THE UNIVERSITY OF BRITISH COLUMBIA
Civil Engineering
Faculty of Applied Science

# British Columbia Streamflow Monitoring Optimization

Dan Kovacek [1] & Steven Weijs [1]

[1]University of British Columbia

## Motivation

The British Columbia Hydrometric Service operates a streamflow monitoring network covering approximately $10^6$ km$^2$. Monitoring density is sparse and unevenly distributed in space. Given the long delay between network design decisions and goals of data collection, how might monitoring network expansion decisions anticipate future information needs?

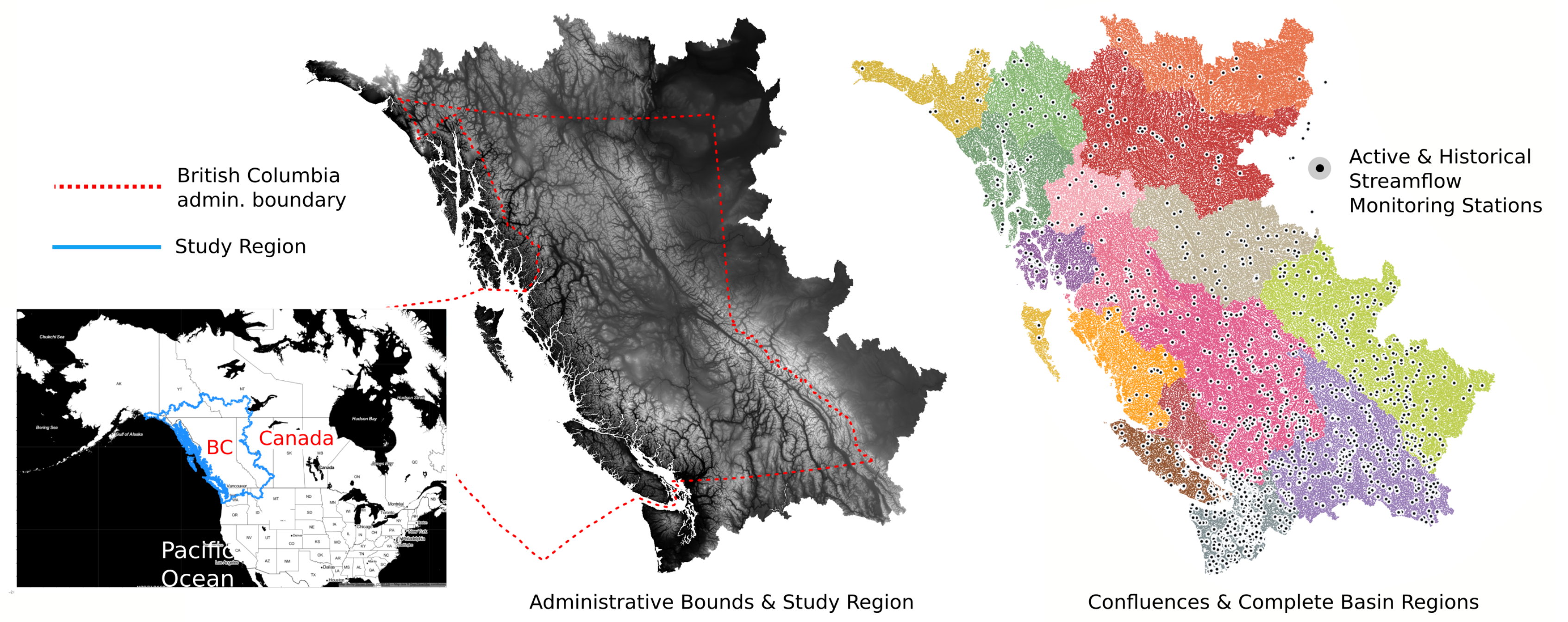
British Columbia admin. boundary
Study Region
BC
Canada
Pacific Ocean
Active & Historical Streamflow Monitoring Stations
Administrative Bounds & Study Region
Confluences & Complete Basin Regions

Figure 1. USGS 3DEP DEM is used to generate basin polygons for $O(10^6)$ river confluences. The study region expands beyond the BC border (red dashed line) to reduce edge selection effects in the optimization problem. Coloured areas (right) correspond to complete watershed regions. There are roughly 1500 active and historical monitoring locations in the study region based on the set of basins included in the HYSETS study [1].

