Review of:

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

July 16, 2021

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

1 Overall comments

Developing a quantitative, explanatory model that links total phosphorus (TP) loads in rivers that discharge into Lake Erie to the TP loads in the lake is useful for both increasing scientific understanding of the processes and for assessing potential management actions. The compilation and "wrangling" of both the lake and river TP measurements and the surface current data was no small task. The state-space model (SSM) framework seems quite appropriate given the time series nature of the data, and the spatial structure of the process model for these data is a crucial feature. The subsequent use of the fitted SSM to conduct "what if" exercises on changes in riverine TP loads shows the work's *potential* utility as a decision support tool.

I have major concerns about the formulation and fitting of the SSM and the explanation of the results.

1. Adjacency matrix: instead of calculating mean daily eastward and northward water velocity (m/sec) for a given node, using the hourly measurements at that node, why not follow the hour by hour trajectory of points across a day? On a per cell basis at hour 1 start with a point at the cell center (the node), say $(Lon_{c,1}, Lat_{c,1})$ and use eq'ns 3 and 4 to advance that point to its longitude and latitude at hour 2, $(Lon_{c,2}, Lat_{c,2})$. Then for whatever cell that point is in, apply eqn's 3 and 4 again to move it to $(Lon_{c,3}, Lat_{c,3})$, and so on until reaching $(Lon_{c,24}, Lat_{c,24})$.

2. A truncated normal distribution for the process model does not make sense. The state component is the "true" log TP concentration and that is not a truncated value. A truncated normal could be used for an observation model, however.

3. Given that year-specific SSMs were fit, it would be useful to compare the posterior distributions for the four process model parameters, β_{mau}, β_{rai}, β_{self}, and β_{lake}. At a minimum there needs to be summaries about the posteriors of these parameters, and some discussion and interpretation of the values are needed.

4. Related to item 3, a more comprehensive approach to fitting these data would be a hierarchical SSM. Model the slope coefficients (the β's) as random variables coming from a generating distribution that reflects between year (or environmental) variation; eg., $\beta_{mau,y} \sim \text{Normal}(\mu_{\beta_{mau}}, \sigma_{\beta_{mau}})$.

5. Nothing has been said about the estimates of the process model precision (Q) and the obs'n model precision (R). How do they vary between years? There are often weak identifiability problems with these kinds of linear Gaussian SSMs (see Auger-Méthé et al. (2016)).

6. Regarding application of the SSM to assess the effect of reducing TP loads in the Maumee River, assuming year-specific SSMs were fit, then the particular SSM that was used needs to be stated. Also related to 3, it would be good to report results for each of the year-specific SSMs to show (some of) the uncertainty in the assessments.

7. While it is good to have included the `R` code and data for fitting the SSM to the 2018 data, something needs to be said about computational time. I ran the code, but it had not completed after 14 hours, and I could not therefore verify any results.

2 Detailed technical comments

p=page, L=line.

1. Introduction, p3, L70: Why test a hypothesis of linearity? A more general aim would be to quantify the nature of a relationship, be it linear or nonlinear.

2. Methods

   (a) p3, L73-79. Is there overlap in the sampling on Lake Erie? For example, do two agencies collect data on the same cells?

   (b) p3, L81. By station is that just referring to the rivers, or does that include the lake? When multiple samples were collected from a "station on a single day, how much variation was there in the measure value? Such information could be used in the observation model of the SSM.

   (c) pp 3-4, L84-100. It would help to create an example figure showing the geometry (or trigonometry) underlying the calculation of a particle's change in position at time t to time $t+1$. For example, draw an x-y plot with points p_t and p_{t+1} with coordinates x=longitude and y=latitude. Draw a right triangle, where the vertical and horizontal sides are parallel to the y-axis and x-axis respectively, and the hypotenuse connects p_t and p_{t-1}. So the vertical side indicates the "northerly" movement, the horizontal side indicates"easterly" movement. (For what it's worth, I find thinking of this as a step direction and step size process—as is done in animal movement modeling.)

   (d) p3, L89. A refinement would be fit a spatially smooth velocity map rather than use values at "nodes", though I'm not sure how much practical effect that would have.

   (e) p4, L101. Does anything need to be handled differently at cells on the perimeter? Such cells could be transferring TP to cells outside the spatial window, presumably.

   (f) p4, L109. References to SSM literature are strongly recommended; e.g., Durbin and Koopman (2012); Shumway and Stoffer (2019). Discussion of the distinction between inference about latent states and fixed parameters would be good.

   (g) p4, L119. Say something about the β's and what they mean. For example, regarding β_{self} and β_{lake}, are they likely less than 1? Is there some loss of TP between a source point k and an end point n from one day to the next? Is any sort of seasonality expected in the river values? This would make the assumption of constant β_{mau} and β_{rai} suspect.

