Supplementary Information for

Scalable, Adaptive, and Risk-Informed Design of Hydrological Sensor Networks

Jeil Oh and Matt Bartos

Contents

1 Mathematical formulation of benchmark sensor placement methods

1.1 Betweenness Centrality

Betweenness centrality quantifies the importance of a node in a graph based on the number of shortest paths passing through it. For a graph $G = (V, E)$, where V is the set of nodes and E is the set of edges, the betweenness centrality $C_B(v)$ of a node $v \in V$ is defined as:

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}, \tag{1}$$

where σ_{st} is the total number of shortest paths between nodes s and t, $\sigma_{st}(v)$ is the number of shortest paths between s and t that pass through node v.

Nodes with higher $C_B(v)$ scores are more central in the network, acting as key bridges for flow or network connectivity. Betweenness centrality is particularly effective for identifying nodes with topological significance in sensor placement.

1.2 Leverage Score Method

The leverage score method identifies critical spatial locations by analyzing the contribution of each location to the principal components of the dataset. Given a data matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$, where rows represent spatial locations and columns represent temporal observations, the singular value decomposition (SVD) of $\mathbf{X}$ is:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top, \tag{2}$$

where $\mathbf{U} \in \mathbb{R}^{m \times m}$ is the matrix of left singular vectors, $\mathbf{\Sigma} \in \mathbb{R}^{m \times n}$ is the diagonal matrix of singular values, and $\mathbf{V} \in \mathbb{R}^{n \times n}$ is the matrix of right singular vectors.

The leverage score ℓ_i for the i-th row (spatial location) of $\mathbf{X}$ is computed as:

$$\ell_i = \|\mathbf{u}_i\|^2, \tag{3}$$

where $\mathbf{u}_i$ is the i-th row of $\mathbf{U}_k$, the truncated matrix containing the first k columns of $\mathbf{U}$ corresponding to the largest k singular values. Expanding this, the leverage score can also be written as:

$$\ell_i = \sum_{j=1}^{k} U_{ij}^2, \tag{4}$$

where U_{ij} represents the (i, j)-th entry of $\mathbf{U}_k$, and k is determined based on the desired rank approximation of $\mathbf{X}$.

Rows with the highest leverage scores correspond to locations that contribute most significantly to the data's variability, making them ideal candidates for sensor placement.

2 Algorithmic Overview

To provide a clearer step-by-step procedure for our proposed sensor placement method, we summarize the framework in Algorithm 1.

Algorithm 1 Sensor Placement using QR Decomposition with Column Pivoting

Require: $\mathbf{X} \in \mathbb{R}^{n \times m}$ (data matrix), number of sensors r, optional weights $\mathbf{W}$, optional fixed column set F

Ensure: Selected column set J corresponding to sensor locations

if weights are provided **then**
    Form $\mathbf{X}_w = \mathbf{X}\mathbf{W}$
else
    $\mathbf{X}_w = \mathbf{X}$
end if
if fixed columns F are specified **then**
    Partition $\mathbf{X}_w$ into $\mathbf{X}_{w,F}$ and $\mathbf{X}_{w,R}$
    Compute $\mathbf{X}_{w,F} = \mathbf{Q}_F\mathbf{R}_F$
    Form $\mathbf{X}'_{w,R} = \mathbf{X}_{w,R} - \mathbf{Q}_F(\mathbf{Q}_F^T\mathbf{X}_{w,R})$
    Apply pivoted QR to $\mathbf{X}'_{w,R}$ to obtain permutation and factorization
    Combine fixed and pivoted columns to get full permutation $\mathbf{P}$
else
    Apply pivoted QR directly to $\mathbf{X}_w$ to obtain $\mathbf{P}$
end if
Let J be the indices corresponding to the first r columns in $\mathbf{P}$
return J

In summary, our method identifies a near-optimal subset of sensor locations by leveraging the rank-revealing properties of the pivoted QR factorization. The approach is flexible, accommodating fixed sensors, weighted priorities, and large-scale datasets.