The network expansion problem is approached as a maximization of total network information, or finding a set of unmonitored locations associated with the greatest expected reduction in surprise for other unmonitored locations. In other words, what locations provide the best regional information transfer model proxies for the greatest number of unmonitored locations?

## Problem Components

The network optimization problem includes data acquisition and pre-processing, model development and validation, and subjective methodology components that are important to address.

- **Decision Space Characterization** A set of candidate monitoring locations (CMLs) is generated from USGS 3DEP DEM [4], along with static attributes describing terrain, soil (GLHYMPS [2]), and land cover (NALCMS [3]), following the HYSETS dataset [1].
- **Simulated vs. Observed Divergence** For all pairs of monitored locations meeting a minimum data concurrency criteria, a model is developed to map basin attribute similarity to divergence of observed and simulated daily streamflow distributions, where the simulated streamflow is generated from a simple area ratio model.
- **Objective Function Development** a baseline $D_{KL}$ is computed for each CML corresponding to the lowest *expected* divergence from all monitored locations. The expected total reduction in surprise from adding a monitoring station is then calculated for each CML as the sum of expected decrease in expected $D_{KL}$ from baseline at all other CMLs.
- **Model Validation** Cross validation is used to test the model on stations left out-of-sample.
- **Questions** i) complexity $O(n^2)$ becomes intractable for large decision spaces, ii) interpretation/validity of the $D_{KL}$-based model, iii) time series discretization, and iv) self similarity in basin representation (input data quality).
- **Uses** While the primary application is monitoring network optimization, the database design is flexible and extensible to support a wide array of questions. A higher-level goal is to develop a database that shortens the feedback loop between generating and testing ideas involving large samples and spatio-temporal comparisons.

## Decision Space Characterization

Static attributes are derived for all CML basins following [1]. The geometry and attribute information is stored in a PostGIS database for efficient spatial querying on multiple geometry types. A minimum basin area of 1km$^2$ is set to support subsequent research questions in stream network accuracy & precision, and it is left to dataset users to justify a lower limit on drainage area.

