OS41A-0160

A statistical model for water quality predictions from a river discharge using coastal observations

Sung Yong Kim†, Eric J. Terrill, and Bruce D. Cornuelle

Scripps Institution of Oceanography, La Jolla, CA 92093-0213

syongkim@mpl.ucsd.edu†

1 Abstract

Understanding and predicting coastal ocean water quality has benefits for reducing human health risks, protecting the environment, and improving local economies which depend on clean beaches. Continuous observations of coastal physical oceanography increase the understanding of the processes which control the fate and transport of a riverine plume which potentially contains high levels of contaminants from the upstream watershed.

A data-driven model of the fate and transport of river plume water from the Tijuana River has been developed using surface current observations provided by a network of HF radar operated as part of a local coastal observatory that has been in place since 2002. The model outputs are compared with water quality sampling of shoreline bacteria indicator, and the skill of an alarm for low water quality is evaluated using the receiver (or relative) operating characteristic (ROC) analysis. In addition, the statistical analysis of beach closures in comparison with environmental variables is also discussed.

2 Observations

The South bay region has been influenced by the contaminated water from several resources: Tijuana River, San Diego Bay, and Los Buenos Creek, and other non-point source storm water runoff [1].

Seventeen water quality (WQ) sampling stations along the coast of south San Diego (South Bay region) are shown in Figure 1, which are chosen from about 110 WQ sampling stations in Southern California. WQ sampling has been conducted manually in every day or week. The archived data cover from January 1996 to March 2007.

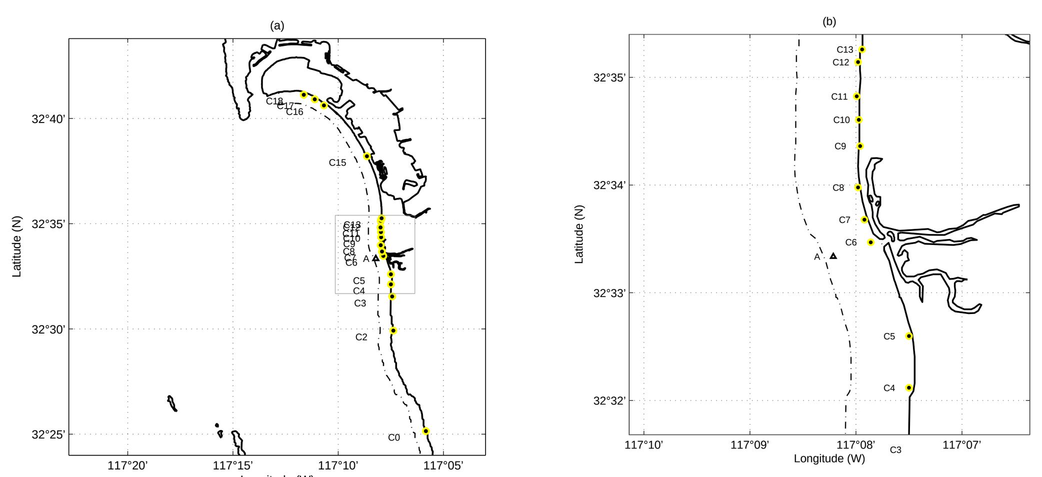

Figure 1: (a) San Diego County water quality shore stations and (b) its closed-up for Tijuana river mouth area. The releasing points of particles in random walk model are the Tijuana river mouth (A), and the particles are counted within 1 km cell from the coastline (dot curve). The water quality sampling stations along the coast of south San Diego (C0, C2-C13, and C15-C18) are indicated.

