Optimization of Sampling Design to Determine the Spatial Distributions of Emerging Contaminants in Estuaries

SETAC 2018

David R. Katz[1], Anne Kuhn[1], Julia C. Sullivan[2], Mark G. Cantwell[1]

[1]US Environmental Protection Agency, ORD, NHEERL, AED, Narragansett, RI; [2]Oak Ridge Institute for Science and Education, Narragansett, RI

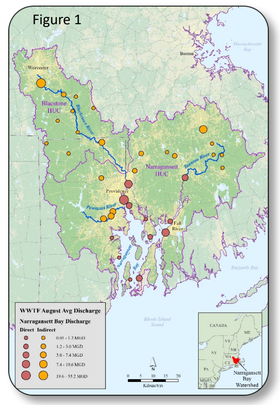
United States Environmental Protection Agency

www.epa.gov

Introduction

- Narragansett Bay (NB) is a coastal plain estuary with an area of 378.6 km^2 and an average depth of 7.6 m (**Figure 1**)
  - Several rivers discharge over 45 m^3/s of freshwater on average
  - Watershed consists of two HUC8 regions (01090003 & 01090004) with an area of 4,799 km^2
  - 12 waste water treatment plants (WWTPs) discharge directly to NB, with 35 total within the watershed
- Most studies to date in NB have not used statistically based sampling designs
- Statistically based study designs are less susceptible to bias
  - Stations are often chosen based on local knowledge or precedent, introducing bias
  - Easier to compare between regions/nationally when studies share same design principle
- Determination of optimal station numbers to maximize results while minimizing costs is hard
  - Too few stations result in high errors in spatial predictions
  - Too many stations wastes limited resources
- Here a statistically based sampling design coupled to a probabilistic spatial model was developed using pseudo-persistent compounds discharged in WWTP effluents

Figure 1



Methods

- Prior measurements of salinity (n = 92) in upper NB were used to evaluate spatial models
  - Models evaluated by cross-validation to the observations (**Figure 2**)
    - Empirical Bayesian kriging (EBK)
    - Diffusion interpolation with barriers (DIB)
    - Ordinary kriging (OK)
    - Log transformed ordinary kriging (LOK)
    - Universal kriging (UK)
- Samples (0.5 L) were extracted by Oasis HLB (Waters Corp) and analyzed by UPLC-MS/MS
- Results modeled at varying spatial densities to test and determine optimal station number
- When optimum station number was determined, stations chosen randomly within tessellated hexagonal grid (**Figure 3**)
- Spatial model error results from cross validation calculations for each compound analyzed by desirability index
  - Desirability index parameters were between 0–1 and calculated from generalized functions with target values and/or upper and lower limits when applicable:
    - Root mean square standardized error (RMSSE) closest to 1
    - Minimum root mean squared error (RMSE)
    - Average standard error (ASE) close to RMSE
    - R^2 closest to 1
  - The desirability index was calculated from the geometric mean of the individual parameters
- Desirability (D) as defined by Derringer and Suich (1980)
  - Routinely across industries when optimizing several different variables to select the most optimal conditions
  - Apply the same logic to evaluate modeling performed at varying spatial densities
  - Assess tradeoffs between high and low density sampling designs represented in cross-validation error analysis
  - Root Mean Squared Standardized Error (RMSSE) closest to one (d_1). Where T_1=1, L_1 = 0.8, U_1 = 1.2, and s=t=2 (**Figure 4a**)
  - Minimal Root Mean Squared Error (RMSE, d_2). Where T_2=0.01, U_2 = 1, and r=1 (**Figure 4b**)
  - Average Standard Error (ASE) close to Root Mean Square Error (RMSE) (d_3). Where T_3 = RMSE, L_3 = (RMSE-0.1), U_3=(RMSE+0.1), s=t=1 (**Figure 4c**)
  - R^2 closest to one (d_4). Where T_4=1, L_4 = 0.60, and r=1 (**Figure 4d**)

