Observability-based sensor placement improves contaminant tracing in river networks

Matt Bartos[1] and Branko Kerkez[1]

[1]University of Michigan

November 22, 2022

Abstract

This study presents a new methodology for identifying near-optimal sensor locations for contaminant source tracing in river networks. To establish a physical basis for the problem, we first derive a linear time-invariant (LTI) model for riverine contaminant transport using the one-dimensional advection-diffusion equation. We then formulate an optimization problem to find the sensor placement that maximizes the observability of the modeled system, and identify two heuristics for efficiently achieving this goal. Evaluating each sensor placement strategy on its ability to reconstruct initial contaminant loads from observed outputs, we find that the best sensor placement is obtained by greedily maximizing the rank of the LTI system's Observability Gramian. In addition to providing the best approximate reconstruction of internal states, this strategy makes it possible to perfectly recover any initial contaminant load while only monitoring a small subset of river branches (˜14%). Our methodology will enable researchers to build sensor networks that better interpolate pollutant loads in ungaged locations, improve contaminant source identification, and inform more effective pollution control strategies.

Observability-based sensor placement improves contaminant tracing in river networks

Matt Bartos
Dept. of Civil Engineering
University of Michigan
mdbartos@umich.edu

Branko Kerkez
Dept. of Civil Engineering
University of Michigan
bkerkez@umich.edu

Abstract

This study presents a new methodology for identifying near-optimal sensor locations for contaminant source tracing in river networks. To establish a physical basis for the problem, we first derive a linear time-invariant (LTI) model for riverine contaminant transport using the one-dimensional advection-diffusion equation. We then formulate an optimization problem to find the sensor placement that maximizes the *observability* of the modeled system, and identify two heuristics for efficiently achieving this goal. Evaluating each sensor placement strategy on its ability to reconstruct initial contaminant loads from observed outputs, we find that the best sensor placement is obtained by greedily maximizing the rank of the LTI system's *Observability Gramian*. In addition to providing the best approximate reconstruction of internal states, this strategy makes it possible to perfectly recover any initial contaminant load while only monitoring a small subset of river branches ($\sim$14%). Our methodology will enable researchers to build sensor networks that better interpolate pollutant loads in ungaged locations, improve contaminant source identification, and inform more effective pollution control strategies.

1 Introduction

Global river health is in a state of crisis. In the United States, roughly 46% of river length is classified as being in poor biological condition [1]. Survey results indicate similar trends worldwide [2, 3]. The main drivers of surface water impairment are nonpoint-source pollutants like sewage, nutrients, and sediments [4]. Sewer overflows release pathogens that are attributable to roughly 3,400-5,600 cases of illness annually in the United States [5]. At the same time, excessive nutrient loads from agriculture result in harmful algal blooms that impair aquatic ecosystems and render water unsafe to drink [6, 7]. Sediment transport results from almost all human land use [8], and leads to destruction of aquatic habitats [9, 10], deposition of toxic chemicals [11, 12], and loss of reservoir capacity [9, 13]. Despite the severity of impacts, nonpoint-source pollutants are often poorly characterized and thus inadequately managed.

To address surface water impairment, new systems for continuous water quality monitoring are desperately needed. In the United States, water quality is heavily regulated but not heavily enforced [14, 15]. Water managers rely on periodic field surveys to measure contaminants of interest. However, water quality measurements are typically only collected on a weekly to monthly basis [16]. At this temporal resolution, it is almost impossible to capture the influence of storm events [16], which exert a dominating effect on water quality [17]. While continuous water quality data are provided by the United States Geological Survey (USGS), there is currently only about one gage per 4,400 km^2 of land area in the United States [18]. Sparse and infrequent data make it difficult to accurately assess river health, identify pollution sources, or even evaluate the effectiveness of existing interventions.

Fortunately, new sensing technologies are enabling researchers and stakeholders to better characterize and respond to waterborne contaminants. Novel optical nitrate sensors [19], continuous-flow PCR

reactors [20], and automated grab samplers [17] are providing new tools to measure pathogen and nutrient loads in-situ. At the same time, *internet of things* technologies are allowing researchers to build larger sensor networks that collect more real-time data than ever before [21]. These advances are enabling researchers to better understand contaminant fate and transport not only at individual sites, but at the scale of entire river basins.

As new technologies make it possible to characterize water quality at unprecedented scales, the question of *where* to place sensors takes on greater significance. Historically, monitoring programs have been initiated on a site-by-site basis in response to known cases of water quality impairment [22]. However, focusing on individual problem sites does not guarantee a suitable sensor placement for system-scale problems like contaminant source identification. Despite recent advances in sensing, water quality monitoring remains costly and labor-intensive. With increasing budgetary constraints being placed on water quality monitoring programs [23], it is imperative that sensor sites be selected efficiently. This study seeks to answer the question: *where should we place sensors to best characterize water quality using the minimum number of sites?*

1.1 Background

Within the field of water resources engineering, the sensor placement problem has been most extensively studied in drinking water distribution systems. In this context, protecting human safety is paramount, and so studies tend to focus on minimizing public health impacts from contamination events [24]. Sensor placement is thus optimized for security-related objectives such as minimizing the time to contaminant detection, minimizing the number of people exposed, or maximizing detection likelihood [25]. To achieve these objectives, studies explore a variety of techniques including expert opinion [26], mixed-integer programming [27–30], genetic algorithms [25, 31–33], graph centrality metrics [34], minimum test cover selection [35], cross-entropy selection [36], and various combinatorial heuristics [37–40]. A recent review catalogs over 90 studies related to sensor placement for contaminant detection in water distribution networks alone [24].

For river networks, the goals of contaminant monitoring are different, and thus different sensor placement strategies are required. River monitoring programs focus less on detecting acute hazards, and more on characterizing long-term human impacts to water quality [41]. In this vein, common applications include wildlife impact assessments [42, 43], contaminant source tracing studies [44, 45], and future predictions of river health based on current trends [46]. This information is vital for water managers seeking to develop effective remediation strategies [47]. However, because monitoring goals differ between river networks and water supply networks, sensor placement techniques designed for one system do not necessarily carry over to the other. Sensor placement strategies for river networks must meet the needs of current monitoring programs, which fundamentally revolve around understanding the present, past and future *states* of the system.

To better characterize surface water quality and its drivers, this study investigates the problem of optimal sensor placement for *state estimation*. In plain language: given a dynamical model of a river network, we seek the sensor placement that enables the best reconstruction of internal states from observed outputs. State estimation informs a number of critically important applications in riverine contaminant fate and transport. First, it enables interpolation of contaminant loads at ungaged locations, which is necessary for assessing impacts in sparsely instrumented basins. Second, state estimation facilitates contaminant source identification, which is essential for tracking down the origin of pollutant releases and taking corrective action. Third, state estimation improves our ability to forecast the movement of contaminant plumes, which helps to coordinate early warnings and interventions. Finally, when paired with real-time control, state estimation enables water managers to actively improve water quality—for instance, by holding back water in retention basins to encourage sedimentation and reduce downstream erosion.

The problem of optimal sensor placement for state estimation has attracted a large body of practical and theoretical research. On the practical side, sensor placement has been investigated with respect to state estimation in electric power grids [48, 49], biochemical reaction networks [50], robotic locomotion and path-planning [51–53], structural health monitoring [54–56], spacecraft [57], weather monitoring [58], and contaminant monitoring in soil [59]. These applications have been supported by recent theoretical advances that make the sensor placement problem amenable to larger and more complex networks. Within this domain, some important contributions include scalable graphical approaches to sensor placement [50], convex relaxations to the sensor placement problem [60], and

approaches based on maximization of "signal energy" [61–63]. While state estimation has long been used in hydrology [64], to our knowledge there are no studies that have examined the problem of optimal sensor placement for riverine contaminant transport.

1.2 This work

Despite numerous practical benefits, there are as of yet no studies that examine the problem of sensor placement for state estimation in river networks. To fill this knowledge gap, this study presents a new methodology for determining optimal locations of riverine water quality sensors. To permit a theoretical treament of the problem, we first derive a linear time-invariant (LTI) model for contaminant transport in river networks based on the one-dimensional advection-diffusion equation. We then propose two heuristics for selecting sensor locations that maximize the *observability* of this LTI system—by maximizing the trace and rank of the system's *Observability Gramian*, respectively. We evaluate each sensor placement heuristic by simulating the system under a large number of random contaminant loads and determining which sensor placement is best able to reconstruct initial contaminant loads from observed outputs. The result is a robust and theoretically-sound method for selecting water quality sensing sites in river basins. Our method will assist practitioners in building more effective sensor networks for riverine contaminant monitoring and also help to identify "blind spots" in existing deployments. While this study primarily focuses on contaminant transport, we discuss how our sensor placement algorithm will also inform better systems for flash flood monitoring and hydrologic state estimation.

2 Mathematical description of transport dynamics

This study focuses on the optimal sensor placement needed to characterize transport of contaminants in river networks. Thus, before we can formulate an algorithm for sensor placement, it is first essential to capture an accurate mathematical representation of riverine transport dynamics. In this section, we derive a linear time-invariant model for contaminant transport in river networks using the one-dimensional advection-diffusion equation. This physically-based model will form the basis for our subsequent sensor placement algorithm.

The transport of a contaminant through a fluid medium is described by the one-dimensional advection-diffusion equation [65]:

$$\frac{\partial c}{\partial t} + \frac{\partial (uc)}{\partial x} - \frac{\partial}{\partial x}\left(D\frac{\partial c}{\partial x}\right) - r(c) = 0 \tag{1}$$

Where c is the quantity of interest per unit length, u is the velocity of flow, D is the diffusion coefficient, $r(c)$ is the endogenous reaction rate, t is time and x is distance.

Discretizing this equation in time and space using an explicit upwind scheme [66], the quantity of interest at an element i can be described as a linear function of the quantities at the upstream element $i-1$ and the downstream element $i+1$:

$$\frac{c_i^{t+\Delta t} - c_i^t}{\Delta t} + \frac{u_i c_i^t - u_{i-1} c_{i-1}^t}{\Delta x_i} - \frac{2D_i}{\Delta x_i}\left[\frac{c_{i+1} - c_i}{\Delta x_i + \Delta x_{i+1}} - \frac{c_i - c_{i-1}}{\Delta x_{i-1} + \Delta x_i}\right] - K c_i^t = 0 \tag{2}$$

For a network of elements, this equation can be generalized as a matrix equation $\mathbf{c}^{t+\Delta t} = A\mathbf{c}^t$, where A is a discrete-time state transition matrix. Expressing the equation in matrix form greatly simplifies the presentation of the sensor placement problem. Thus, in the following sections we derive the advective, diffusive, and reactive components of the state transition matrix, and present a general linear time-invariant state space model for contaminant transport in river networks.

