creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

공학박사 학위논문

Sediment Load Estimation Based on Optimized Parameters and Clusters with Hydro-acoustic Backscatter

초음파 산란도를 활용한 최적 인자 및 군집 기반 유사량 산정 기법

2024년 2월

서울대학교 대학원

건설환경공학부 건설환경공학전공

노 효 섭

초음파 산란도를 활용한 최적 인자 및 군집 기반 유사량 산정 기법

Sediment Load Estimation Based on Optimized Parameters and Clusters with Hydro-acoustic Backscatter

지도교수 박 용 성

이 논문을 공학박사 학위논문으로 제출함

2023 년 12 월

서울대학교 대학원

건설환경공학부 건설환경공학전공

노 효 섭

노효섭의 공학박사 학위논문을 인준함

2023 년 12 월

위 원 장 _____ (인)

부 위 원 장 _____ (인)

위 원 _____ (인)

위 원 _____ (인)

위 원 _____ (인)

Abstract

Sediment Load Estimation Based on Optimized Parameters and Clusters with Hydro-acoustic Backscatter

Hyoseob Noh

Department of Civil and Environmental Engineering

Civil and Environmental Engineering Major

The Graduate School of Seoul National University

Sediment transport in natural rivers holds significant importance for civil and environmental engineering. However, limitations arise in enhancing the spatial and temporal resolutions of sediment monitoring due to the labor-intensive nature of the traditional sediment measurement method, which relies on sample analysis. Recent efforts were made to improve the temporal resolution of suspended sediment concentration (SSC) monitoring by fitting the backscattering signal of horizontal acoustic Doppler current profilers (H-ADCPs) with measured SSC. Although the H-ADCP-based monitoring increases the data acquisition rate significantly, there are limitations in SSC estimation accuracy due to nonlinearity in the backscattering signal and SSC. This study primarily aims to improve a sediment load estimation method using hydro-acoustic

backscatter by the newly proposed parameter optimization methods for clustering and support vector regression (SVR) techniques. To consider nonlinearity in the SSC-backscattering relationship, additional hydraulic variables, in addition to backscattering signal and machine learning techniques, are employed. Model Selection by Global Optimization for SVR (MOSGO-SVR), which fine-tunes hyperparameters and input variables, is proposed in order to enhance SSC predictability efficiently. It simultaneously determines the combination of input variables and hyperparameters of SVR using global optimization. An iterative clustering method is also employed to find the optimal clustering model. These techniques are applied to H-ADCP data to enhance predictability and measurability based on the procedure to determine the SSC monitoring model with H-ADCP signal using MOSGO-SVR, incorporating the combination of input variables. Notably, the proposed sediment monitoring procedure includes simultaneous total load estimation and SSC monitoring. Next, the iterative clustering method is used to classify sediment monitoring stations in South Korea. Based on clustering analysis, the application strategy of the H-ADCP-based SSC monitoring method to sediment ungauged stations is discussed. Third, a new hydraulic model for total load estimation is derived from the suspended-to-total load fraction using SVR and symbolic regression methods to complement the hydraulic aspect of the proposed sediment monitoring procedure. As a result, this study presents a sediment load assessment framework using H-ADCP, integrating the obtained results. This con-

tribution enhances the performance of sediment estimation, allowing for systematic estimation of total loads, even in sediment ungauged stations. The application of these findings is expected to advance our understanding of river sediment management and sediment transport mechanisms, improving the predictability and measurability of sediment monitoring.

Keywords: Total sediment load, Sediment transport, Sediment monitoring, H-ADCP, Acoustic backscatter, Optimization, Machine learning regression, Clustering

Student Number: 2019-38726

Contents

List of Figures

List of Tables

Notation

$\boldsymbol{\mu}$, Σ	Mean and covariance matrix of Gaussian distribution
$\mathcal{N}$	Gaussian distribution
$\vec{D}$	Diagonal matrix of the covariance matrix of the complex
$\vec{m}_i$	i-th SOM node
$\vec{w_p}$	Weight vector without p-th vector component
$\vec{w}$	Weight vector
$\vec{x_j}^{(-p)}$	$\vec{x_j}$ without p-th feature
$\vec{X}_c, \vec{X}_r, \vec{X}_{oc}, \vec{X}_{ic}$	The centroid, reflected, outside contraction, and inside contraction parameter points in global optimization
A	Cross-sectional area
a_t	Transducer radius
A_T,B_T,C_T,D_T	Coefficients of the Turowski model
A_{drain}	Drainage area
a_{SL},b_{SL}	Regression coefficients for suspended load rating curves
a_{TL},b_{TL}	Regression coefficients for total load rating curves
AIC	Akaike information criterion
b	Offset of the linear SVR
BIC	Bayesian information criterion
c	Wave speed
$C(\cdot)$	Sediment concentration
C_1,C_2	Regression coefficient

C_g	Curvature coefficient
c_i	the centroid of the cluster S_i
c_p	Ranking criterion in RFE
C_u	Uniformity coefficient
C_v	Volumetric sediment concentration
C_{ref}	Reference sediment concentration
C_{SVR}	Regularization coefficient of SVR
$C_{w,t}$, $C_{w,b}$	Total load and bedload concentrations by weight
C_w, C_{ppm}	Sediment concentration by weight and parts per million
d_*	Dimensionless grain size
d_{84},d_{50},d_{16}	Sediment particle sizes of the 84%, 50%, and 16% of the material by weight
$d_{ce}(Q_k, Q_l)$	Distance between clusters Q_k and Q_l
d_s	Characteristic sediment particle size diameter
DBI	Davies-Boulding index
E	Ratio of bed layer thickness to flow depth
e_B	Bagnold coefficient
$f(\cdot)$	Arbitrary function
F_{sus}	Suspended-to-total load fraction
$Fr = \frac{U}{\sqrt{gh}}$	Froude number
$Fr_d = \frac{U}{\sqrt{g(G_s-1)d_{50}}}$	Densimetric Froude number
G_s	Specific gravity of sediment
$Gr = \frac{1}{2}(\frac{d_{84}}{d_{50}} + \frac{d_{50}}{d_{16}})$	Gradation coefficient
I_v	Optimization flag designating the input data columns
J_1,J_2,J_1',J_2'	Einstein integral components

K	The number of subsets in clustering or cross-validation
k	Wavenumber
k, l	Location index in the two-dimensional SOM grid
$K(x, x)$	Kernel function
k^*, l^*	Location index of the winning node in the two-dimensional SOM grid
k_Q	Coefficent of the general discharge relationship
k_s	Roughness height
LL	Log likelihood
MB	Measured backscatter
N	The number of dataset
n_i^*	The sorted rank of the i-th individual
N_p	The number of parametes
N_w, N_h	Width and height of SOM grid
n_{comp}	The number of individuals in the complex
$p(\cdot)$	Probability density function
P_i	i-th bin's size fraction
Q	Flow discharge
Q_n	Steady-uniform flow discharge
Q_{bm}	Bed material load
Q_m	Measured sediment load
Q_{TL}, Q_{SL}, Q_{BL}	Sediment discharge (total, suspended, bedload)
q_{TL}, q_{SL}, q_{BL}	Unit sediment discharge (total, suspended, bedload)
Q_{um}	Unmeasured sediment load
Q_{wl}	Wash load
QE	Quantization error

r	Sound wave travel distance
R^2	Coefficient of determination
R^2_{CV}	Coefficient of determination from cross-validation
$r_* = r\lambda/(\pi a_t^2)$	Dimensionless sound wavelength
$Re_* = \frac{u_* h}{\nu}$	Shear Reynolds number
$Re_h = \frac{Uh}{\nu}$	Flow Reynolds number
$Re_w = \frac{w_s d_{50}}{\nu}$	Falling particle Reynolds number
$Re_{d*} = \frac{u_* d_{50}}{\nu}$	Particle shear Reynolds number
$Re_{d50} = \frac{U d_{50}}{\nu}$	Particle Reynolds number
Ro	Rouse number
S_C	Distance between the center of clusters and data points
S_i	i-th cluster
SCB	Sediment corrected backscatter
SL	Source level
T	Temperature
T_K	Temperature in Kelvin
t_k, μ_k, σ_k	k-th Gaussian weight, mean matrix, covariance matrix on the Giussian mixture
TE	Topological error
TL	Transmission loss
TS	Target strength
U	Cross-section averaged streamwise velocity
u	Streamwise velocity at a point
u_*	Shear velocity
U_{cr}	Critical velocity at incipient particle motion

u_{TE}	Topoligical error function
W	Channel width
w_i	Particle falling velocity of of i-th bin
w_s	Sediment falling velocity
$w_{k^*l^*}$	Wining node
$w_{s*} = w_s/\sqrt{(G_s-1)gd_s}$	Dimensionless falling velocity
WCB	Water corrected backscatter
x, y, z	Coordinates (streamwise, transverse, vertical)
Y_{obs},Y_{est}	Observed and estimated values
z_n	Minimum height of the suspended sediment sampler nozzle
α,α^*	Lagrangian multiplier
α_w,α_s	Backscatter correction coefficients for water and sediment
β	Ratio of the turbulent mixing coefficient of sediment to the momentum exchange coefficient
δ_b	Bedload layer thickness
ϵ	Margin width of SVR
γ_w,γ_s	Specific weight of water and sediment
γ_{RBF}	inverse of the influence radius of the RBF kernel
κ	von Karman coefficient
κ_s^2	Backscatter strength coefficient
λ	Wavelength
λ_n	Neighborhood function
ν	Kinematic viscosity of water
ω	Frequency
ψ	Irregular sound diffusion correction factor

ρ_w,ρ_s	Density of water and sediment
$\sigma_g = (\frac{d_{84}}{d_{16}})^{1/2}$	Gradation of the sediment mixture
σ_s	Backscatter strength of sound wave
τ	Shear stress
$\tau_* = \frac{u_*^2}{g(G_s-1)d_{50}}$	Shields number
τ_0	Bed shear stress
τ_i', τ_{ci}	Tractivee and critical tractive forces of i-th bin
τ_{xz}	Turbulent shear stress at distance z above the bed
ξ,ξ^*	Slack variables of SVR
ζ_s	Normalized attenuation coefficient
$\overline{Y_{(obs)}}$	Mean observed value
$\vec{x_i}$,$\vec{x_j}$	i-th and j-th input data point vector
A_{M3}	Coefficient of the MGGP3 model
A_{M5},B_{M5}	Coefficients of the MGGP5 model
A_{O3},B_{O3},C_{O3},D_{O3},E_{O3}	Coefficients of the Operon3 model
A_{O5},B_{O5}	Coefficients of the Operon5 model

Chapter 1. Introduction

1.1 Necessity and background

Sediment yield observations in rivers are crucial for river management. The sediment deposition significantly affects flood control capabilities, the design life of hydraulic structures, and water quality, including turbidity. However, direct sampling methods for river sediment observation are labor-intensive, resulting in a very low data acquisition rate.

Additionally, simultaneous flow rate and sediment concentration measurements are necessary to assess suspended sediment loads accurately. However, due to the time required for conducting both suspended sediment concentration (SSC) and flow rate measurements, carrying out these observations presents practical challenges.

Due to the time-consuming and labor-intensive sediment sampling process, practitioners often develop and apply simple power-law-type flow rate-sediment loads rating curves. While this method is practically useful, Rajaee et al. (2011) pointed out that it lacks the ability to reproduce complex sediment behavior occurring in natural rivers, e.g., during storm events. Furthermore, constituents of suspended sediment at low flows differ from those at high flows, so the rating coefficients are not constant over the flow rate range (Hoffmann et al., 2020). Noh et al. (2023b) demonstrated that relying on such curves for sediment budget evaluations during flooding events can

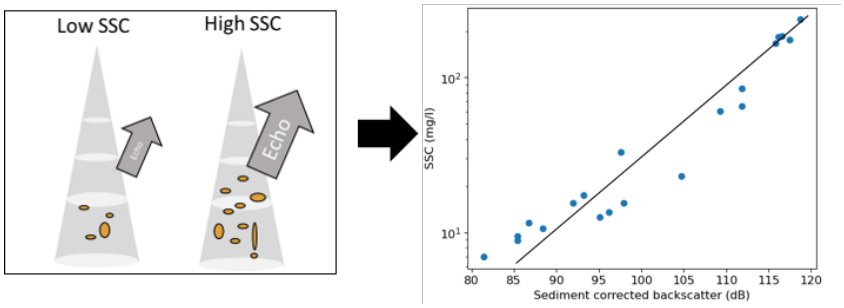


Figure 1.1 Schematic example of the increase of the backscattering signal along SSC and correlating it to SSC

yield incorrect outcomes with errors reaching up to 10,000 t/d.

Recently, attempts have been made to observe sediment yield using acoustic Doppler current profilers (ADCPs). This approach utilizes the backscatter values of the ADCP signal, which are correlated with the concentration of suspended particles in the water (Urick, 1948, 1975) as shown in Figure 1.1. Studies have demonstrated the feasibility of real-time suspended sediment observations using horizontal ADCP (H-ADCP; Topping et al. 2006, 2007; Landers 2012; Landers et al. 2016; Guerrero et al. 2016; Haught et al. 2017; Guerrero and Di Federico 2018; Szupiany et al. 2019; Aleixo et al. 2020; Son 2021; Noh et al. 2022, 2023b,c). This approach enables the concurrent monitoring of SSC and flow rate, facilitating the simulation of hysteresis in sediment transport. Consequently, utilizing H-ADCP at automated flow measurement stations in sediment monitoring can reduce the temporal gap in sediment yield observations.

In South Korea, 62 automatic flow measurement stations are currently equipped with H-ADCP, in operation for monitoring river water flow in Korea (MoE, 2019b).

H-ADCPs provide flow data at 10-minute intervals and offer a significant advantage in measuring sediment concentration with fewer limitations than traditional sampling methods, as long as the ADCP remains submerged beneath the water surface. Therefore, if H-ADCPs in monitoring stations are used for sediment monitoring, worker safety can be ensured, and temporal resolution can be dramatically improved. This approach enables the concurrent monitoring of SSC and flow rate, facilitating the simulation of hysteresis in sediment transport, thereby advancing flow and sediment monitoring accuracy.

1.1.1 Enhancing the performance of H-ADCP-based SSC monitoring

In H-ADCP-based SSC monitoring, cross-section averaged SSC values are derived using the sediment-corrected backscatter (SCB), obtained from a small portion of the ensonified volume. This method is called the index concentration method, which is akin to the index velocity method employed in flow rate monitoring. The typical estimation of sediment concentration using ADCPs follows the relationship:

$$\log_{10}(\text{SSC}_V) = C_1 \cdot \text{SCB} + C_2, \tag{1.1}$$

where C_1 and C_2 denote the regression coefficients, and SCB represents sediment-corrected backscatter, a concept to be discussed in the next chapter. However, the current H-ADCP-based SSC monitoring method faces several challenges in achieving

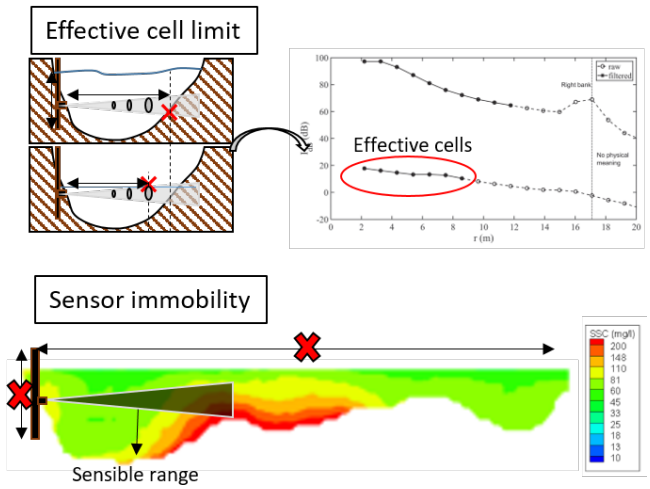


Figure 1.2 A schematic diagram of the H-ADCP sensible area limitation

accurate SSC estimation.

Firstly, accurately determining SCB is challenging due to its nonlinear dependence on sediment particles and water temperature (Landers, 2012; Landers et al., 2016; Guerrero et al., 2016; Guerrero and Di Federico, 2018; Aleixo et al., 2020). For instance, suspended sediment particle size distribution (SSPSD) vary during rainfall events (Landers and Sturm 2013;MoE, 2019b 2019; Hoffmann et al. 2020). As emphasized by Hoffmann et al. (2020), a nonlinear regime shift occurs in the SSC-flow rate rating curves, resulting in a break in sediment rating. Addressing this nonlinearity becomes necessary, and this can be achieved through empirical relationships, such as employing a nonlinear method or incorporating multiple ratings in one station.

An additional challenge arises from the constrained coverage area of H-ADCP in both depth and transverse directions. H-ADCP captures signals from the transmitter

to the opposite bank through numerous ensonified cells. However, unwanted sonar reflections from the bottom and water surface can occur, introducing noise in the backscattering signals. Consequently, SCB values must be derived from only a few cells near the H-ADCP's soundwave transmitter. Coupled with H-ADCP's vertical immobility, these coverage limitations lead to the incapability of collecting vertical and transverse information (Figure 1.2. These constraints adversely affect H-ADCP's capability to simulate average cross-sectional SSC.

Specifically, H-ADCP-based SSC estimation with the limited coverage area is grounded in an assumption of uniform SSC across a cross-section, despite the widely accepted theory of a vertically varying SSC profile as per the Rousean profile (Rouse, 1937). Recent studies utilizing down-looking ADCP for cross-sectional SSC mapping (Guerrero et al., 2013; Pomázi and Baranya, 2022; Chalov et al., 2022) have illustrated two-dimensional SSC variation. This implies that relying on a small number of cells can exacerbate inherent errors in SSC estimation.

One method to resemble such limitations, including unsteadiness and immobility of H-ADCP, is to consider additional variables such as water level and flow rate besides SCB (Son, 2021). In recent efforts, machine learning techniques have been successfully applied to address these non-linearities (Nagy et al., 2002; Melesse et al., 2011; Rajaee et al., 2011; Noh et al., 2023c). By leveraging the strengths of both approaches, estimation performance enhancement is expected by considering

additional variables and adopting machine learning techniques.

To address limitations such as the unsteadiness and immobility of H-ADCP, one approach is to consider additional variables like water level and flow rate alongside SCB (Son, 2021). Recent endeavors have successfully utilized machine learning techniques to manage these nonlinearities (Nagy et al., 2002; Melesse et al., 2011; Rajaee et al., 2011; Noh et al., 2023c). By leveraging the strengths of both approaches, an enhancement in estimation performance is anticipated, which is achieved by considering additional hydraulic variables and applying machine learning techniques.

1.1.2 Presenting extended application strategy of H-ADCP-based SSC monitoring for sediment ungauged stations

Accurate sediment transport measurement data is essential for developing an H-ADCP-based sediment monitoring model. Specifically, the model is established by fitting the recorded H-ADCP backscattering signals acquired during sediment sampling with the measured SSC. The problem is that only 18 out of 62 automated flow observation stations are equipped to monitor sediment loads. Consequently, the H-ADCP-based sediment monitoring model is infeasible to derive in 70% of the flow monitoring stations. This monitoring subject disparity is illustrated in Figure 1.3 by a Venn diagram.