43 3 Supplementary Figures

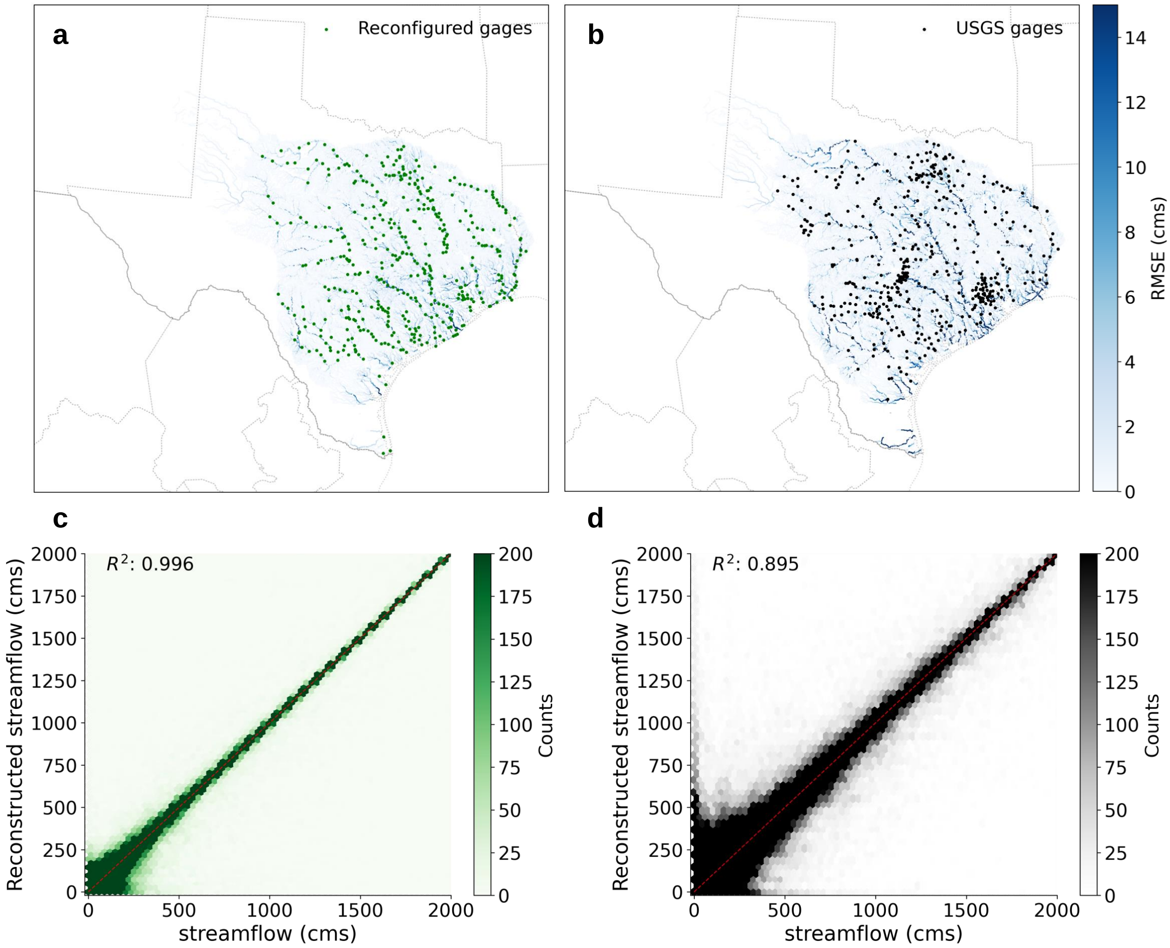

Figure 1: Comparison of streamflow reconstruction performance between the QR-based sensor placement framework and the current U.S. Geological Survey (USGS) gauge network.(a) Reconfigured gauge network derived from the QR-based placement, with the background map displaying RMSE errors for each stream (darker colors indicate higher errors). (b) Existing USGS gauge network, with corresponding RMSE errors shown on the same scale for comparison. (c) Scatter plot comparing reconstructed streamflow from the QR-based sensor placement to the National Water Model (NWM) retrospective data, achieving an R-squared value of 0.996. (d) Scatter plot comparing reconstructed streamflow from the USGS gauge network to NWM retrospective data, yielding an R-squared value of 0.895.