The water quality indicator (g) is a binary value for the contamination of the sampling area – $\mathcal{C}$ (clean) or $\mathcal{D}$ (contaminated), and is defined as based on the observations:

$$g = g(c_1, c_2, c_3, t_d)\,, \tag{1}$$

where c_1, c_2, and c_3 are water quality control criteria. c_1 is *Total Coliform* (CFU: colony forming units), c_2 is *Fecal Coliform* (CFU), and c_3 is *Enterococcus* (MPN - most probable number of colony forming units). t_d is the duration that the water quality sampling is valid, and will be determined later. The criteria of the $\mathcal{D}$ condition are

$$g = \{\, g \mid c_1 > 10000, c_2 > 400, c_3 > 104, \left(\frac{c_2}{c_1} > 0.1\right) \cap (c_1 > 1000)\}, \tag{2}$$

where c_1, c_2, and c_3 are the amount in 100 (ml).

The water quality sampling data for four years (April 2003-March 2007) are shown in Figure 2(a). All available sampling data are indicated with black triangle, and the contaminated conditions ($\mathcal{D}$) are presented with red one. $\mathcal{D}$ condition is highly correlated with the wet weather season (November-March).

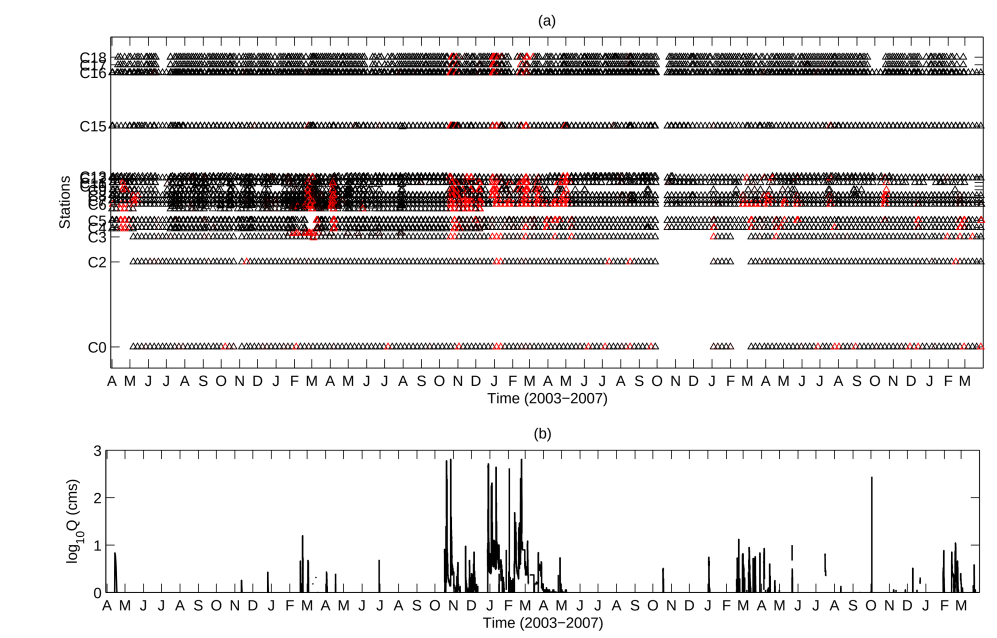

Figure 2: (a) Water quality sampling data at stations along the south San Diego. The red color indicates the contaminated condition ($\mathcal{D}$), and the black one means clean ($\mathcal{C}$) with the criteria in Equation 2. (b) River flow flux (log scale).

The surface currents are objectively mapped based on the covariance matrix of the four year hourly data [2]. The uncertainty of the surface current field due to the regularization is 8.6 cms^{-1}. The surface currents within 1 km cell are projected along the shoreline, which are referred as to the along-shore currents here.

The CTD cast as a part of the South Bay outfall observation have conducted nearly monthly: temperature, salinity, density, and chlorophyll, and some of location are shown as dots in Figure 3(a). As an example of Tijuana river plume during a heavy rain in January 2005, the linearly interpolated salinity at surface is shown in Figure 3(a). Lower salinity water (less than 32.8 permiles) spread at surface as a jet shape and extended near 6 km from the coastline. The satellite observation images of the Chlorophyll-a in the observation domain on January 6 of 2005 is shown in Figure 3(b). The higher density tongue stretched as a jet aligned with the salinity surface map.