Fortunately, it has been reported that the backscatter of H-ADCP depends on regional sediment characteristics, such as the diameter of the suspended sediment

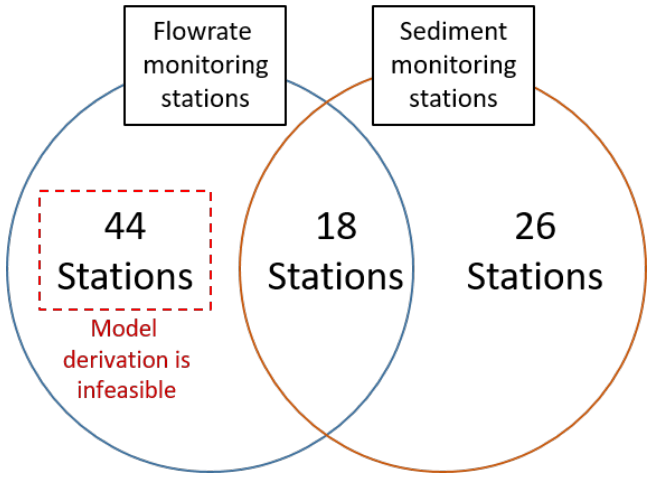


Figure 1.3 Disparity between the automated flow monitoring station and sediment monitoring stations

(Urick, 1975; Topping et al., 2007; Landers et al., 2016; Guerrero and Di Federico, 2018; Aleixo et al., 2020). This dependence suggests that the relationship of backscatter models can be similar where such sediment characteristics are similar. Therefore, it would be possible to apply the same calibrated model in places where spatial sediment characteristics, such as particle size distributions of suspended sediment and bed material, appear similar. Therefore, exploring areas exhibiting homogeneous sediment transport characteristics for application is necessary.

1.1.3 Enabling simultaneous estimation of total load using SSC

On the other hand, fluvial sediment transport is attributed to total load, not only suspended load, which can be estimated using H-ADCP. The total sediment load Q_{TL}, which is regarded as the sum of the suspended Q_{SL} and bed Q_{BL} loads. In

particular, monitoring bed loads is costlier than monitoring suspending loads. Alternative methods to monitor suspended sediment have been proposed that utilize various equipment, such as optical sensors (Agrawal and Pottsmith, 2000) and hyperspectral cameras (Kwon et al., 2022a,a; Gwon et al., 2023), enabling high spatiotemporal resolution monitoring in the simplified monitoring process. Technological advances in the monitoring of bed loads are comparatively slower than those achieved for suspended loads, owing to difficulties in access both physically and optically. Specifically, suspended loads can be easily calibrated with optical features using turbidity or reflectances, which are readily measured remotely.

For these reasons, the total loads are frequently estimated using suspended loads (Turowski et al., 2010). One popular approach is the modified Einstein procedure (MEP) (Colby and Hembree, 1954), which estimates the total load using suspended sediment transport information and its computer program implementation called the Bureau of Reclamation Automated MEP (Holmquist-johnson, 2006) is available. However, MEP has problems, such as arbitrarily defined terms, physically impossible results ($Q_{SL} > Q_{TL}$), and Rouse number (Ro) tuning. Thus, because of some improbable results and estimation difficulty in using MEP, it has been revised to the series expansion MEP (SEMEP) for depth-integrating samplers (Shah-Fairbank et al., 2011) and point-integrating samplers (Shah-Fairbank and Julien, 2015), respectively. Although analytically driven MEP-based methods are theoretically sound,

their application range is limited to sand-bed streams (Shah-Fairbank and Julien, 2015; Yang and Julien, 2019).

Another solution for the total load estimation is to invert the relationship defined by the fraction of suspended load to total load $F_{sus} = Q_{SL}/Q_{TL}$. The suspended-to-total load ratio formulation can be derived from the relationship between Q_{TL} and Q_{SL} of SEMEP (Shah-Fairbank and Julien, 2015; Yang and Julien, 2019). However, as pointed out by (Shah-Fairbank and Julien, 2015), the applicability of SEMEP is limited to the coarse armored bed condition with high wash load and grain size smaller than 2 mm. On the other hand, (Turowski et al., 2010) furnished a profound investigation of F_{sus} using the measured data from various natural rivers and proposed the empirical equations for short-term sediment having a form $Q_{BL} = AQ_{SL}^{B}$, where A and B are the regression coefficients obtained without hydraulics-related factors. Accordingly, there is a need to design a field data-driven empirical model for F_{sus} that contains physical information.

1.2 Objectives of the study

This study proposes a method for efficiently monitoring real-time sediment concentration and flow rate using the H-ADCP installed at automated flow monitoring stations. The research further derives an SVR model for estimating SSC and suspended loads from the H-ADCP backscattering signal. To develop the SVR model,

this work presents and applies a method that involves determining input variables and hyperparameter tuning through the GO algorithm. Additionally, the wider applicability of H-ADCP-based SSC monitoring is presented, including its potential for total load monitoring and its application in sediment measuring stations with missing data.

Three main objectives drive this study:

- Enhancing the performance of H-ADCP-based SSC monitoring
- Presenting extended application strategy of H-ADCP-based SSC monitoring for sediment ungauged stations
- Enabling simultaneous estimation of total load using SSC

The specific sub-objectives are succinctly outlined in the subsequent subsections.

1.2.1 Enhancing the performance of H-ADCP-based SSC monitoring

- To present methods that involve efficiently determining input variables and model hyperparameter settings.
- To propose a method for efficiently monitoring real-time sediment concentration using the H-ADCP installed at automated flow monitoring stations, considering the inclusion of factors such as SCB, flow rate, and stage for SSC monitoring performance.
- To suggest a total load estimation protocol in a systematic connection with the

SSC monitoring method

1.2.2 Presenting extended application strategy of H-ADCP-based SSC monitoring for sediment ungauged station

- To classify the sediment monitoring stations concerning sediment transport characteristics.
- To propose a method determining H-ADCP-based SSC monitoring stations.
- To explore the feasibility of applying calibrated SSC estimation models.

1.2.3 Enabling simultaneous estimation of total load using SSC

:

- To develop a field data-driven empirical model for the suspended-to-total load ratio (F_{sus}) to enhance total sediment load predictions using SVR and symbolic regression techniques.
- To assess the applicability of the developed models and discuss using F_{sus} for total load estimation.
- To provide the inference of the relationships between F_{sus} and input dimensionless hydraulic variables based on the clustering analysis.

These objectives aim to propose a sediment load estimation method using hydro-acoustic backscatter aided by the newly proposed parameter optimization meth-

ods for clustering and support vector regression. Figure 1.4 briefly shows the three pillars of this study and detailed sub-objectives.

1.3 Overview

This dissertation is structured into seven chapters. Chapter 2 introduces the background theories relevant to the subsequent sections, while the following chapter outlines the primary methodologies employed. Chapter 4 delves into the enhancement of estimation performance in the H-ADCP-based sediment monitoring method by applying the proposed SVR model selection method. Chapter 5 explores the utilization of the H-ADCP-based SSC monitoring model in sediment ungauged stations through an analysis of homogeneous sediment monitoring stations' characteristics using the iterative GMM. Chapter 7 presents a novel method for estimating total loads using SSC, focusing on the suspended-to-total sediment load fraction and discussing its dependency on hydraulic variables. Consequently, Chapter 8 suggests a new total sediment load estimation framework incorporating the major results throughout Chapters 3–8. Chapter 9 provides a comprehensive summary of this dissertation and concluding remarks.

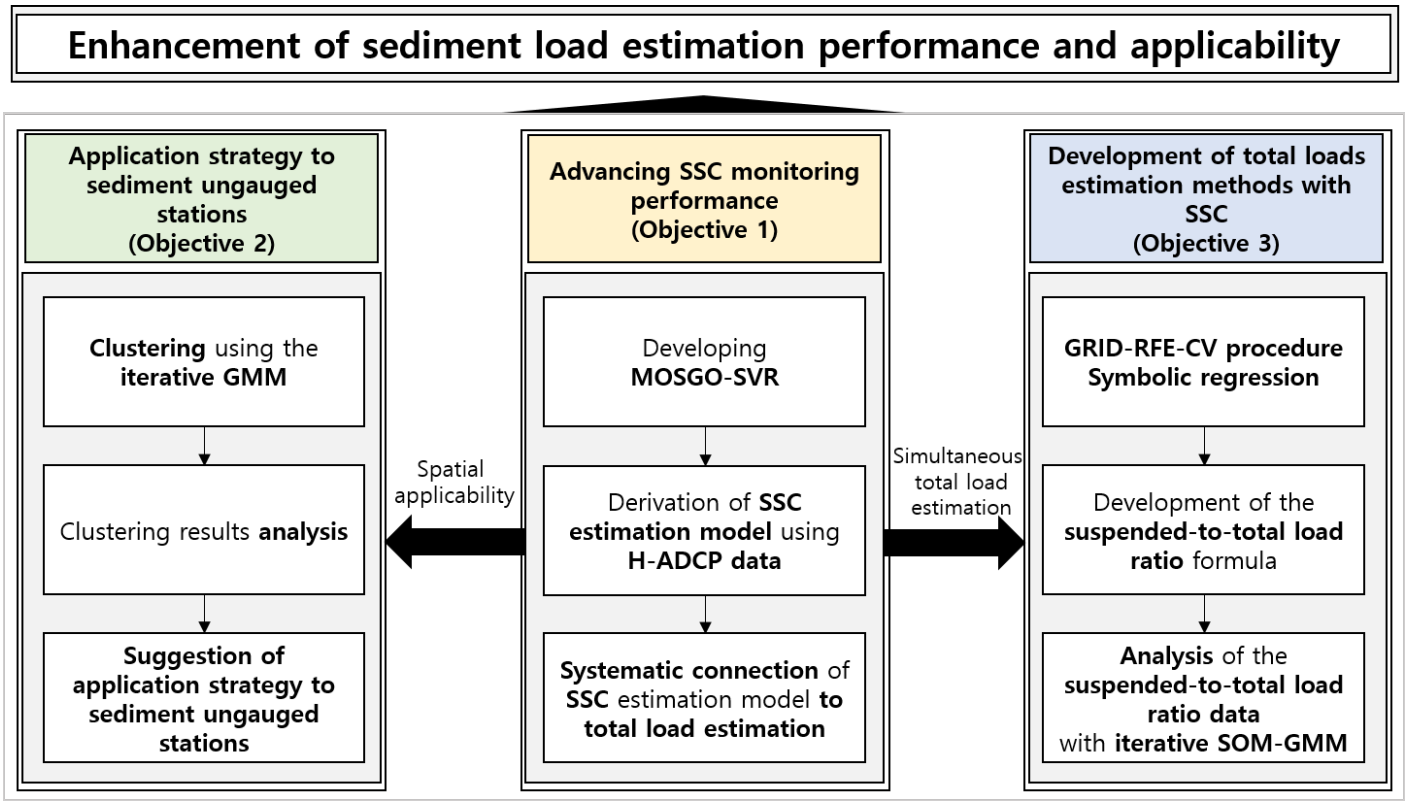


Figure 1.4 Detailed objectives of this study

Chapter 2. Theoretical backgrounds

This chapter consists of descriptions of the methods, including their background theories. In terms of the purposes, the chapter is divided into the following four subjects: a review of total sediment load theory, measurement techniques, soft computing regression, clustering methods, and a brief review of metaheuristic optimization algorithms.

2.1 Total sediment transport review

Sediment is the material detached from the rocks (earth's crust) by the physical and chemical fragmentation (Van Rijn, 1993). Sediment particles are transported by fluid having various shapes and sizes.

The sediment particle motions are often classified by three modes: (1) sliding and rolling, (2) saltating (or hopping), and (3) suspended. Regarding the type of particle movement, the first mode and regular saltating particles are considered bedloads. The particles transported in suspension due to sufficiently strong shear stress and turbulence are suspended loads. With these definitions, The total sediment load Q_{TL} in a stream is considered the summation of suspended load Q_{SL} and bedload Q_{BL} transports.

In addition to the suspended-bedload classification, the sediment transports are categorized in several ways regarding movement type, measurement method, and

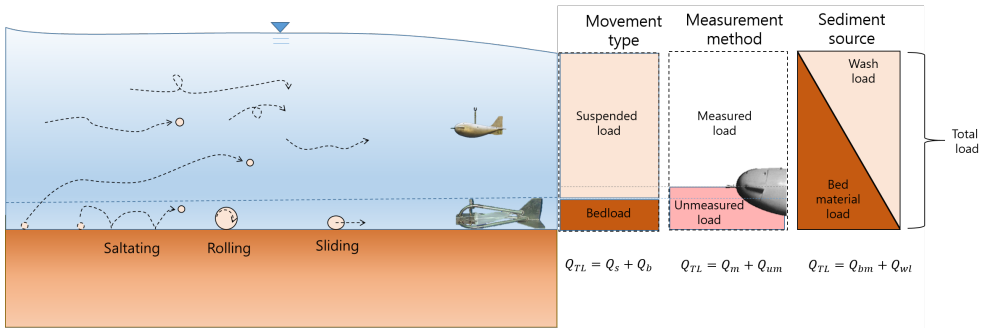


Figure 2.1 Schematic diagram of sediment transport mechanism and classification

sediment source. Figure 2.1 portrays a simple schematic diagram of the sediment particle motions and classifications.

The measurement-based classification is due to the geometric shape of samplers. Taking an example of suspended sediment samplers, suspended sediment samplers have a fish-like shape with the nozzle at 3 to 7 inches upward from the bottom (Edwards et al., 1999). Hence, at least 3 inches of non-measured depth exist using suspended sediment samplers. Let h and z_n be the nozzle height and water depth, respectively. Then, the measured suspended sediment load Q_m is defined as an integration of sediment flux from z_n to h.

The third classification distinguishes the source of passing sediment particles into the bed material load Q_{bm} and wash load Q_{wl}. The bed material load is the amount of particles originating from bed materials by local flows according to the channel capacity. The wash load is part of suspended loads but has significantly smaller particles than the bed material size. The wash load is from the catchment and

rarely deposited.

2.1.1 Total sediment estimation using hydraulic parameters

In this subsection, the total load estimation formulae are briefly reviewed. This section introduces the landmark concept models. The branch of the landmark models will be introduced in section 6.1.

Einstein procedure

As discussed in section 2.1, the unit total load q_{TL} can be obtained by summation of the unit suspended load q_{SL} and unit bedload q_{BL}. q_{SL} can be considered as the integration of sediment concentration flux along the elevation z. Accordingly, Einstein (1950) describes the total load q_{TL} can be obtained by:

$$q_{TL} = q_{BL} + \int_{\delta_b}^{h} u(z)C(z)dz \tag{2.1}$$

where $C(z)$ is the sediment concentration at the vertical distance z from the bed, δ_b is the bed layer thickness, which can be defined as $2d_s$, $u(z)$ is the streamwise velocity at z. $u(z)$ in Equation (2.1) can be assumed by Keulegan's velocity profile (Keulegan, 1938):

$$u(z) = \frac{u_*}{\kappa} \ln(\frac{30z}{k_s}) \tag{2.2}$$

where $u(z)$ is the measure velocity at z; $u_* = \sqrt{\tau_0/\rho}$ is the shear velocity; τ is the shear stress; ρ_w is the water density; κ is the von Karman constant; k_s is the roughness height. Similarly, the $C(z)$ can be taken from the Rousean concentration profile (Rouse, 1937) as given in Equation (2.3).

$$C(z) = C_{ref}(\frac{h-z}{z}\frac{\delta_b}{h-\delta_b})^{Ro} \tag{2.3}$$

where C_{ref} is the reference concentration ($= C(\delta_b)$); Ro is the Rouse number ($= w_s/(\beta\kappa u_*)$; β is the ratio of the turbulent mixing coefficient of sediment to the momentum exchange coefficient (assumed to be 1); and w_s is the falling velocity of sediment particles. Subsequently, with substitutions of $u(z)$ and $C(z)$, the total unit bed sediment discharge is given below.

$$q_{TL} = q_{BL} + \int_{2d_s}^{h} C_{ref}\frac{u_*}{\kappa}(\frac{h-z}{z}\frac{\delta_b}{h-\delta_b})^{\frac{w_s}{\beta_s\kappa u_*}}\ln(\frac{30z}{d_s})dz \tag{2.4}$$

where $C_{ref} = q_{BL}/au(\delta_b)$ is the reference concentration at the bedload layer. The bedload layer thickness δ_b, is assumed to be twice the characteristic particle diameter $2d_s$. Rearranging Equation (2.4),

$$q_{TL} = q_{BL} + 0.216q_{BL}\frac{E^{Ro-1}}{(1-E)^{Ro-1}}\{\ln(\frac{30h}{d_s})J_1 + J_2\} \tag{2.5}$$

J_1 and J_2 are the Einstein integrals and are defined as follows.

$$J_1 = \int_E^1 (\frac{1-z}{z})^{Ro} dz \tag{2.6}$$

and

$$J_2 = \int_E^1 \ln z (\frac{1-z}{z})^{Ro} dz \tag{2.7}$$

where E is the ratio of bed layer thickness to flow depth, which is commonly used in the form $2d_{50}/h$; is. The Einstein integrals can be estimated by monographs given in Einstein (1950) or by numerical integration.

The resultant q_{TL} is obtained by the Einstein procedure (EP). In EP, the q_{SL} and q_{BL} are computed for bins of the SSPSD and bed material particle size distribution (BMPSD), respectively. Ro is obtained by trial and error.

Modified Einstein procedure

Since EP was proposed for hydraulic design, hydraulic variables are determined by formulae. In contrast, the modified Einstein procedure (MEP) was originally proposed by Colby and Hembree (1954) to estimate the total load from measured sediment load from depth-integrated suspended sediment samplers. Therefore, it requires measured mean velocity and measured sediment sample. The basic idea is to estimate the unmeasured load by extrapolation of the Rousean profile using the measured load.

Although the MEP was presented as a simpler modification of the Einstein

procedure, its computational procedure demands substantial experience, engineering sense, and time due to its very complex process that uses more than 30 equations (Holmquist-johnson, 2006). For example, Ro is still estimated by trial and error for the dominating particle size bin and determined by the power of 0.7 equation for the other bins. Thus, reliable estimation and consistent reproduction of total load for multiple users are challenging.

To improve the MEP's accuracy and procedure, additional modifications were proposed over the years (Colby and Hubbell, 1961; Lara, 1966; Burkham and Dawdy, 1980; Shen and Hung, 1983). For practical application Holmquist-johnson (2006) presented the MEP computation software, Bureau of Reclamations Automated Modified Einstein Procedure (BORAMEP) adopting the re-modification of Lara (1966), which revised the relationship between Ro and the falling velocity (w_s). A detailed computation process of MEP can be found in Holmquist-johnson (2006).

Series expansion of the modified Einstein procedure

Shah-Fairbank (2009) observed the following problems: (1) Ro being not computable when no particles are observed in the measured zone, (2) a negative relationship between Ro and w_s, the existence of overlapping bins, (3) suspended loads greater than total loads. Shah-Fairbank (2009) proposed the series expansion of the modified Einstein procedure (SEMEP) to complement the problems of MEP.

The distinguished improvements of SEMEP compared to MEP are as follows:

1. The calculation of total sediment discharge relies on the median suspended sediment grain size (d_{50ss}), eliminating the need for the PSD bins.

2. There is no regression fitting of Ro based on data from overlapping bins.

3. Ro is evaluated based on the ratio of settling velocity (w_s) to shear velocity (u_*), assuming $\beta_s = 1$ and $\kappa = 0.4$.

4. q_{BL} is computed using the measured sediment discharge, eliminating the need to favor Einstein's bed load equation or arbitrarily divide the bed load intensity by two, and it ensures $q_{TL} \geq q_m$.

5. The series expansion of Guo and Julien (2004) is employed to obtain J_1' and J_2'.

In SEMEP, the Einstein integrals (J_1 and J_2) are computed based on the integration algorithm proposed by Guo and Julien (2004). Contrary to MEP employing q_{SL} to estimate q_{TL}, the SEMEP computes q_{TL} using the measured unit suspended sediment load q_m.

$$q_m = 0.216 q_{BL} \frac{E^{Ro-1}}{(1-E)^{Ro-1}} \{\ln(\frac{30h}{d_s}) J_1' + J_2'\}, \tag{2.8}$$

For the integration of the measurable area, the corresponding integrals J_1' and J_2' can be computed by substituting E with z_n/h (for example, $J_1' = \int_a^1 (\frac{1-z}{z})^{Ro} dz$). The

unit bedload, q_{BL}, can be determined using the unit measured load, q_m (Equation (2.8)).

Basically, the MEP uses sediment samples from the depth-integrated suspended sediment samplers. For the point-integrated suspended sediment sampler users, Shah-Fairbank and Julien (2015) proposed SEMEP for point measurements (SEMEPP).

