Ensemble modeling of *E. coli* in the Charles River, Boston, Massachusetts, USA

F.L. Hellweger
Department of Civil and Environmental Engineering, Northeastern University, 400 Snell Engineering Center, Boston, Massachusetts, USA (E-mail: ferdi@coe.neu.edu)

Abstract A case study of ensemble modeling of *Escherichia coli* (*E. coli*) densities in surface waters in the context of public health risk prediction is presented. The output of two different models, mechanistic and empirical, are combined and compared to data. The mechanistic model is a high-resolution, time-variable, three-dimensional coupled hydrodynamic and water quality model. It generally reproduces the mechanisms of *E. coli* fate and transport in the river, including the presence and absence of a plume in the study area under similar input, but different hydrodynamic conditions caused by the operation of a downstream dam and wind. At the time series station, the model has a root mean square error (RMSE) of 370 CFU/100mL, a total error rate (with respect to the EPA-recommended single sample criteria value of 235 CFU/100mL) (TER) of 15% and negative error rate (NER) of 30%. The empirical model is based on multiple linear regression using the forcing functions of the mechanistic model as independent variables. It has better overall performance (at the time series station), due to a strong correlation of *E. coli* density with upstream inflow for this time period (RMSE = 200 CFU/100mL, TER = 13%, NER = 1.6%). However, the model is mechanistically incorrect in that it predicts decreasing densities with increasing Combined Sewer Overflow (CSO) input. The two models are fundamentally different and their errors are uncorrelated ($R^2 = 0.02$), which motivates their combination in an ensemble. Two combination approaches, a geometric mean ensemble (GME) and an "either exceeds" ensemble (EEE), are explored. The GME model outperforms the mechanistic and empirical models in terms of RMSE (190 CFU/100mL) and TER (11%), but has a higher NER (23%). The EEE has relatively high TER (16%), but low NER (0.8%) and may be the best method for a conservative prediction. The study demonstrates the potential utility of ensemble modeling for pathogen indicators, but significant further research is needed to establish the approach for the Charles River, as outlined in the paper.

Keywords Charles River; consensus model; empirical model; ensemble model; *Escherichia coli* (*E. coli*); mechanistic model; pathogen indicator

Introduction Effective management of public health risk at primary and secondary contact areas requires accurate prediction of the potential for fecal contamination, which is typically quantified using indicator organisms like *Escherichia coli* (*E. coli*). Present laboratory methods for measuring bacterial indicators require 24 hours of incubation, and there can be a poor correlation between the density on the sampling day and the next day when the results are available (Boehm *et al.*, 2002; Olyphant and Whitman, 2004). Modeling constitutes a potential alternative to monitoring, and numerous case studies of mechanistic and empirical models are available in the literature. Connolly *et al.* (1999) used a three-dimensional hydrodynamic and water quality model to simulate pathogen fate and transport in Mamala Bay. McCorquodale *et al.* (2004) used a three-dimensional hydrodynamic model to simulate fecal coliform densities in Lake Pontchartrain. Liu *et al.* (2006) used a two-dimensional finite-element hydrodynamic and water quality model to simulate *E. coli* and enterococci at the Indiana shoreline of Lake Michigan. Olyphant and Whitman (2004) relate *E. coli* at a Chicago beach to wind direction and speed, rainfall, insolation, lake stage, water temperature and turbidity. Nevers and Whitman
(2005) relate *E. coli* at a number of south Lake Michigan beaches to wind direction, wave height and period, precipitation, lake chlorophyll and turbidity. Eleria and Vogel (2005) relate fecal coliform to numerous variables, including antecedent rainfall, hydrologic and meteorologic variables in the lower Charles River (also this study area).

Since models, by definition, are simplified approximations of the real system, all will have some deficiencies and associated errors. For two or more models that are fundamentally different and the errors are uncorrelated, a better prediction may be attainable using an ensemble of the models. Ensemble modeling has become well accepted and is now common practice in weather, climate and hurricane track prediction (see review below). The adoption of this approach for pathogen indicators is therefore a natural progression. This paper presents a case study of ensemble modeling of *E. coli* in Boston’s Charles River. First, some background on ensemble modeling is presented. Then the study site, data and models (mechanistic, empirical, ensemble) are presented followed by the conclusions.

**Ensemble modeling**

**Theory**

“Ensemble modeling” is used in the literature to describe various approaches, including varying the initial conditions, forcing functions or parameterizations of one model (e.g. Monte Carlo simulations), as well as combining the output of different models. In this paper, ensemble modeling refers to the later, which is also refereed to as multi-model ensemble modeling. Ensemble modeling is based on the fact that “two or more inaccurate but independent predictions of the same future events may be combined in a very specific way to yield predictions that are, on the average, more accurate than either or any of them taken individually” (Thompson, 1977). The theoretical basis for this was presented by Thompson (1977), who also presents theoretical equations to predict the increase in model skill that can be expected by combining models. Additional theory is presented by Fraedrich and Leslie (1987).