   (h) p5 L126-128: The joint prior described for β_{mau} and β_{rai} does not exactly match what the `R` code indicates:

$$
\begin{aligned}
\mu_{\beta_{river}} &\sim \text{Normal}(0, 1/\sqrt{0.01}) \\
\tau_{\beta_{mau}} &\sim \text{Gamma}(0.001, 0.001) \\
\tau_{\beta_{rai}} &\sim \text{Gamma}(0.001, 0.001) \\
\beta_{mau} &\sim \text{Normal}\left(\mu_{\beta_{river}}, 1/\sqrt{\tau_{\beta_{mau}}}\right) \\
\beta_{rai} &\sim \text{Normal}\left(\mu_{\beta_{river}}, 1/\sqrt{\tau_{\beta_{rai}}}\right)
\end{aligned}
$$

   where the 2nd parameter in the normal is the standard deviation.

   (i) p5, L137-138: Why is cross-validation needed to make comparisons of goodness of fit across years? Doesn't the Bayesian R^2 do that? Cross-validation is more often used for model selection, which is the focus of both the referenced Vehtari et al and the Piironen and Vehtari papers.

   (j) p6,L155-164: I cannot tell what is being done here. Is this necessary?

   (k) p6, L171-174: Write down an equation for deflection, $d_{n,y}$, and for the normalized estimate.

(l) p6, L179-185: Are 252 regressions being fit each year? Write down what effective load, $\tilde{l}_{n,y}$ means—is it an average? Why is there a subscript n if it is measured on the Maumee? I'd be curious about identifiability issues/posterior correlations with the parameters in eq'n 11, too—seems overly complicated. Would it make sense to take average "raw" y and take the log of that?

(m) R code: The priors for precision to the Lake, Maumee River, and River Raisin are calculated as the inverse of standard deviation; shouldn't that be the variance? Also using the data to set priors is questionable.

(n) Data for R code: why does the adjacency object `use` have 4 values for the "next" cell, at time $t+1$? Only the first is used in the code.

(o) Sensitivity analysis for the priors needs to be conducted.

3. Results

(a) p7, around L185. Say something about the ranges and averages of northerly and easterly velocities (m/sec), of TP concentrations (distinguishing between Maumee River, River Raisin, and western Lake Erie), and of calculated distances moved in a single day.

(b) p7, L195-199. In addition to the summaries in Table 1, show some plots of posterior mean values for some of the cells across the 136 day period against corresponding observations. Show an example histogram (or two) of the distribution of predicted log concentrations with the observation. Presumably these R^2 are only calculated on the cells with observations. I don't understand what the cross-validated measure is doing (relates to earlier comment in methods).

(c) p7, L200-205. As mentioned above need to report out results on parameter estimates for the β (each of the 4 for all 11 years) and report out the process and observation model standard deviations ($1/\sqrt{Q}$ and $1/\sqrt{R}$, based on `R` code).

Minor editorial remarks

1. Throughout, consider using the word "cell" instead of "node" as a node is usually interpreted as a point. Node could then refer to cell center.

2. Section 2 Methods.

(a) pp 3-4. Consider creating a new subsection for the material in the first paragraph (L84-107) of the Model Description subsection, maybe naming it Construction of an Adjacency matrix.

(b) p4, L109. Emphasize that 11 different SSMs will be fit.

(c) p4, L119. The observations are *modeled* with a normal distribution; they are not *estimated* with a normal. As mentioned previously, a truncated normal (perhaps just on the left) could be used to account for measurement limitations.

(d) p5, L139. Consider renaming Section 2.2.2 "Fitting the SSM" as "SSM Fit" could be interpreted as a result not a method.

(e) p5,L133: Perhaps move text beginning with "The model was run.." into Section 2.2.2. Could delete/move material in the sentence beginning "The efficacy of.." as it is redundant with material on L140-151.

(f) p5, L136. "efficacy" seems an odd choice, why not goodness of fit?

(g) p5,L137: The Vehtari, et al., 2017 paper does not refer to Bayesian R^2. Give a mathematical definition of the R^2 here: not clear to me what resolved and residual variances mean.

(h) p5,L138 What does "utility" mean? And what value of K was used?

(i) p5,L145: Cross-validation is meant for CV not coefficient of variation?

(j) p5, L154: What does preferentially mean?

(k) p6, L165: Instead of Model Experimentation, wouldn't Model Usage or Application make more sense here (and elsewhere)?

(l) 6, L166: As said previously, the SSM for which year was used?

3. Section 3 Results.

   (a) p7, L187-193. Re: the degree of missing data, I think it would be easier to follow by first saying how many space-time cells there are and then give the number with data: "For the Lake, there 252*11*136 = 376,992 cells of which 1218 had data, and for the two rivers, there were 2*11*136 = 2992 cells, of which 2258 had observations". Save the discussion of the inference for cells without data till later.