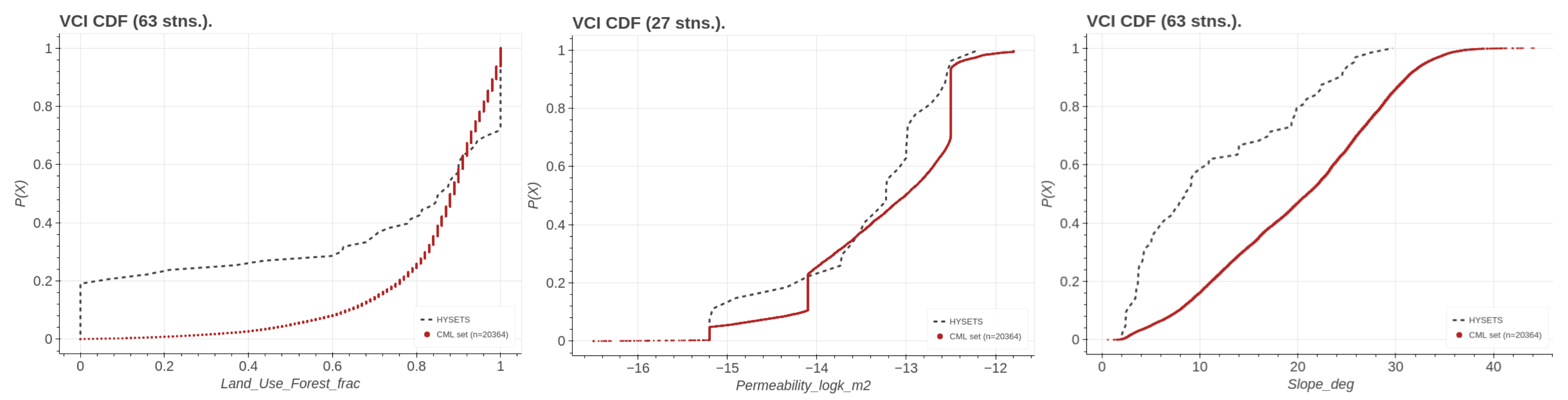

Figure 2. The monitoring network can be compared to the region it represents based on attributes. For example on Vancouver Island, the set of monitored basins over-represents unforested basins (left), reasonably represents the range of permeability (middle), and underrepresents steep basins (right).

## Simulated vs. Observed Model

The study region contains just over 1500 active and historical streamflow monitoring stations. 483 of these stations have at least 20 years of concurrent record with another station, yielding a sample of just over 116K basin pairs, or fewer if we impose a spatial distance limit between basin centroids. The "distance" between each basin pair is the L1 norm $L_{ij} = \sum_{k=1}^{n} |l_{ik} - l_{jk}|$. The divergence between observed streamflow time series distribution $P$ and the simulated distribution $Q$ is the Kullback-Leibler divergence $D_{KL}(P \parallel Q) = \sum_{x \in \chi} P(x) log\left(\frac{P(x)}{Q(x)}\right)$ where the simulated daily streamflow series is simply a function of the area ratio and the proxy observed series: $X_s^i = X_o^j \frac{A_i}{A_j}$.