**Examples from other disciplines**

Ensemble modeling has been applied successfully in several other areas, including weather, climate and tropical cyclone (TC) track prediction. A non-exhaustive but illustrative collection of studies is discussed here. Sanders (1973) found that a simple average of daily temperature and precipitation predictions for Boston made by students, faculty and staff at the Massachusetts Institute of Technology (MIT) generally outperformed predictions made by individuals. Fraedrich and Leslie (1987) combined a mechanistic and an empirical model of daily precipitation using an optimal linear combination (OLC) technique. The performance of the ensemble model was better than that of either of the individual models. Vislocky and Fritsch (1995) combined two mechanistic models of temperature, wind, cloud cover and precipitation using a simple average (OLC showed no meaningful improvement). The consensus model was found to be superior for all variables. Interestingly, the authors quantify the improvement in terms of equivalent scientific advancement (2–8 years). Leslie and Fraedrich (1990) combined a mechanistic and an empirical model of TC tracks in the Australian tropics using their OLC technique, which allows for assigning different weights to different models at different times. The coefficients obtained using standard multiple linear regression techniques and a training data set put more weight on the empirical model early in the prediction and less later in the prediction. The ensemble model outperformed the two individual models and two other prediction models on an independent data set. Goerss (2000) found that a simple average of two or three hurricane track models performs better than the individual models. Krishnamurti et al. (1999) constructed a “superensemble” model by combining...
seven weather and climate models using weight coefficients based on individual model performance. The ensemble was able to predict weather, seasonal climate and TCs better than all individual models and an ensemble based on simple averaging. Rajagopalan et al. (2002) developed a Bayesian methodology for combining predictions from different models, used it to combine three global climate models (GCMs), and showed that the ensemble performs better than individual models and a straight average ensemble for temperature and precipitation.

**Illustrative example**
Consider that the objective is to predict the density of *E. coli* at a certain location, and that model performance is quantified using the root mean square error (RMSE) over the time series. Note that various other skill measures, like negative, positive and total error rates (see below), sensitivity and specificity (Morrison et al., 2003), or mean relative deviation (Hellweger et al., 2004) are also sometimes used to quantify model performance. Assume that there are two models that have equal skill and are uncorrelated (Figure 1a&b). Model 1 gets the first, but misses the second peak. Model 2 gets the second, but misses the first peak. The RMSE of both models is \((9 \times 0^2 + 1 \times 400^2)/10\)^{0.5} = 126 (all densities are presented in CFU/100mL). Now consider a simple linear combination of the two models into an ensemble:

\[ C_{ME} = a_{M1}C_{M1} + a_{M2}C_{M2} \]  

where \( C_{ME}, C_{M1} \) and \( C_{M2} \) are the densities from the ensemble, models 1, 2, respectively, and \( a_{M1} \) and \( a_{M2} \) are weighing coefficients (taken here as 0.5, i.e. simple average). The ensemble RMSE is \((8 \times 0^2 + 2 \times 200^2)/10\)^{0.5} = 89, indicating that the model performance improved (Figure 1c).

However, public health risk is often judged in a binary manner, meaning density above or below a criteria value (the EPA-recommended single sample criteria value of 235 CFU/100mL is used here). With respect to this metric the ensemble model is clearly worse than the individual models (Figure 1). More specifically, the individual models each have one false negative (Type II, model < criteria and data > criteria) error, and corresponding negative error rates (NER), which is the number of false negatives per number of positives (data > criteria), of 1/2 = 50%. The ensemble model has a negative

![Figure 1](https://iwaponline.com/wst/article-pdf/56/6/39/437885/39.pdf)

**Figure 1** Ensemble modeling illustration. (a,b) Models 1 and 2 are of equal skill with uncorrelated errors. (c) The ensemble model is based on a simple average (Equation 1). Density values are in CFU/100mL.
error rate of 2/2 = 100%. A different method of combining the models is based on either model exceeding the criteria:

$$v_{ME} = \max\{v_{M1}, v_{M2}\}$$  \hspace{1cm} (2)

where $v_{ME}$ is the ensemble exceedance (binary, 0 = below, 1 = above criteria), and $v_{M1}$ and $v_{M2}$ are the exceedances predicted by models 1 and 2, respectively. This ensemble model does not produce a density and thus can not be evaluated using the RMSE metric. However, its NER is now 0/2 = 0%, which is better than the individual models. This example illustrates that the utility of using an ensemble model will depend on the metric used for assessing performance and the approach used to construct the ensemble.