Advection: Advection refers to the transport of a contaminant due to the bulk motion of a carrier fluid. To represent advection in the form of a matrix expression, we first define a directed weighted adjacency matrix U to represent the inward flux into each element, with entries defined as:

$$U_{ij} = \begin{cases} \frac{u_{ji}\phi_{ji}\Delta t}{\Delta x_i} & \text{if there exists a directed path from } v_j \text{ to } v_i, \text{ and } v_i \neq v_j \\ 0 & o/w \end{cases} \tag{3}$$

Where i is the row index, j is the column index, v_j is the source node, v_i is the destination node, and $\phi_{ji} \in [0, 1]$ is the fraction of flow from v_j that enters v_i. Each node in the network corresponds to an individual stream segment. Next, we define a weighted adjacency matrix W to represent the outward flux from each node:

$$W_{ij} = \begin{cases} \frac{u_{ij}\phi_{ij}\Delta t}{\Delta x_i} & \text{if there exists a directed path from } v_i \text{ to } v_j, \text{ and } v_i \neq v_j \\ 0 & o/w \end{cases} \tag{4}$$

Finally, let V be the diagonal matrix $V = \text{diag}(W\mathbf{1})$. Thus, the advective component of the equation can be represented by the matrix expression $(U - V)\mathbf{c}^t$, which yields the net advective transport through each node i at time t:

$$((U - V)\mathbf{c}^t)(i) = \sum_{j \in \mathcal{U}_i} \frac{c_j^t u_{ji}\phi_{ji}\Delta t}{\Delta x_i} - \sum_{k \in \mathcal{D}_i} \frac{c_i^t u_{ik}\phi_{ik}\Delta t}{\Delta x_i} \tag{5}$$

Where $\mathcal{U}_i$ is the set of nodes upstream of node v_i, and $\mathcal{D}_i$ is the set of nodes downstream of node v_i.

Diffusion: Diffusion represents the transport of a contaminant from regions of high concentration to low concentration due to random movements within the carrier fluid (e.g. turbulence). To express the diffusion component as a matrix expression, we first define the following undirected weighted adjacency matrix X, with elements defined as follows:

$$X_{ij} = \begin{cases} \frac{2}{\Delta x_i + \Delta x_j} & \text{if } v_i \text{ adjacent to } v_j, \text{ and } v_i \neq v_j \\ 0 & o/w \end{cases} \tag{6}$$

Thus, each nonzero element represents the inverse of the average distance between the center of node v_i and the center of adjacent node v_j. Let Z be the diagonal matrix:

$$Z = \text{diag}\left(\frac{D_1\Delta t}{\Delta x_1}, \ldots, \frac{D_n\Delta t}{\Delta x_n}\right) \tag{7}$$

Finally, let Y be a diagonal matrix $Y = \text{diag}(X\mathbf{1})$. Then the diffusive component of the advection-reaction-diffusion equation can be represented by the matrix expression $Z(Y - X)$, which yields the diffusion from node v_i to neighboring nodes at time t.

$$(Z(Y - X)\mathbf{c}^t)(i) = \frac{2D_i\Delta t}{\Delta x_i} \sum_{j \in \mathcal{N}_i} \frac{c_i^t - c_j^t}{\Delta x_i + \Delta x_j} \tag{8}$$

Where $\mathcal{N}_i$ is the set of nodes adjacent to (i.e. either upstream or downstream of) node v_i. Note that the expression $Z(Y - X)$ is equivalent to a rescaled graph Laplacian.

Reaction: The reaction component represents the portion of contaminant lost or gained due to interactions with the surrounding environment (e.g. chemical reactions, sedimentation, etc.). To express the reaction term, define a diagonal matrix of reaction coefficients $R = \text{diag}(K_1\Delta t, K_2\Delta t, \ldots K_n\Delta t)$. Then the reaction rate at node i is expressed as:

$$(R\mathbf{c}^t)(i) = c_i^t K_i \Delta t \tag{9}$$

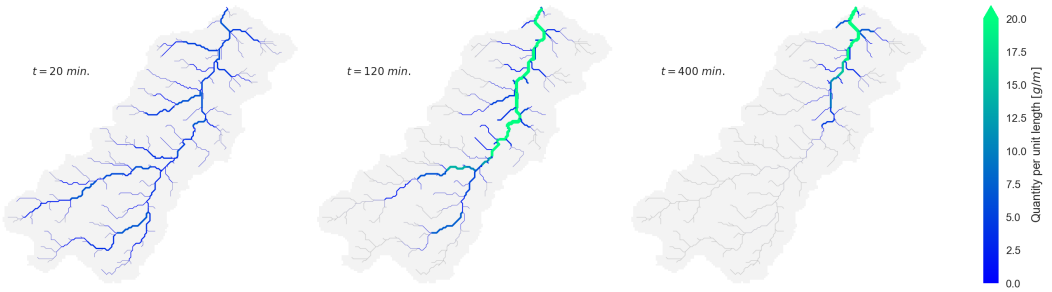

Figure 1: Contaminant progression over time, with contaminant quantity per unit length (in nominal units of g/m) indicated by both color and width. The watershed contributing area is shown in light gray, while the channel network is shown in dark gray.

State space equation: Following the previous derivations, we may define the state transition matrix A that carries the system states from time t to $t + \Delta t$:

$$A = (U - V) - Z(Y - X) + R + I \tag{10}$$

Thus, assuming no exogenous input, the state transition from time t to $t + \Delta t$ for node v_i is:

$$\begin{aligned}(\mathbf{c}^{t+\Delta t})(i) = (A\mathbf{c}^t)(i) = \sum_{j \in \mathcal{U}_i} \frac{c_j^t u_{ji} \phi_{ji} \Delta t}{\Delta x_i} - \sum_{k \in \mathcal{D}_i} \frac{c_i^t u_{ik} \phi_{ik} \Delta t}{\Delta x_i} + \\ \frac{2 D_i \Delta t}{\Delta x_i} \sum_{j \in \mathcal{N}_i} \frac{c_i^t - c_j^t}{\Delta x_i + \Delta x_j} + c_i^t K_i \Delta t + c_i^t\end{aligned} \tag{11}$$

Figure 1 shows the transport of a contaminant in a river network using the above formulation. Using the state transition matrix A, the advection-diffusion dynamics on the network can be formulated as a discrete-time linear time-invariant (LTI) state space system:

$$\mathbf{c}(t + \Delta t) = A\mathbf{c}(t) + B\mathbf{u}(t) \tag{12}$$

$$\mathbf{y}(t) = C\mathbf{c}(t) \tag{13}$$

Where A is the $n \times n$ state transition matrix defined previously, B is the $n \times p$ input matrix, C is the $m \times n$ observation matrix, $\mathbf{c}(t)$ is the $n \times 1$ vector of system states at time t, $\mathbf{u}(t)$ is the $p \times 1$ exogenous input vector, and $\mathbf{y}(t)$ is the $m \times 1$ observed output vector of the system at time t.

The state equation (12) describes the evolution of system states over time, while the output equation (13) describes the observable output of the system (e.g. sensor readings). With respect to the sensor placement problem, the rows of C represent sensor locations. Independent sensors located on each river reach correspond to an observation matrix equal to the identity matrix ($C = I$). If we consider the case of independent sensors located on some subset of reaches in the network, then C would be a matrix formed from a subset of the rows of the identity matrix. In the context of sensor placement, one can represent different sensor strategies by changing the rows of the observation matrix C. In the following section, this formulation will enable us to define algorithms for computing optimal sensor placement strategies.

3 Sensor placement methodology

We define an optimal sensor placement as one that enables the best reconstruction of a system's internal states from its measured outputs. This definition follows from the concept of *observability* in control theory, wherein a system is said to be *completely observable* if all past states can be perfectly inferred from its sensor outputs [67]. In the following section, we will quantify what it means to "best

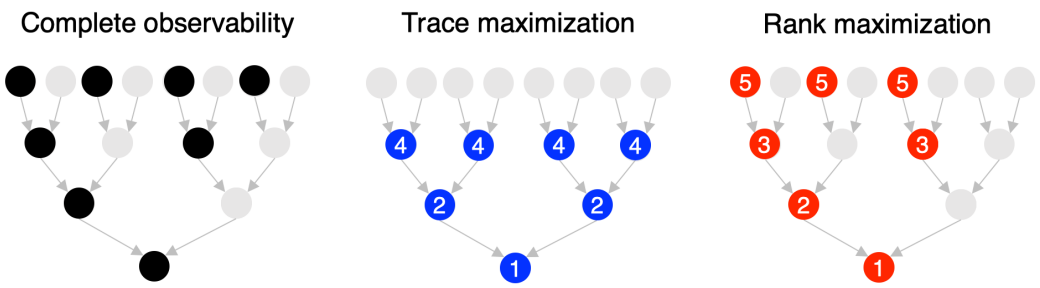

Figure 2: Conceptual sensor placements for a convergent binary tree graph. Left: sensor placement needed for complete observability (sensors in black). Center: incremental sensor placement by trace maximization of W_o. Right: incremental sensor placement by rank maximization of W_o. Sensor placements are ordered with a left-handed bias.

reconstruct internal states", and propose algorithms for finding the sensor placement that achieves this goal.