Bagnold's stream power concept

The most popular concept is the stream power concept developed by Bagnold (1966). The basic idea is that the power of the flow excites the bedload transport, supplying sufficient energy.

$$\tau_0 U = \rho_w g U h S_f = \gamma_w q S_f \tag{2.9}$$

where, τ_0 is the bed shear stress; U is the cross-section averaged flow velocity; ρ_w and γ_w are the density and specific weight of water, respectively; g is the gravitational acceleration; and S_f is the friction slope.

$$q_{TL} = q_{BL} + q_{SL} = \frac{\tau_0 U}{G_s - 1}(e_B + 0.01\frac{U}{w_s}) \tag{2.10}$$

where τ_0 is the bed shear stress; e_B is the Bagnold coefficient; and $G_s = \gamma_s/\gamma_w$ is the specific gravity of sediment with γ_s being the specific weight of sediment.

Note that the concept of incipient motion is not considered in the stream power

approach so that the transport rate does not reduce to zero even in low velocity with large grain size (Julien, 2010).

Yang's the unit stream power concept (on energy dissipation)

The unit stream power (on energy dissipation) concept, presented by Yang (1979), is one of the stream power-based formulae. Dividing Equation 2.9 by gh leads to stream power per unit weight of water, US_f.

$$\frac{dz}{dt} = \frac{dx}{dt}\frac{dz}{dx} = US_f \tag{2.11}$$

From the Rousean profile, Yang (1979); Yang and Molinas (1982) deduced the vertical sediment concentration distribution related to turbulence energy production.

$$\frac{C(z)}{C_{ref}} = [\frac{\tau_{xy}\frac{dU_x}{dz}}{(\tau_{xz}\frac{du(x)}{dz})_{z=\delta_b}}]^{Ro} \tag{2.12}$$

where, τ_{xz} is the turbulent shear stress at distance z above the bed. The total sediment concentration C_{TL} can be obtained by integration of the $C(z)$ profile.

$$C_{ppm} = A(\frac{US_f}{w_s})^B \text{ or } \log C_{TL} = A_Y + B_Y \log(\frac{US_f - U_{cr}S_f}{w_s}) \tag{2.13}$$

where U_{cr} is the critical velocity at incipient particle motion. As expressed above, the unit stream power is often represented in the dimensionless form, US_f/w_S. Based

on this analysis, Yang (1979) proposed empirical equations separating sand bed and gravel bed streams. For example, the dimensionless relationship for sand bed streams is given below.

$$\begin{aligned} C_{ppm} =& 5.435 - 0.286 \log \frac{w_s d_{50}}{\nu} - 0.457 \log \frac{U_*}{w_s} \\ &+ (1.799 - 0.409 \log \frac{w_s d_{50}}{\nu} - 0.314 \log \frac{U_*}{w_s}) \log(\frac{U S_0}{w_s} - \frac{U_{cr} S_0}{w_s}) \end{aligned} \tag{2.14}$$

where C_{ppm} is the total load concentration in parts per million; and ν is the kinematic viscosity of water. In the equations, U_{cr}/w_s can be obtained by Equation (2.15).

$$\frac{U_{cr}}{w_s} = \begin{cases} \frac{2.5}{\log(\frac{U_* d_{50}}{\nu}) - 0.06} + 0.66 & \text{for } 1.2 < \frac{U_* d_{50}}{\nu} < 70.0 \\ 2.05 & \text{for } 70 \leq \frac{U_* d_{50}}{\nu} \end{cases} \tag{2.15}$$

Tractive force concept

Another concept uses the tractive force (shear stress) to describe the particle motions. Inspired by the fact that particles move if the tractive force exceeds a critical value, Laursen (1958) developed a semi-empirical model as the following equation.

$$C_w = 0.01 \gamma_s \sum_i P_i (\frac{d_i}{h})^{7/6} (\frac{\tau_i'}{\tau_{ci}} - 1) f(\frac{u_*}{w_i}) \tag{2.16}$$

where P_i is the i-th bin's size fraction; τ_i' and τ_{ci} are the tractive and critical tractive forces of i-th bin, respectively; and w_i is the particle falling velocity of of i-th bin. In

this model, the tractive forces can be obtained by $\tau_i' = \frac{\rho_w U^2}{58}(\frac{d_{50}}{h})$ and $\tau_{ci} = Cd_{50}$, where ρ_w is the water density; and d_{50} is the sediment particle sizes of the 50% of the material by weight. $f(u_*/w_i)$ can be determined by a plot given in Laursen (1958).

2.1.2 Using suspended load to estimate total load

2.1.2.1 Models using $F_{sus} = Q_{SL}/Q_{TL}$

Not using detailed suspended sediment sampling results, the total load can be calculated by using the suspended-to-total load fraction (or the bedload fraction). Taking $F_{sus} = Q_{SL}/Q_{TL}$ as an example, if Q_{SL} is directly sample, Q_{TL} can be calculated by $Q_{TL} = Q_{SL}/F_{sus}$.

In this study, total load estimation methods are proposed that assist the H-ADCP-based suspended sediment monitoring system. With the readily obtainable suspended loads, using F_{sus} is practically useful. For instance, the H-ADCP- or remote sensing-based monitoring systems only provide suspended load. In this manner, this study introduces several efforts to estimate F_{sus}.

Rule of thumb: rough estimation using table

Another approach for estimating the total load is to use rough estimations. A method is assumming F_{sus} from 0.6 to 0.9 (Turowski et al., 2010). The reference of this assumption can be found in the estimation tables as shown in 2.1 (Maddock and Borland, 1950; Lane and Borland, 1951).

Table 2.1 F_{sus} rough estimation table (modified after Turowski et al. (2010)

Bed type	Concentration (ppm)	Suspended-to-total load fraction in percent	
		Maddock and Borland (1950)	Lane and Borland (1951)
Sand	$C_{ppm} < 1000$	50 to 100	40 to 80
	$1000 \leq C_{ppm} < 7500$	80 to 90	75 to 90
	$7500 \leq C_{ppm}$	80 to 90	87 to 95
Gravel	$C_{ppm} < 1000$	95	90 to 95
	$1000 \leq C_{ppm} < 7500$	90 to 95	90 to 95
	$7500 \leq C_{ppm}$	92 to 98	92 to 98

Indeed, F_{sus} models from Maddock and Borland (1950); Lane and Borland (1951) are simple and practical. However, F_{sus} determination depends on the engineer's intuition, and there is no clear reason for the numbers (Turowski et al., 2010).

Turowski et al. (2010) method

Turowski et al. (2010) developed empirical models to elucidate the solid basis to estimate F_{sus}. The long-term and short-term F_{sus} estimation models were derived separately, using a dataset with suspended load and total load measured at the same time.

The short-term model has a power-law form as:

$$Q_b = \begin{cases} A_T Q_{SL}^{B_T} & \text{for } Q_{SL} \leq (A_T/C_T)^{1/(D_T - B_T)} \\ C_T Q_{SL}^{D_T} & \text{otherwise} \end{cases} \tag{2.17}$$

Equation (2.17) was repeatedly fitted with the 25th, 50th, and 75th percentiles of the dataset. The fitted coefficients A_T, B_T, C_T, and D_T are given as Table 2.2.

Table 2.2 Coefficients of Equation (2.17)

Coefficients	25 percentile	50 percentile	75 percentile
A_T	0.131 ± 0.007	0.833 ± 0.052	0.653 ± 0.594
B_T	1.340 ± 0.125	1.340 ± 0.079	1.1425 ± 0.092
C_T	0.241 ± 0.131	0.437 ± 0.210	1.473 ± 0.518
D_T	0.588 ± 0.062	0.647 ± 0.076	0.590 ± 0.052
$Q_{SL} = (A_T/C_T)^{1/(D_T - B_T)}$	2.249 kg sec^{-1}	0.394 kg sec^{-1}	0.345 kg sec^{-1}

The long-term estimation model was derived for the gravel bed streams using the drainage area, A_{drain}.

$$F_{sus} = 0.55 + 0.040 \ln(A_{drain}) \tag{2.18}$$

A long-term model for the sand bed streams was not explicitly suggested. Despite this, they pointed out that sandy bed streams F_{sus} are less insignificant in a given drainage area.

Note that the existing F_{sus} models proposed in Maddock and Borland (1950); Lane and Borland (1951); Turowski et al. (2010) are simple and easy to use. However, the existing F_{sus} models do not consider hydro-geomorphic variables, such as velocity, water depth, and particle sizes. Hence Turowski et al. (2010) pointed out that the relationships have to be carefully used since F_{sus} may vary considerably in a given Q_{SL}.

Series expansion of the modified Einstein procedure

In Equation (2.5), q_{SL} can be derived simply by subtracting q_{BL}. Then, F_{sus} can be derived as follows.

$$F_{sus}(Ro, h, d_s) = \frac{0.216\frac{E^{Ro-1}}{(1-E)^{Ro-1}}\{\ln(\frac{30h}{d_s})J_1' + J_2'\}}{1 + 0.216\frac{E^{Ro-1}}{(1-E)^{Ro-1}}\{\ln(\frac{30h}{d_s})J_1 + J_2\}} \tag{2.19}$$

Shah-Fairbank et al. (2011) deduced this equation to described the influences of h/d_s and u_*/w_s. They explained that u_*/w_s is primarily influential to F_{sus} and suggested the thresholds $u_*/w_s = 1$ and $u_*/w_s = 2.5$ for $F_{sus} = 0$ and $F_{sus} = 1$, respectively. Despite the physical basis of SEMEP, estimating total loads using Equation (2.8) is ideal.

2.1.3 Machine learning models in total load estimation

Withstanding the popularity of machine learning, machine-learning techniques are applied to drive total sediment load estimation models. The application directions of machine learning can be divided into two. One is a model that learns the time series data and their time lags, and the other is a model that analyzes or calculates sediment load using physical variables.

In recent decades, a number of machine learning applications to time series sediment discharge were introduced using the artificial neural networks (ANNs) and support vector machine (SVM) (Rajaee et al., 2011; Kim and Seo, 2015; Kim et al.,

2017b,a; Yadav et al., 2018; Riahi-Madvar and Seifi, 2018; Choubin et al., 2018; Torabi and Dehghani, 2018; Meshram et al., 2020; Zounemat-Kermani et al., 2020; Jung et al., 2021). The majority of studies have concentrated on the utilization of ML models for simulating Q_{SL}, compared to Q_{BL} or Q_{TL}. Specifically, in cases where studies encompassed total loads, separate models for Q_{SL} and Q_{BL} were developed Zounemat-Kermani et al. (2020).

There is less research on physics-based sediment load estimation models than on time-series analysis. Yang et al. (2009) conducted pioneering research on estimating Q_{TL} using ANN. Tayfur et al. (2013) applied the principal component analysis (PCA) to identify influential hydraulic variables for explaining total load transport. In addition, they presented an ANN model and nonlinear equations calibrated by GA. Pektaş and Doğan (2015); Pektaş (2015) developed an ANN model with input variables determined by PCA and clustering analysis for Q_{SL} and Q_{BL} separately.

Among them, most machine learning approaches rely on black box models such as SVM and ANNs, so it is difficult to provide physical insight. Thus, creating a model with explicit equations can contribute more to the understanding of sediment transport and subsequent research (Okcu et al., 2016). In this manner, contemporary machine learning called symbolic regression methods prove valuable as it provides explicit equations as outputs (Harun and Ab. Ghani, 2020; Harun et al., 2021).

Despite numerous machine learning applications, there is no machine learning

application for understanding F_{sus}.

2.1.4 Hysteresis in time-series data

In time series sediment monitoring data, measured sediment concentration data shows hysteretic behavior, that different curves are drawn in rising and falling limbs. It is known that the unsteady river flows accompany the counterclockwise hysteresis loop hysteresis in stage-flow rate graphs. The hysteresis phenomenon is observed also in the flow rate-sediment concentration relationship. This section phenomenologically reviews the possible circumstances with hysteresis concerning Williams (1989); Chaudry (2008); Gellis (2013); Julien (2018).

2.1.4.1 Stage-flow rate hysteresis

Flow analysis is conducted in steady flow assumption, comprising a simple monotonic relationship between stage and flow rate. However, the most common flow type is the unsteady flow in rivers. In unsteady flows, the stage-flow rate relation presents a looping phenomenon, hysteresis, when the natural flood waves propagate during rainfall events or weir operations (Kim et al., 2016; Muste et al., 2020, 2022a,b).

Hysteresis in flow can be interpreted with the unsteady one-dimensional shallow water equation (St. Venant equation). Assumptions of the St. Venant equation are as follows:

1. Hydrostatic pressure distribution.

2. Small bottom slope so that velocity measurements in vertical and normal to bottom directions are the same.

3. Uniform flow velocity of over entire cross-section.

4. Uniform prismatic cross-section over distance.

5. The head losses in unsteady flow can be simulated by using the steady-state resistance laws (e.g. Manning equation).

Being with underlying assumptions, the continuity equation is given by

$$\frac{\partial A}{\partial t} + \frac{\partial Q}{\partial x} = q_l, \tag{2.20}$$

where A is the cross-sectional area; t is the time; Q is the flow rate; x is the streamwise coordinate; and q_l is the lateral input (or output) flow rate. The momentum equation is

$$\frac{\partial U}{\partial t} + g\frac{\partial}{\partial x}(\frac{U^2}{2g} + y) = g(S_0 - S_f), \tag{2.21}$$

where S_f is the friction slope; S_0 is the bed slope; and y is the water level. The rearrangement of the momentum equation in terms of the friction slope yields the following relationship.

$$S_f = S_0 - \frac{\partial}{\partial x}(\frac{U^2}{2g} + y) - \frac{1}{g}\frac{\partial U}{\partial t} \tag{2.22}$$

The first term on the right-hand side is the channel bottom slope, and $S_f \approx S_0$ in steady-uniform flow. The second and third terms describe the convective and local accelerations, respectively. The acceleration terms are active in the nonuniform flow. In unsteady flows, $\partial U/\partial t$ has to be considered.

On the other hand, the general flow rate in unsteady flow can be expressed with channel parameters as Equation (2.23).

$$Q = k_Q A R_h^m \sqrt{S_f} \tag{2.23}$$

where k_Q is the coefficient. In steady-uniform flow, S_f is equal to S_0. If denoting steady-uniform flow rate in the same channel parameters as Q_n, the equation yields $Q_n = k_Q A R_h^m \sqrt{S_0}$ which is followed by a simple stage-flow rate rating curve. Accordingly, we have $Q = Q_n \sqrt{\frac{S_f}{S_0}}$. Substituting Equation 2.22, we obtain

$$Q = Q_n \sqrt{1 - \frac{1}{S_0}\frac{\partial y}{\partial x} - \frac{1}{S_0}\frac{U}{g}\frac{\partial U}{\partial x} - \frac{1}{S_0}\frac{1}{g}\frac{\partial U}{\partial t}}. \tag{2.24}$$

Equation (2.24) explains how the hysteresis phenomenon is observed in unsteady flows.

When a simple flood wave propagates through the one-dimensional channel, the wave first arrives at the upstream end of a reach, control volume, increasing stage, and later at the downstream end. That is, $\partial y/\partial x$ becomes negative. In this case,

the upstream flow velocity is greater than the downstream flow velocity, resulting in $\partial U/\partial x < 0$ and $\partial U/\partial t > 0$. It is known that the scale of local acceleration, $\partial U/\partial t$, is smaller than the other terms. Thus, the square root exceeds 1 and $Q > Q_n$ in the rising stage. In the same manner, $\partial y/\partial x$ and $\partial U/\partial x$ are positive, so $Q_n > Q$, in the falling stage (Table 2.3). As a result, the stage and flow rate graph shows a counterclockwise loop, as shown in Figure 2.2.

Table 2.3 Sign of each term in Equation 2.24 for stage variation

		$\partial y/\partial x$	$\partial U/\partial x$	$\partial U/\partial t$	$S_f - 1$	$Q - Q_n$
Sign	Rising	-	-	+	+	+
	Falling	+	+	-	-	-

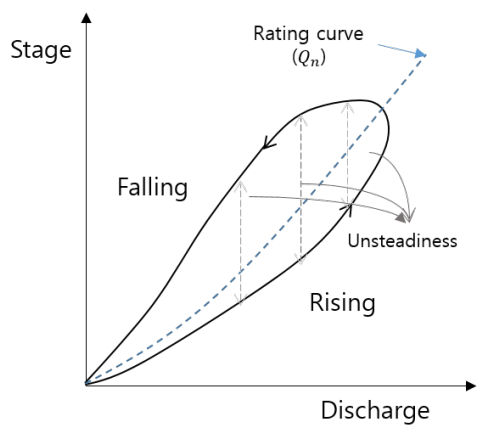


Figure 2.2 Schematic diagram of hysteresis in stage-flow rate relationship

2.1.4.2 Hysteresis in sediment transport

In sediment transport, various types of hysteresis curves are possible (Williams, 1989; Gellis, 2013). The sediment discharge and flow rate curve and corresponding time

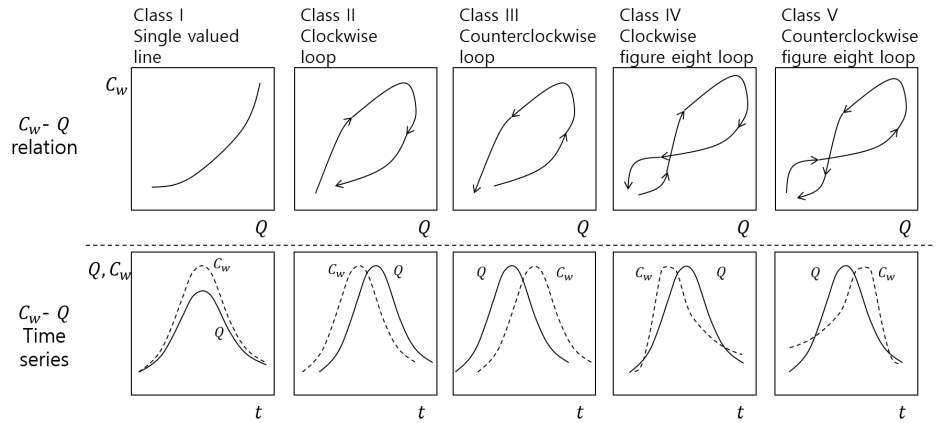


Figure 2.3 Schematic diagram of possible sediment transport hysteresis classes (modified after (Williams, 1989; Gellis, 2013))

series in five situations are exemplified in Figure 2.3.

Fundamentally, the sediment hysteresis can be interpreted with shear stress in unsteady flow. Shear stress, strongly related to sediment transport, can also be expressed similarly to flow rate using the simplified momentum equation of the St. Venant equation.

$$\tau_0 = \gamma_s R_h S_f = \gamma_s R_h (S_0 - \frac{\partial y}{\partial x} - \frac{U}{g}\frac{\partial U}{\partial x} - \frac{1}{g}\frac{\partial U}{\partial t}) \tag{2.25}$$

where R_h is the hydraulic radius. The above equation evidences that shear stress shows a similar trend to flow rate, representing larger stress in the rising stage than in the falling stage. Since shear stress fundamentally drives sediment transport, hysteresis in bed material sediment discharge can be analyzed with this relationship, simply replacing Q with Q_{TL}. However, S_f square times contribute to τ_0 than to Q, increase

and decrease.

If there is no interruption of sediment supply, $\tau_0/Q \propto \sqrt{S_f}$, then peaks of flow rate and sediment discharge are simultaneously observed. In this situation, the Class 1 curve can be observed.

On the other hand, according to Equation (2.25), Julien (2018) explains how Class 2 hysteresis is probable in upland areas, whereas it does not occur in wash load-dominated streams. Williams (1989) describes that the clockwise hysteresis loop (Class 2), the most common mode, occurs in case of sediment source exhaustion. The progressive wetting of soil, armoring, and bank erosion are potential depletion of deposited sediments. An increase in base flow can lead to dilution of sediment concentration having the Class 2 hysteresis. Raises of water depth and bottom slope before the peak flow lead to an early shear stress peak, resulting in an early sediment concentration peak.