   (b) p7, around L188. It would help to see an example plot that shows the spatial dist'n of cells, in a given year, that had at least one observation (see Figure 1).

   (c) Also a plot showing "source" cells and "end" cells would be good to show the adjacency. (Note: in the `R` code, `use` object has 4 values in the second dimension, but it appears that only the 1st value is used as the adjacency matrix—what are the other 3 values for?)

   (d) p7, around L188. Add a plot or two of the log(TP) concentrations. See Figure 2.

   (e) p7, L195: "efficiency" here, but "efficacy" in Methods, but would model quality or goodness-of-fit be more appropriate?

   (f) p7, L197: typo: Table 1 not Table 2.

   (g) p7, L202: Say that Figure 2 shows 2018 and Figure B.1 in Appendix B shows 2006. The captions in those figures need to indicate that the black dots are observed values.

4. p8,L222 (and p10, L294): "amending" seems an odd choice: to amend would mean to modify data in such a way that the modified data are an improvement.

5. p8, L231: Was the notation k for the derived adjacency matrix used before? k was just the "source" cell from time t which feed a "sink" cell at time $t + 1$?

References

Auger-Méthé, M., Field, C., Albertsen, C. M., Derocher, A. E., Lewis, M. A., Jonsen, I. D., and Flemming, J. M. (2016). State-space models' dirty little secrets: even simple linear gaussian models can have estimation problems. *Scientific reports*, 6(1):1–10.

Durbin, J. and Koopman, S. J. (2012). *Time series analysis by state space methods*. Oxford university press.

Shumway, R. and Stoffer, D. (2019). *Time series: a data analysis approach using R*. CRC Press.

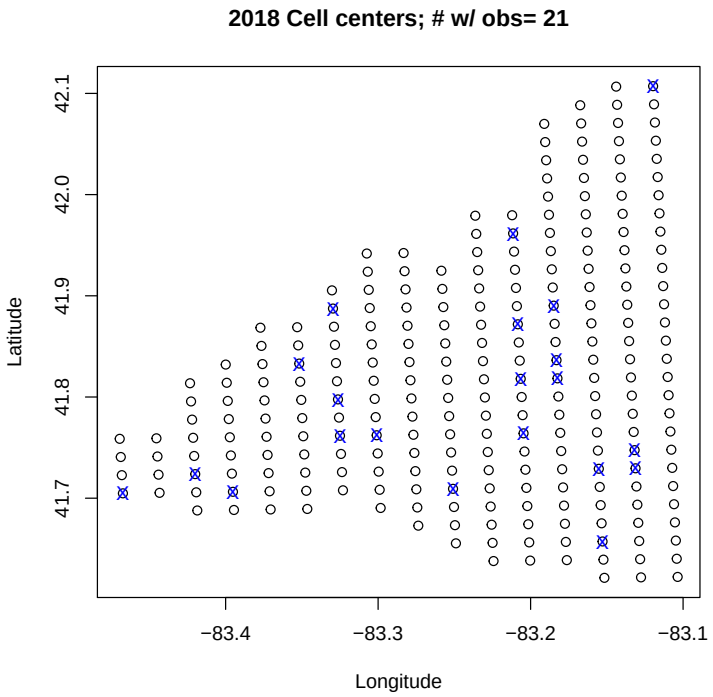

Figure 1: Cells with any TP data in 2018 marked in blue.

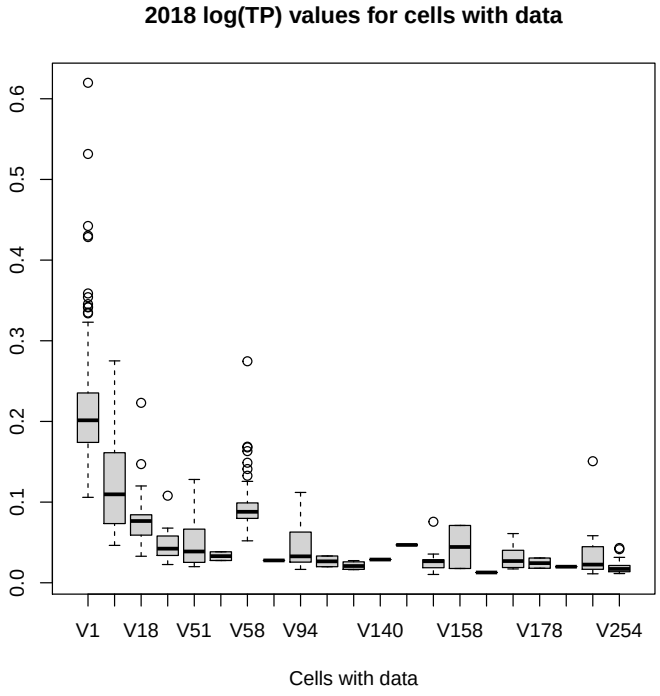

Figure 2: TP data for 2018 by cell.