The most general test for the observability of a deterministic linear system is given by the Observability Gramian. For a discrete-time LTI system, the Observability Gramian $W_o(t_f, t_0)$ is a matrix-valued function defined over the time interval t_0 to t_f [67]:

$$W_o(t_f, t_0) = \sum_{t=t_0}^{t_f} (A^T)^t C^T C A^t \tag{14}$$

Where t_0 and t_f are the start and end times of the observation interval, respectively. A linear system is *completely observable* over this time interval if and only if the Observability Gramian is *full rank*. Under this condition, the system state at any initial time t_0 can be recovered from the observed outputs $\mathbf{y}(t)$ using the least-squares formulation [67]:

$$\mathbf{x}(t_0) = W_o^{-1}(t_f, t_0) \sum_{t=t_0}^{t_f} (A^T)^t C^T \mathbf{y}(t) \tag{15}$$

While the Observability Gramian is defined with respect to a particular time interval, one can characterize the general observability of the system by letting t_f approach infinity. In this case, the infinite time-horizon Observability Gramian W_o is given by the solution to the discrete-time Lyapunov equation [67]:

$$A^T W_o A - W_o + C^T C = 0 \tag{16}$$

If W_o is full rank, then the history of internal states can be reconstructed from some arbitrarily long sequence of sensor outputs, and the system is completely observable in the general case. The infinite time-horizon Observability Gramian provides a way to determine the observability of a system that may take an indefinitely long time to reach a steady state (which is often the case when diffusion effects are present).

For dendritic systems with convergent flow, the conditions for complete observability are difficult to achieve in practice. In general, for every branch in which p upstream nodes converge to 1 downstream node, sensors must be placed on both the downstream node as well as on $p - 1$ of the upstream nodes (see Figure 2 for an illustration) [50]. Thus, assuming a bifurcating branching pattern, a river network would require roughly half of all tributaries to be monitored in order to be completely observable.[1]

However, in addition to determining whether a linear system is completely observable, the Observability Gramian also indicates the system's relative observability. Some intuition for this concept is

[1]Note that this statement is only true for the case where all edge weights are equal. For a branching network with non-symmetric edge weights, complete observability can be achieved with many fewer sensors.

given by the relationship between the Observability Gramian and the estimation error covariance. Consider an LTI system in which each sensor is corrupted by Gaussian white noise. In this case, the estimation error covariance $P(t_f, t_0)$ gives the expected error associated with reconstructing the system's state at some previous time t_0:

$$P(t_f, t_0) = E[(\mathbf{x}(t_0) - \hat{\mathbf{x}}(t_0))(\mathbf{x}(t_0) - \hat{\mathbf{x}}(t_0))^T] \tag{17}$$

Where $\mathbf{x}(t_0)$ is the initial state at time t_0, and $\hat{\mathbf{x}}(t_0)$ is the minimum variance estimate of the initial state given a sequence of sensor outputs from t_0 to t_f. The diagonal elements of P are the variances of the estimation errors for each state. If the measurement error at each sensor is independent and identically distributed with variance σ^2, it can be shown that the estimation error covariance P is inversely proportional to the Observability Gramian [63]:

$$P(t_f, t_0) = \sigma^2 W_o^{-1}(t_f, t_0) \tag{18}$$

This relationship reveals that "maximizing" the Observability Gramian will in some sense "minimize" the estimation error. However, one of the central challenges in sensor placement is deciding what it means to "maximize" a matrix-valued function. Within the literature, various methods have been proposed, including maximizing the trace, the determinant, the rank, or the minimum eigenvalue of the Observability Gramian [63]. Some of these metrics—such as the determinant—are only valid for completely observable systems.

For systems that are not completely observable, two natural candidates emerge for minimizing the estimation error. The first option is maximizing the trace of the Observability Gramian, which corresponds to minimizing the average estimation variance over all observable modes. The second option is maximizing the rank of the Observability Gramian, which corresponds to maximizing the number of observable subspaces. These two approaches have different strengths and limitations in terms of the types of signals they are best able to detect.

3.1 Sensor placement algorithms

In the following discussion, we formalize the sensor placement problem in terms of maximizing the trace and rank of the Observability Gramian, respectively. For the application of contaminant monitoring, we will assume that it is possible to place a sensor at each river reach on the network. In this case, the observation matrix C is a subset of the rows of the identity matrix I. Moreover, we will assume that the number of desired sensors is known and given by N. Under these conditions, sensor placement can be formulated as an optimization problem for each objective function of interest.

Trace optimization: Maximizing the trace of the Observability Gramian has an intuitive justification in terms of minimizing the average estimation variance. The variances of the estimation errors in each direction of the state space are given by the eigenvalues of the estimation error covariance P. Moreover, from Equation (18), we know that the Observability Gramian is the inverse of the estimation error covariance. Therefore, because the trace is equal to the sum of the eigenvalues, maximizing the trace of the Observability Gramian is equivalent to minimizing the sum of the estimation variances.

To solve the optimization problem efficiently, we use an interesting correspondence between the Observability Gramian and its dual, the Controllability Gramian. Specifically, it can be shown that the N sensor locations that maximize the trace of the Observability Gramian correspond to the N largest diagonal indices of the corresponding Controllability Gramian (see Theorem 1 in Appendix Section A1). Thus, we can state the trace-optimized sensor placement strategy as the following optimization problem:

$$\begin{aligned} \max_{K \in \mathcal{K}} \quad & \sum_{k \in K} e_k^T W_c e_k \\ \text{s.t.} \quad & AW_cA^T - W_c + I = 0 \\ & |K| = N \end{aligned} \tag{19}$$

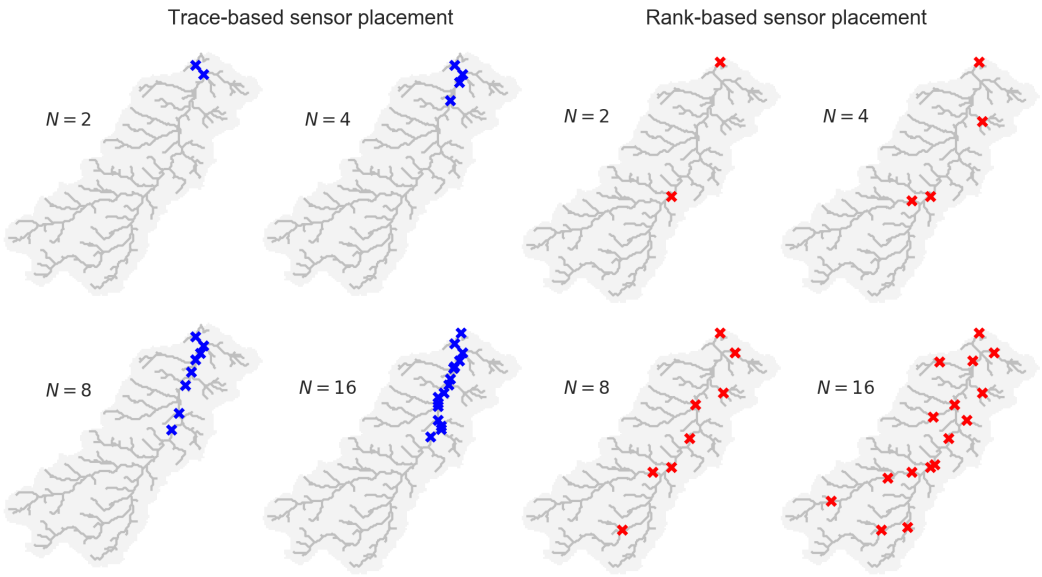

Figure 3: Sensor placement progression for rank- and trace- optimized strategies from $N = 2$ to $N = 16$ sensors.

Where K is a set of sensor locations with elements $\{k \in \mathbb{N} | 1 \leq k \leq n\}$, $\mathcal{K}$ is the set of all possible sets of sensor locations, and $e_k \in \mathbb{R}^n$ is the k^{th} canonical basis vector (i.e. the k^{th} column of the identity matrix). This optimization problem can be solved quickly by computing W_c, and then finding the indices of the N largest diagonal elements. The computational complexity of this approach is approximately $\mathcal{O}(n^3)$—(see Appendix Section A2 for details).

Rank optimization: Compared to maximizing the trace, maximizing the rank of the Observability Gramian enables better identification of spatially localized signals. While optimizing the trace maximizes the overall "signal energy", it provides no guarantees on the number of observable subspaces. In the context of river networks, this means that the trace-optimized sensor placement will often struggle to detect which branch in the network a contaminant load originated from. Selecting the rank as the objective function ensures more equal representation of each subspace, enabling better characterization of smaller tributaries. We can express the rank-optimized sensor placement strategy as the following optimization problem:

$$\begin{aligned} \max_{K \in \mathcal{K}} \quad & \text{rank}(W_o) \\ \text{s.t.} \quad & A^T W_o A - W_o + \sum_{k \in K} e_k e_k^T = 0 \\ & |K| = N \end{aligned} \tag{20}$$

With all variables defined previously. Prior work has shown that the rank of the Observability Gramian is a submodular function with respect to the columns of the observation matrix [62]. This property means that a greedy optimization algorithm is capable of efficiently generating a near-optimal solution—with the worst-case performance guaranteed to be at least 63% of the optimal value [62]. In practice, greedy optimization will usually perform much better than this lower bound [62]. Our greedy rank-optimized sensor placement algorithm is described formally in Algorithm 1. The computational complexity of this approach is approximately $\mathcal{O}((N+1)n^4)$—(see Appendix Section A3 for details).

Visual comparison of algorithms: Visualization reveals that the rank- and trace-based sensor placement strategies produce markedly different outcomes. Figure 3 shows the sensor placements obtained through rank-based (left) and trace-based (right) optimization of the Observability Gramian, for $N = 2$ to $N = 16$ sensors. From this figure, it can be seen that the trace-based strategy

Data:
State transition matrix $A \in \mathbb{R}^{n \times n}$;
Desired number of sensors N;
Result: Set of rank-optimal sensor locations K;

$K \leftarrow \emptyset$;
$G_1 \leftarrow \{W_i \mid A^T W_o A - W_o + e_i e_i^T = 0,\ i \in [1, n]\}$;
$G_2 \leftarrow \emptyset$;
while $|K| < N$ **do**
  $k \leftarrow \underset{i}{\text{argmax}} \left\{\text{rank}(W_i + \sum_{W_j \in G_2} W_j)\right\},\ \forall\, W_i \in G_1$;
  $K \leftarrow K \cup \{k\}$;
  $G_1 \leftarrow G_1 \setminus \{W_k\}$;
  $G_2 \leftarrow G_2 \cup \{W_k\}$;
end

Algorithm 1: Greedy rank-optimized sensor placement algorithm.

concentrates all sensors along the river mainstem, whereas the rank-based strategy attempts to divide the watershed into roughly equally-sized subcatchments. In general, the rank-based strategy optimizes for *exploration* of the state space by distributing sensors evenly throughout the network, while the trace-based strategy optimizes for *exploitation* of the total signal energy by placing all sensors in the most information-dense region.