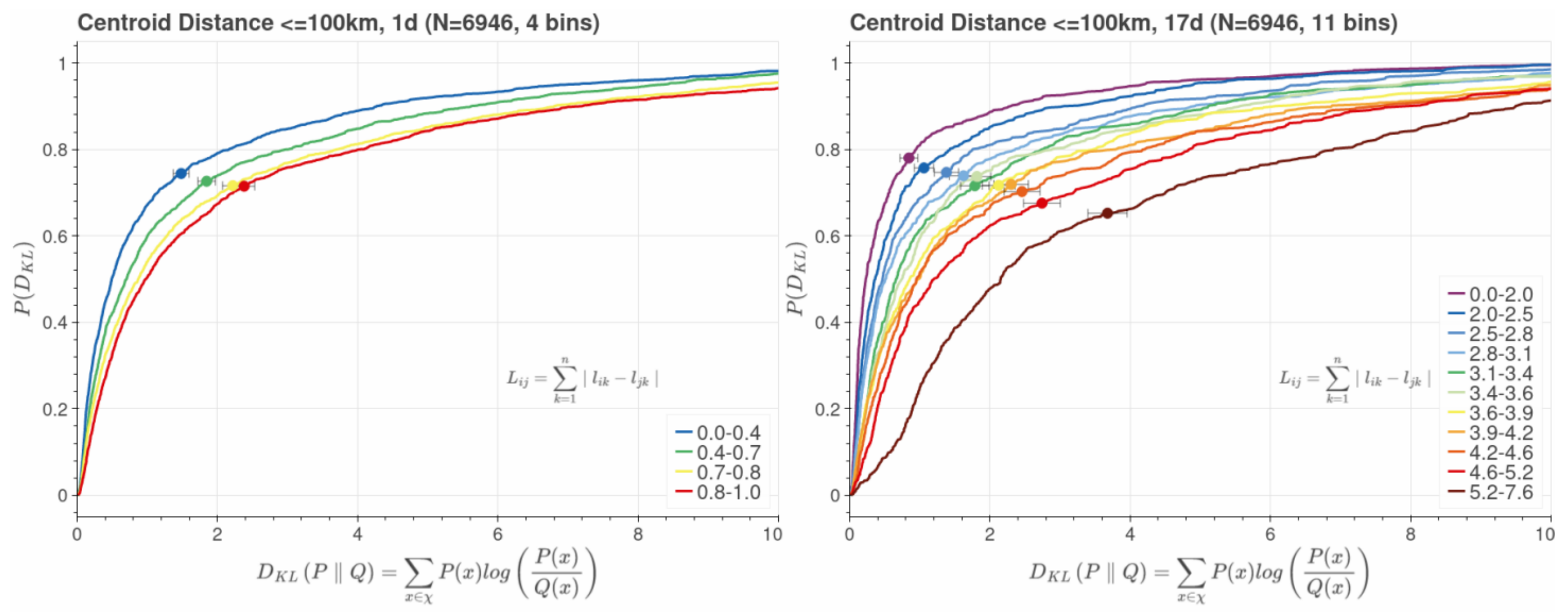

Figure 3. The total sample is recursively divided into as many attribute distance ($L_{ij}$) bins as the data support statistically different expected values of $D_{KL}$. At left, $L_{ij}$ is defined only by the spatial distance between basin centroids, while at right $L_{ij}$ is the L1-norm of the 17-d vector of static attributes, suggesting that static attributes contain information about the divergence of long-term flow distribution.

The model mapping attribute space distance to expected divergence between two locations is a piece-wise function:

$$f(L_{ij}) = \begin{cases} E_1[D_{KL}(P_i||Q_j)] & \text{if } 0 \leq L_{ij} < b_1 \\ E_2[D_{KL}(P_i||Q_j)] & \text{if } b_1 \leq L_{ij} < b_2 \\ \vdots & \\ E_n[D_{KL}(P_i||Q_j)] & \text{if } b_{n-1} \leq L_{ij} \leq b_n \end{cases}$$

## Method & Objective Function

The objective function for the greedy network expansion problem is a maximization of the total reduction of expected divergence from baseline. The steps are outlined as follows:

1. Baseline divergence is first computed for each of $n$ CMLs as the minimum expected divergence from $m$ monitored locations: $\mathbf{D}'_i = \min_{k=1}^{m} \mathbb{E}(D_{KL}^{ik}) \quad \forall i \in \{1, 2, \ldots, n\}$.
2. Setting each CML in turn as the target location, the expected divergence of all other CMLs ($j > i$) from the target ($i$) is computed as: $\mathbf{D}_i = \mathbb{E}(D_{KL}^{ij}) \quad \forall i, j \in \{1, 2, \ldots n\} : j > i$.
3. The expected reduction in divergence from expanding the network to location $i$ is $\mathbf{\Delta}_i = \mathbf{D}'_i - \mathbf{D}_i$ where negative values are set to zero since they do not represent an improvement over the baseline.
4. The optimal location for network expansion $i^*$ is one that maximizes the combined reduction in divergence at all unmonitored locations (CMLs).

The objective function is then:

$$i^* = \text{argmax}_{i \in \{1,2,\ldots n\}} \sum_{j=1, j>i, \Delta_{ij}>0}^{n} \Delta_{ij}$$

The probabilistic model used to estimate expected divergence does not produce a unique solution, but yields grades, or classes describing the quality of a candidate monitoring location for providing information about other unmonitored locations. This result provides important flexibility for network expansion decisions.