Study site and data
The study site is the lower Charles River in Boston, which extends nine miles from the Watertown Dam to the New Charles River Dam. The study focuses on a specific location, Community Boating, which is located in the heavily used (sailing and windsurfing) area between the Massachusetts Avenue and Longfellow Bridges (Figure 2a). Because of the dam at the downstream end, the basin exhibits hydrodynamic characteristics closer to a lake or reservoir than a river. Two major tributaries, Stony Brook and Muddy River, discharge just upstream of the Massachusetts Avenue Bridge on the Boston side, and previous research suggests that they constitute a major source of E. coli (Hellweger and Masopust 2007). Two other recent studies focused on the water quality of the river (Eleria and Vogel, 2005; Hellweger et al., 2006). The E. coli density at Community Boating was observed during two temporal surveys over 7 days at hourly intervals ($N = 336$). Samples were collected using an ISCO autosampler as described in detail by Hellweger and Masopust (2007). The data reveal significant structure at scales not resolved by typical monitoring programs. On 4/20/05, for example, the density rises above the criteria value and then sinks below it, within a time-span of significantly less than 24 h (Figure 3b).

Mechanistic model
The details of the mechanistic model are presented in Hellweger and Masopust (2007). Briefly, the model is a time-variable, three-dimensional coupled hydrodynamic and water quality model. The computational grid consists of 3,066 grid cells with average length...
dimension of 25 m (Figure 2a). Forcing functions include upstream and downstream boundary conditions, Stony Brook and Muddy River CSO and non-CSO discharge and wind. The model generally reproduces the observed patterns, including the presence and absence of a plume in the study area under similar input, but different hydrodynamic conditions caused by operation of the New Charles River Dam and wind, which is illustrated in Figure 2b. For the Community Boating time series data, the model has an RMSE = 370 CFU/100mL, total error rate (TER) = 15% and NER = 30% (Figure 3).

Empirical model

The empirical model is based on multiple linear regression. The general model equation is:

\[ C_E = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + e \]  

where \( C_E \) is the \( E. \ coli \) density, \( \beta_0 \) is a constant, \( \beta_1 \) through \( \beta_n \) are the regression coefficients, \( x_1 \) through \( x_n \) are the independent variables, and \( e \) is the residual error. Bacterial indicator densities can vary over a wide range of time-scales due to a number of factors including storm and sanitary sewer infrastructure changes, climate change and variability (e.g. ENSO), rainfall and wind events, including conditions preceding rainfall (antecedent rainfall), tide (diurnal and spring/neap), waves (height and period, and sunlight (including sky condition) (e.g. Boehm et al., 2002; Morrison et al., 2003; Nevers and Whitman, 2005; Eleria and Vogel, 2005). To allow for a fair comparison with the mechanistic model, the same variables that are used as forcing functions in the mechanistic model are considered, including inflow rate at Watertown Dam, Stony Brook and Muddy River non-CSO and CSO flow rate, flow rate at the New Charles River Dam and wind speed and direction. The regression coefficients for this linear model were optimized using MS Excel. The best-fit equation for \( E. \ coli \) density at Community Boating is:

\[ C_E = -420 + 23(Watertown \ Q) + 59(SB + MR \ non-CSO \ Q) - 49(SB + MR \ CSO \ Q) + 0.51(Dam \ Q) + 4.9(Wind \ speed) + 0.29(Wind \ dir.) + e \]  

The model has an \( R^2 = 0.60 \), which is comparable to the performance of the model of Eleria and Vogel (2005) at the nearby Longfellow Bridge (Figure 2a) (\( R^2 = 0.56 \)). The correlation is mostly due to Watertown flow, which by itself correlates to \( E. \ coli \) with an \( R^2 = 0.56 \). The other independent variables all have \( R^2 < 0.1 \). The model captures the long-term trend in \( E. \ coli \), but misses some of the small-scale features (like the peak on 4/20, see Figure 3b), which are due to processes that are too complex to be captured by the simplified regression equation. Also, because this empirical model is not

![Figure 3](https://iwaponline.com/wst/article-pdf/56/6/39/437885/39.pdf)
constrained by fundamental laws (i.e. mass balance) it produces negative densities at some times. The negative regression coefficient for the Stony Brook and Muddy River CSO flow indicates decreasing E. coli with increasing CSO flow, which is inconsistent with a mechanistic interpretation of the system. As shown in Table 1, the RMSE = 200 CFU/100mL, which is better than the mechanistic model. The PER (20%) is higher than the mechanistic model, which is mostly due to overpredicting densities towards the end of time series 1 (Figure 3a). However, the NER (1.6%) is lower and the TER (13%) is slightly lower than that for the mechanistic model. The TER is on the high side, but within the range of other empirical models (TER = 2.0 – 14, Table 1). As suggested by the $R^2$ values, correlating E. coli density to Watertown flow produces similar performance metrics (Table 1).