4 Evaluation of sensor placement

We evaluate the trace- and rank-based sensor placement algorithms by assessing how well each sensor placement strategy reconstructs the location and magnitude of a series of initial contaminant loads. In terms of real-world applications, this assessment corresponds to the problem of *contaminant source identification*, in which one seeks to trace the origin of a contaminant of interest. In theoretical terms, this test is similar to the fundamental test of observability in a linear system—except that in this case we seek the sensor placement that enables the best approximation of an initial state, as opposed to a sensor placement that enables its perfect reconstruction.

The evaluation procedure takes place in three main steps. First, we generate a series of randomized initial contaminant loads. For each sensor placement and initial load, we then propagate the advection-diffusion model forward in time and collect a corresponding sequence of observed outputs. We then reconstruct the least-squares estimate of the initial state from the observed outputs and compute the mean squared error associated with each sensor placement. This procedure yields the relative reconstruction error associated with each sensor placement.

Sensor placement strategies are evaluated on their ability to reconstruct a series of initial contaminant loads from observed outputs. Because a sparse sensor placement generally does not result in a completely observable system, the initial contaminant load cannot be perfectly reconstructed and must instead be approximated with a low-rank representation. For a given sparse sensor placement K, the low-rank estimate of the initial contaminant load is given by:

$$\hat{\mathbf{x}}(t_0, K) = W_o^+(t_f, t_0, K) \sum_{t=t_0}^{t_f} (A^T)^t C^T \mathbf{y}(t) \tag{21}$$

Where $W_o^+(t_f, t_0, K)$ is the Moore-Penrose pseudoinverse of the Observability Gramian associated with sensor placement K [68]. Loosely speaking, the pseudoinverse is the "closest one can get" to the true matrix inverse W_o^{-1} for the case where W_o is singular. The corresponding mean squared error (MSE) of this low-rank estimate is thus:

$$MSE(s) = ||\hat{\mathbf{x}}(t_0, K) - \mathbf{x}(t_0))||_2^2 \tag{22}$$

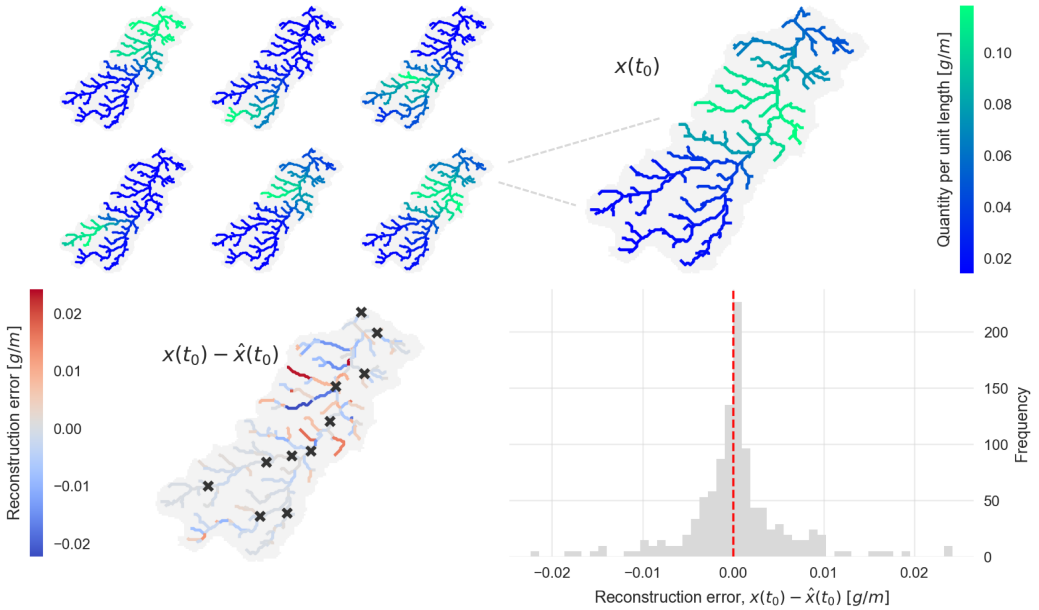

Figure 4: Overview of evaluation procedure. Top-left: initial contaminant loads generated using the "heat kernel" method. Top-right: An example initial contaminant load $\mathbf{x}(t_0)$. Bottom-left: Map of reconstruction error with $N = 11$ sensors. Bottom-right: Histogram of reconstruction errors.

The mean squared reconstruction error provides an intuitive metric for comparing the effectiveness of different sensor placement strategies, with a lower MSE indicating that the sensor placement is better able to reconstruct an initial contaminant load.

To ensure a robust evaluation, it is important to generate a large number of initial contaminant loads at different locations and with differing spatial extents. To accomplish this task, we use the "heat kernel" method for generating random signals on graphs [69]. Specifically, for each trial we generate a heat kernel G in the graph spectral domain, and then center the kernel around a vertex v_j using the generalized graph translation operation to produce a contaminant load g:

$$G(\lambda_\ell) = e^{-\rho\lambda_\ell} \tag{23}$$

$$g(i) = \sum_{\ell=0}^{n-1} G(\lambda_\ell) u_\ell^*(j) u_\ell(i) \tag{24}$$

Where λ_ℓ and u_ℓ are the ℓ^{th} eigenvalue/eigenvector pair of the graph Laplacian associated with the river network's adjacency matrix. This operation effectively "localizes" the contaminant load around a vertex v_j with a spatial spread defined by ρ. Figure 4 (top) shows a series of initial contaminant loads generated using this procedure. For the purposes of evaluation, we generate 200 randomized initial contaminant loads by letting the heat kernel parameter ρ correspond to a normally distributed random variable with a mean and standard deviation of 100. For each trial, the vertex v_j around which the kernel is localized is selected uniformly at random from the set of river reaches. To ensure consistency between trials, all initial contaminant loads are normalized to be unit-norm.

In addition to comparing the trace- and rank-based sensor placement strategies against one another, we also compare these two strategies to a baseline strategy in which the locations of sensors are selected uniformly at random from all river reaches in the network. For this assessment, 10 different randomized sensor placement trials are chosen, and the associated reconstruction error is compared against the trace- and rank-optimized strategies.

For our test case, we focus on the Sycamore Creek watershed in the Dallas–Forth Worth metroplex. This 83 km^2 urban creekshed is the subject of a long-term sensor deployment led by the authors [70,

71]. In constructing the dynamical model, we assume that the bulk velocity of flow is approximately constant throughout the watershed—an assumption that has been empirically validated in previous studies [72]. Note that for a chosen discrete timestep, the system is completely parameterized by the basin Peclet number (the ratio of advection to diffusion). Based on values from the literature, we assume that the basin Peclet number is $Pe = 10$ [72]. The contaminant is assumed to be a conservative tracer (i.e. with a reaction constant of zero). We extract the channel network from digital elevation model (DEM) data by selecting an accumulation threshold that visually agrees with channel extents predicted by the National Hydrography Dataset [73].

4.1 Summary of evaluation procedure

Given a set of sensor placements $\mathcal{K}$ to test, we evaluate each sensor placement K as follows:

1. Generate a randomized set of initial contaminant loads $\mathcal{X} = \{\mathbf{x}_1(t_0) \dots \mathbf{x}_M(t_0)\}$.
2. For each combination of sensor placement and initial state $(K, \mathbf{x}(t_0)) \in \mathcal{K} \times \mathcal{X}$:
   (a) Propagate the advection-diffusion model given by Equations (12) and (13) forward in time from t_0 to t_f, collecting a sequence of observed outputs $\{\mathbf{y}(t_0), \dots \mathbf{y}(t_f)\}$.
   (b) Use Equation (21) to generate a low-rank approximation of the initial state $\hat{\mathbf{x}}(t_0, K)$ from the sequence of observed outputs.
   (c) Compute the mean squared error between the true initial state $\mathbf{x}(t_0)$ and the estimated initial state $\hat{\mathbf{x}}(t_0, K)$ using Equation (22).

This evaluation procedure yields the mean squared error for each sensor placement with respect to each initial contaminant load. The overall effectiveness of each sensor placement is evaluated as the average mean squared error over all initial conditions tested.

4.2 Evaluation results

Among all methods considered, the rank-optimized sensor placement strategy produces the best reconstruction of initial contaminant loads. Figure 5 (left) shows the mean squared error (MSE) vs. the number of sensors for each placement strategy, averaged over all 200 initial contaminant loads tested. In this case, the rank-optimized sensor placement achieves the lowest reconstruction error regardless of how many sensors are used. The benefit of the rank-optimized strategy becomes even more apparent when viewed on a log-scale (Figure 5, right). Here, it can be seen that the rank-based sensor placement not only achieves the lowest MSE, but that the marginal advantage of the rank-based strategy improves as the number of sensors increases.

The trace-optimized sensor placement exhibits comparatively poor performance. Indeed, the mean-squared error associated with the trace-optimized strategy is generally larger than the error associated with randomized sensor placements. This result suggests that the trace-optimized sensor placement does not encourage sufficient exploration of the state space. As noted in Figure 3, the trace-optimized algorithm places all sensors along the river mainstem, making it difficult to disambiguate the location in the network from which any particular signal originates. In general terms, this result implies that placing sensors evenly throughout the network (i.e. exploration) does more to reduce the overall reconstruction error than focusing all sensors on the most information-dense regions (i.e. exploitation).

The rank-based sensor placement strategy also converges to full observability faster than either the randomized or trace-based strategies. For our test case, rank-based sensor placement requires only 32 sensors to achieve complete observability of the system. Given that there are 223 channel segments, this means that only 14% of reaches need to be observed to perfectly reconstruct all internal states. By comparison, the trace-based strategy requires 213 sensors to achieve complete observability, while the randomized placements require 141 sensors on average (corresponding to 96% and 63% coverage of stream reaches, respectively). This result highlights the critical importance of site selection in the deployment of sparse sensor networks.