In Class 3, having a counterclockwise loop, the sediment concentration peak arrives later than the flow rate peak. It implies delayed sediment transport. The late sediment arrival is justified by distant sediment sources, a tributary for example, or bank erosion on the falling limb. Contrary to Class 2, concentration observation is comparatively higher than observed concentration during an earlier storm (Sidle and Campbell, 1985).

Eder et al. (2010) reported the figure eight curve with an initial clockwise

loop (Class IV). It was explained that initial sediment is flushed away from the vicinity, likewise Class II. Afterward, the contributions from sub-catchments provide sediments of higher concentration than base concentration.

Seeger et al. (2004) investigated the last hysteresis class. It was reported that it occurs under dry soil moisture conditions with a low hydraulic conductivity of the soil. The authors of Seeger et al. (2004) interpreted this event as a partial sequential of Classes II and III. Initially, near sediment flush increases concentration, saturating soils with macropores. As macropores are saturated, all catchment areas become contributing areas with Hortonian flow, providing high sediment concentration. During flood recession, the contributing areas rapidly decrease, resulting in limiting sediment sources.

2.2 Measurement techniques

One of the sub-objectives of this study is to propose a technique for monitoring the total sediment load. This section discusses the monitoring methods employed according to Edwards et al. (1999). The total sediment load is determined by aggregating independently collected suspended sediment samples and bedload samples. Figure 2.4 displays field sampling photos for suspended load and bedload measurements.

Fish-shaped isokinetic sediment samplers, designed to minimize differences between nozzle inlet and ambient velocities, are utilized to monitor suspended sedi-

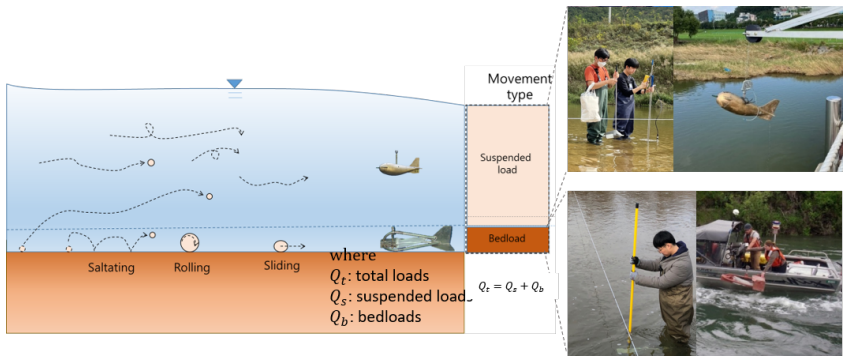


Figure 2.4 Photographs on suspended sediment and bedload sampling

ment load. These samplers, called point- and depth-integrated samplers, collect turbid water at a specific point or along a vertical line. Subsequently, the suspended sediment samples undergo further analysis in the laboratory using filtration or evaporation methods, effectively separating sediment particles from water. The resulting weights per volume yield the SSC.

To measure bedload, various types of samplers are employed, with the Helley-Smith sampler being the most widely used. Unlike suspended samples collected into a bottle as a water sample, bedload samplers capture sediment particles transported through a duct-connected meshed sample bag placed at the river bottom. During bedload sample collection, measurements of channel width and the specific time the sampler was on the riverbed are taken to calculate bedload discharge passing through a cross-section. Using the mass of the dried bedload samples analyzed in a laboratory, unit bedload discharge at a vertical can be determined by dividing the mass by time, with the width correction involving the ratio of unit width over sampler width. It is

recommended to conduct bedload sampling at more than 40 verticals, varying the distances between them.

Recently, surrogate measurement techniques have been employed to measure the suspended sediment concentration using acoustic or optic backscatter signals, enhancing the monitoring efficiency. The expected costs of the measurement methods, including the modern backscattering-based techniques, are shown in Table 2.4 for comparing the pros and cons. Traditionally, samples from conventional samplers

Table 2.4 The costs of the riverine suspended sediment monitoring methods

Methods	Cost (USD)	Time of survey (s)	Personnel	Discharge	Sample analysis
Sampler	1,000 + (crane)	Very high	2∼4	X	Every sample
H-ADCP	60,000	Low	1	O	Derivation set
LISST-SL2	50,000 + (crane)	High	2∼4	O	Calibration set
LISST-200X	60,000 + (crane)	High	2∼4	X	Calibration set

are considered ground truth, but as mentioned earlier, they are not the most efficient method. In particular, flow needs to be measured independently, and the time required for a single cross-section is the longest among all methods. Additionally, the flow velocity needs to be measured separately, and further analysis is required for all samples. Such sample analysis is also a time-consuming process that may take up to several weeks. On the other hand, once installed, H-ADCP can continuously receive flow data and measure suspended sediments through scattering. However, initial sample analysis is required for calibration purposes. Conversely, laser in situ Scattering and transmissometery (LISST) devices from Sequoia Scientific Inc. can measure suspended sediments without the need for additional sample analysis. An

additional advantage of LISST-200X and LISST-SL2 devices is that they can measure the size of suspended sediments. However, they require crane-based measurements on cross-sections, which are time-consuming and subject to weather constraints. Therefore, for long-term real-time monitoring, H-ADCP is a desirable method that takes into account rainfall events.

On the other hand, bedload transport poses challenges in measurement, leading to a reliance on empirical formulas such as rating curves (Willis and Griggs, 2003; Boateng et al., 2012), given the slower technological advancements in surrogate measurement methods. The complexities associated with bedload sampling result in a predominant focus on measuring suspended load, with total loads being estimated. For instance, in South Korea, the estimation of total sediment load (suspended + bedload) discharge relies on the use of MEP, while SSC and flow are directly measured (Ministry of Environment, 2019). Consequently, accessible data for total load measurement, including bedload, is limited in South Korea. Although rating curves for suspended and total loads are developed annually for practical applications, they cannot accurately reflect the hysteresis that occurs in actual rivers, leading to significant errors (Rajaee et al., 2011; Zounemat-Kermani et al., 2020). This study aims to propose an efficient and accurate monitoring technique. Therefore, the following chapters will focus on utilizing H-ADCP as a surrogate for suspended load measurement and total load estimation.

2.2.1 Measurement of suspended sediment concentration using H-ADCP signal

Measurement of suspended sediment concentration using H-ADCP signal is performed with analysis of the backscattered sound wave on the suspended particles. The measured backscatter (MB) corresponding to the source level (SL) transmitted from a transducer of ADCP can be expressed by the simplified sonar equation (Urick, 1948), which is given by:

$$MB + 2TL = SL + TS \tag{2.26}$$

where, TL is an abbreviation for transmission loss, which refers to the loss of sound waves due to various attenuation factors such as sound diffusion, viscosity, and scattering attenuation that occur during the propagation of sound waves. Since TL occurs both when the sound wave is transmitted and reflected, it is sometimes expressed as a two-way transmission loss ($2TL$) by multiplying the one-way loss value by two, as shown in the equation above. The last term in the sonar equation, TS, stands for target strength, which in this study refers to the degree to which ultrasound emitted from an ADCP is reflected by particles present in the path of ultrasound propagation within the medium. $2TL$ is calculated taking into account the following three major attenuation factors: (1) attenuation due to the scattering pattern of the sound waves emitted from the transducer, (2) attenuation due to the viscosity of the fluid, ionic

relaxation effects, and other factors, (3) attenuation due to scattering caused by the viscosity and shape of the boundary surface of the suspended particles.

2.2.1.1 Water and sediment corrected backscatters

When modeling sound diffusion geometrically as spherical, the two-way transmission loss of sound waves can be expressed as $20log_{10}(r)$ depending on the distance r that the sound wave travels. However, there is a characteristic of irregular diffusion of ultrasound in the initial area adjacent to the transducer. Therefore, Downing et al. (1995) proposed a correction factor ψ, such as $20log_{10}(\psi r)$, using a dimensionless number $r_* = r\lambda/(\pi a_t^2)$ based on the wavelength λ of the sound wave and the radius a_t of the ultrasound sensor to account for diffusion loss. Here, ψ is calculated by the following equation.

$$\psi = \frac{1 + 1.35r_* + (2.5r_*)^{3.2}}{1.35r_* + (2.5r_*)^{3.2}}, \text{ where } r_* = \frac{r\lambda}{\pi a_t^2} \tag{2.27}$$

Here, λ can be obtained from the relationship between the wave speed, c, and the frequency, f, $c = \lambda f$. c can be computed as given by:

$$\begin{aligned} c =& 1.402385 \cdot 10^3 + 5.38813T \\ & - 5.799136 \cdot 10^{-2}T^2 + 3.287156 \cdot 10^{-4}T^3 \\ & - 1.398845 \cdot 10^{-6}T^4 + 2.787860 \cdot 10^{-9}T^5 \end{aligned} \tag{2.28}$$

The attenuation caused by the fluid and the suspended particles can be calculated as $2r\alpha_w + 2r\alpha_s$ by multiplying the sound propagation distance r with the correction coefficients α_w for water and α_s for particles. The relationship between the left-hand side of Eq. (1) and the water-corrected backscatter (WCB) and the sediment-corrected backscatter (SCB), which is additionally corrected for particles, can be summarized as follows.

$$\begin{aligned} MB + 2TL &= MB + 20\log_{10}(\psi r) + 2r\alpha_w + 2r\alpha_s \\ &= WCB + 2r\alpha_s = SCB \end{aligned} \tag{2.29}$$

As a result, the volumetric SSC, C_{ppm}, can be measured by deriving a relationship between C_{ppm} and SCB. Usually, the formula has a form of $\log_{10}(C_{ppm}) = C_1 \times SCB + C_2$, where C_1 and C_2 are the regression coefficients. SCB is influenced by physical conditions such as SSPSD and water temperature. It has been reported that since these physical conditions vary owing to locality, the regression coefficients C_1 and C_2 of the corrected equation show a negative correlation and vary from site to site (Noh et al., 2022). To improve the accuracy of the equation, α_s (Landers et al., 2016) or water level measured at the observation site (Son, 2021) can also be introduced as input variables in the regression equation.

2.2.1.2 Computation of sediment corrected backscatter

Practically, when calculating WCB, the attenuation of ultrasonic energy in the fluid is corrected by calculating the attenuation coefficient as a function of salinity and temperature, which is caused by the relaxation effect of magnesium sulfate (MgSO4) ions due to their binding and decomposition, as well as the attenuation due to the viscosity of the medium. In the case where the effect of ions and salinity can be neglected, such as in river water, and only the attenuation due to the viscosity of water is considered, the value of WCB can be estimated using the following relationship by Schulkin and March (1962).

$$\alpha_w = 8.69 \frac{3.38 \cdot 10^{-6} f^2}{21.9 \cdot 10^{6-1520/(T+273)}} \tag{2.30}$$

On the other hand, attenuation due to suspended particles is affected by complex interactions with the viscosity effect between small particles and the scattering effect that occurs when ultrasound collides with particles. The viscosity effect is dominated by the surface area of suspended particles, the frequency of ultrasound, fluid viscosity, and the specific gravity of particles. Given a volume concentration of particles in suspension, as particle diameter decreases, the surface area per unit volume increases, causing attenuation to increase. Conversely, as particle diameter increases, attenuation decreases due to shear and viscosity (Landers et al., 2016).

The scattering attenuation effect is closely related to particle circumference,

ϕd_s, rather than surface area. When the wavelength is much larger than the circumference, scattering attenuation increases rapidly. However, when the circumference and wavelength of the particle are similar, attenuation behavior becomes complicated (Urick, 1948; Flammer, 1962). Urick (1948) developed and verified an empirical equation for the attenuation coefficient as a function of particle concentration for the two particle suspension mechanisms mentioned earlier. Later, Sheng and Hay (1988) presented another scattering attenuation empirical formula by finding the maximum value of the attenuation coefficient for particle suspension that was not reflected in Urick (1948), based on a more in-depth study of particle scattering attenuation by Flammer (1962).

To complement the two aforementioned attenuation models, Landers (2012) proposed the hybrid Urick-Sheng-Hay formula (Eq 2.31) by replacing the scattering attenuation term in the formulae from Urick (1948) and Sheng and Hay (1988).

$$\alpha_s = C_{ppm}[k(G_s-1)^2(\frac{s}{s^2+(G_s+\tau)^2}) + \frac{k^4 d_s^3}{5(1+1.3k^2d_s^2+0.24k^4d_s^2)}]4.34 \tag{2.31}$$

where, k is the wavenumber of a wave with wavelength λ in units of cm (i.e., $k = 2\pi/\lambda$); $s \equiv \frac{9}{4\beta d_s}(1+\frac{1}{\beta d_s})$; $\tau \equiv 0.5 + \frac{9}{4\beta d_s}\left(1+\frac{1}{\beta d_s}\right)$; and $\beta \equiv \sqrt{\omega\pi/\nu}$, where ν is the kinematic viscosity of water. Figure 2.5 shows the variation of α_s with d_s, where $\omega = 3$ MHz and $C_{ppm} = 1{,}000$ ppm.

On the other hand, not only the characteristic particle size but also the standard

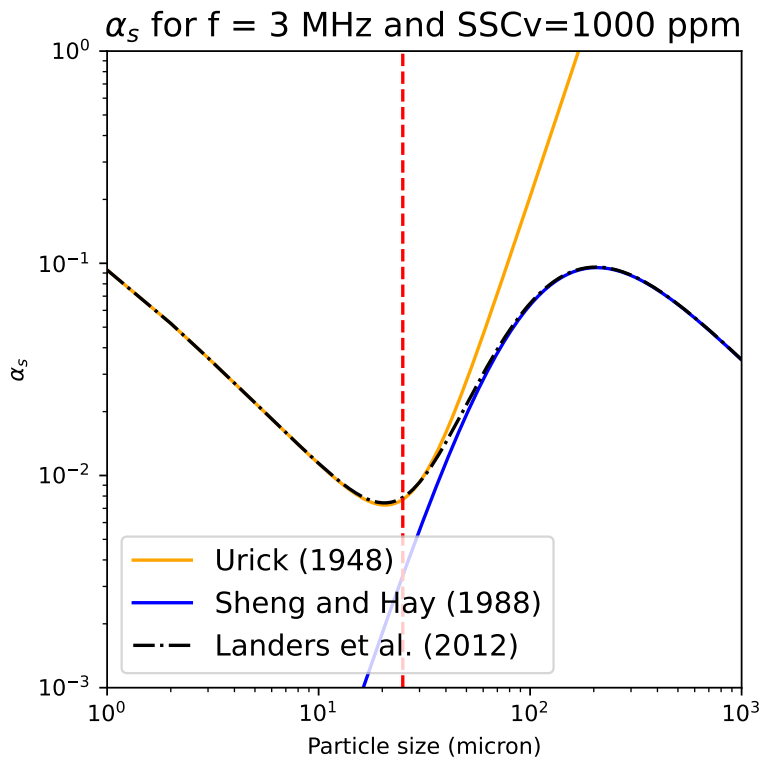


Figure 2.5 Coefficients varying particle size of the estimation models (modified after Landers et al. (2016))

deviation (STD) of PSD affect viscous and scattering attenuations (Guerrero et al., 2016; Guerrero and Di Federico, 2018; Aleixo et al., 2020). Guerrero et al. (2016) reported that the maximum viscous attenuation decreases and minimum scattering attenuation increases with the increase of the PSD standard deviation. With this observation, Guerrero and Di Federico (2018) proposed a SSC_V monitoring method accounting for particle diameter and STD, and Aleixo et al. (2020) proved applicability in long-term sediment monitoring in rivers. Their method uses the following models: backscatter $\sigma_s^2 = \kappa_s^2 SSC_V$ and the sediment attenuation coefficients $\alpha_s = \zeta_s SSC_V$. Guerrero et al. (2016) suggested attenuation to backscatter ratio considering the observation of Moore et al. (2012):

$$\text{ABR} = \frac{\zeta_s \text{SSC}_V}{\kappa_s^2 \text{SSC}_V} = \frac{\zeta_s}{\kappa_s^2} \tag{2.32}$$

where ζ_s is the normalized attenuation coefficient, which is identical to the parenthesis term in Equation (2.31); κ_s^2 is the backscatter strength coefficient. This model is based on the concentration relationships to backscatter, $\sigma_s^2 = \kappa_s^2 SSC_V$, and attenuation, $\alpha_s = \zeta_s SSC_V$.

In accordance with Eq 2.31, α_s estimate requires a lot of assumptions since its parameter uncertainty and nonlinear behavior with respect to d_s. On the other hand, Topping et al. (2007) demonstrated that SCB from WCB can be computed by assuming a constant concentration and SSPSD using a multi-cell ADCP and

thus reduced uncertainty in α_s calibration (Landers et al., 2016). If the particle concentration and size distribution are assumed to be constant along the observation cells of the H-ADCP, the last two terms of Eq 2.29 become zero when differentiated with respect to the acoustic path length.

$$\frac{d}{dr}(SCB) = \frac{d}{dr}(WCB + 2r\alpha_s) = 0 \tag{2.33}$$

The solution of the above equations is $\alpha_s = -0.5\frac{d}{dr}WCB$. Subsequently, α_s and SCB can be obtained (Landers, 2012). Substitution of Eq 2.29, in decibel scale, leads to the same relationship to the model in (Guerrero et al., 2016; Guerrero and Di Federico, 2018; Aleixo et al., 2020):

$$\alpha_s = -\alpha_w - \frac{1}{2r} - \frac{1}{40\log(e)}\frac{d(MB)}{dr}. \tag{2.34}$$

2.2.1.3 Estimation of suspended sediment concentration

Recalling $\alpha_s = \zeta_s C_v$, ζ_s can be determined using a relationship between ζ_s and ABR with respect to SSPSD proposed by Guerrero et al. (2016); Guerrero and Di Federico (2018). By computing α_s using Equation (2.34), and ζ_s, the volumetric concentration can be calculated by

$$C_v = \frac{1}{\zeta_s}(-\alpha_w - \frac{1}{2r} - \frac{1}{40\log(e)}\frac{dI_{dB}}{dr}). \tag{2.35}$$

To effectively implement the ABR method, a thorough examination of both particle size and the corresponding backscattering signal is crucial. When confronted with a non-stationary PSD during rainfall events, determining ζ_s becomes challenging unless particle sizes are monitored simultaneously, for instance, using instruments like LISST. Unfortunately, automated flow monitoring stations lack particle size analyzers, making direct ζ_s estimation unfeasible. One could potentially use the PSD from the calibration step for ζ_s determination, assuming stationary PSD. However, this assumption shares the limitation of using only one linear fitting model in Equation (1.1) for a station, as aforementioned in the Introduction. Recognizing this limitation, this study addresses the nonlinearity inherent in unsteady PSD during SSC estimation by employing a machine learning technique—specifically, support vector regression.

In Son (2021), Equation (1.1) was modified to address the nonlinearity arising from estimating SSC from SCB. Specifically, the paper focused on correcting errors due to hysteretic behavior, incorporating the water stage of the observation point. Rather than simply applying multiple linear regression, the interaction between SCB and water stage was considered, adding $SCB \cdot h$ as an additional variable in the equation. The modified equation used in this context is as follows.

$$\log_{10}(\text{SSC}_V) = C_1 \cdot \text{SCB} + C_2 \cdot h + C_3(\text{SCB} \cdot h) + C_4 \tag{2.36}$$

As mentioned earlier, hysteresis is influenced not only by the depth and flow

rate but also by their derivatives. Although Son (2021) showed an improvement in accuracy with the mentioned equation, there is potential for further enhancement by incorporating temporal information.