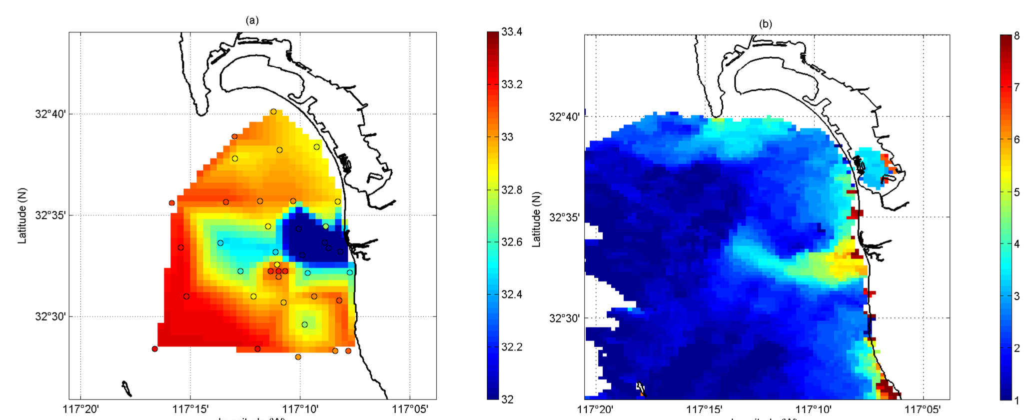

Figure 3: (a) Linearly interpolated salinity (permiles) at surface during the one of heavy rain events on January 2-6, 2005. (b) Chlorophyll-a image on January 6 1939Z 2005 (GMT).

3 Theoretical backgrounds

3.1 Statistical trajectory model

The number of the particles (f) arriving within the sampling bin of the model is a function of several parameters:

$$f = f(l, t_p, w\,;\, n, \mathbf{x}_0, \mathbf{u}, \epsilon), \tag{3}$$

where l is the lifetime of particles, t_p is the specific time period to be considered as the independent events (e.g., river flood period or wet season), w is the width of the sampling bin along the coastline, n is the number of particles to be released at a unit time, $\mathbf{x}_0$ is the initial position where particles are released, $\mathbf{u}$ is the surface vector current field, and ϵ is the diffusion parameter in random walk model. The location of the particle releasing of the random walk model (A in Figure 1) and the imaginary coastal boundary cell 1 km apart from the coastline (A dot line in Figure 1) are indicated.

Lagrangian time integration using the surface current field based on the random walk is

$$\mathbf{x}(t) = \int_{t_0}^{t} \left(\mathbf{u}(t') + \epsilon e^{i\theta}\right) \mathrm{d}t' + \mathbf{x}(t_0) = \sum_k \left(\mathbf{u}(t'_k) + \epsilon e^{i\theta}\right) \Delta t'_k + \mathbf{x}(t_0), \tag{4}$$

where $\mathbf{x}(t) = x(t) + iy(t)$, $\mathbf{u}(t) = u(t) + iv(t)$, and $\theta = \theta(t)$ is the variable to follow the uniform distribution ($0 \le \theta \le 360$ degrees). The random walk model (RWM) is chosen for the similarity of the Lagrangian statistics in the coastal region [3]

A snapshot of the particle trajectory model and the concentration of the particle within the near coast cell is is shown in Figure 4.