The rank-based sensor placement algorithm is also robust to parameter uncertainty in the underlying dynamical system. We test robustness to parameter uncertainty by evaluating the performance of the sensor placement algorithm when the basin Peclet number (Pe) is uncertain. Specifically, we generate sensor placements assuming $Pe = 10$ and then run the evaluation procedure for systems with a

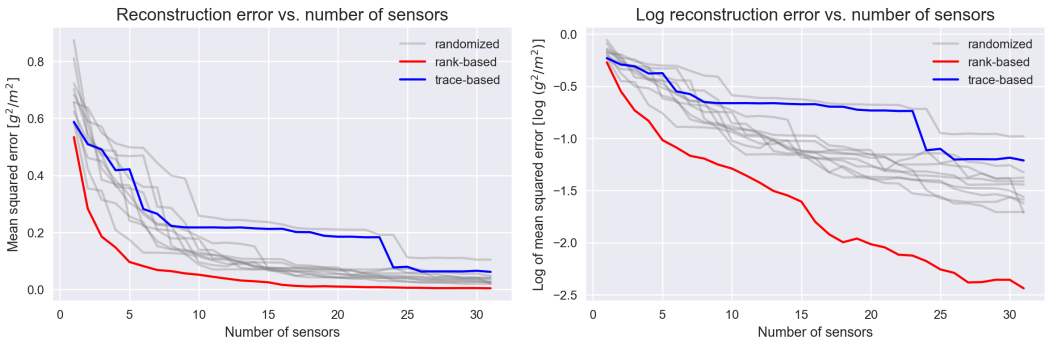

Figure 5: Reconstruction error for different sensor placement strategies averaged over all 200 initial contaminant loads. Left: reconstruction error as measured by the mean-squared error. Right: reconstruction error as measured by the log of the mean-squared error.

true parameter value of $Pe \in \{7.5, 15\}$ to see how well the sensor placement is able to reconstruct initial states. Figures A1 and A2 in the Appendix show the resulting reconstruction errors. While the relative performance is slightly degraded in the high-advection case, the rank-based sensor placement strategy still shows the lowest MSE and the fastest error convergence among all sensor placement strategies considered.

5 Discussion

The rank-optimized sensor placement algorithm provides a powerful tool for improving contaminant source identification in river networks. From a management perspective, information on contaminant sources is a vital prerequisite to the development of pollution control strategies [47]. By tracking down sources of nitrogen and phosphorus, for instance, water managers can prevent eutrophication by enforcing targeted limits on fertilizer application or installing tile drains to halt nutrient seepage [74]. Similarly, source tracing of coliform bacteria may help to locate sanitary sewer leakages and prevent disease outbreaks [75]. However, accurate contaminant source identification depends heavily on the placement of sensors. Currently, little is known about the effectiveness of existing sensor placements or sampling strategies. Our methodology will help fill this gap by providing water managers with a clear procedure for placing water quality sensors to ensure more effective contaminant source tracing.

The sensor placement methodology presented in this paper will also strengthen emergency response during acute contamination events by improving our ability to delineate and forecast contaminant spread. Detection and mitigation of chemical spills in surface water is a central concern for water managers. When the extent of contamination is known, water managers may deploy countermeasures like pipe flushing and no-drink/no-use orders to avert health impacts [76]. For instance, in response to the Elk River chemical spill of 2014, authorities tracked the progression of the contaminant plume and were able to limit exposure by shutting off water intakes at treatment plants until after the plume had passed [76]. These interventions are most effective when the extent and progression of contamination is well-characterized [76]. Our methodology will help improve rapid response to contamination events by enabling water managers to better pinpoint the location, magnitude and trajectory of contaminant plumes.

Finally, our sensor placement methodology will bolster the development of *smart* stormwater systems that use real-time control to improve urban water quality [21]. Recent studies have highlighted the potential for real-time control of stormwater systems to restore ailing urban waterways by managing nutrient and sediment loads [77–81]. Retrofitting retention basins with actuated valves, for instance, makes it possible to strategically hold back water during storm events [70]. Strategic retention of stormwater helps to reduce contaminant loads by enhancing sedimentation and reducing downstream erosion [21]. Real-time control of sewer systems may also improve the treatment capacity of receiving wastewater treatment plants by removing suspended solids and limiting inflows to sustainable levels [82]. While real-time control promises to help reverse the impacts of urbanization on river health, effective control is predicated on an accurate understanding of system states. Put simply, one needs to know how a system is behaving in order to control it. Towards this goal, our sensor placement

methodology will facilitate more effective water quality control by enabling better real-time estimates of contaminant concentrations in urban watersheds.

5.1 Limitations and future work

One potential limitation of our sensor placement methodology is that it does not provide a mechanism for incorporating uncertainty about the underlying dynamical system. While the algorithm is reasonably robust to parameter uncertainty in its current state, one could potentially improve performance by integrating information about the distribution of uncertain parameters into the algorithm itself. In the case of contaminant transport, the most important uncertain parameter is the Peclet number, which describes the ratio of advective to diffusive transport. Although it is possible to characterize the probability distribution of this uncertain parameter through tracer studies, the process of combining sensor placements associated with each realization of this parameter into a single representative sensor placement is not straightforward, but rather subject to interpretation and design goals. In real-world applications, however, system parameters such as the Peclet number will generally incorporate some degree of uncertainty. Thus, future research should extend our sensor placement methodology to account for parameter uncertainty in the underlying dynamical system.

Further work should also investigate ways to extend our sensor placement algorithm to larger watersheds. The time complexity of the rank-based sensor placement algorithm is approximately $\mathcal{O}((N+1)n^4)$, with n corresponding to the number of candidate sensor sites, and N corresponding to the number of desired sensors. Thus, for very large river networks (on the order of thousands or millions of stream reaches), the algorithm may prove computationally intractable. One simple and intuitive way to handle this issue is to simply limit the number of stream reaches under consideration by restricting the dynamical model to streams above a specified size. For stream networks delineated from DEM data, this goal is readily accomplished by setting the accumulation threshold to an appropriate user-specified value. Future research should also investigate simplified sensor placement algorithms that do not require computing the full Observability Gramian of the system. Sensor placement algorithms based purely on network topology for instance [83], could provide a scalable alternative to our theoretically-motivated approach.

In addition to its implications for water quality monitoring, our sensor placement methodology may also inform more effective flood and streamflow monitoring networks. Although primarily used for contaminant fate and transport, studies have shown that the advection-diffusion equation also approximately characterizes the hydrodynamic response of river networks [72]. Specifically, when the transport velocity u is taken to be the mean kinematic wave celerity and the diffusion coefficient D is taken to be the hydrologic dispersion coefficient, the state space system given by Equation 12 yields the approximate streamflow response to a runoff perturbation [72].[2] Thus, if the advection-diffusion equation serves as a first-order dynamical model for streamflow routing, then it follows that the rank-optimized sensor placement strategy will also enable better estimation of hydraulic states like depth and discharge. Using the sensor placement methodology proposed in this study, hydrologists may one day deploy sensor networks that better detect localized flash floods, characterize basin-scale water balances, or enable inverse modeling of rainfall from streamflow measurements.

6 Conclusion

This study investigates the problem of optimal placement of water quality sensors in river networks. Specifically, we focus on the problem of optimal sensor placement for state estimation, which has major implications for pollutant source tracing, forecasting of contaminant transport, and real-time data assimilation. To motivate a theoretical treatment of the problem, we first derive a linear time-invariant model of contaminant transport in river networks using the one-dimensional advection-diffusion equation. Drawing on this model, we propose two heuristics for selecting sensor locations that maximize the *observability* of the system—specifically, by maximizing the rank and trace of the system's *Observability Gramian*. To evaluate each heuristic, we simulate the system under a large number of randomized contaminant loads, and measure the extent to which each sensor placement algorithm is able to reconstruct the system's initial state from observed outputs. Based on this assessment, we find that the rank-based sensor placement heuristic results in the lowest

[2] Within this theoretical framework, the response to an impulsive runoff input is sometimes called the *geomorphological impulse unit hydrograph*.

reconstruction error, and is also able to achieve perfect observability with the smallest number of sensors. Our general-purpose methodology will help practitioners to deploy more effective riverine sensor networks for both scientific and practical applications. By enhancing our ability to characterize riverine contaminants, our method will enable better stewardship of surface water systems and inform more effective policies for restoring impaired waterways.

Acknowledgments

Support for this project was provided by the National Science Foundation Smart and Connected Communities program (Grant 1737432), the National Science Foundation EarthCube initiative (Grant 1639640), and the J. Robert Beyster Computational Innovation Graduate Fellowship.

Declaration of interest

Declaration of interest: none.

Data and code availability

Code and data for this project can be found at:

`https://github.com/klabUM/sensor-placement`

References

[1] U.S. Environmental Protection Agency, "National Water Quality Inventory: Report to Congress," Tech. Rep. EPA 841-R-16-011, 2017.

[2] Ministry of Environmental Protection of the People's Republic of China, "2008 Report on the State of the Environment in China," Tech. Rep., 2009.

[3] C. J. Vörösmarty, P. B. McIntyre, M. O. Gessner, D. Dudgeon, A. Prusevich, P. Green, S. Glidden, S. E. Bunn, C. A. Sullivan, C. R. Liermann, and P. M. Davies, "Global threats to human water security and river biodiversity," *Nature*, vol. 467, no. 7315, pp. 555–561, Sep. 2010. doi: 10.1038/nature09440

[4] J. G. Rowny and J. R. Stewart, "Characterization of nonpoint source microbial contamination in an urbanizing watershed serving as a municipal water supply," *Water Research*, vol. 46, no. 18, pp. 6143–6153, Nov. 2012. doi: 10.1016/j.watres.2012.09.009

[5] U.S. Environmental Protection Agency, "Report to Congress on Impacts and Control of Combined Sewer Overflows and Sanitary Sewer Overflows," Tech. Rep. EPA 833-R-04-001, 2004.

[6] Y. Zhou, D. R. Obenour, D. Scavia, T. H. Johengen, and A. M. Michalak, "Spatial and temporal trends in Lake Erie hypoxia, 1987–2007," *Environmental Science & Technology*, vol. 47, no. 2, pp. 899–905, Jan. 2013. doi: 10.1021/es303401b

[7] D. Scavia, M. A. Evans, and D. R. Obenour, "A scenario and forecast model for Gulf of Mexico hypoxic area and volume," *Environmental Science & Technology*, vol. 47, no. 18, pp. 10 423–10 428, Sep. 2013. doi: 10.1021/es4025035

[8] G. M. Carr and J. P. Neary, *Water quality for ecosystem and human health*. UNEP/Earthprint, 2008.