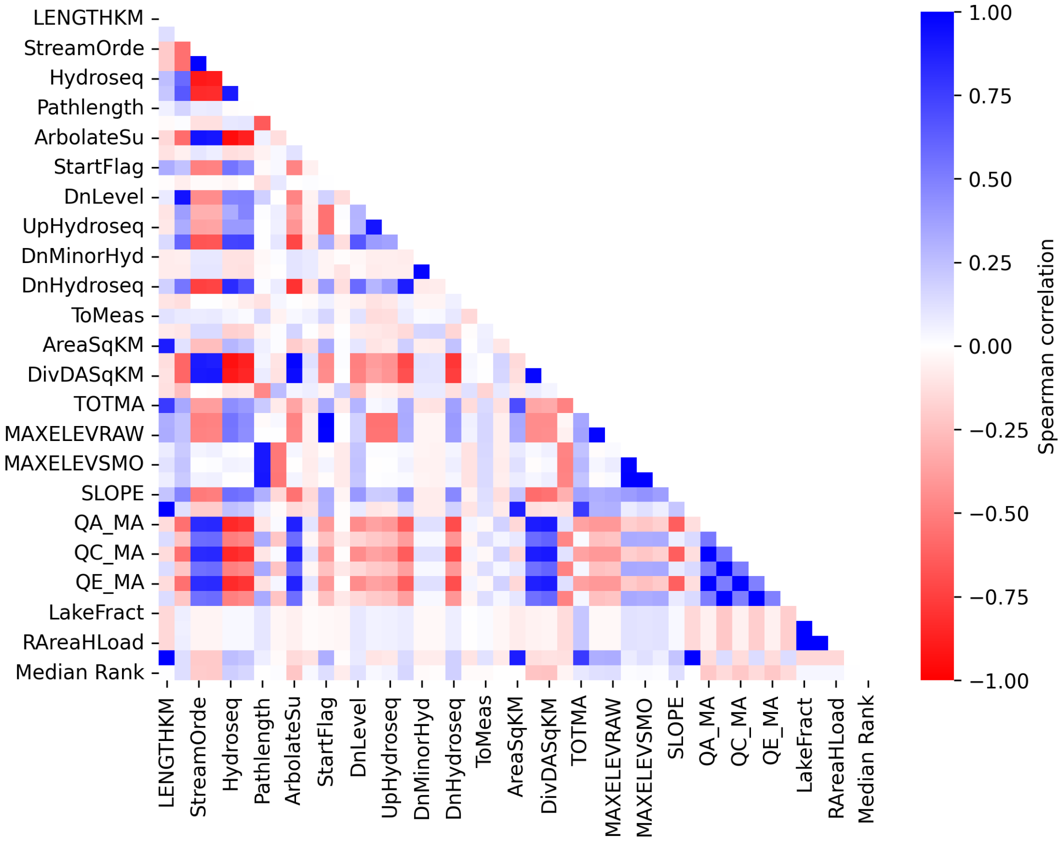

Figure 2: Spearman correlation plot between sensor rankings and key attributes from the National Hydrography Dataset (NHD), such as slope, drainage area, and stream order.

4 Supplementary Tables

Table 1: NHDPlus Variables and Descriptions

Variable Name	**Description**
LengthKM	Length of the flowline in kilometers
StreamLeve	Stream level; distinguishes main stream paths from tributaries
StreamOrde	Stream order; Modified Strahler Stream Order
StreamCalc	Stream Calculator
Hydroseq	Hydrologic sequence number; places flowlines in hydrologic order
LevelPathI	Level Path Identifier–Hydrologic sequence number of most downstream NHDFlowline feature in the level path
PathLength	Distance to the terminal NHDFlowline feature downstream along the main path
TerminalPa	Terminal Path Identifier - Hydrologic sequence number of terminal NHDFlowline feature
ArbolateSu	Arbolate Sum - Kilometers of stream upstream of the bottom of the NHDFlowline feature
Divergence	Indicates flowline divergence: 0 = none, 1 = main, 2 = minor path
StartFlag	Start flag; 1 = start of flowline, 0 = not start
TerminalFl	Terminal flag; 1 = terminal, 0 = not terminal
DnLevel	Streamlevel of main stem downstream NHDflowline feature
UpLevelPat	Upstream mainstem level path identifier
UpHydroSeq	Upstream mainstem hydrologic sequence number
DnLevelPat	Downstream mainstem level path identifier
DnMinorHyd	Downstream minor hydrologic sequence number
DnDrainCou	Count of NHDFlowline features immediately downstream
DnHydroSeq	Downstream mainstem hydrologic sequence number
FromMeas	ReachCode route measure (m-value) at bottom of NHDFlowline feature
ToMeas	ReachCode route measure (m-value) at top of NHDFlowline feature
RtnDiv	Returning Divergence Flag; 0 = no upstream divergences return at the top of this NHDFlowline feature, 1 = one or more upstream divergences returned to the network at the top of this NHDFlowline feature
AreaSqKM	Area of the supercatchment in square kilometers
TotDASqKM	Total Upstream Cumulative Drainage Area, in square kilometers, at the downstream end of the NHDFlowline feature
DivDASqKM	Divergence-routed Cumulative Drainage Area, in square kilometers, at the downstream end of the NHDFlowline feature
Tidal	Indicates tidal influence; 1 = tidal, 0 = not tidal.
TOTMA	Mean Annual Time of Travel (days)
HWNodeSqKM	Catchment area in square kilometers that drains to the headwater node of the NHDFlowline feature
MaxElevRaw	Maximum elevation (unsmoothed) in centimeters
MinElevRaw	Minimum elevation (unsmoothed) in centimeters
MaxElevSmo	Maximum elevation (smoothed) in centimeters
MinElevSmo	Minimum elevation (smoothed) in centimeters
Slope	Slope of flowline (meters/meters) based on smoothed elevations
SlopeLenKm	NHDFlowline feature length (kilometers) used to compute slope.
QA_MA	Mean Annual Flow from runoff (cfs)
VA_MA	Mean Annual Velocity for QA (fps)
QC_MA	Mean Annual Flow with Reference Gage Regression applied to QB (cfs)
VC_MA	Mean Annual Velocity for QC (fps)
QE_MA	Mean Annual Flow from gage adjustment (cfs)
VE_MA	Mean Annual Velocity from gage adjustment (fps)
LakeFract	Fraction of lake assigned to Flowline
SurfArea	Lake surface area assigned to flowline in square meters
RAreaHLoad	Reciprocal area hydraulic loads assigned to flowline in days/meter