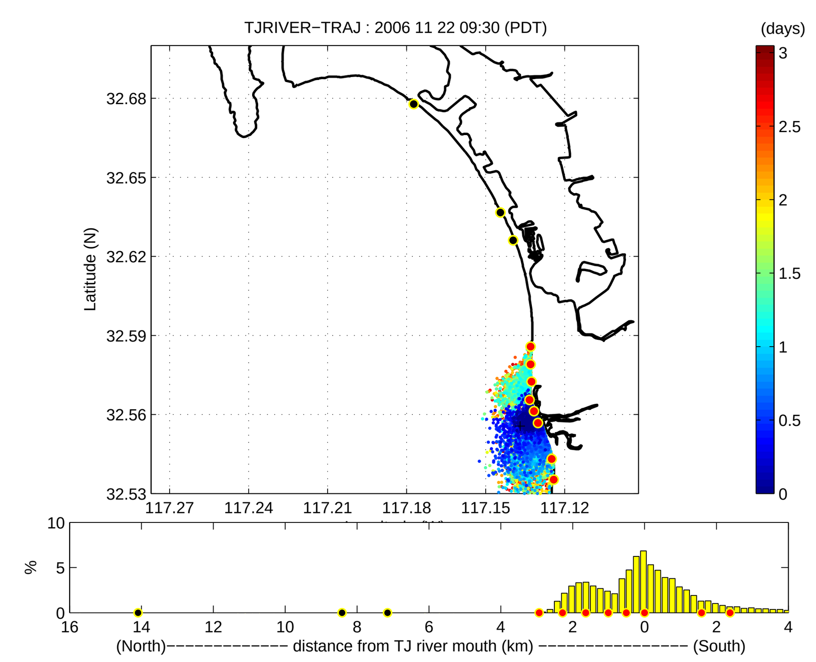

Figure 4: A snapshot of the particle track model and the histogram of the particle concentration within the near coast cell. The particles have 3 days life time, and 50 particles are released at every hour. The near real-time particle track model has been in operation since March 2006. http://sdcoos.ucsd.edu/data/particles/IB/

3.2 Receiver Operating Characteristic (ROC) analysis

The binary classifier system about the beach closure is evaluated with the diagnosis based on the RWM and the indicator of the water quality sampling data ($\mathcal{C}$ or $\mathcal{D}$). A contingency table for four cases is shown in Figure 5 [4]. The positive and negative are considered as the signal and noise events. True-Positive (TP) and False-Negative (FN) occur when the diagnosis and the event agree, and False-Positive (FP) and True-Negative (TN) occur when the diagnosis and the event disagree.

		Event	
		Positive	Negative
Diagnosis	Positive	True-Positive (TP)	False-Positive (FP)
	Negative	False-Negative (FN)	True-Negative (TN)

Figure 5: A contingency table for two alternative events and two diagnosis. The positive and negative represent $\mathcal{D}$ and $\mathcal{C}$ conditions, respectively, in the Receiver Operating Characteristic (ROC) analysis.

The true-positive proportion (sensitivity, α) and the false-positive proportion (1-specificity, β) are defined with a given threshold (λ):

$$\alpha(\lambda) = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = P(g \equiv \mathcal{D} \mid f \le \lambda), \tag{5}$$

$$\beta(\lambda) = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} = P(g \equiv \mathcal{C} \mid f \le \lambda). \tag{6}$$

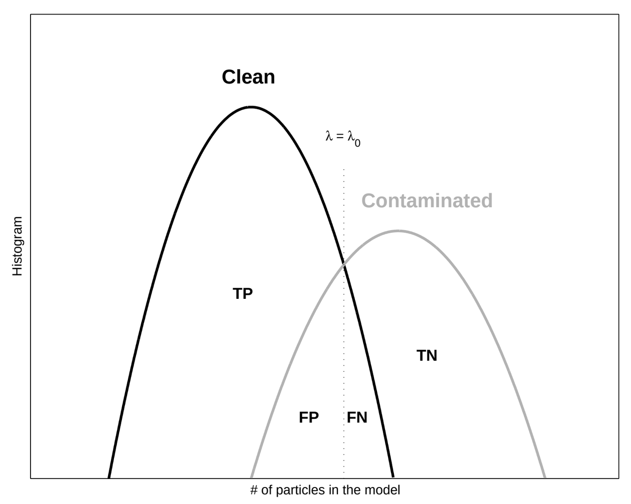

Figure 6: An ideal curve to show variables in the Receiver Operating Characteristic (ROC) analysis. λ_0 is a threshold value, and TP, FP, FN, and TN are a function of this threshold (λ).