2.3 Regression methods

2.3.1 Support vector regression (SVR)

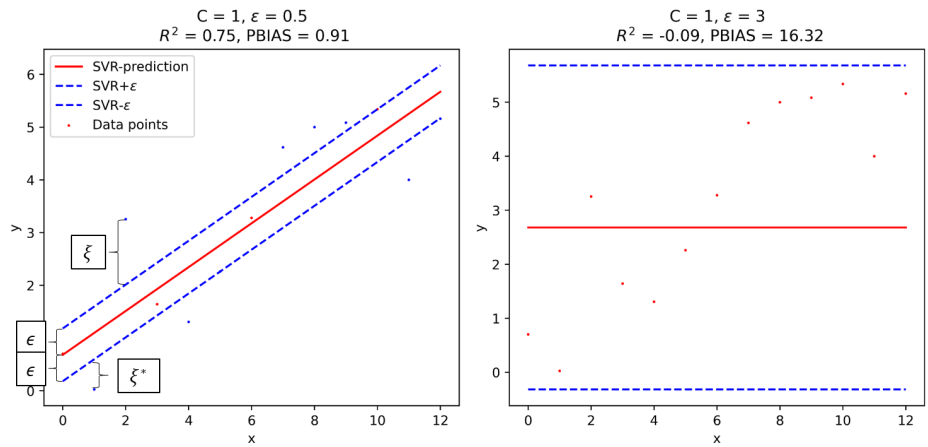


Figure 2.6 Schematic examples of the linear SVR's training rule. The figure depicts data points generated from a noisy sinusoidal signal. The red and blue points represent inside- and outside-margin points, respectively. The thick red line represents the exact SVR prediction, while the dashed blue line denotes the margin boundary.

Support vector regression (SVR) is a branch of an SVM Drucker et al. (1996). In the classification problem, SVM (or support vector classification) separates data classes from the decision boundary by maximizing the margin, which is the distance between two parallel hyperplanes expanded from the decision boundary. In particular, the ϵ-insensitive SVR achieves regression by placing target data points within the fixed-

width margin of 2ϵ width and constructing the flattest regression function possible unless the points are outside of the margin Awad et al. (2015); Kazemi et al. (2021). Usually, the real-world data are non-linearly distributed with errors so there are cases of the margin not being able to contain all data points. To consider the possible errors, SVR allows the upper and lower offsets from the margin demarcation by introducing slack variables (ξ and ξ^*) and the regularization coefficient (C_{SVR}). Figure 2.6 illustrates a schematic example of two SVR fitting cases with the linear kernel to help understand the training rule of SVR. In the figure, the tube consisting of the two blue dashed lines is the margin, and the width between the blue dashed lines is 2ϵ. However, setting too large ϵ is not favorable since the prediction is too flat, resulting in deteriorated prediction accuracy.

C-SVR is trained by the optimization process of the following primal problem:

$$\begin{aligned} \min_{\vec{w},b} \quad & \tfrac{1}{2}||\vec{w}||^2 + C_{SVR}\textstyle\sum_{i=1}^{n} F(\xi_i) + C_{SVR}\sum_{i=1}^{n} F(\xi_i^*) \\ \text{subject to} \quad & (\vec{w}^T\vec{x_i} + b) - y_i \leq \epsilon + \xi_i \\ & y_i - (\vec{w}^T\vec{x_i} + b) \leq \epsilon + \xi_i^* \\ & \xi_i, \xi_i^* \geq 0 \\ \text{for} \quad & i = 1, 2, ..., n, \end{aligned} \tag{2.37}$$

where C_{SVR} is the regularization cost coefficient; $F(\xi)$ is the arbitrary cost function

for ξ. SVR solves the Lagrangian dual problem in Equation 2.39. By setting the cost function l-1 $F(\xi) = \xi$, the Lagrangian dual problem can be set as follows:

$$\begin{aligned} \max_{\alpha,\alpha^*} \quad & -\tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(\vec{x}_i, \vec{x}_j) \\ & + \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)y_i - \sum_{i=1}^{n}(\alpha_i\epsilon + \alpha_i^*\epsilon^*) \\ \text{subject to} \quad & \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0 \\ & 0 \leq \alpha_i, \alpha_i^* \leq C_{SVR} \\ & \text{for } i, j = 1, ..., n, \end{aligned} \tag{2.38}$$

(2.39)

where α and α^* are Lagrangian multipliers and $K(x, x)$ is the kernel function. The kernel function maps the dot product $\vec{x_i}^T\vec{x_j}$ to a higher dimension such that SVR is likely to find the appropriate predictive function. When no kernel is applied, it is equal to the linear kernel, which has the functional form $K(\vec{x_i}, \vec{x_j}) = \vec{x_i}^T\vec{x_j}$. Another popular kernel is the radial basis function (RBF) kernel, which is defined as:

$$K(\vec{x_i}, \vec{x_j}) = \exp[-\gamma_{RBF}||\vec{x_i} - \vec{x_j}||^2], \tag{2.40}$$

where γ_{RBF} is the inverse of the influence radius of the samples.

Notably, the above Lagrangian dual problem is quadratic programming with respect to α and α^*, that is, the convex optimization rule is applicable. Furthermore, this problem satisfies the Karush-Kuhn-Tucker conditions, which guarantee that the solution to the dual problem coincides with that of the primal problem. Thus, SVR always yields a unique optimum solution when the target data and parameter combinations are provided. The fact that SVR always converges to a unique optimum solution benefits SVR. In contrast, neural networks are prone to converge to local optima because of parameter setting, learning rate, and noise in the data (Smola and Schölkopf, 2004).

In this study, the Python machine learning library Scikit-learn (Pedregosa et al., 2011) was exploited to train SVR models.

2.3.1.1 Parameter tuning using the grid search

Indeed, hyperparameter tuning by trial-and-error referring to the fitness score, such as R^2, is the way to tune the hyperparameters. However, it's not a clever way to guarantee the accuracy of the trained model. The most favorable way is using local or global optimization methods, but it may consume tremendous computation resources. The much simpler way to find the hyperparameter combination with high accuracy is the grid search.

The idea of grid search is very simple (rather similar to the manual trial-and-error approach). The difference between trial and error is that it covers the user-defined

feasible parameter space. The grid search is to evaluate fitness scores for all user-defined parameter combinations based on a grid. For example, if the user defines the parameter grid $C_{SVR} = 1, 2, 3$ and $\epsilon = 0.01, 0.1, 1$, the method evaluates 9 scores for ([C_{SVR}, ϵ] = [[1,0.01],[1,0.1],[1,1],[2,0.01],[2,0.1],[2,1],[3,0.01],[3,0.1],[3,1]]) and the user finally chose the parameter combination that the fittest case among all possible combinations.

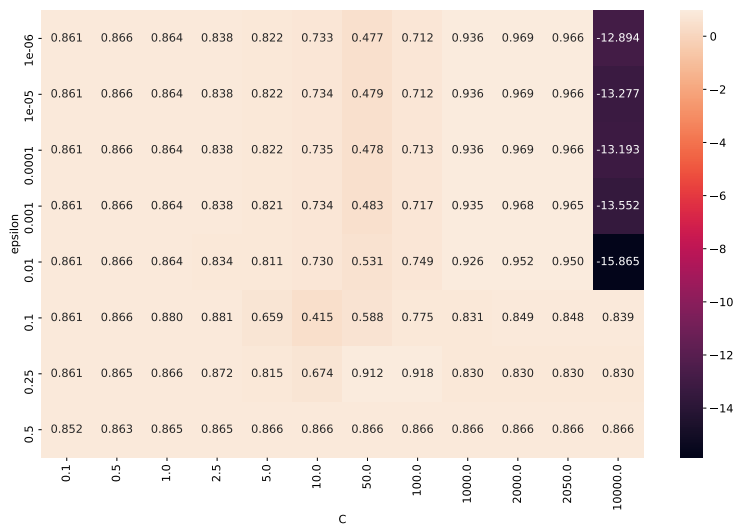


Figure 2.7 Example of the evaluated score by the grid search

2.3.1.2 Implementation of the K-fold cross validation

One important feature of the good model is generalization. For example, the well-generalized model that is globally applicable would be said to be better than the model that can only be applicable to a particular condition. In a more detailed example, if the regression model, which is only correct for the trained condition, is applied to the

different conditions, it will produce wrong predictions, resulting in wrong decision-making. In other words, we want to find a robust model. Conventionally, the given dataset is divided into two subsets: the training set and the test set. After, the model is fitted only using the training set, and the scores are evaluated using both subsets.

The question can be expanded to whether the hyperparameter combination has a high score only for the particular separation of the given dataset, having consistently low scores for the other subset combinations. The problem is also related to over-fitting. The K-fold cross validation (K-CV) is the most popular way to check the generalization ability Berrar (2019).

The idea of the K-CV is simple. First, the target dataset is separated into K subsets. There are many methods to sample the subsets (e.g., sequentially and randomly), but all the methods have the same purpose of dividing the dataset into the same size of K-subsets. After the dataset is divided, the fitness score is evaluated for K times regarding each subset as a test set.

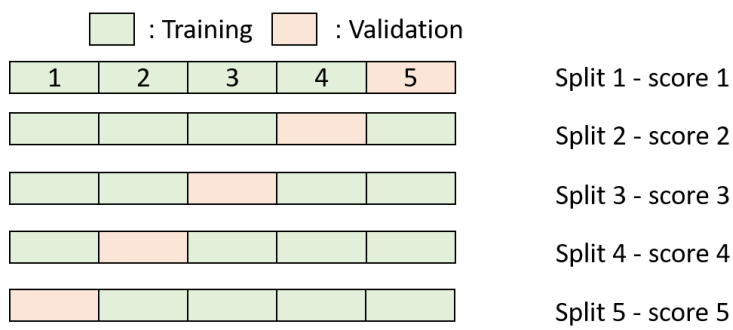


Figure 2.8 Example of the K-fold cross validation ($K = 5$)

Figure 2.8 shows an example of the K-CV where $K = 5$. The final score of the model with a certain hyperparameter combination is obtained by averaging the fitness scores (2.41).

$$\text{Score}_{CV} = \frac{1}{K} \sum_{i=1}^{K} \text{Score}_i \tag{2.41}$$

where, Score_{CV} and Score_i are the averaged K-CV score and the i-th trial's score. As a result, the case with high scores for every split can have a high cross-validation score since the score of the case with over-fitting will be suppressed due to poor predictions of the rest split.

On the other hand, the produced model with optimal hyperparameter based on the K-CV has a general feature over the given dataset. Therefore, the final model to be used in a practical sense can be derived using the whole dataset. This approach is helpful to deal with the lack of data. Even though the model using the whole dataset has a generalized ability, the performance comparison has to be conducted using the fitness from the K-CV.

The K-CV and the grid search can be applied independently. Hence, the user can conduct the K-CV for each score evaluation of the grid search. As a result, the user can find the robust model, which has the optimal hyperparameter set.

2.3.1.3 Recursive feature elimination for SVR (RFE-SVR)

The extraction of the governing feature to express the empirical relationship was performed by recursive feature elimination for SVR (RFE-SVR). RFE-SVR is a feature-selection technique for the SVM problem suggested by Guyon et al. (2002). In RFE-SVR, the importance of each feature is updated according to the ranking criterion. For the linear SVM, the ranking criterion c_p is $\vec{w_p}^2$, which is the p-th weight vector component corresponding to the p-th feature. As a generalization of nonlinear kernel applications, the ranking criterion of the p-th feature c_p can be computed as:

$$c_p = \frac{1}{2}|\sum_{i,j=1}^{N}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(\vec{x_j}, \vec{x_j})) - \sum_{i,j=1}^{N}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(\vec{x_j}^{(-p)}, \vec{x_j}^{(-p)}))|, \tag{2.42}$$

where $\vec{x_j}^{(-p)}$ is $\vec{x_j}$ without the p-th feature. The update step eliminates the smallest feature importance c_p. Subsequently, SVM is trained using the input data of the reduced features. The training-elimination sequence continues until the features remain in the user-defined feature size. Figure 2.9 presents the flowchart of the RFE-SVR algorithm.

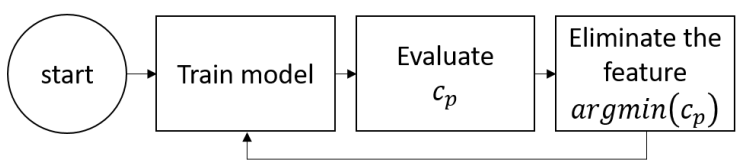


Figure 2.9 Schematic of RFE-SVR

Optimizing the model based on the combination of input variables can be achieved by repeatedly training the model on possible variable combinations and

selecting the combination that yields the most accurate predictions. However, trying all possible combinations requires a significant amount of resources. For instance, if there are N_p candidate input variables, $2^{N_p}-1$ training iterations are needed. However, RFE-SVM has the advantage of significantly reducing the number of required training iterations from $2^{N_p}-1$ to N_p for determining the variables in SVM.

2.3.1.4 Incorporation of CV with RFE (RFE-CV)

Additionally, cross-validation for each iteration of RFE-SVR was performed to assess the model fitness. CV provides information about the generalized performance of the model with minimized overfitting risk. The so-called K-fold CV method divides the entire dataset into K subsets and repeats the model fitting K times. For the i-th model fitting, the i-th subset is regarded as a test set, and the model is fitted to the remaining K-1 subsets. By repeating the training for each subset, the average test-set fitness score is considered the CV score. In RFE-SVR incorporated with CV, the algorithm evaluates the CV scores at every feature elimination step. CV signifies that the model with a certain parameter setting (e.g., input variable, hyperparameters of SVM) predicts not only the training set but also other datasets as well as the CV score.

2.3.2 Genetic programming (GP)

Considering that SVR is a black-box model lacking explicit details about the underlying physics, practitioners unfamiliar with machine learning may encounter difficulties.

To address this, developing explicit equations driven by symbolic regression methods can provide insights into the physical structure and make it more accessible for practitioners to use.

Genetic programming (GP), introduced by Koza (1992), is a symbolic regression technique that exploits the learning rule of the GA in the empirical formulation. Unlike SVR, MGGP is a gray-box model because it produces explicit estimation equations where the machine finds the final equations (strictly, the regression function of SVR can be computed using α and α^*).

The individuals of the population are the genes in GP, as well as in GA. Every GP gene has a tree structure consisting of terminally connected branches. In the tree structure, functional operators, such as $+, -, \times, \div, \sqrt{\cdot}$, comprise a terminal, and the input variables are at the branches. Each gene becomes an equation by combining the variables according to the adjoint functional terminals, and regression performance measures are adopted as an objective function of the GP.

Because the GA concept is implemented in GP, the two representative GA operators, namely, mutation and crossover, are under the user-defined mutation and crossover probabilities. These GA operators modify the functional terminals of the population genes in every evolution of the selected gene. Mutation reproduces the offspring by changing the mathematical operators of the terminals. Two genes are required for the crossover operation. The crossover exchanges the terminals of the

chosen genes to breed offspring. Examples of the two GP operations are illustrated in Figure 2.10, where the mutation and crossover are differentiated using colors.

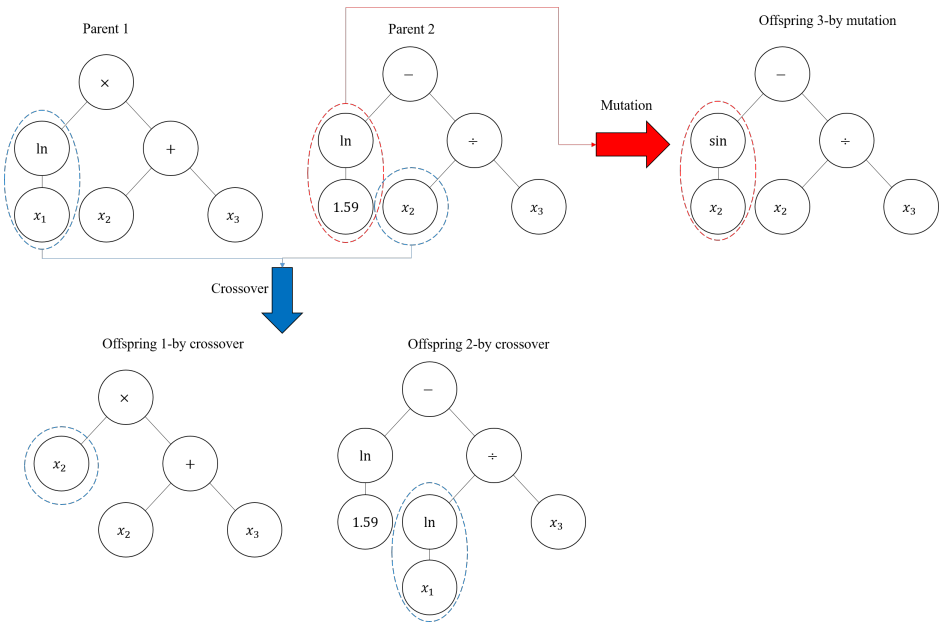


Figure 2.10 Examples of the GP operations (modified from Noh et al. (2020)). The blue and red markers indicate the crossover and mutation operations, respectively.

As a result of repeated evolutions, the population comprises various forms of equations. The best-fit equation in the last evolution is selected as the final product.

2.3.2.1 Multi-Gene Genetic Programming (MGGP)

MGGP is an advanced GP model. MGGP produces equations with multiple genes (terms of equations) for each solution (produced equation) to enhance variability without increasing the depth of the tree. Figure 2.11 shows an example of the gene expression of MGGP [tree depth = 3 and the number of trees = 2]. Additionally, GA

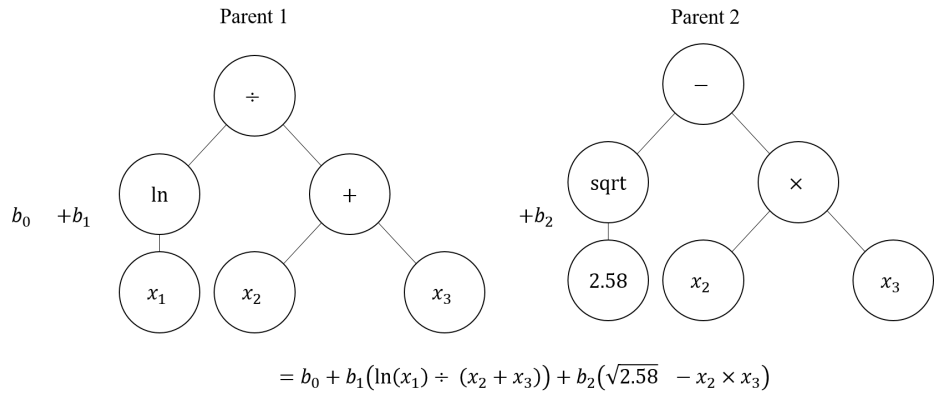


Figure 2.11 Example of MGGP formulation with trees multiplied by arbitrary regression coefficients b_0, b_1, and b_2 (modified from Noh et al. (2020))

operators operate in the MGGP. In MGGP, mutation and crossover events occur not only at the under-gene level but also at the gene-by-gene level. The former and latter operations are called high- and low-level operations for differentiation, respectively. For example, the high-level crossover exchanges the sub-genes of the two selected gene trees.

GA operations only formulate the structure of each formula in the population in MGGP. The regression coefficients (b_0, b_1, and b_2 in Figure 2.11) remain unknown. The least squares rule determines the regression coefficients. Finally, individuals in the population acquire a fully functional structure that can evaluate the target variable.

The MATLAB MGGP library genetic programming toolbox for the identification of physical systems (GPTIPS) was used, which yields Pareto solutions, as proposed by Searson (2015). The other advantage of GPTIPS is that it provides multiple independent runs, and thus, the initialization effect decreases. GPTIPS is available

at https://sites.google.com/site/gptips4matlab

2.3.2.2 Operon

The main challenge of symbolic regression is to enhance the accuracy of the produced equations by modifying the GP algorithm, for instance, the adoption of the high-level GA operation in MGGP. Recently, La Cava et al. (2021) compared the performance of cutting-edge symbolic regression methods and black-box machine-learning models using several benchmark problems. The benchmark analysis includes the accuracy and equation complexity of each symbolic regression method. The benchmark test result indicated that Operon (Burlacu et al., 2020) was a Pareto front model that considered accuracy and model complexity and was a state-of-the-art method with respect to accuracy (La Cava et al., 2021).

