# Review of the Revision of:

# Spatially Referenced Bayesian State-Space Model of Total Phosphorus in western Lake Erie

November 11, 2021

Referee: Ken Newman, Biomathematics & Statistics Scotland, and School of Mathematics, University of Edinburgh

## 1 Overall comments

The manuscript remains generally well organized and written. The authors have responded fairly thoroughly to the comments made in the first review, and I appreciate them checking out what happens with hourly step sizes to construct the adjacency matrix. Several substantive concerns remain, however.

### 1.1 All $\beta$'s in process model near 1 and consequence

That the posterior distributions for all four $\beta$'s in the process model are so concentrated near 1 (across all years) seems remarkable. That coincidence aside, assuming that the latent states are the logarithms of the true TP value, the effects on the expected TP transfer from a "source" node (denoted $k$) to a corresponding "sink" or recipient node (denoted $n$) implies an increase in TP at the sink node on the raw scale. Letting $z_{n,t,y}$ be the raw scale TP value, thus $x_{n,t,y,} = \ln(z_{n,t,y})$, then the expected TP value in the sink node

$$E[z_{n,t,y}|z_{k,t,y}] = \exp\left(\beta + \ln(z_{k,t,y}) + \tau^2/2\right)$$

Substituting $\beta$=1 and $\tau^2$=$0.2^2$ (a rough average from Table C1):

$$E[z_{n,t,y}|z_{k,t,y}] = \exp\left(1 + \ln(z_{k,t,y}) + 0.2^2/2\right) = 2.77 z_{k,t,y}$$

### 1.2 Posterior summaries

Another concern is with regard to the posteriors for the process and obs'n standard deviations. In Figure C1, the ranges of the joint priors for the pairs $(\beta_{rai}, \beta_{mau})$, $(\beta_{self}, \beta_{lake})$, and $(\ln(\sigma_y), \ln(\tau_x))$ (my added subscripts) are denoted by the red convex hulls. The black polygons denote posterior fitted values (for all 11 years).

- Are fitted values posterior means?
- I don't understand the sentence in the figure caption about the fitted values for each year not overlapping. Based on Table C1 there is considerable similarity in the $\beta$'s: they are all very close to 1.
- With Figure C1.c it is disconcerting to show posterior means outside the support of the prior; that should not be so.
- Based on the `R` code the priors for the precision for the process and obs'n models ($1/\tau^2$ and $1/\sigma^2$) are Gamma(0.001, 0.001). The ranges for $\tau$ and $\sigma$ are the positive real numbers (which in practice, see below, can be 0 and $\infty$), and the subsequent range for $\ln(\tau)$ and $\ln(\sigma)$ is much larger than Figure C1.c indicates (roughly -6 to 11).

```
set.seed(301)
n <- 10000
Q     <- rgamma(n,0.001,0.001)  # state
sd.Q <- 1/sqrt(Q)
ok.Q <- sd.Q != 0 & !is.infinite(sd.Q)
summary(log(sd.Q[ok.Q]))
#   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# -3.705  66.940 148.888 159.015 243.251 368.766
```

## 1.3 Identifiability of $\sigma^2$ and $\tau^2$

Potential identifiability issues for $\tau$ and $\sigma$ have not been addressed. What needs to be examined is the correlation between $\tau$ and $\sigma$. Scatterplots of sampled pairs from the posterior distribution, at a minimum, need to be examined (and shown). Figure 1 shows how the lower and upper bounds of the PIs relate, suggesting a negative association.

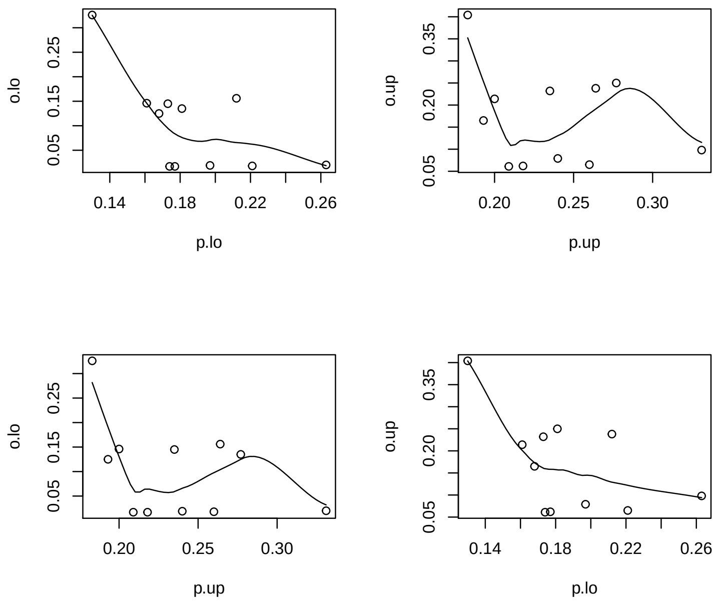

Figure 1: Scatterplots of lower and upper bounds of 95% PIs for $\tau$ and $\sigma$. p.lo and p.up are lower and upper endpoints for the process standard deviation; similarly, o.lo and o.up for observation.

