the forecasting system can improve through time.

The structure of the BN was developed using: (1) Expert knowledge; (2) Statistical exploration of the data, including feature selection techniques. A variety of temporal aggregations of the data were explored (monthly, 3-monthly, growing season); and (3) Cross-validation of a variety of BN model structures. A Gaussian BN was chosen, where continuous variables are used for the nodes. BN parameters (conditional probabilities) were then set by training on observed lake data from the period 1981-2019 (1996-2019 for cyanobacteria).