$$D = (d_1 d_2 d_3 d_4)^{1/4}$$

Figure 3

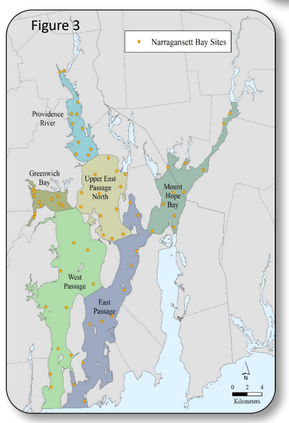
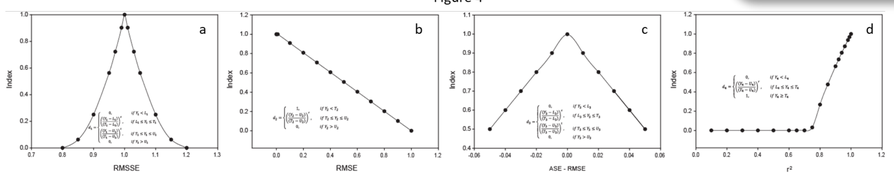


Figure 4



Results

Figure 6

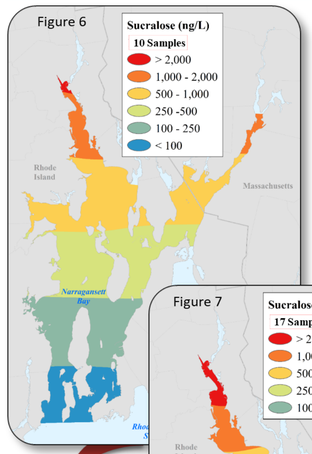


Figure 7

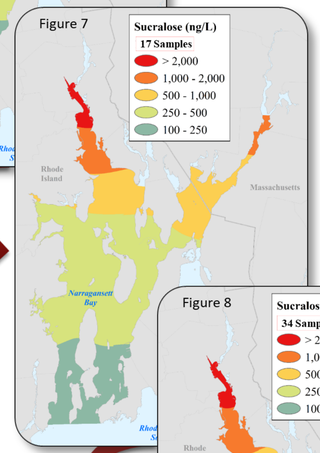


Figure 8

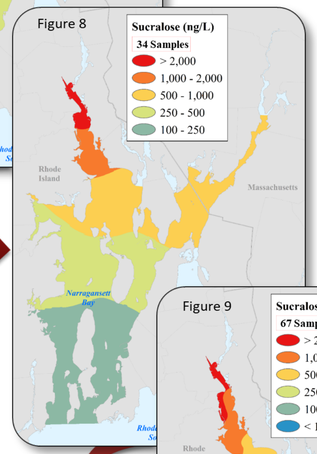


Figure 9

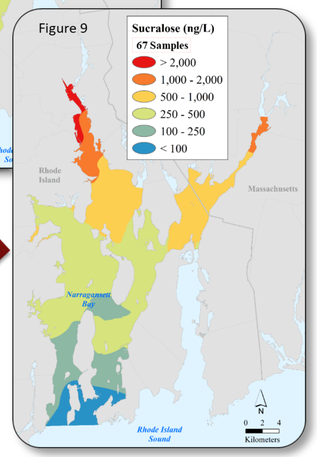
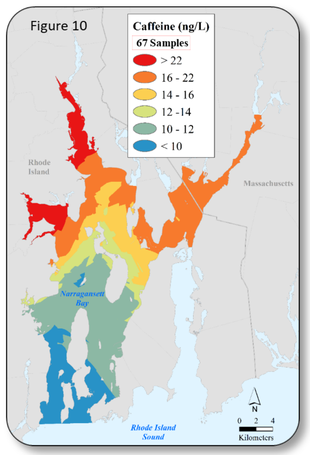
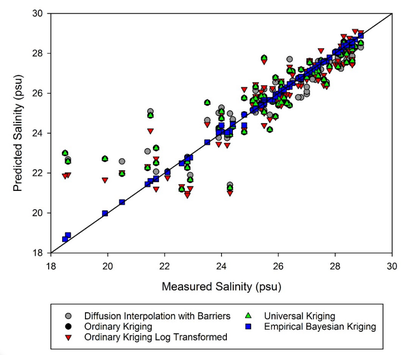