### Ensemble model

As expected, neither the mechanistic nor the empirical model is perfect, but they appear to complement each other. The empirical model has a better performance for the long-term trends, whereas the mechanistic model captures more of the short-term dynamics. The overall error may be reduced when an ensemble of the two models is used. Of course, if the two models would have the same deficiencies (e.g. overpredict density during CSO events), this would not result in better model performance. However, the models are fundamentally different and their errors are uncorrelated ($R^2 = 0.02$). First, a geometric mean ensemble (GME) is constructed using:

$$\log(C_{ME}) = a_M \log(C_M) + a_E \log(C_E)$$

with $a_M = a_E = 0.5$. The GME model densities are compared to data and the other models in Figure 3. The GME model has an RMSE = 190 CFU/100mL, TER = 11%, and NER = 23%. The RMSE and TER suggest better performance than the individual mechanistic and empirical models (Table 1). However, the NER is significantly worse than the empirical model, which may be of concern if a conservative prediction is desired. To provide a more conservative prediction, an “either exceeds” ensemble (EEE) is constructed based on Equation 2. That model has a TER = 16% and NER = 0.8%. The low NER, and corresponding higher TER, indicate the conservative nature of the

### Table 1 Summary of model performance metrics(a)

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE (CFU/100mL)</th>
<th>PER (Type I) (%)</th>
<th>NER (Type II) (%)</th>
<th>TER (Type I + II) (%)</th>
<th>Total Correct (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>This study</td>
<td>370</td>
<td>6.7</td>
<td>30</td>
<td>15</td>
<td>85</td>
</tr>
<tr>
<td>Mechanistic</td>
<td>200</td>
<td>20</td>
<td>1.6</td>
<td>13</td>
<td>87</td>
</tr>
<tr>
<td>Empirical</td>
<td>(210)</td>
<td>(20)</td>
<td>(1.6)</td>
<td>(13)</td>
<td>(87)</td>
</tr>
<tr>
<td>Geometric mean ensemble (GME)</td>
<td>190</td>
<td>3.4</td>
<td>23</td>
<td>11</td>
<td>89</td>
</tr>
<tr>
<td>Either exceeds ensemble (EEE)</td>
<td>–</td>
<td>26</td>
<td>0.8</td>
<td>16</td>
<td>84</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Other studies</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Eleria and Vogel (2006) (b)</td>
<td>–</td>
<td>0.72</td>
<td>67</td>
<td>2.1</td>
</tr>
<tr>
<td>Morrison et al. (2003)(b)</td>
<td>–</td>
<td>28</td>
<td>29</td>
<td>28</td>
</tr>
<tr>
<td>Nevers and Whitman (2006)</td>
<td>–</td>
<td>0.0</td>
<td>45</td>
<td>2.0</td>
</tr>
<tr>
<td>Francy (2006)</td>
<td>–</td>
<td>9.1</td>
<td>38</td>
<td>14</td>
</tr>
</tbody>
</table>

(a) RMSE = root mean square error, PER = positive error rate, NER = negative error rate, TER = total error rate. (b) ‘Best’ logistic regression model at Longfellow Bridge, Table 8-9 in Eleria (2002). (c) For 48-h rainfall > 0.2 in; all beaches; validation dataset.
EEE model and may make this model most useful in a management program designed to protect public health.

Conclusions

Mechanistic and empirical models of \textit{E. coli} density in the Charles River were combined into an ensemble, which outperformed the individual models. This demonstrates the potential utility of ensemble modeling for \textit{E. coli} in the Charles River. Even though the ensemble approach can provide better or more conservative predictions, which will be useful in a management context, it does not contribute directly to the scientific–mechanistic understanding of the system. However, the study did highlight a shortcoming in the mechanistic model (long-term dynamics) and it suggested a cause (upstream inflow).

Further research is needed to establish ensemble modeling for \textit{E. coli} in the Charles River. This includes further calibration over a longer time span encompassing a wider range of environmental conditions (i.e. year-round). Based on the observed fluctuations in \textit{E. coli} density, the sampling frequency will need to be significantly higher than daily (e.g. hourly), but it may be reduced once the confidence in the model(s) and their ability to interpolate densities between observations increases. The inclusion of additional models into the ensemble, including the previous day’s \textit{E. coli} density (i.e. \( C_t = C_{t-1} \)) and an artificial neural network (ANN), should be explored. Different metrics (e.g. probability of exceedance) and ensemble approaches (e.g. Bayesian; Rajagopalan et al., 2002) should be evaluated. Also, a validation of the various models in a forecast scenario with explicit acknowledgement and consideration of uncertainty is critical. For this, the mechanistic model will need to incorporate forecasts of management decisions, like the operation of dams (i.e. New Charles River Dam) and gates and pumps in the sewer system (i.e. Cottage Farm CSO facility, Boston Gatehouses).

References