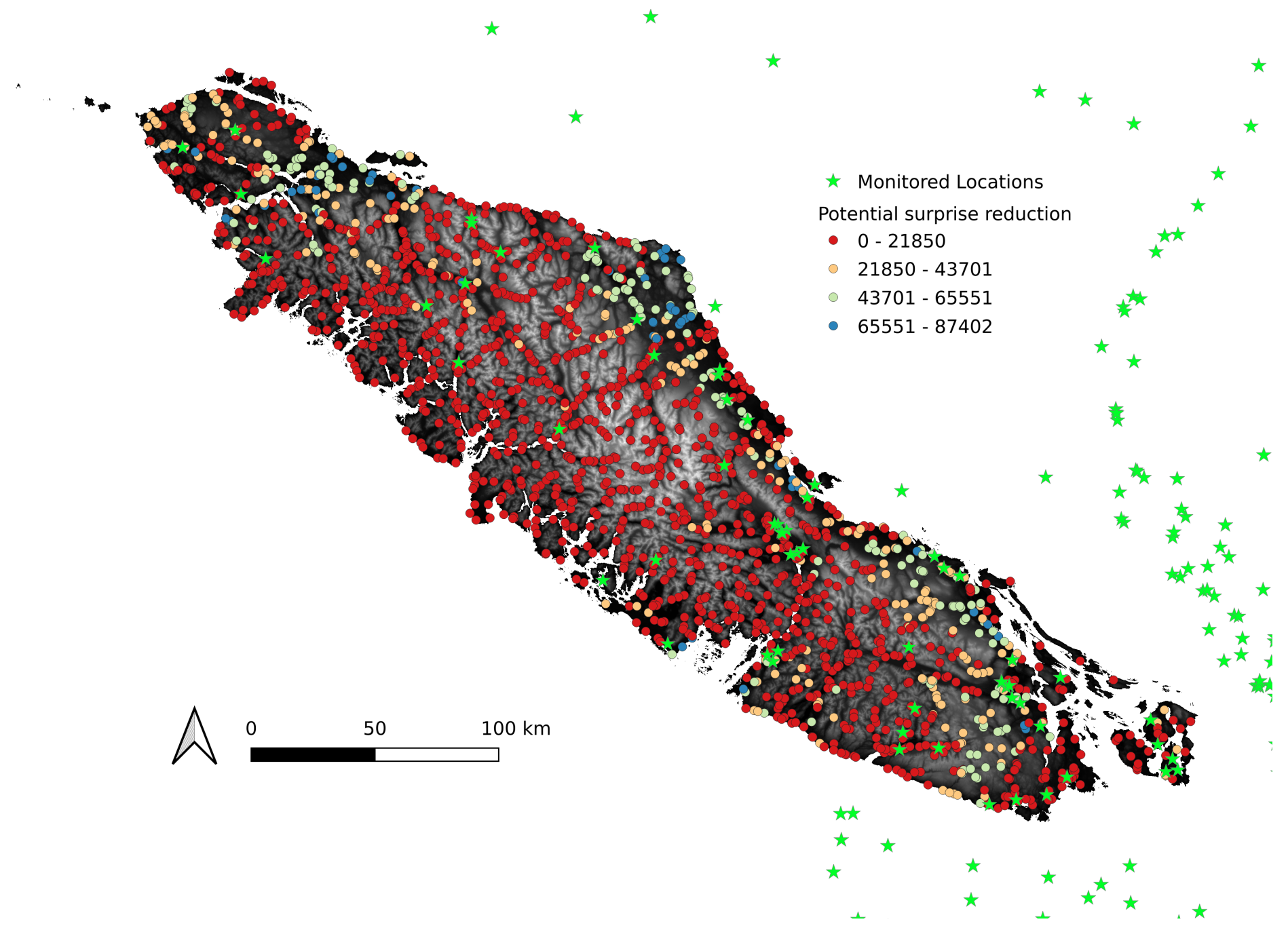

Figure 4. Vancouver Island is shown as an example of mapping the expected reduction in divergence (surprise) in space to inform network expansion decisions.

## References

[1] Richard Arsenault, François Brissette, Jean-Luc Martel, Magali Troin, Guillaume Lévesque, Jonathan Davidson-Chaput, Mariana Castañeda Gonzalez, Ali Ameli, and Annie Poulin.
A comprehensive, multisource database for hydrometeorological modeling of 14,425 north american watersheds.
*Scientific Data*, 7(1):1–12, 2020.

[2] Tom Gleeson.
GLobal HYdrogeology MaPS (GLHYMPS) of permeability and porosity, 2018.

[3] R Latifovic, C Homer, R Ressl, D Pouliot, SN Hossain, RR Colditz Colditz, I Olthof, C Giri, and A Victoria.
North american land change monitoring system (nalcms).
*Remote sensing of land use and land cover: principles and applications. CRC Press, Boca Raton*, 2010.

[4] U.S. Geological Survey.
Usgs 3d elevation program digital elevation model, 2020.
[Online; accessed 3 March 2022].

https://geoinfotheory.org/

Information Theory in the Geosciences Workshop, Garmisch-Partenkirchen, Germany, 10-14 Sept. 2023

dkovacek@mail.ubc.ca