Burlacu et al. (2020) suggested a new tree initialization algorithm to ensure the population diversity with the linear tree encoding and implemented it in Operon. In the linear tree encoding of Operon, an exemplar tree $(a + b) \cdot (a + c) \cdot (b + c)$ is represented as $[a, b, +, a, c, +, b, c, +, \times]$. The tree initialization algorithm randomly samples the root, the highest level operator $\times$ in the example. Then, fill the functional and terminal sets according to the function's arity limits $[a'_{min}, a'_{max}]$, which are newly computed at each iteration. In addition, the functional variability was increased by adjusting a'_{min} to zero according to the user-specified probability.

In the tree evaluation, Operon enhances the efficiency by the data level paral-

lelism and vectorized computation. Operon determines the coefficients (such as b_0) of the symbolic inputs using a local search algorithm based on the nonlinear least squares method, which is supported by automatic differentiation, whereas the coefficients are determined by the GA operation in MGGP. The local search fine-tunes the coefficients of the individual equations, thereby increasing the accuracy of the final formulae. Operon accepts the failure event during offspring generation by returning *maybe* type (*optional* type) signal when the offspring does not meet the configurable criteria, terminating the offspring selection. In addition, Operon's encoding and offspring generation strategies reinforce strong parallelism and low memory demand.

The Operon code was originally developed in the C++ environment. In this work, we utilized the PyOperon library, a Python binding of Operon, to develop equations. PyOperon is available at https://github.com/heal-research/pyoperon

2.4 Clustering analysis

This section explores the methodology of clustering analysis, a powerful technique employed for various essential purposes. Clustering analysis serves as a key to unlocking the inherent structure within datasets, enabling the extraction of valuable insights, detection of anomalies, and identification of significant features. Beyond its exploratory role, clustering is pivotal for establishing natural classifications, revealing the degrees of similarity among different entities. Additionally, clustering acts as an

efficient means of data compression, allowing us to organize and summarize complex information through the identification of cluster prototypes. This study leverages clustering techniques to classify sediment characteristics and provide insight into the physical sediment transport process. This section outlines the fundamental principles and techniques employed in clustering analysis, shedding light on its diverse applications and significance in data exploration.

2.4.1 K-means

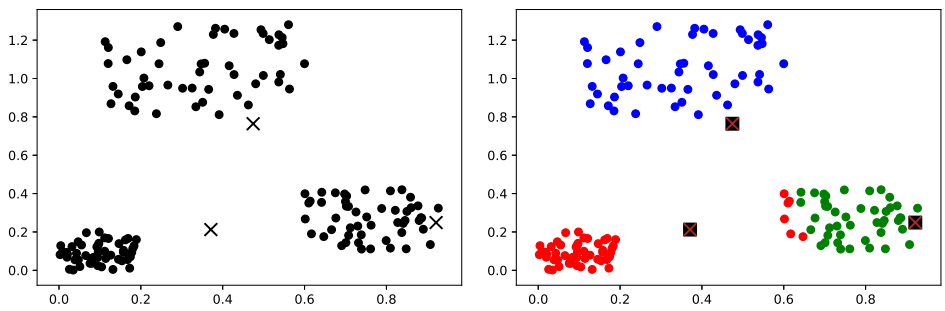
$$Var = \sum_{i=1}^{K} \sum_{\vec{x_j} \in S_i} |\vec{x_j} - c_i|^2 \tag{2.43}$$

in which, S_i is the ith cluster; $\vec{x_j}$ is the jth data point; c_i is the centroid of the cluster S_i.

The most popular algorithm for K-means clustering is the Lloyd algorithm (Lloyd, 1982), and a simple illustration of the algorithm is depicted in Figure 2.12. At first, the K number of centroids is initialized, randomly distributed. Then nearest points around centroids are grouped as K number of clusters (equation (2.44)).

$$Q_i^{(t)} = \{x_p : |x_p - c_i^{(t)}|^2 \leq |x_p - c_j^{(t)}|^2 \forall j, 1 \leq j \leq K\} \tag{2.44}$$

In the next step, centroids of newly grouped clusters are calculated using Equation (2.45). The algorithm repeats the sequence until components of each cluster no longer change.



(a) Centroid initialization

(b) Labelling

(c) Centroid update

(d) Re-labelling

Figure 2.12 Training process of K-means by the Lloyd algorithm. The X markers indicate the initial centroids and the square markers indicate the updated centroid

$$c_i^{(t+1)} = \frac{1}{|Q_i^{(t)}|} \sum_{\vec{x_j} \in Q_i^{(t)}} \vec{x_j} \tag{2.45}$$

2.4.2 Gaussian mixture model (GMM)

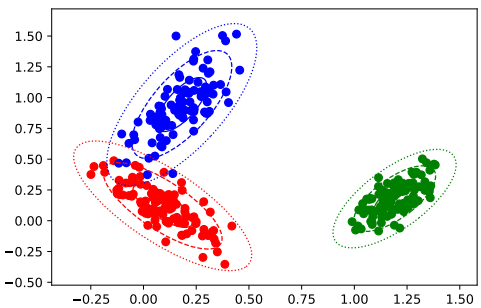


Figure 2.13 Gaussian mixture model mapping example on an arbitrary two-dimensional dataset ($K = 3$). The dots are randomly generated points using three artificial Gaussian distributions. Each trained Gaussian model is displayed with a colored ellipse, and assigned points are denoted by the colors of ellipses.

In natural cases, many datasets have statistical distributions. The Gaussian mixture model (GMM) assumes the data distribution as a mixture of K multi-variate Gaussian distributions, which is represented as

$$\mathcal{N}(x|\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp(-\frac{1}{2}(x - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(x - \boldsymbol{\mu})), \tag{2.46}$$

where x denotes the input data point, $\boldsymbol{\Sigma}$ denotes the covariance matrix, D denotes the number of dimensions, and $\boldsymbol{\mu}$ denotes the mean matrix. Figure 2.13 depicts how the three Gaussian distributions are mapped using GMM. By mapping data space

into several Gaussian superpositions according to weight, probabilities of the data points for each Gaussian can be calculated. Let t_k be the k-th Gaussian weight on the Gaussian mixture and μ_k and σ_k be the mean and covariance matrices, respectively; then, the probability density function of the trained GMM is calculated using Equation 2.47.

$$p(x) = \sum_{k=1}^{K} t_k \mathcal{N}(x|\boldsymbol{\mu}_k \boldsymbol{\Sigma}_k) \tag{2.47}$$

The probability of certain data can be viewed as the membership of K clusters.

The most common method used for training the GMM is the expectation-maximization (EM) algorithm Dempster et al. (1977). The EM algorithm repeats the expectation and maximization steps until it converges with the log-likelihood objective function. In the expectation step, it calculates the membership of the data points in k-th Gaussian distribution according to the following equation:

$$p(z_k = 1|x) \equiv \frac{p(z_k = 1)p(x|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(x|z_j = 1)} = \frac{t_k \mathcal{N}(x|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})}{\sum_{j=1}^{K} \tau_j \mathcal{N}(x|\boldsymbol{\mu_j}, \boldsymbol{\Sigma_j})} \tag{2.48}$$

In the maximization step, the algorithm maximizes the log-likelihood of the Gaussian mixture. Once the $p(z_k = 1|x)$ values are obtained, the maximization step updates

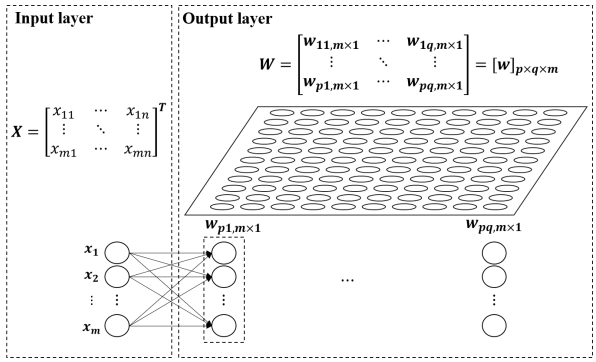


Figure 2.14 Schematic diagram of SOM networks.

the parameters μ, Σ, and τ as follows:

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}) = \sum_{n=1}^{N} [\sum_{j} \tau_j \mathcal{N}(x_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)] \tag{2.49}$$

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_n \tag{2.50}$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(x_n - \boldsymbol{\mu}_k)(x_n - \boldsymbol{\mu}_k)^T \tag{2.51}$$

$$t_k = \frac{N_k}{N} \tag{2.52}$$

Here, N is the quantity of data.

A detailed derivation of Equations 2.48 – 2.52 can be found in Bishop (2006).

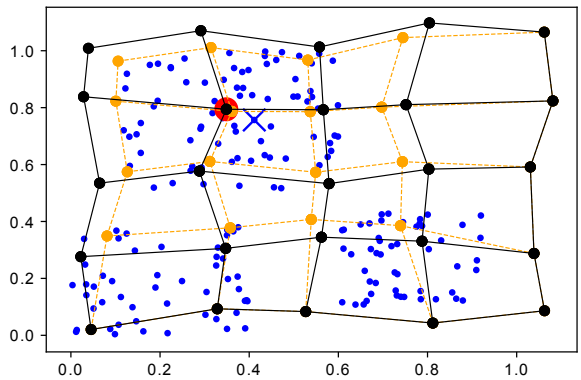


Figure 2.15 A simple SOM update example for 5 × 5 network for an iteration where σ = 3: the blue X marker is the target data point and the red dot is the winning node corresponding to the X marker.

2.4.3 Self-organizing map (SOM)

Self-organizing maps (SOMs) are simple models that map a data space to a lower-dimensional manifold. The primal SOM was introduced by Kohonen (1990).

The update rule of the primal SOM involves pulling the best matching unit (BMU), which is the closest grid node, to a randomly selected data point and adjacent nodes. The batch learning SOM (Kohonen, 2012) learns the dataset in a statistical sense such that simultaneously updating BMUs for all data points is identical to updating each selected data point at least once. Let $\vec{m}_i$ be the i-th node and $\vec{x}_j$ be the j-th data point; then, the batch SOM finds the BMU of all data points according to the following equation:

$$c(\vec{x}_j) = \arg\min_i (d[\vec{x}_j, \vec{m}_i]), \tag{2.53}$$

$$\vec{m}_i = \frac{\sum_j \lambda_n(c(\vec{x}_j), i), \vec{x}_j)}{\sum_j \lambda_n(c(\vec{x}_j), i)}, \tag{2.54}$$

where, $\lambda_n(c(\vec{x}_j), i)$ is the neighborhood function describing the grid node-wise distance (e.g., $\lambda_n(c(\vec{x}_j), i) = \exp(c(\vec{x}_j) - i)$) and $d[\vec{x}_j, \vec{m}_i]$ is the Euclidian distance between $\vec{x}_j$ and $\vec{m}_i]$.

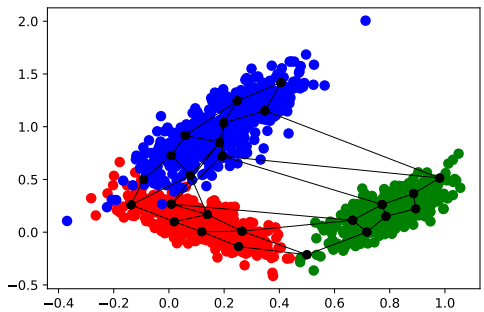


Figure 2.16 An example of 10×10 grid mapping of three Gaussian distributions by a planar self-organizing map. The randomly generated points under three Gaussian distributions are marked in red, blue, and green colored dots, respectively. The black dots and their connections are the trained self-organizing map grid components for entire points.

Figure 2.16 shows the 10×10 planar rectangular SOM grid mapped on random data points generated using three Gaussian distributions. SOM mimics the data distribution using the SOM map as black grids in Figure 2.16. Each grid point quantizes (summarizes) the data.

As the SOM map nodes are connected in a grid shape, the SOM map resembles the links between the quantized points. The advantageous feature of the SOM map is depicted in Figure 2.17. The hexagonal grid contours correspond to the x and y axes in Figure 2.16. The green dot cluster takes the place of the low y and the highest x.

The upper right side of the SOM map projects the green cluster such that the grid nodes are bright and dark in 2.17 (a) and (b), respectively.

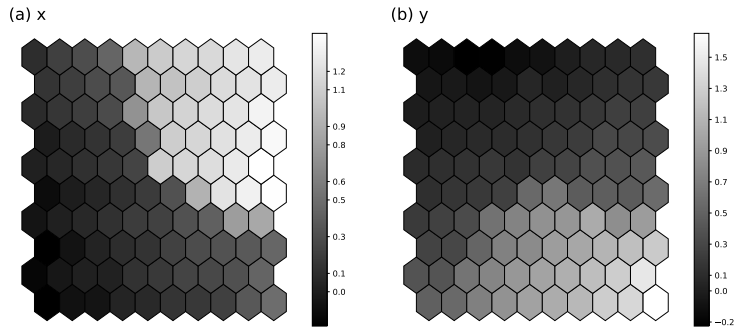


Figure 2.17 Component planes of the planar SOM depicted in Figure 2.16 for (a) x and (b) y. The face color of each hexagon denotes corresponding (a) x and (b) y values.

2.4.4 Clustering quality criteria

SOM, K-means, and GMM suffer from local extrema problems so the clustering result can change for every trial. Furthermore, a user of K-means has to specify the number of clusters K in order to perform K-means clustering. Hence, quantization error (QE) (Kohonen, 2012), topographic error (TE) (Kiviluoto, 1996), and Davies-Bouldin index (DBI) (Davies and Bouldin, 1979) were adopted to determine the performance of partitioning. QE and TE are the popular SOM performance measures on data density of the winning nodes and topographic preservation, respectively. The QE can be obtained by averaging distances between each data point and its winning node. The TE captures topographic continuity by counting whether the first and second winning

nodes are adjacent in the ordered weight vector grid. Two performance measures can be given as follows, respectively:

$$QE = \frac{1}{n}\sum_{j=1}^{n} ||\vec{x_j} - w_{k^*l^*}|| \tag{2.55}$$

where $w_{k^*l^*}$ is the winning node corresponding to the j-th data point $\vec{x_j}$.

$$TE = \frac{1}{n}\sum_{j=1}^{n} u_{TE}(\vec{x_j}), \text{ where } \begin{cases} 1, \text{ first- and second-winning nodes non-adjacent} \\ 0, \text{ otherwise} \end{cases} \tag{2.56}$$

The DBI measures how clustered data points are well-divided by calculating each cluster's variance and within-cluster distance. The mathematical expression of the DBI is given by

$$DBI(Q) = \frac{1}{K}\sum_{K=1}^{K} \max_{k \neq l}\left(\frac{s_C(S_k) + s_C(S_l)}{d_{ce}(S_k, S_l)}\right) \tag{2.57}$$

in which, s_C is the distance between the center of clusters and data points ($= \Sigma_i ||X_i - c_k||/N_k$); d_{ce} is the distance between clusters S_k and S_l (e.g., $= ||c_k - c_l||$). The numerator, S_C, of equation (2.57) has a negative relationship with homogeneity and the denominator is proportional to the dissimilarity between two clusters. Hence, the cluster number of the smallest DBI is the optimal number, since the dissimilarity of

clusters means that clusters are well classified.

The fitness of the GMM can be evaluated using model criteria. The Akaike information criterion (AIC) Akaike (1974) and Bayesian information criterion (BIC) Schwarz (1978) are popular examples of GMM fitness measures. AIC and BIC are defined by Equations (2.58) and (2.59), respectively.

$$AIC = -2LL + 2N_p, \tag{2.58}$$

$$BIC = -2LL + N_p log(N)., \tag{2.59}$$

where LL is the log-likelihood of the fitted model and N_p is the number of parameters of the fitted model. A model with a small AIC and BIC is considered good.

2.5 Metaheuristic global optimization

In hyperparameter and input variable tuning for SVR, traditional methods like grid search or RFE can be limited by predefined grids and variable nominees, leading to varied results. Grid search, in particular, demands an extensive grid setting for fine-tuning, resulting in a drastic increase in computational costs Stenger and Abel (2022). Consequently, there arises a necessity to explore more advanced optimization techniques.

Two widely utilized approaches are Bayesian optimization (BO) and meta-

heuristic algorithms. While BO excels in low-dimensional spaces, its efficiency diminishes in higher dimensions, especially in the context of nonlinear mixed-integer programming Garrido-Merchán and Hernández-Lobato (2020). On the other hand, metaheuristic optimization algorithms, leveraging evolutionary algorithms, operate without the need for gradients, handle discontinuities adeptly, are computationally efficient, and demonstrate robustness to noise, making them a suitable choice for diverse optimization challenges.

Studies comparing various optimization techniques for Support Vector Regression (SVR) models, such as the work by Malik et al. (2020), reveal that metaheuristic algorithms consistently outperform BOAs in terms of model fitting. Alibrahim and Ludwig (2021) conducted a comparative analysis of Bayesian optimization, GA, and grid search for tuning Artificial Neural Networks (ANNs). The results highlighted the superior speed of GA, with more than a twofold improvement in producing optimal solutions compared to Bayesian optimization. Furthermore, the inherent parallelizability of metaheuristic algorithms contributes to their efficiency and cost-effectiveness.

In our specific study, the goal extends beyond tuning SVR hyperparameters to developing a model capable of tuning an arbitrary number of input variables. Efficiently reducing costs in such scenarios, particularly dealing with high dimensions, makes metaheuristic algorithms highly appealing (Garrido-Merchán and Hernández-Lobato, 2020). Thus, our focus lies on introducing the Shuffled complex evolution-

University of Arizona (SCE-UA)-based metaheuristic algorithm embedded in the developed library. This method is known for its promise in solving hydrological problems involving unknown high dimensions (Naeini et al., 2019). This section will delve into the evolutionary algorithms integrated into our developed library, emphasizing their parallelizability and efficiency in dealing with complex optimization challenges.

2.5.1 Shuffled complex evolution (SCE)

As the name suggests, the SCE-UA algorithm (Duan et al., 1992, 1993) centers around the fundamental concept of shuffling subsets, termed 'complexes,' to evolve parameters in a global optimization algorithm iteratively. Figure 2.18 provides an overview of the optimization strategy employed by SCE-UA. Initially, it initializes the population within a predefined parameter space. Subsequently, individuals in the population are grouped into subsets known as complexes. The algorithm then employs an evolutionary algorithm (EA) to explore new offspring points for each complex. At each step, it assesses the fitness of the objective function. The offspring generated from the complexes replace individuals with lower fitness values. This optimization process repeats until the predefined stopping criteria are satisfied.

SCE-UA exhibits notable flexibility in the parameter update step, allowing for the application of various EAs during the complex updating process. Taking advantage of this adaptability, Naeini et al. (2018) explored the implementation of diverse EAs, including Competitive Complex Evolution (CCE; Duan et al., 1993), Modified CCE

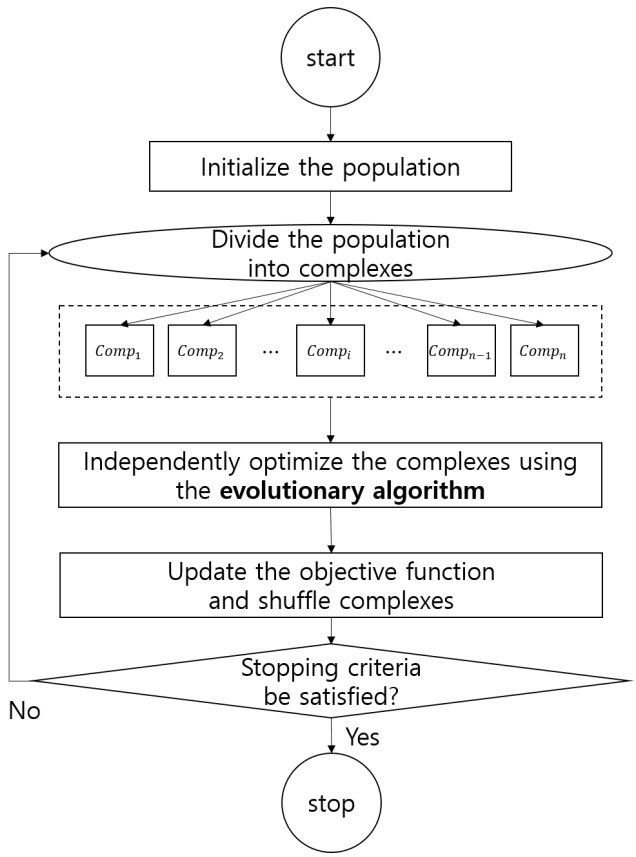


Figure 2.18 Flowchart of the shuffled complex evolution optimization structure

(MCCE; Chu et al., 2011), Frog Leaping (FL; Eusuff and Lansey, 2003), Greywolf Optimizer (GO; Mirjalili et al., 2014), and Differential Evolution (DE; Storn and Price, 1997). This chapter focuses on the description of CCE and MCCE, which serve as fundamental EAs for SCE.