[9] P. N. Owens, R. J. Batalla, A. J. Collins, B. Gomez, D. M. Hicks, A. J. Horowitz, G. M. Kondolf, M. Marden, M. J. Page, D. H. Peacock, E. L. Petticrew, W. Salomons, and N. A. Trustrum, "Fine-grained sediment in river systems: environmental significance and management issues," *River Research and Applications*, vol. 21, no. 7, pp. 693–717, 2005. doi: 10.1002/rra.878

[10] C. P. Newcombe and D. D. MacDonald, "Effects of suspended sediments on aquatic ecosystems," *North American journal of fisheries management*, vol. 11, no. 1, pp. 72–82, 1991. doi: 10.1577/1548-8675(1991)011<0072:eossoa>2.3.co;2

[11] N. Warren, I. Allan, J. Carter, W. House, and A. Parker, "Pesticides and other micro-organic contaminants in freshwater sedimentary environments—a review," *Applied Geochemistry*, vol. 18, no. 2, pp. 159–194, Feb. 2003. doi: 10.1016/s0883-2927(02)00159-2

[12] M. G. Macklin, P. A. Brewer, D. Balteanu, T. J. Coulthard, B. Driga, A. J. Howard, and S. Zaharia, "The long term fate and environmental significance of contaminant metals released by the January and March 2000 mining tailings dam failures in Maramureş County, upper Tisa Basin, Romania," *Applied Geochemistry*, vol. 18, no. 2, pp. 241–257, Feb. 2003. doi: 10.1016/s0883-2927(02)00123-3

[13] I. Foster and J. Lees, "Changing headwater suspended sediment yields in the LOIS catchments over the last century: a paleolimnological approach," *Hydrological Processes*, vol. 13, no. 7, pp. 1137–1153, 1999. doi: 10.1002/(sici)1099-1085(199905)13:7<1137::aid-hyp794>3.0.co;2-m

[14] J. W. Boyd, "The New Face of the Clean Water Act: A Critical Review of the EPA's Proposed TMDL Rules," *Duke Environmental Law & Policy Forum*, vol. 11, 2000. doi: 10.2139/ssrn.215149

[15] L. Roberts, "Is the gun loaded this time? EPA's proposed revisions to the total maximum daily load program," *The Environmental Lawyer*, vol. 6, p. 635, 2000.

[16] J. W. Kirchner, X. Feng, C. Neal, and A. J. Robson, "The fine structure of water-quality dynamics: the(high-frequency) wave of the future," *Hydrological Processes*, vol. 18, no. 7, pp. 1353–1359, Apr. 2004. doi: 10.1002/hyp.5537

[17] B. P. Wong and B. Kerkez, "Adaptive measurements of urban runoff quality," *Water Resources Research*, vol. 52, no. 11, pp. 8986–9000, 2016. doi: 10.1002/2015WR018013

[18] United States Geological Survey, "National water information system: Water-quality data for the nation," https://waterdata.usgs.gov/nwis/qw, 2020.

[19] B. A. Pellerin, B. D. Downing, C. Kendall, R. A. Dahlgren, T. E. C. Kraus, J. Saraceno, R. G. M. Spencer, and B. A. Bergamaschi, "Assessing the sources and magnitude of diurnal nitrate variability in the San Joaquin River (California) with an in-situ optical nitrate sensor and dual nitrate isotopes," *Freshwater Biology*, vol. 54, no. 2, pp. 376–387, Feb. 2009. doi: 10.1111/j.1365-2427.2008.02111.x

[20] A. C. Hatch, T. Ray, K. Lintecum, and C. Youngbull, "Continuous flow real-time PCR device using multi-channel fluorescence excitation and detection," *Lab Chip*, vol. 14, no. 3, pp. 562–568, 2014. doi: 10.1039/c3lc51236c

[21] B. Kerkez, C. Gruden, M. Lewis, L. Montestruque, M. Quigley, B. Wong, A. Bedig, R. Kertesz, T. Braun, O. Cadwalader, A. Poresky, and C. Pak, "Smarter stormwater systems," *Environmental Science & Technology*, vol. 50, no. 14, pp. 7267–7273, 2016. doi: 10.1021/acs.est.5b05870

[22] R. A. Smith, R. B. Alexander, and M. G. Wolman, "Water-Quality Trends in the Nation's Rivers," *Science*, vol. 235, no. 4796, pp. 1607–1615, Mar. 1987. doi: 10.1126/science.235.4796.1607

[23] Environmental Law & Policy Center, "EPA Region 5 clean water enforcement declines," Tech. Rep., 2020.

[24] W. E. Hart and R. Murray, "Review of sensor placement strategies for contamination warning systems in drinking water distribution systems," *Journal of Water Resources Planning and Management*, vol. 136, no. 6, pp. 611–619, Nov. 2010. doi: 10.1061/(asce)wr.1943-5452.0000081

[25] A. Ostfeld and E. Salomons, "Optimal layout of early warning detection stations for water distribution systems security," *Journal of Water Resources Planning and Management*, vol. 130, no. 5, pp. 377–385, Sep. 2004. doi: 10.1061/(asce)0733-9496(2004)130:5(377)

[26] G. B. Trachtman, "A "Strawman" Common Sense Approach for Water Quality Sensor Site Selection." in *Water Distribution Systems Analysis Symposium 2006*. Cincinnati, Ohio, United States: American Society of Civil Engineers, Mar. 2008. doi: 10.1061/40941(247)106. ISBN 978-0-7844-0941-1 pp. 1–13.

[27] B. H. Lee, R. A. Deininger, and R. M. Clark, "Locating monitoring stations in water distribution systems," *Journal - American Water Works Association*, vol. 83, no. 7, pp. 60–66, Jul. 1991. doi: 10.1002/j.1551-8833.1991.tb07180.x

[28] J.-P. Watson, H. J. Greenberg, and W. E. Hart, "A multiple-objective analysis of sensor placement optimization in water networks," in *Critical Transitions in Water and Environmental Resources Management*. American Society of Civil Engineers, Jun. 2004. doi: 10.1061/40737(2004)456

[29] J. Berry, W. E. Hart, C. A. Phillips, J. G. Uber, and J.-P. Watson, "Sensor placement in municipal water networks with temporal integer programming models," *Journal of Water Resources Planning and Management*, vol. 132, no. 4, pp. 218–224, Jul. 2006. doi: 10.1061/(asce)0733-9496(2006)132:4(218)

[30] L. Sela and S. Amin, "Robust sensor placement for pipeline monitoring: Mixed integer and greedy optimization," *Advanced Engineering Informatics*, vol. 36, pp. 55–63, Apr. 2018. doi: 10.1016/j.aei.2018.02.004

[31] J. Guan, M. M. Aral, M. L. Maslia, and W. M. Grayman, "Optimization model and algorithms for design of water sensor placement in water distribution systems," in *Water Distribution Systems Analysis Symposium 2006*. American Society of Civil Engineers, Mar. 2008. doi: 10.1061/40941(247)103

[32] Z. Y. Wu and T. Walski, "Multi-objective optimization of sensor placement in water distribution systems," in *Water Distribution Systems Analysis Symposium 2006*. American Society of Civil Engineers, Mar. 2008. doi: 10.1061/40941(247)105

[33] A. Preis and A. Ostfeld, "Multiobjective sensor design for water distribution systems security," in *Water Distribution Systems Analysis Symposium 2006*. American Society of Civil Engineers, Mar. 2008. doi: 10.1061/40941(247)107

[34] L. Perelman and A. Ostfeld, "Application of graph theory to sensor placement in water distribution systems," in *World Environmental and Water Resources Congress 2013*. American Society of Civil Engineers, May 2013. doi: 10.1061/9780784412947.060

[35] L. S. Perelman, W. Abbas, X. Koutsoukos, and S. Amin, "Sensor placement for fault location identification in water networks: A minimum test cover approach," *Automatica*, vol. 72, pp. 166–176, Oct. 2016. doi: 10.1016/j.automatica.2016.06.005

[36] G. Dorini, P. Jonkergouw, Z. Kapelan, F. di Pierro, S. T. Khu, and D. Savic, "An efficient algorithm for sensor placement in water distribution systems," in *Water Distribution Systems Analysis Symposium 2006*. American Society of Civil Engineers, Mar. 2008. doi: 10.1061/40941(247)101

[37] A. Kessler, A. Ostfeld, and G. Sinai, "Detecting accidental contaminations in municipal water networks," *Journal of Water Resources Planning and Management*, vol. 124, no. 4, pp. 192–198, Jul. 1998. doi: 10.1061/(asce)0733-9496(1998)124:4(192)

[38] A. Kumar, M. L. Kansal, G. Arora, A. Ostfeld, and A. Kessler, "Discussion of detecting accidental contaminations in municipal water networks," *Journal of Water Resources Planning and Management*, vol. 125, no. 5, pp. 308–310, Sep. 1999. doi: 10.1061/(asce)0733-9496(1999)125:5(308)

[39] J. Uber, R. Janke, R. Murray, and P. Meyer, "Greedy Heuristic Methods for Locating Water Quality Sensors in Distribution Systems," in *Critical Transitions in Water and Environmental Resources Management*. Salt Lake City, Utah, United States: American Society of Civil Engineers, Jun. 2004. doi: 10.1061/40737(2004)481. ISBN 978-0-7844-0737-0 pp. 1–9.