Figure 10



- EBK model clearly out-performed all other models (Figure 2)
- Optimum station number (n = 67) was estimated from this data and sampling stations were chosen (Figure 3)
- Prediction error statistics calculated from cross-validation of the spatial model to the data, examples for sucralose and caffeine shown (**Tables 1 and 2**)
- Carbamazepine concentration ranges were quite different from data collected the previous year (**Figure 5**)

Figure 2



- A desirability index used to interpret the error statistics (**Table 3**) showed peak performance of the model is approximately 34 stations, with 6 of 11 compounds having the highest desirability index values
- Sucralose, the most recalcitrant of the compounds used, showed the best model behavior (**Figures 6-9**)
- Labile compounds, i.e. caffeine, performed less well (Table 3 and **Figure 10**)

Table 1. Sucralose cross validation errors and desirability index parameters

# stations	Root Mean Square Standardized prediction error (RMSSE)	Root Mean Squared (RMSE)	Average Standard error (ASE)	R^2	RMSSE closest to One	Minimal RMSE	ASE close to RMSE	R^2 closest to 1	Desirability Geo Mean
67	1.037	0.282	0.263	0.892	0.66	0.73	0.81	0.73	0.73
50	1.071	0.219	0.181	0.928	0.42	0.79	0.62	0.82	0.64
34	0.973	0.183	0.188	0.951	0.75	0.83	0.95	0.88	0.85
25	1.115	0.511	0.444	0.679	0.18	0.49	0.33	0.20	0.28
17	1.002	0.297	0.322	0.829	0.98	0.71	0.75	0.57	0.74
10	0.958	0.46	0.431	0.82	0.62	0.55	0.71	0.55	0.60

Table 2. Caffeine cross validation errors and desirability index parameters

# stations	Root Mean Square Standardized prediction error (RMSSE)	Root Mean Squared (RMSE)	Average Standard error (ASE)	R^2	RMSSE closest to One	Minimal RMSE	ASE close to RMSE	R^2 closest to 1	Desirability Geo Mean
67	0.968	0.508	0.531	0.467	0.71	0.50	0.77	0.00	0.00
34	0.957	0.352	0.376	0.612	0.62	0.65	0.76	0.03	0.31
17	0.981	0.561	0.57	0.081	0.82	0.44	0.91	0.00	0.00
10	0.973	0.581	0.599	0.001	0.75	0.42	0.82	0.00	0.00

Figure 5

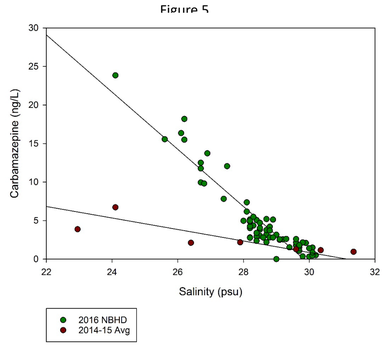


Table 3. Desirability index for all compoundsta

# stations	Salinity	Sucralose	Caffeine	Metoprolol	Valsartan	Carbamazepine	Sulfamethoxazole	Atenolol	Trimethoprim	Diltiazem	Losartan
67	0.66	0.73	0.00	0.67	0.90	0.79	0.64	0.41	0.69	0.59	0.81
34	0.69	0.85	0.31	0.79	0.83	0.72	0.51	0.63	0.64	0.66	0.60
17	0.47	0.74	0.00	0.35	0.51	0.63	0.38	0.48	0.54	0.00	0.66
10	0.49	0.60	0.00	0.00	0.18	0.24	0.00	0.54	0.00	0.00	0.66

Conclusions

- EBK was the best performing spatial model.
- Optimum station numbers for all of NB was approximately 34 for sucralose, the recalcitrant example compound used.
- Decrease is spatial accuracy minimal for recalcitrant compounds with fewer than 34 stations.

Acknowledgments

The authors thank Mrs. Patricia DeCastro for poster preparation and Mr. Michael Charpentier for preparation of GIS figures (GDIT). Mention of trade names does not constitute endorsement or recommendation for use. The views expressed in this poster are those of the authors and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency.

GENERAL DYNAMICS Information Technology