2.5.1.1 Competitive complex evolution (CCE)

As EA, SCE-UA employs the CCE to update the population. The CCE algorithm sorts the parameter points according to their fitness and selects the reference points according to the triangular probability assigned by rank. Then, it generates offspring points by the simplex method, originally proposed by Nelder and Mead (1965). The pseudo-code for one epoch of CCE is outlined as follows:

1. Assign triangular probability to individuals based on the fitness value according to $p = \frac{2(n_{comp}+1-n_i^*)}{n_{comp}(n_{comp}+1)}$ where n_{comp} is the number of individuals in the complex, and n_i^* is the sorted rank of the i-th individual

2. Select N_p individuals from the complex, including parameter value and fitness

3. Generate offspring using the Nelder-Mead simplex algorithm from the selected individuals

4. Replace the worst individual in the complex with the newly generated offspring

Within the offspring generation, the following sub-steps are executed:

1. Find the centroid point $\vec{X}_c$

2. Reflection step: Compute the reflected point $\vec{X}_r$ with the worst individual $\vec{X}_w$, and $\vec{X}_c$ according to $\vec{X}_r = 2\vec{X}_c - \vec{X}_w$,

3. If fitness is improved, store $\vec{X}_r$ as offspring, and go to step (6); otherwise, do the expansion step: compute $\vec{X}_e = 2\vec{X}_r - \vec{X}_c$

4. If fitness is improved, store $\vec{X}_e$ as offspring, and go to step (6); otherwise, do the contraction step: compute $\vec{X}_{oc} = (\vec{X}_r + \vec{X}_c)/2$

5. If fitness is improved, store $\vec{X}_{oc}$ as offspring, and go to step (6)

6. Otherwise, generate a random point

7. Generate offspring

8. End the offspring generation process

2.5.2 Shuffled complex evolution with principal component analysis

SCE-UA relies on the n-dimensional Nelder–Mead simplex scheme for searching and evolving. Despite the effectiveness of the simplex method in reproducing qualified offspring, it introduces a critical issue: population degeneration Chu et al. (2010). This phenomenon arises as the offspring particles, produced through a series of simplex processes, converge into a subspace with lower dimensionality than the original search space. Subsequent evolutions then remain restricted within this subspace, hindering the recovery to the full parameter space.

To address the population degeneration challenge, Chu et al. (2011) introduced the Shuffled complex evolution with principal component analysis (SP-UCI) method. Unlike the original SCE-UA, SP-UCI incorporates the principal component analysis (PCA) to identify and search along dimensions not spanned by the sample population. This innovation ensures that the particle population can search the full space during every epoch.

SP-UCI integrates four key concepts expressed by individual algorithmic modules: (a) The complex shuffling scheme; (b) Population dimensionality monitoring and restoration using PCA; (c) Modified competitive complex evolution (MCCE) strategy; (d) Multinormal resampling.

2.5.2.1 Modified competitive complex evolution (MCCE)

In Chu et al. (2011), the CCE algorithm underwent slight modifications to enhance parameter searchability and dimension preservability. These adjustments include the incorporation of an additional inside contraction and the adoption of multi-modal resampling. In between steps (5) and (6) of offspring generation in CCE, the algorithm identifies the inside contraction point, $\vec{X}ic$, through the formula $\vec{X}oc = (\vec{X}w + \vec{X}c)/2$. In cases where no better point is identified through the reflection, expansion, and contraction operations, a random sample is generated, serving as the offspring in step (7), regardless of its fitness. Contrary to SCE-UA, SP-UCI takes a distinct approach by generating a random point from a multinormal distribution with a mean

of $\vec{X}c$ and a covariance matrix $\vec{D_m} = 2(\vec{D} + \text{mean}(\vec{D}))$, where $\vec{D}$ represents the diagonal matrix of the covariance matrix of the complex.

Chapter 3. Model parameter and input variable optimization

3.1 Necessities of the parameter and variable optimization techniques

The field measurement data related to sediment in this study may exhibit noise, and the dataset size may be insufficient for the application of complex models, such as artificial neural networks. Additionally, there exists domain knowledge regarding the physical relationships between SSC and backscatters or tractive force and sediment transport. Hence, a preferred approach involves primarily leveraging domain knowledge and utilizing machine learning techniques to address nonlinearities. In this context, an ideal machine learning method should be capable of accommodating noise and adhere to a training policy that minimizes model complexity. Consequently, SVR, which aligns with these recommendations, was employed to model the nonlinearity in SSC estimation using H-ADCP signals.

On the other hand, SVR needs fine-tuning of hyperparameters, such as margin parameter, regularization parameter, and kernel parameter. The grid search approach with cross-validation is commonly used to find the best hyperparameter tuning. However, the optimal solution depends on user-defined grid point setting and prior information (Raghavendra. N and Deka, 2014). In recent years, global optimization (GO)

approaches have been utilized for tuning the parameters of SVR (Lin et al., 2006; Liu et al., 2018; Tikhamarine et al., 2019; Kazemi et al., 2021). This method is more efficient for the higher-dimensional dataset. As an example of tuning four variables, using grid search with each grid point having 20 elements requires 160,000 evaluations of the objective function. On the other hand, GO can produce more precise values with 80 populations and 100 generations.

Input variable selection is also important in training SVR. The RFE-SVR, proposed by Guyon et al. (2002), efficiently determines the input variables combination, especially for high-dimensional datasets. It reduces iteration for input variables determination from $2^{N_d} - 1$ to N_d, where N_d is the dimension of the data (see, e.g., Noh et al. 2023c). Yan and Zhang (2015) highlighted instances where important variables were eliminated early, possibly resulting in their exclusion from the final model decision. During model calibration procedures in this study, it was suggested that the best feature set may vary based on the initial feature set, underscoring the limitations of relying solely on RFE for determining the optimal input variable set.

Meanwhile, parameter and model optimization are also crucial in clustering analysis. Specifically, two key challenges need to be addressed: (1) the convergence to different results in each iteration due to initialization dependency (local optima problem), and (2) the determination of the number of clusters problem. To tackle these issues, clustering metrics such as DBI and AIC have been developed, applied,

and employed for evaluating clustering models. However, a systematic resolution to the aforementioned two problems is still necessary.

Given that both SVR and clustering analysis demand careful optimization of parameters and models, this section elucidates the methodologies employed to fine-tune and enhance the performance of these models. The grid search with RFE and Cross-Validation (RFE-CV) presents an approach for simultaneously determining SVR hyperparameters and variables using a global optimization technique. Additionally, robust methods for determining the number of clusters, resilient to initialization challenges, are detailed.

3.2 SVR parameter and variable set optimization technique

3.2.1 Grid search with RFE and CV for SVR (Grid-RFE-CV)

In each hyperparameter combination of the grid-CV sequence, RFE-SVR was additionally performed, hereafter referred to as GRID-RFE-CV. In this GRID-RFE-CV system, the user can determine the hyperparameter values and input variables of the model with a generalized capability supported by the cross-validation score.

3.2.2 MOdel Selection with Global Optimization for SVR (MOSGO-SVR)

This study introduces a novel method for selecting an SVR model by simultaneously optimizing the three SVR-related hyperparameters, namely ϵ, C_{SVR}, and γ_{RBF}, along

with input variables. The proposed method utilizes the GO algorithm to fine-tune the SVR hyperparameters. Additionally, the GO algorithm is employed to determine the optimal combination of input variables, removing the need for feature elimination techniques such as RFE. Adopting this approach reduces the algorithmic complexity of model derivation, achieving hyperparameter fine-tuning and input variable determination through a single optimization technique.

This method adopts a cross-validation score for fitness evaluation to avoid overfitting and assure the robust performance of the resulting models. The hyperparameters and input variables can be optimized by the mixed integer nonlinear programming (MINLP), and the objective function of the proposed method is set to be:

$$\begin{aligned} \min_{C_{SVR},\epsilon,\gamma,I_v} \quad & 1 - R^2_{CV}(y, SVR(X[:, I_v])) \\ s.t. \quad & 10^{-9} \leq C_{SVR} \leq 10000 \\ & 10^{-9} \leq \epsilon \leq 10 \\ & 10^{-9} \leq \gamma_{RBF} \leq 10 \\ & I_v(i) \in \{0, 1\}, \text{for } i = 1, 2, ..., n_v. \end{aligned} \tag{3.1}$$

In the expression, y represents the target-dependent variable, and $R^2_{CV}(y, SVR(X[:, I_v]))$ denotes the cross-validation score of SVR estimation using the input variable set $X[:, I_v]$. Here, I_v serves as a flag designating a column of input variables, with each variable defined by a Boolean integer. This allows the identification of input

variable combinations with high scores during the optimization process. Given that C_{SVR} spans several decades, ranging from 10^{-9} to 10^{4}, a logarithmic mapping of the search space is adopted.

The decision variables for MOSGO-SVR include SVR's hyperparameters (C_{SVR}, ϵ, and γ) and flags denoting the usage of the i-th variable ($I_v(i)$). Optimization of these three hyperparameters occurs within a continuous real-valued search space using the GO technique. Simultaneously, the matrix I_v undergoes optimization through the GO algorithm within a range of 0 to 1. The real-number values obtained are rounded to 0 or 1, rendering them integer values. Then, the matrix I_v is composed of 0 or 1 integer components, corresponding to the number of input variable nominees. Each component serves as a variable use flag: setting $I_v(i)$ to 1 implies the inclusion of the i-th variable in model training, while 0 implies exclusion. For instance, in a scenario with three variables to optimize, consider $I_v = [1, 0, 1, 0]$. In this case, the first and third variables are included in the training, while the second and fourth variables are excluded. Consequently, the dimension of the optimization problem becomes 6 (three for hyperparameters and four for input variables).

As all model input variables and parameter decisions depend on the global optimization technique, it was named MOdel Selection with Global Optimization for SVR (MOSGO-SVR). It was implemented MOSGO-SVR with the Python library for Global Optimization and SHallow learning (pyGOSH). The pyGOSH is available at

https://github.com/hyoddubi1/pyGOSH. Note that any GO algorithm accepting the above objective function can be employed for MOSGO-SVR.

In addressing MINLP problems, the choice between Bayesian optimization and metaheuristic algorithms depends on the nature of the objective function. Bayesian optimization excels in optimizing expensive black-box objectives with uncertainty analysis capabilities, particularly for deep neural networks. In cases where the objective function, like the SVR used in this study, is not excessively expensive, metaheuristic algorithms offer a computationally efficient alternative for MINLP problem-solving (Eghbal et al., 2011; Garrido-Merchán and Hernández-Lobato, 2020). Also, for MINLP, Bayesian optimization might not be as effective in searching integer variables (Garrido-Merchán and Hernández-Lobato, 2020). For SVR model selection, a shuffled complex evolution with principal components analysis–University of California at Irvine (SP-UCI; Chu et al. 2011), which is designed to address the issue of losing dimensionality in the local population, was utilized, thereby improving the searchability of the parameter space.

The SP-UCI algorithm begins by initializing a population, which is then divided into subsets known as complexes. Each complex evolves its solution using a Modified Competitive Complex Algorithm (MCCE). In this process, the MCCE algorithm selects reference points within a complex based on their fitness, as determined by a triangular probability density function. These points are subsequently updated

Table 3.1 Fine-tuning ability and computational costs of the SVR optimization approaches

Features	Grid search+brute force	GRID-RFE-CV	GO-RFE-CV	MOSGO-SVR
Hyperparameter fine-tuning	X	X	O	O
Input variable selection	X	O	O	O
Computation cost order	$O((2^{n_v}-1)n_{grid}^3)$	$O(n_v n_{grid}^3)$	$O(3n_v n_{comp} n_{evol})^{(1)}$	$O(3n_{comp} n_{evol})^{(1)}$
Exemplar cost	50,625$^{(2)}$	13,500$^{(2)}$	48,000$^{(3)}$	12,000$^{(3)}$

$^{(1)}$: The computational order depends on the algorithm
$^{(2)}$: $n_{grid}^3 = 15$ and $n_v = 4$
$^{(3)}$: The SP-UCI algorithm with 20 complexes of 100 population and 200 evolutions

using the Nelder-Mead simplex method (Nelder and Mead, 1965). Afterward, all points are shuffled and re-sampled. SP-UCI employs PCA in each evolution step to maintain population dimensionality. If the dimensionality is reduced, the algorithm triggers a resampling. The SP-UCI process continues until certain stop criteria, such as objective function improvements, are met. The detailed algorithm of SP-UCI can be found in Chu et al. (2011).

3.2.3 Comparison of the SVR optimization approaches

To assess the fine-tuning capabilities of the SVR optimization approaches, Table 3.1 provides a summary of these methods. Alongside GRID-RFE-CV and MOSGO-SVR, the comparison includes GO-RFE-CV, which employs global optimization exclusively for hyperparameter optimization with the RFE-determined input variable set. The table presents the order and exemplar amount of computational costs to illustrate the efficiency of the proposed algorithm.

Grid search-based methods incur high computational costs due to the curse of dimensionality. For instance, considering SVR's three hyperparameters as an example,

the computational cost scales with n_{grid}^3. As the number of hyperparameters increases, the computational cost escalates exponentially. GRID-RFE-CV can significantly reduce the computational cost with systematic determination of the input variable set, but it shares the disadvantage of grid search. While GO-RFE-CV can fine-tune hyperparameters, applying the GO algorithm exclusively for hyperparameters may require a computational cost comparable to brute force methods. MOSGO-SVR holds an advantage in optimizing both hyperparameters and input variable combinations simultaneously, offering a more cost-effective alternative compared to grid search-based approaches.

Despite its efficacy in identifying optimal hyperparameters and input variable sets, the MOSGO-SVR method employed in this study has a limitation. While increasing the number of variables or model complexity may enhance accuracy, it is undesirable due to the risk of overfitting. Although cross-validation can address the issue of generalization, i.e., overfitting, there is no regulation on the number of input variables, potentially leading to the inclusion of more variables than necessary.

To address this limitation, MOSGO-SVR could be enhanced by introducing a penalty for model complexity, similar to how the BIC penalizes the number of parameters relative to the data size. However, if the penalty weight is too large, the model may only utilize a few variables, resulting in suboptimal performance. Therefore, finding an appropriate modification for the objective function remains an

open challenge. Another potential solution involves providing a Pareto front solution set, as adopted in MGGP. This approach offers accuracy for every possible number of variables, providing users with flexibility. While effective in scenarios without objective function saddle points concerning the number of variables, it may not be user-friendly, particularly for practitioners seeking a deterministic model.

Considering both model performance and computational cost, MOSGO-SVR emerges as the preferred choice. However, its effectiveness may diminish when the number of variable nominees increases, as MOSGO-SVR optimizes the input variable combination by searching through the possible combinations. Therefore, in such cases, GRID-RFE-CV proves to be a practical alternative.

3.3 Iterative SOM–GMM algorithm

The incorporation of the Self-Organizing Map (SOM) in this study was instrumental for efficiently summarizing intricate information within the sediment dataset. The SOM's unique feature, the component planes, proved particularly valuable in minimizing the analytical complexity, especially when confronted with datasets of high dimensionality. Furthermore, the SOM's exceptional capability to map data using a grid manifold enhanced its effectiveness across various complex data distributions, allowing for a comprehensive understanding of the underlying patterns within the sediment data.

The two-stage clustering method is commonly used to apply SOM by incorporating an additional clustering approach. In general, a trained SOM network is further divided using K-means Li et al. (2018); Noh et al. (2021) or hierarchical clustering methods Alvarez-Guerra et al. (2008); Kim et al. (2020). K-means clustering is a more intuitive and simpler model than other models, but it has certain disadvantages because of the assumption that the data points are distributed in spherical clusters. This assumption can lead to misclassification when non-spherically distributed data are used. Moreover, K-means is a hard clustering method that assigns one label to one data point; therefore, it is not appropriate to manipulate datasets when data regions of different classes overlap (Heil et al., 2019). This hard separation feature renders K-means sensitive to noise or outliers (Jain, 2010; Oyelade et al., 2016). A fuzzy c-means clustering (FCM) was introduced by Bezdek et al. (1984) as an alternative to overcome the problem of hard division by fuzzifying K-means directly. However, FCM is limited to hyperspherical clustering.

However, GMM assumes a fuzzy mixture of multi-variate Gaussians with varying cross-correlations, which is an advantage of GMM over K means and FCM. From another perspective, the expectation of K-means can be reproduced when the user sets the covariance matrix of GMM to be spherical (i.e., $\mathbf{\Sigma}_k = \sigma_k \boldsymbol{I}$). These characteristics of GMM make it more reliable than K-means in data classification in general. Regime shifts of the sediment transport mechanism in natural rivers might not

be clearly divided and spherically distributed but rather composed of thin ellipses. The Gaussian shape mapping rule of GMM that allows cross-correlation is advantageous for summarizing the sediment transport dataset. Therefore, GMM was selected as the secondary clustering method in this study. Hereafter, the two-stage clustering algorithm using SOM and GMM is referred to as SOM-GMM.

Two challenges of SOM-GMM must be considered: (1) the prerequisite of the predefined number of clusters K (and grid size $p \times q$) and (2) local optima followed by initialization. Different strategies were applied at each stage to address these challenges.

There are several ways to update SOM: updating point by point in the order of the data, randomly selecting data for updates, and using batch computing, which updates the entire dataset based on the grid's position in each learning step, considering adjacent multiple data points (e.g., Kohonen, 2013). The first method may converge to different results when the order of the data changes, while the second method may lead to different results in each iteration. To minimize the impact of such initial conditions or dataset ordering, this study applies batch computing, training the entire dataset simultaneously in every step.

For the SOM stage, the grid size was determined according to the relationship $p \times q = 5\sqrt{n}$ (Vesanto et al., 2000). The location of each grid point, comprising a two-dimensional grid, was initialized by linearly spanning the grid over the two

largest principal components following the PCA of the target dataset (Kohonen, 2012, 2013). This PCA-based grid initialization strategy always yields the same training results unless the training epochs and dataset change. To optimize the SOM training, the training epoch was optimized, minimizing both QE and TE (Equations (2.55) and (2.56)).

The final two-stage GMM partitioning result was selected using an iterative method that was similar to a method used previously (Noh et al., 2021). The GMM was essentially trained over the possible number of clusters K. Because GMM is prone to converge to the local optimum solution depending on the initial state, it is iteratively retrained for each K. For example, the SOM-GMM procedure runs 200 times when the possible K values are in the range of 2–11, and 20 independent iterations are specified. AIC and BIC can be computed such that the clustering quality can be evaluated for every iteration. Finally, the case with the minimum $AIC + BIC$ was selected as the best clustering result produced by the SOM-GMM procedure (Figure 3.1). The flowchart of the iterative SOM-GMM algorithm is illustrated in Figure 3.2.

3.4 pyGOSH

This study proposes the Python library for Global Optimization and SHallow machine learning (pyGOSH). It is a comprehensive Python library that facilitates global optimization and shallow machine-learning tasks. This library encompasses a suite of

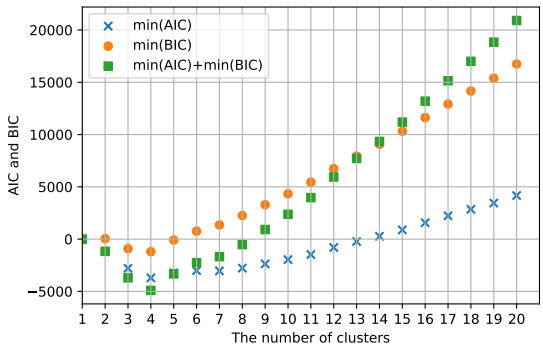


Figure 3.1 Example of the AIC and BIC evaluations with respect to the number of clusters.

functionalities, including global optimization algorithms, clustering techniques, and support vector regression-based algorithms, as introduced in this study.