[40] A. Krause, J. Leskovec, C. Guestrin, J. VanBriesen, and C. Faloutsos, "Efficient sensor placement optimization for securing large water distribution networks," *Journal of Water Resources Planning and Management*, vol. 134, no. 6, pp. 516–526, Nov. 2008. doi: 10.1061/(asce)0733-9496(2008)134:6(516)

[41] E. G. Stets, L. A. Sprague, G. P. Oelsner, H. M. Johnson, J. C. Murphy, K. Ryberg, A. V. Vecchia, R. E. Zuellig, J. A. Falcone, and M. L. Riskin, "Landscape drivers of dynamic change in water quality of U.S. rivers," *Environmental Science & Technology*, vol. 54, no. 7, pp. 4336–4343, Mar. 2020. doi: 10.1021/acs.est.9b05344

[42] J. E. Norman, B. J. Mahler, L. H. Nowell, P. C. V. Metre, M. W. Sandstrom, M. A. Corbin, Y. Qian, J. F. Pankow, W. Luo, N. B. Fitzgerald, W. E. Asher, and K. J. McWhirter, "Daily stream samples reveal highly complex pesticide occurrence and potential toxicity to aquatic life," *Science of The Total Environment*, vol. 715, p. 136795, May 2020. doi: 10.1016/j.scitotenv.2020.136795

[43] P. M. Bradley, C. A. Journey, D. T. Button, D. M. Carlisle, B. J. Huffman, S. L. Qi, K. M. Romanok, and P. C. V. Metre, "Multi-region assessment of pharmaceutical exposures and predicted effects in USA wadeable urban-gradient streams," *PLOS ONE*, vol. 15, no. 1, p. e0228214, Jan. 2020. doi: 10.1371/journal.pone.0228214

[44] L. N. Christian, J. L. Banner, and L. E. Mack, "Sr isotopes as tracers of anthropogenic influences on stream water in the Austin, Texas, area," *Chemical Geology*, vol. 282, no. 3-4, pp. 84–97, Mar. 2011. doi: 10.1016/j.chemgeo.2011.01.011

[45] A. C. Gellis, C. C. Fuller, P. C. V. Metre, B. J. Mahler, C. Welty, A. J. Miller, L. A. Nibert, Z. J. Clifton, J. J. Malen, and J. T. Kemper, "Pavement alters delivery of sediment and fallout radionuclides to urban streams," *Journal of Hydrology*, vol. 588, p. 124855, Sep. 2020. doi: 10.1016/j.jhydrol.2020.124855

[46] D. A. Saad, G. E. Schwarz, D. M. Argue, D. W. Anning, S. A. Ator, A. B. Hoos, S. D. Preston, D. M. Robertson, and D. Wise, "Estimates of long-term mean daily streamflow and annual nutrient and suspended-sediment loads considered for use in regional SPARROW models of the conterminous united states, 2012 base year," 2019.

[47] D. Walling, "Tracing suspended sediment sources in catchments and river systems," *Science of The Total Environment*, vol. 344, no. 1-3, pp. 159–184, May 2005. doi: 10.1016/j.scitotenv.2005.02.011

[48] K.-S. Cho, J.-R. Shin, and S. H. Hyun, "Optimal placement of phasor measurement units with GPS receiver," in *2001 IEEE Power Engineering Society Winter Meeting. Conference Proceedings (Cat. No.01CH37194)*. IEEE. doi: 10.1109/pesw.2001.917045

[49] D. Dua, S. Dambhare, R. Gajbhiye, and S. Soman, "Optimal multistage scheduling of PMU placement: An ILP approach," *IEEE Transactions on Power Delivery*, vol. 23, no. 4, pp. 1812–1820, Oct. 2008. doi: 10.1109/tpwrd.2008.919046

[50] Y.-Y. Liu, J.-J. Slotine, and A.-L. Barabasi, "Observability of complex systems," *Proceedings of the National Academy of Sciences*, vol. 110, no. 7, pp. 2460–2465, Jan. 2013. doi: 10.1073/pnas.1215508110

[51] D. Moreno-Salinas, A. M. Pascoal, and J. Aranda, "Optimal sensor placement for underwater positioning with uncertainty in the target location," in *2011 IEEE International Conference on Robotics and Automation*. IEEE, May 2011. doi: 10.1109/icra.2011.5980152

[52] J. Perez-Ramirez, D. K. Borah, and D. G. Voelz, "Optimal 3-D landmark placement for vehicle localization using heterogeneous sensors," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 7, pp. 2987–2999, Sep. 2013. doi: 10.1109/tvt.2013.2255072

[53] M. Rafieisakhaei, S. Chakravorty, and P. R. Kumar, "On the use of the observability gramian for partially observed robotic path planning problems," in *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*. IEEE, Dec. 2017. doi: 10.1109/cdc.2017.8263868

[54] G. W. van der Linden, A. Emami-Naeini, R. L. Kosut, H. Sedarat, and J. P. Lynch, "Optimal sensor placement for health monitoring of civil structures," in *Proceedings of the 2011 American Control Conference*. IEEE, Jun. 2011. doi: 10.1109/acc.2011.5991121

[55] M. Weickgenannt, S. Neuhaeuser, B. Henke, W. Sobek, and O. Sawodny, "Optimal sensor placement for state estimation of flexible shell structures," in *Proceedings of 2011 International Conference on Fluid Power and Mechatronics*. IEEE, Aug. 2011. doi: 10.1109/fpm.2011.6045791

[56] M. Safizadeh and I. Z. M. Darus, "Optimal location of sensor for active vibration control of flexible square plate," in *10th International Conference on Information Science, Signal Processing and their Applications (ISSPA 2010)*. IEEE, May 2010. doi: 10.1109/isspa.2010.5605515

[57] Z. Yu, P. Cui, and S. Zhu, "Observability-based beacon configuration optimization for Mars entry navigation," *Journal of Guidance, Control, and Dynamics*, vol. 38, no. 4, pp. 643–650, Apr. 2015. doi: 10.2514/1.g000014

[58] C. Guestrin, A. Krause, and A. P. Singh, "Near-optimal sensor placements in gaussian processes," in *Proceedings of the 22nd international conference on Machine learning - ICML 05*. ACM Press, 2005. doi: 10.1145/1102351.1102385

[59] C. C. Castello, J. Fan, A. Davari, and R.-X. Chen, "Optimal sensor placement strategy for environmental monitoring using wireless sensor networks," in *2010 42nd Southeastern Symposium on System Theory (SSST 2010)*. IEEE, Mar. 2010. doi: 10.1109/ssst.2010.5442825

[60] S. Joshi and S. Boyd, "Sensor selection via convex optimization," *IEEE Transactions on Signal Processing*, vol. 57, no. 2, pp. 451–462, Feb. 2009. doi: 10.1109/tsp.2008.2007095

[61] G. Yan, G. Tsekenis, B. Barzel, J.-J. Slotine, Y.-Y. Liu, and A.-L. Barabási, "Spectrum of controlling and observing complex networks," *Nature Physics*, vol. 11, no. 9, pp. 779–786, Aug. 2015. doi: 10.1038/nphys3422

[62] F. L. Cortesi, T. H. Summers, and J. Lygeros, "Submodularity of energy related controllability metrics," in *53rd IEEE Conference on Decision and Control*. IEEE, Dec. 2014. doi: 10.1109/cdc.2014.7039832

[63] Brian T. Hinson, "Observability-Based Guidance and Sensor Placement," Ph.D. dissertation, University of Washington, 2014.

[64] Y. Liu, A. H. Weerts, M. Clark, H.-J. H. Franssen, S. Kumar, H. Moradkhani, D.-J. Seo, D. Schwanenberg, P. Smith, A. I. J. M. van Dijk, N. van Velzen, M. He, H. Lee, S. J. Noh, O. Rakovec, and P. Restrepo, "Advancing data assimilation in operational hydrologic forecasting: progresses, challenges, and emerging opportunities," *Hydrology and Earth System Sciences*, vol. 16, no. 10, pp. 3863–3887, Oct. 2012. doi: 10.5194/hess-16-3863-2012

[65] J. L. Martin, S. C. McCutcheon, and R. W. Schottman, *Hydrodynamics and Transport for Water Quality Modeling*. CRC Press, May 2018.

[66] S. V. Patankar, *Numerical Heat Transfer and Fluid Flow*. CRC Press, Oct. 2018.

[67] C.-T. Chen, *Linear System Theory and Design*, 3rd ed., ser. The Oxford Series in Electrical and Computer Engineering. New York: Oxford University Press, 1999. ISBN 978-0-19-511777-6 978-0-19-511778-3

[68] A. J. Laub, *Matrix Analysis for Scientists and Engineers*. SIAM, Jan. 2005.

[69] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 83–98, May 2013. doi: 10.1109/msp.2012.2235192

[70] M. Bartos, B. Wong, and B. Kerkez, "Open storm: a complete framework for sensing and control of urban watersheds," *Environmental Science: Water Research & Technology*, vol. 4, no. 3, pp. 346–358, 2018. doi: 10.1039/c7ew00374a

[71] H. Habibi, I. Dasgupta, S. Noh, S. Kim, M. Zink, D.-J. Seo, M. Bartos, and B. Kerkez, "High-resolution hydrologic forecasting for very large urban areas," *Journal of Hydroinformatics*, vol. 21, no. 3, pp. 441–454, Feb. 2019. doi: 10.2166/hydro.2019.100

[72] I. Rodriguez-Iturbe and A. Rinaldo, *Fractal river basins: chance and self-organization*. Cambridge University Press, 2001.

[73] United States Geological Survey, "National hydrography geodatabase," https://viewer.nationalmap.gov/viewer/nhd.html?p=nhd, 2013.

[74] I. S. University, "Iowa science assessment of nonpoint source practices to reduce nitrogen and phosphorus transport in the Mississippi river basin," Tech. Rep., 2012.