Each point of the ROC curve represents a pair of α and β in a given threshold (λ). The area (A, AOC) under the ROC curve represents how well the diagnosis system distinguish between the positive and negative cases as the discrimination.

$$A = \int_0^1 \alpha \,\mathrm{d}\beta \approx \sum_k \alpha(\beta_k) \Delta\beta \tag{7}$$

4 Results

The ROC analysis is applied to several rain events during wet season. The number of particle at the near sampling station and the water quality indicator are sorted, then the FP and TP are calculated as a function of the threshold (λ) (Figure 6).

The particle concentration profile in time, the rainfall flux, and the ROC curve are shown as an example in Figure 7. The area below the ROC is about 74%. Using four years water quality samplings and the RWM during several wet seasons (rain events), the average accuracy of the alarm model is about 70%, which is a reasonable classification.

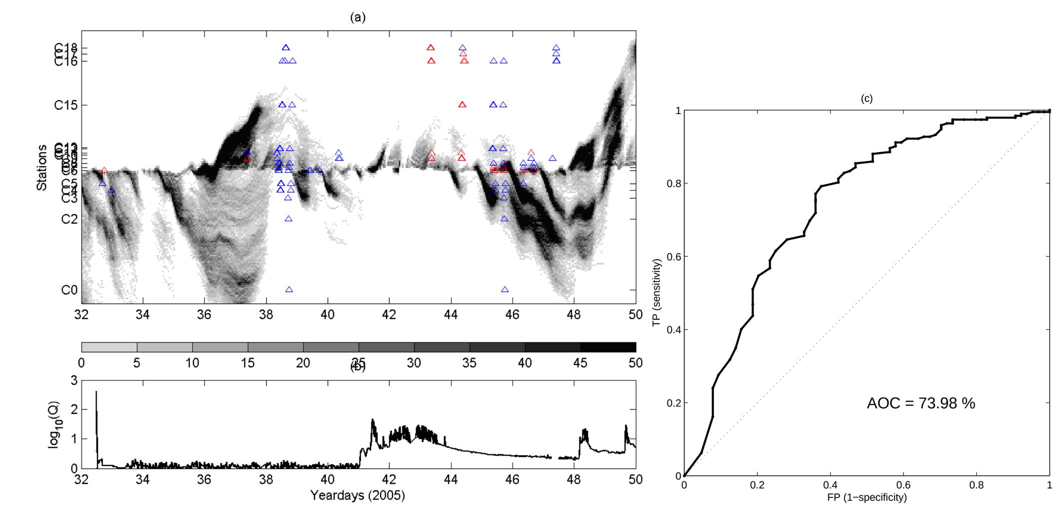

Figure 7: (a) The particle concentration within 1 km cell and the water quality sampling indicator. (b) Rainfall flux (log scale). (c) ROC curve.

5 Discussions

The binary indicator by the water quality samplings and the coastal observations – surface currents, rain fall measurements, CTD cast data, and satellite images – are used to build a data driven statistical model for the coastal water quality prediction. The random walk model using the surface current observation provide the spatial probability map of the fate of the contaminants. Four years data are examined, and the accuracy of the particle tracking model during several rain events is estimated as about 70%.

The representational error of of the sparse water quality sampling data should be considered consistently with the other oceanographic observations in their spatial and temporal resolution.

Acknowledgement

We would like to acknowledge the generous sponsors of this work. SYK is supported through a graduate student fellowship provided by the California Institute of Telecommunication and Information Technology (CALIT2). EJT is partially supported through funding provided by the Office of Naval Research (ONR), NOAA, the State of California, Coastal Ocean Currents Monitoring Program (COCMP). BDC is supported by both NOAA and COCMP.

References

[1] D. James, A. Groce (California and the World Ocean 2006, 2006).

[2] S. Y. Kim, E. J. Terrill, B. D. Cornuelle, *J. Geophys. Res.* **112** (2007).

[3] A. Griffa, K. Owens, L. Piterbarg, B. Rozovskii, *J. Mar. Res.* **53**, 371 (1995).

[4] J. A. Swets, *Science* **240**, 1285 (1988).