## 1.4 Bayesian $R^2$ calculations

My understanding of the Bayesian $R^2$ calculations based on Gelman et al (2019, eq'n 3) is that the calculated value of $var_{res}$ would be the estimated value of the observation variance terms, $(\sigma^2)^s$ for sample $s$ from the posterior. What is written in the text seems to match the Gelman et al eq'n 2 definition for *non*-Bayesian $R^2$. How much difference that makes needs to be examined—the fact that the majority of the average Bayesian $R^2$ are greater than 0.98 or so seems remarkable.

### 1.5 Cross-validation calculation and subsequent modelling of scores

The description of the K-Fold cross-validation score in eq'n 8 is unclear. The summation index is $d$ but the last value of the index is also $d$, the term $d$ does not appear in the values being summed, and the conditioning notation $\tilde{y}_{n,t,y}$ is is ambiguous. I am assuming that the eq'n is patterned after equation 3 in Piironen and Vehtari (2017) and conditioning would be on the set excluding the $y_{n,t,y}$ values; something like they used $D_{-y_{n,t,y}}$ might be more understandable.

The cross-validation scores are then regressed on $\mu_{o,n}$, $\mu_n$'s (eq'n 9), or $\mu_{o,y}$ and $\mu_y$ (eq'n 10), where these regressors are then given priors. The terms on the left-hand sides of eq'ns 9 and 10 need subscripting. More critically, are the regressors pure random effects and what are called priors actually the probability distributions for these random effects? I tried to find where in Kruschke (2014; that should be 2015 I believe) such a procedure was used—chapter or page numbers need to indicated.

## 2 Detailed technical comments

p=page, L=line.

1. Methods

   (a) p3, L93: the 2.2 heading "Model description" still seems misleading, as the real model is the SSM. Subsection 2.2 is about creating the adjacency matrix, which is *not* a model, per se.

   (b) Constraints on State process values: Need to state that $x_{n,t,y}$ is the logarithm of TP: this is also important as the authors' reply about TP not being negative applies to the raw scale values, not the logarithm. While I can see the argument for constraining the range of $x_{n,t,y}$ based on expert opinion, I don't think that the observations in a given year should be used to determine those constraints: the prior needs to be independent of the data, nor does it make sense to me that measuring instrument limits constrain the true value. Why not simply say that our prior opinion is that log of TP lies between $a$ and $b$ and use the same values for each year?

   I am possibly misinterpreting the scale for the obs'ns but based on the R code: `concs = log(concs*10`$^3$`)`, summaries of raw and logged TP values for the Maumee River, Raisin River, and the lake are the following:

| | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | NA's |
|---|---|---|---|---|---|---|---|
| Maumee R | 77.33 | 173.38 | 218.07 | 261.20 | 334.25 | 804.80 | 21.00 |
| Raisin R | 28.00 | 65.95 | 85.15 | 107.12 | 109.35 | 458.20 | 65.00 |
| Lake | 5.50 | 20.60 | 33.20 | 51.05 | 63.81 | 275.07 | 91789.00 |
| log Maumee R | 4.35 | 5.16 | 5.38 | 5.44 | 5.81 | 6.69 | 21.00 |
| log Raisin R | 3.33 | 4.19 | 4.44 | 4.51 | 4.69 | 6.13 | 65.00 |
| log Lake | 1.70 | 3.03 | 3.50 | 3.62 | 4.16 | 5.62 | 91789.00 |

   The authors' reply that 0.7% were less than 10 $ug/l$ is presumably referring to the lake samples; I found one out of 191 of the lake values less than 10. Note that the above means differ from those reported in the Results (lines 214-215), but maybe I've made mistakes.