[75] B. Sercu, L. C. V. D. Werfhorst, J. L. S. Murray, and P. A. Holden, "Sewage exfiltration as a source of storm drain contamination during dry weather in urban watersheds," *Environmental Science & Technology*, vol. 45, no. 17, pp. 7151–7157, Sep. 2011. doi: 10.1021/es200981k

[76] A. J. Whelton, L. McMillan, C. L.-R. Novy, K. D. White, and X. Huang, "Case study: the crude MCHM chemical spill investigation and recovery in West Virginia USA," *Environmental Science: Water Research & Technology*, vol. 3, no. 2, pp. 312–332, 2017. doi: 10.1039/c5ew00294j

[77] E. Gaborit, D. Muschalla, B. Vallet, P. A. Vanrolleghem, and F. Anctil, "Improving the performance of stormwater detention basins by real-time control using rainfall forecasts," *Urban Water Journal*, vol. 10, no. 4, pp. 230–246, Aug. 2013. doi: 10.1080/1573062x.2012.726229

[78] J. R. Middleton and M. E. Barrett, "Water quality performance of a batch-type stormwater detention basin," *Water Environment Research*, vol. 80, no. 2, pp. 172–178, Feb. 2008. doi: 10.2175/106143007x220842

[79] C. Jacopin, E. Lucas, M. Desbordes, and P. Bourgogne, "Optimisation of operational management practices for the detention basins," *Water Science and Technology*, vol. 44, no. 2-3, pp. 277–285, Jul. 2001. doi: 10.2166/wst.2001.0780

[80] J. F. Carpenter, B. Vallet, G. Pelletier, P. Lessard, and P. A. Vanrolleghem, "Pollutant removal efficiency of a retrofitted stormwater detention pond," *Water Quality Research Journal*, vol. 49, no. 2, pp. 124–134, Dec. 2013. doi: 10.2166/wqrjc.2013.020

[81] A. Mullapudi, M. Bartos, B. Wong, and B. Kerkez, "Shaping streamflow using a real-time stormwater control network," *Sensors*, vol. 18, no. 7, p. 2259, Jul. 2018. doi: 10.3390/s18072259

[82] S. C. Troutman, N. G. Love, and B. Kerkez, "Balancing water quality and flows in combined sewer systems using real-time control," *Environmental Science: Water Research & Technology*, vol. 6, no. 5, pp. 1357–1369, 2020. doi: 10.1039/c9ew00882a

[83] M. Bartos and B. Kerkez, "Hydrograph peak-shaving using a graph-theoretic algorithm for placement of hydraulic control structures," *Advances in Water Resources*, vol. 127, pp. 167–179, May 2019. doi: 10.1016/j.advwatres.2019.03.016

[84] G. Golub, S. Nash, and C. V. Loan, "A Hessenberg-Schur method for the problem AX + XB= C," *IEEE Transactions on Automatic Control*, vol. 24, no. 6, pp. 909–913, Dec. 1979. doi: 10.1109/tac.1979.1102170

[85] C. A. R. Hoare, "Algorithm 64: Quicksort," *Communications of the ACM*, vol. 4, no. 7, p. 321, Jul. 1961. doi: 10.1145/366622.366644

[86] G. Golub and W. Kahan, "Calculating the singular values and pseudo-inverse of a matrix," *Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis*, vol. 2, no. 2, pp. 205–224, Jan. 1965. doi: 10.1137/0702016

Appendix

A1 Efficient optimization of the trace of the Observability Gramian

Theorem 1: The indices of the N columns of the identity matrix I that maximize the trace of the Observability Gramian are given by the indices of the N largest diagonal elements of the Controllability Gramian, when the input matrix $B = I$.

Proof: Maximizing the trace with respect to the columns of the observation matrix is given by the optimization problem:

$$\begin{aligned} \max_{K \in \mathcal{K}} \quad & \text{Tr}(W_o) \\ \text{s.t.} \quad & A^T W_o A - W_o + \sum_{k \in K} e_k e_k^T = 0 \\ & |K| = N \end{aligned} \tag{25}$$

By the definition of the Observability Gramian this is equivalent to:

$$\begin{aligned} \max_{K \in \mathcal{K}} \quad & \text{Tr}\left(\sum_{t=0}^{\infty} \sum_{k \in K} (A^T)^t e_k e_k^T A^t \right) \\ \text{s.t.} \quad & |K| = N \end{aligned} \tag{26}$$

Because the trace is a linear operator, the summation can be rearranged as:

$$\begin{aligned} \max_{K \in \mathcal{K}} \quad & \sum_{t=0}^{\infty} \sum_{k \in K} \text{Tr}\left((A^T)^t e_k e_k^T A^t \right) \\ \text{s.t.} \quad & |K| = N \end{aligned} \tag{27}$$

Using the cyclic property of the trace:

$$\begin{aligned} \max_{K \in \mathcal{K}} \quad & \sum_{t=0}^{\infty} \sum_{k \in K} \text{Tr}\left(e_k^T A^t (A^T)^t e_k \right) \\ \text{s.t.} \quad & |K| = N \end{aligned} \tag{28}$$

Note that the quantity within the trace is a scalar. Because the trace of a scalar quantity is equal to that quantity itself, the above expression is equivalent to:

$$\begin{aligned} \max_{K \in \mathcal{K}} \quad & \sum_{t=0}^{\infty} \sum_{k \in K} e_k^T A^t (A^T)^t e_k \\ \text{s.t.} \quad & |K| = N \end{aligned} \tag{29}$$

Rearranging the sums:

$$\begin{aligned} \max_{K \in \mathcal{K}} \quad & \sum_{k \in K} e_k^T \left(\sum_{t=0}^{\infty} A^t (A^T)^t \right) e_k \\ \text{s.t.} \quad & |K| = N \end{aligned} \tag{30}$$

Inserting the identity matrix:

$$\begin{aligned} \max_{K \in \mathcal{K}} \quad & \sum_{k \in K} e_k^T \left(\sum_{t=0}^{\infty} A^t I I^T (A^T)^t \right) e_k \\ \text{s.t.} \quad & |K| = N \end{aligned} \tag{31}$$

Note that the quantity inside the parentheses is the infinite time-horizon Controllability Gramian, when the input matrix B is equal to I. Thus, the columns of the identity matrix that maximize the trace of the Observability Gramian are given by the N largest diagonal indices of the corresponding Controllability Gramian:

$$\begin{aligned} \max_{K \in \mathcal{K}} \quad & \sum_{k \in K} e_k^T W_c e_k \\ \text{s.t.} \quad & A W_c A^T - W_c + I = 0 \\ & |K| = N \end{aligned} \tag{32}$$

A2 Computational complexity of optimizing the trace of the Observability Gramian

This section derives the computational complexity of the trace-based sensor placement algorithm. We will assume a linear time-invariant system defined by the state transition matrix $A \in \mathbb{R}^{n \times n}$. First, note the following known time complexities:

- The time complexity of solving a Lyapunov equation using the Bartels-Stewart algorithm is approximately $\mathcal{O}(n^3)$ [84].
- The time complexity of sorting a list of real numbers using quicksort is $\mathcal{O}(n \log n)$ [85].

The trace-based sensor placement algorithm requires the following steps:

Procedure	Time complexity
Solve the Lyapunov equation $AW_cA^T - W_c + I = 0$	$\mathcal{O}(n^3)$
Sort the diagonal elements of W_c	$\mathcal{O}(n \log n)$
Index the N largest elements	$\mathcal{O}(N)$

Ignoring the lower-order terms, the approximate computational complexity is thus $\mathcal{O}(n^3)$.

A3 Computational complexity of optimizing the rank of the Observability Gramian

This section derives the computational complexity of the rank-based sensor placement algorithm. We will assume a linear time-invariant system defined by the state transition matrix $A \in \mathbb{R}^{n \times n}$. First, note the following known time complexities:

- The time complexity of solving a Lyapunov equation using the Bartels-Stewart algorithm is approximately $\mathcal{O}(n^3)$ [84].
- The time complexity of computing the rank of a square matrix using the singular value decomposition is $\mathcal{O}(n^3)$ [86].

The rank-based sensor placement algorithm requires the following steps:

Procedure	Time complexity
Solve n Lyapunov equations: $A^T W_o A - W_o + e_i e_i^T = 0, \;\; i \in [1, n]$	$\mathcal{O}(n^4)$
Solve $k_1 = \underset{i}{\text{argmax}} \left\{\text{rank}(W_i)\right\}, \; \forall \, i \in [1, n]$	$\mathcal{O}(n^4)$
Solve $k_2 = \underset{i}{\text{argmax}} \left\{\text{rank}(W_i + W_{k_1})\right\}, \; \forall \, i \in [1, n] \setminus \{k_1\}$	$\mathcal{O}((n-1)n^3)$
Solve $k_3 = \underset{i}{\text{argmax}} \left\{\text{rank}(W_i + W_{k_1} + W_{k_2})\right\}, \; \forall \, i \in [1, n] \setminus \{k_1, k_2\}$	$\mathcal{O}((n-2)n^3)$
$\vdots$	$\vdots$
Solve $k_N = \underset{i}{\text{argmax}} \left\{\text{rank}(W_i + \sum_{j=1}^{N-1} W_{k_j})\right\}, \; \forall \, i \in [1, n] \setminus \{k_1, k_2, \ldots, k_{N-1}\}$	$\mathcal{O}((n-N+1)n^3)$

Thus, the total computational complexity is $\mathcal{O}(n^4 + n^3 \sum_{i=0}^{N-1}(n-i))$. Note that for a sparse sensor placement ($N \ll n$), the approximate computational complexity is $\mathcal{O}((N+1)n^4)$.

A4 Robustness to parameter uncertainty

Figure A1: Reconstruction error for different sensor placement strategies averaged over all 200 initial contaminant loads when sensor placements are constructed assuming $Pe = 10$, but true model has $Pe = 7.5$. Left: reconstruction error as measured by the mean-squared error. Right: reconstruction error as measured by the log of the mean-squared error.

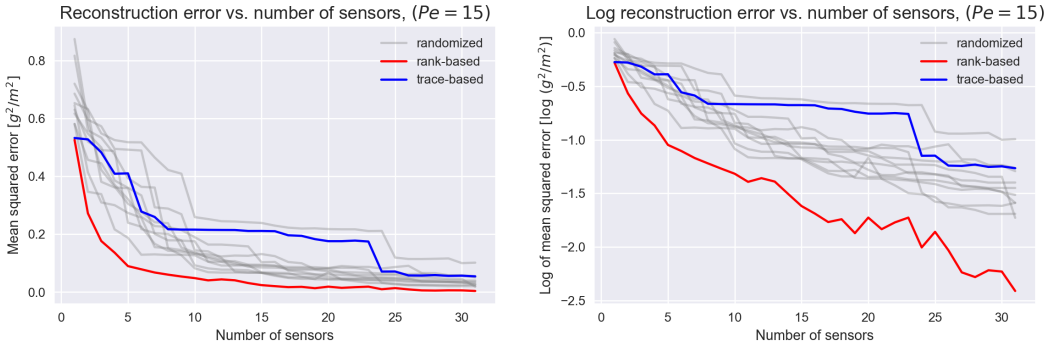

Figure A2: Reconstruction error for different sensor placement strategies averaged over all 200 initial contaminant loads when sensor placements are constructed assuming $Pe = 10$, but true model has $Pe = 15$. Left: reconstruction error as measured by the mean-squared error. Right: reconstruction error as measured by the log of the mean-squared error.