Analysis of *Escherichia coli* and enterococci concentrations patterns in a Pennsylvania creek using empirical orthogonal functions

Dong Jin Jeon¹*, Yakov Pachepsky¹, Dana Harriger,², Rachael Picard², Cary Coppock¹, Edward Wells², Eun-Mi Hong³

¹USDA-ARS, Environmental Microbial and Food Safety Laboratory, Beltsville, Maryland, USA
²Wilson College, Division of Integrated Sciences, Chambersburg, Pennsylvania, USA
³Kangwon National University, School of Natural Resources and Environmental Science, Chuncheon, Republic of Korea

*Corresponding author, e-mail Dongjin.Jeon@ars.usda.gov
Supplemental material 1

Spatial anomalies are calculated by subtracting the spatial average value of logarithms of microbial concentrations at the same time from all monitoring stations. If $c_i(t)$ is the logarithms of microbial concentrations at station $i$ ($i=1,\ldots,m$) and time $t$ ($t=1,\ldots,n$), the spatial anomaly ($z_i(t)$) is calculated as:

$$z_i(t) = c_i(t) - \frac{1}{m} \sum_{j=1}^{m} c_j(t)$$

Where $j$ is an index of monitoring stations and $m$ is the number of monitoring stations. The spatial anomalies have the same average shading because the average value of logarithms of microbial concentrations for each monitoring date was removed. $z_i(t)$ is arranged into a matrix $Z$ ($n$ by $m$: $n$ monitoring times and $m$ monitoring locations) to analyze the spatial variability of the data.

The next step is to calculate the covariance matrix $V$ of the matrix $Z$ of spatial anomalies. The spatial covariance $v_{t,\tau}$ at two times $t$ and $\tau$ is calculated as:

$$v_{t,\tau} = \frac{1}{m} \sum_{i=1}^{m} z_i(t)z_i(\tau)$$

This equation sums the products of the spatial anomalies at times $t$ and $\tau$ across all stations. Then the covariance matrix $V$ ($n$ by $n$) can be found:

$$V = \frac{1}{m} Z^t Z$$

Where the subscript $t$ indicates a matrix transpose. The spatial covariance can examine how spatial anomalies are correlated between times.

The next step is to diagonalize the covariance matrix to find the eigenvectors and eigenvalues.
\[ \text{VE} = \text{LE} \] (4)

where \( L \) (n by n) is a matrix containing the associated eigenvalues \( l_{\tau,t} \) along the diagonal and zeros at off-diagonals, and each eigenvalue represents the variance explained by each EOF. \( E \) (n by n) contains the eigenvectors \( e_{\tau,t} \) of \( V \) in the column vectors, representing ECs. For more details on the procedure see (Perry and Niemann 2007, Korres et al. 2010).

The eigenvectors in matrix \( E \) are used as the weights applied to each component in matrix \( V \) to diagonalize the \( V \). This transformation rotates the original coordinate axes in a multi-dimensional space to align the data along a new set of orthogonal axes in the direction of the largest variance. Thus, the first axis or eigenvector identified by the diagonalization is oriented in the direction that explains the most covariance in the spatial anomaly dataset. The subsequent axes are constrained to be orthogonal to the other axes and explain the most remaining covariance. After diagonalization of \( V \), \( E \) and \( L \) are arranged accordingly to keep the eigenvalues in \( L \) sorted in descending order. Therefore, the portion of the variance, \( P_j \), that the jth EOF explains is:

\[ P_j = \frac{\lambda_{jj}}{\sum_{k=1}^{n} \lambda_{kk}} \] (5)

Where \( \lambda_{jj} \) and \( \lambda_{kk} \) are the eigenvalues corresponding to the jth and kth EOF. This equation was used to judge the importance of each new axis in describing the variability of the spatial anomalies by the magnitude of the associated eigenvalue.

The EOF pattern (m by n) can be calculated by projecting the \( Z \) onto the matrix \( E \) as:

\[ F = ZE \] (6)

Where the columns of \( F \) matrix represent the EOF pattern.
Supplemental Table S1. Land use in drainage areas of the stream reaches between sampling locations.

<table>
<thead>
<tr>
<th>Land use</th>
<th>Upstream of TP</th>
<th>TP to I81</th>
<th>I81 to SS</th>
<th>SS to SG</th>
<th>SG to SD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Area (ha)</td>
<td>%</td>
<td>Area (ha)</td>
<td>%</td>
<td>Area (ha)</td>
</tr>
<tr>
<td>Urban</td>
<td>694</td>
<td>7.1</td>
<td>1,565</td>
<td>18.8</td>
<td>601</td>
</tr>
<tr>
<td>Crop/pasture</td>
<td>6.4</td>
<td>0.1</td>
<td>2,054</td>
<td>24.6</td>
<td>889</td>
</tr>
<tr>
<td>Grasses</td>
<td>63</td>
<td>0.6</td>
<td>1,078</td>
<td>12.9</td>
<td>557</td>
</tr>
<tr>
<td>Forested</td>
<td>8,809</td>
<td>89.4</td>
<td>3,415</td>
<td>40.9</td>
<td>1,798</td>
</tr>
<tr>
<td>Total</td>
<td>9,853</td>
<td>97.2</td>
<td>8,344</td>
<td>97.2</td>
<td>3,876</td>
</tr>
</tbody>
</table>