   (c) p4, L122-135: I find the explanation of the SSM awkward. Need to state somewhere that the latent states are log transformed (only by checking the R code did that become clear). Consider: "The SSM consists of two models, one for the data ($y$) called the observation model and one for latent states ($x$) called the process model." The observations were modelled as follows.

$$y_{n,t,y} \sim \text{Normal}\left(x_{n,t,y}, \sigma^2\right) \tag{1}$$

   where $y_{n,t,y}$ is the natural logarithm of the measured TP concentration at node $n$ on day $t$ of year $y$, $x_{n,t,y}$ is latent (unobserved) true log TP concentration, and $\sigma^2$ is the observation standard deviation. The process model is first order Markov, only depending on the value of the node at time $t-1$ which transported to TP to node $n$ at time $t$. That source node is denoted $k$ and for nodes in the river, $k$=$n$, and for nodes in the lake, $k$ is determined from the time $t$ adjacency matrix.

$$x_{n,t,y} \sim \text{Truncated Normal}\left(f(x_{k,t-1,y}), \tau^2\right) \quad I(a \le x_{n,t,y} \le b), \tag{2}$$

where

$$f(x_{k,t-1,y}) = \begin{cases} \beta_{mau} * x_{k,t-1,y} & \text{if } n\text{=Maumee River node} \\ \beta_{rai} * x_{k,t-1,y} & \text{if } n\text{=River Raisin node} \\ \beta_{self} * x_{k,t-1,y} & \text{if } n\text{=same lake node} \\ \beta_{lake} * x_{k,t-1,y} & \text{if } n\text{=a different lake node} \end{cases} \tag{3}$$

The process standard deviation is $\tau$ and the values for $x_{n,t,y}$ are restricted to $[a, b]$
Also please write the eq'ns for the priors for the $\beta$'s to ease comprehension for the reader. It's not clear to me what the sentence at lines 136-137 is saying. Only by examining the R code could I tell that:

$$\begin{aligned} \beta_{self}, \beta_{lake} &\overset{iid}{\sim} \text{Normal}\,(0, 10,000) \\ \beta^* &\sim \text{Normal}\,(0, 10,000) \\ \tau_{mau}, \tau_{rai} &\overset{iid}{\sim} \text{Gamma}\,(0.001, 0.001) \\ \beta_{mau} &\sim \text{Normal}\left(\beta^*, \tau_{mau}^{-2}\right) \\ \beta_{rai} &\sim \text{Normal}\left(\beta^*, \tau_{rai}^{-2}\right) \end{aligned}$$

Also write the priors for $1/\tau^2$ and $1/\sigma^2$.

2. Results. Major comments about the results were given in Section 1. One other point is the regression of mean TP concentration on effective Spring TP load: doesn't distance of the lake nodes from the Maumee River have an effect on the relationship? Distance is not included in the regression.

3. Discussion. Not clear what in Auger-Methe, et al (2021) is being referred to regarding identifiability and visual determination of priors dominating. Is what is meant that the posterior and prior will not differ much for parameters that are unidentiable or weakly identifiable?

## Minor editorial remarks

1. Abstract

   (a) p1, L22: "estimated that, in the absence of the Maumee River load, lake concentrations..."

2. Introduction

   (a) p2, L46: "affects"

   (b) p2, L52: here refer to soluble reactive phosphorous (SRP), but later (starting with p3, L80) refer to TP, without defining what TP means. I'm no expert on water chemistry, but my understanding is that TP, total phosphorous, includes SRP. TP needs to be defined, and if SRP is not referred to again, perhaps do not add the abbreviation.

   (c) p2, L60: perhaps "Bayesian inference" instead of "Bayesian frameworks". This paragraph is more about SSMs than about Bayesian inference, and it might be better to make a statement about SSMs first (the topic sentence), e.g., use the 2nd sentence without the adjective Bayesian: "State-space models (SSMs) have been used...". Then discuss the application areas and then add a sentence or two about Bayesian SSMs.

(d) p3, L68: “While spatial models”

(e) p3, L71: “incorporate concentration data”

(f) p3, L80: define TP

3. Methods

(a) p4, L100-110. Perhaps: “Hourly northward and eastward transport … was expressed in radians:

$$dLat_t = \frac{dN_t}{R}$$
$$dLon_t = \frac{dE_t}{R\cos\left(\pi \frac{Lat_t}{180}\right)}$$

Then add “The latitude and longitude at time $t+1$, given the latitude and longitude at time $t$ and the above derivatives, was calculated as follows:

$$Lat_{t+1} = Lat_t + dLat_t * \left(\frac{180}{\pi}\right)$$
$$Lon_{t+1} = Lon_t + dLon_t * \left(\frac{180}{\pi}\right)$$

I don’t think Eqs 1-4 ever get referred to and they do not need to be numbered. I found the use of 0 and 1 confusing as this procedure is carried out at every time step.

(b) p5, L145. The material beginning “The model was run” belongs in a section labelled Fitting the SSM, not in the SSM model description section. It might be more appropriate to name Section 2.2.2 SSM fitting and diagnostics.