The global optimization algorithm based on SCE was incorporated into pyGOSH. Harnessing the benefit of SCE's complex-wise update structure, this library facilitates parallel computing for global optimization. Through parallel computing, there was a notable enhancement in computational efficiency.

Notably, pyGOSH stands out by offering an advanced integration of machine learning algorithms, featuring key methodologies such as GRID-RFE-CV, MOSGO-SVR, and iterative SOM-GMM. This integration enhances the synergy between various algorithms, fostering efficient and coherent workflows. Furthermore, pyGOSH seamlessly integrates with popular machine learning models from the Scikit-learn library (Pedregosa et al., 2011), including SVR and scalers. This integration ensures compatibility and ease of use, empowering researchers and practitioners to leverage

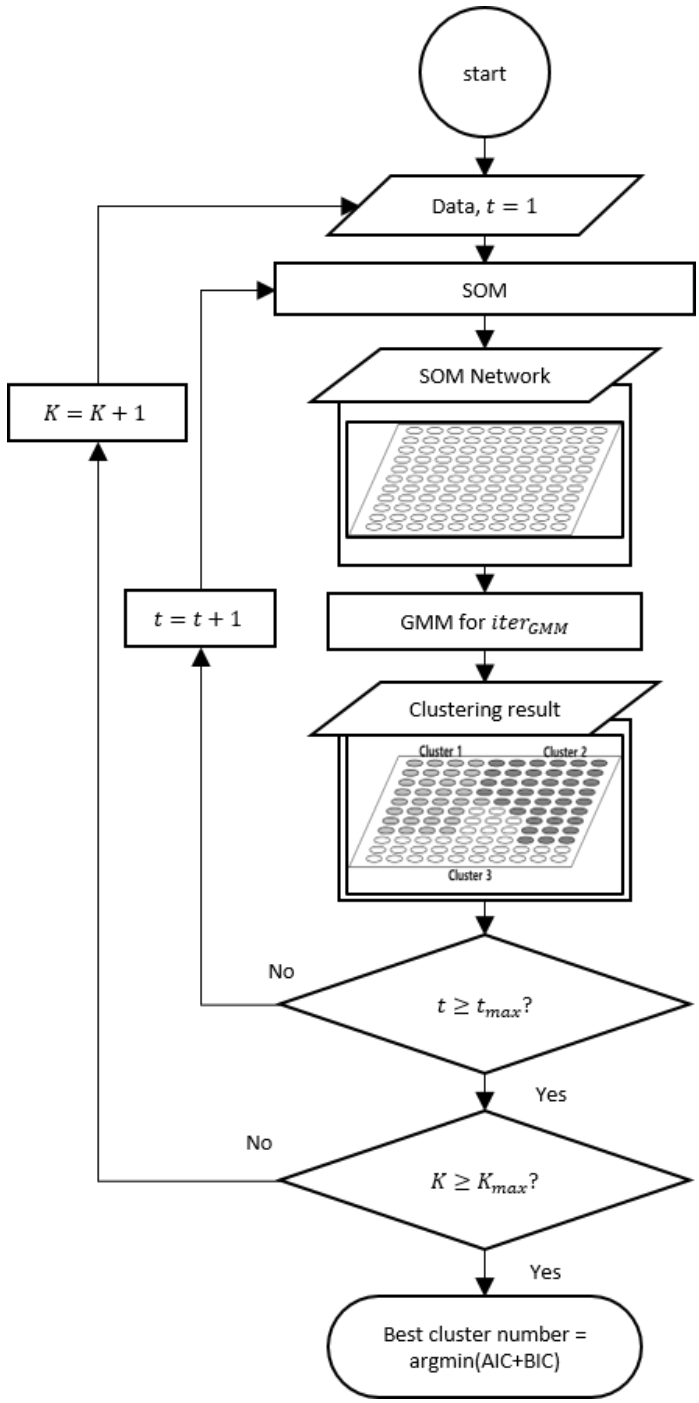


Figure 3.2 A flowchart of the SOM–GMM algorithm

the diverse capabilities of global optimization and shallow machine learning in a unified Python environment.

Chapter 4. Advancing H-ADCP-based real-time sediment load monitoring system using MOSGO-SVR and hydraulic variables

This chapter contains material that is partially reproduced from a manuscript currently under revision: Noh, Son, Kim & Park (2024, accepted) *Adv. Water Resour*. Note that the paper is currently accepted, and the information provided is subject to potential changes during the peer-review process. The inclusion of this partially reproduced content in the dissertation is based on the manuscript in its current form.

4.1 Dataset

4.1.1 Study sites and data acquisition

This study used the data observed in 2018 and 2019 at eight observation stations in South Korea. These stations were selected at points where the South Korean national gauging stations and H-ADCP installations overlap. Table 4.1 shows the observa-

Table 4.1 Data acquisition conditions of the study sites. Bridge, Weir, River, and Creek are marked in the table as B., W., R., and C., respectively.

Stations	Catchment	H-ADCP frequency (kHz)	Distance to sediment sampling location (km)	Distance to weir gate (km)	Data period
Geukrak B.	Yeongsan R.	1,200	-0.64	11.35	2019.07.01-08.31
Gumi B.	Nakdong R.	300	0	11.37	2019.07.01-10.31
Gyenae-ri	Nakdong R.	300	0.44	8	2018.07.01-08.31
Hoguk B.	Nakdong R.	300	2.23	-1.75	2018.06.01-10.31
Ipo W. upstream	Han R.	300	0	0.3	2019.06.01-10.31
Jijeong B.	Seom R.	300	-0.45	-21.1	2019.06.01-10.31
Naju B.	Yeongsan R.	300	-0.14	-4.64, 15.62	2019.07.10-08.31
Nampyeong B.	Jiseok C.	600	-0.24	1.10	2019.07.01-10.31

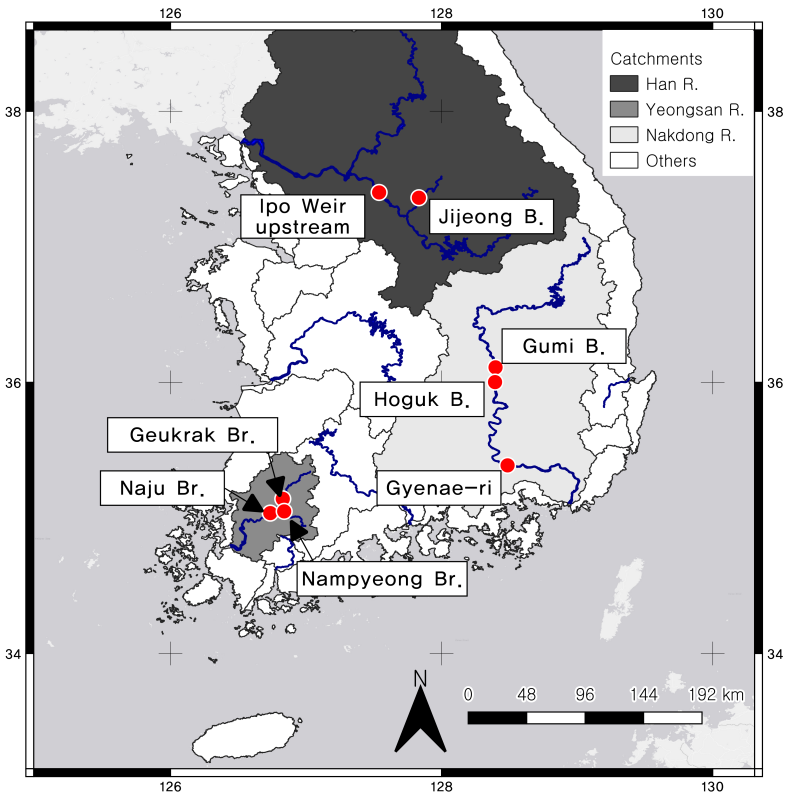


Figure 4.1 Geographical locations of the study sites

tion conditions, including H-ADCP frequencies, sampling location, and monitoring period. Figure 4.1 depicts the locations of the stations. All observation stations are located on major rivers, including Nakdong River, Han River, and Yeongsan River. Specifically, the Jijeong Bridge is on the tributary of the Han River, called the Seom River, and the Nampyeong Bridge is on the tributary of the Yeongsan River, called Jiseok Creek. The official water level and the flow rate data are accessible through the following URLs: https://www.hrfco.go.kr/ (Han River), https://www.nakdongriver.go.kr (Nakdong River), https://www.yeongsanriver.go.kr/ (Yeongsan River). Note that all data acquired in this study is from the existing monitoring stations.

Table 4.2 summarizes the measurement data, where d_{50bm} and d_{50ss} are the median particle sizes of the bed materials and suspended sediment, respectively. σ_{bm} is the standard deviation of the bed material, and it can be computed by:

$$\sigma_{bm} = \sqrt{d_{84bm}/d_{16bm}} \tag{4.1}$$

where, d_{84bm} and d_{16bm} are the 84% and 16% of the material finer by weight, respectively. σ_{ss} can be similarly obtained by the above relationship.

As explained in Chapter 2, the shear stress in unsteady flow is influenced by $\frac{\partial y}{\partial x} - \frac{U}{g}\frac{\partial U}{\partial x} - \frac{1}{g}\frac{\partial U}{\partial t}$. For a rectangular channel, the term $\frac{\partial U}{\partial t}$ can be expressed as $\frac{\partial}{\partial t}\left(\frac{Q}{Wh}\right)$. Moreover, spatial derivatives can be transformed into temporal derivatives

Table 4.2 Field measurements data summary of the study sites

Stations	The number of samples	Q (cms)			Fr			SSC (mg/l)			d_{50bm} (mm)	σ_{bm}	d_{50ss} (μm)	σ_{ss}
		Min	Mean	Max	Min	Mean	Max	Min	Mean	Max				
Geukrak	20	15.20	104.81	267.25	0.08	0.19	0.27	11.8	59.5	207.2	4.10	5.77	12.36	6.89
Gumi	19	92.44	802.37	4687.70	0.01	0.03	0.10	2.6	44.1	493.3	0.79	1.81	43.11	9.02
Gyenaeri	18	134.93	2272.57	4265.48	0.01	0.14	0.45	6.0	100.3	234.5	1.54	4.60	10.36	8.63
Hoguk	16	31.03	1517.23	3253.79	0.00	0.11	0.21	1.3	42.0	167.6	6.51	14.80	9.06	4.43
Ipo	17	97.79	246.62	689.67	0.01	0.03	0.09	4.0	31.9	172.3	1.21	6.66	385.79	4.25
Jijeong	13	19.48	132.19	255.94	0.03	0.12	0.19	7.4	68.2	228.8	1.07	4.69	22.50	5.73
Naju	15	39.99	265.30	530.66	0.02	0.09	0.16	10.2	64.8	212.0	4.39	9.48	42.56	11.55
Nampyeong	20	5.71	234.20	782.06	0.01	0.12	0.29	2.6	68.3	315.8	4.10	5.77	9.42	5.01

by multiplying $\partial x/\partial t$ by a chain rule. These transformations underscore the significance of Q, y, and their temporal variations in the context of unsteady shear stress. Additionally, Landers and Sturm (2013) highlighted that even at the same flow rate, SSPSD, a crucial parameter for SCB estimation, can vary before and after rainfall. Consequently, temporal derivatives of Q and y were included in the analysis to capture these variations.

SCB, water level (y), and flow rate (Q) obtained at 10-minute intervals from the automatic gauging station were used as input variables in the model derivation. Note that the 10-minute interval data represents the average of values measured at a 1 Hz sampling rate for 7 minutes and 30 seconds. Moreover, dy/dt and dQ/dt were computed to incorporate temporal information concerning flood waves. The time derivatives were determined by calculating the first derivatives of the fitted cubic spline curves. To develop the prediction model, the dataset of suspended sediment samples from the *Annual Hydrological Report of Korea* (MoE 2018; 2019) was compiled and used as ground-truth observations.

The suspended sediment concentrations were measured using a D-74 depth-integrating suspended sediment sampler. For streams over 300 m width, seven verticals were sampled, whereas five verticals were sampled for less than 300 m width. Following that, cross-section-averaged SSC was employed for the analysis. The suspended sediment samples were analyzed by the filtration method to obtain concentration. In

addition to that, the particle size distribution of the suspended sediment samples was measured using a laser diffraction device (Mastersizer3000). The SSPSD variations were observed in storm events.

Unfortunately, according to the hydrological survey report (MoE, 2019b 2019), there are cases where real-time water level observation station locations, H-ADCP installation points, and suspended sediment gauging station locations do not coincide with the distance between them being up to over 2 km. Additionally, Korea's major rivers have weir gates that control water level and flow rate. The operation of these weirs can cause significant fluctuations in water levels. In particular, MoE, 2019b (2019) notes that water level changes due to weir gates frequently occurred at the Gyenae-ri, Hoguk Bridge, Ipo Weir upstream, and Nampyeong Bridge stations. Moreover, when the weirs are not fully open, backwater can occur.

Figure 4.2 depicts the water level-flow rate graph of all monitoring stations listed in Table 4.1. In Gyenae-ri, Hoguk Bridge, and Ipo Weir upstream stations, the water level-flow rate loop presents significant fluctuations, and 8-shaped loops are observed due to the weir operation. The relationship between the water level and flow rate in Nampyeong Bridge shows a simple curve, whereas Naju Bridge does not follow a power-law trend.

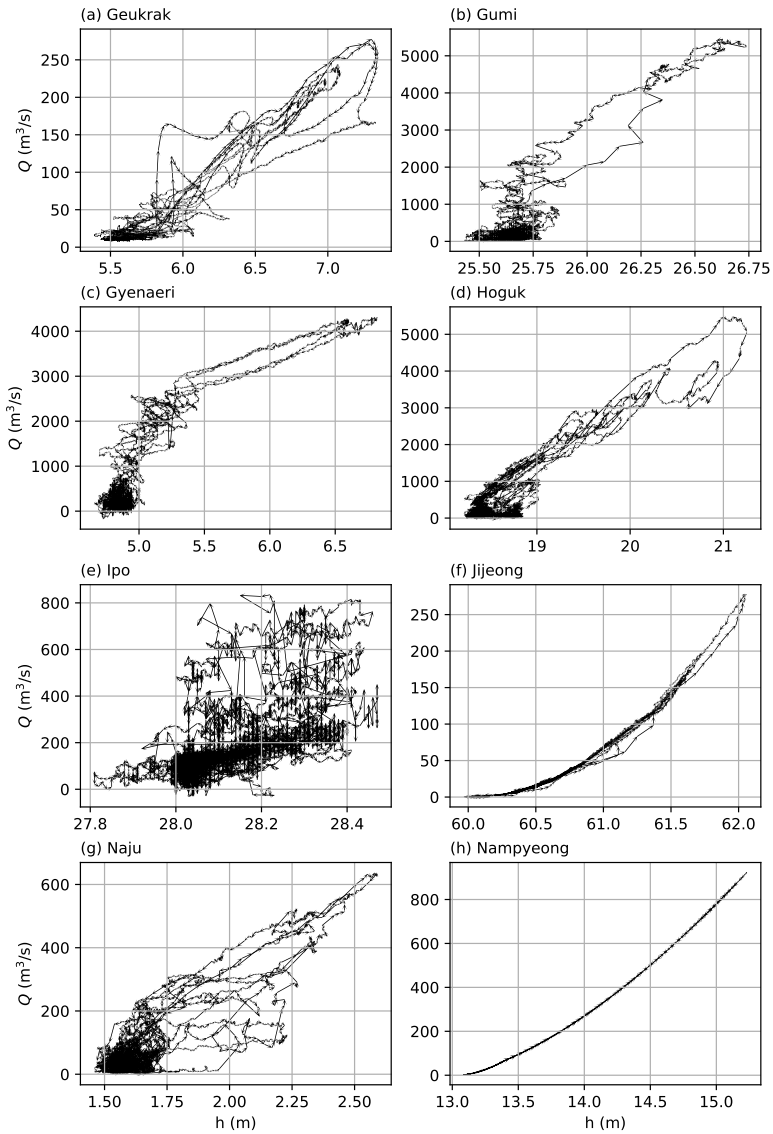


Figure 4.2 The water level-flow rate graph

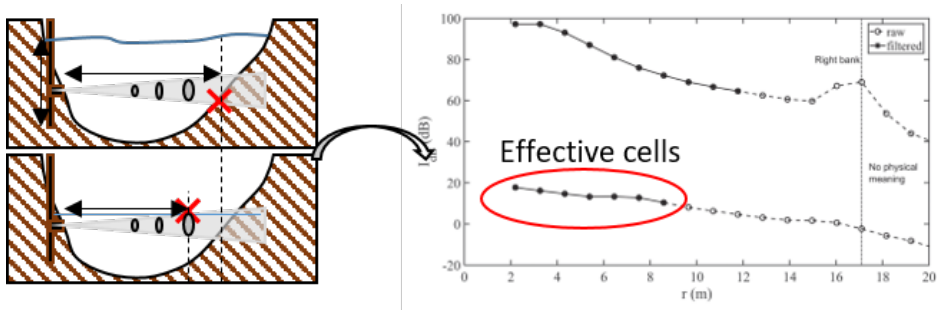


Figure 4.3 Schematic diagram of the effective cell decrease due to unwanted acoustic reflectances

4.1.2 H-ADCP signal processing

The H-ADCP signal returns measurements from the pre-defined cell setting. Prior to analysis, cells showing a low correlation between concentrations and echo intensity values were trimmed. By this step, the water surface or bottom reflected cells are trimmed. The number of trimmed cells varies with the water surface stage where the signal is reflected so that some valid cells in high stages may show erratic results in low stages (Figure 4.3. Therefore, from a conservative and practical point of view, the same number of cells was analyzed over the analysis period as the minimum number of cells from the transducer.

Equation (2.33) implies that the spatial derivatives of the echo intensity lead to the sediment-corrected backscatter. In some cells, the sign of derivatives can be positive, locally, resulting in negative SCB. However, the average derivative over the analysis region should be related to SSC under the uniform concentration assumption. Instead of calculating the derivative in each cell, the spatially averaged dI/dr

was obtained from the linear fitting of the intensity profile. Subsequently, the relationship between SCB and SSC was derived from the cross-section averaged index concentration concept.

Guerrero and Di Federico (2018)'s approach is useful in understanding a profound relationship with particle size distribution. However, PSD keeps changing, especially stream responses to distance sediment sources in the falling limb. In those cases, their model is particularly good with PSD analyzer's assistance, such as LISST-200X. However, this study is based on field sampling analysis without real-time PSD monitoring. In our cases, PSD variations between samplings are unknown.

In this context, the simple relationship was adopted without PSD information. Instead, this study adopts SVR with kernel to consider the nonlinearity due to the PSD effect in concentration evaluation.

4.2 Instream application performance comparison of the SVR model determination methods in the Nampyeong Bridge station test case

In the Nampyeong Bridge station, a comparison is conducted among various SSC monitoring techniques, including the rating curve, SVR optimization methods, and linear models. Table 4.3 provides an overview of the SVR model optimization methods and their respective calibration approaches, along with the optimal input variables

Table 4.3 Optimal hyperparameters determined by various SVR optimization approaches in the Nampyeong Bridge station

Methods	Model calibration		Variables	C_{SVR}	ϵ	γ_{SVR}
	Variables	Hyperparameters				
GRID-RFE-CV	RFE	Grid search	SCB	$2.048 \cdot 10^{3}$	$4 \cdot 10^{-1}$	1
GO-RFE-SVR	RFE	GO	SCB	$5.042 \cdot 10^{3}$	$2.02 \cdot 10^{-1}$	$3.95 \cdot 10^{-1}$
MOSGO-SVR	GO	GO	SCB, Q	$7.363 \cdot 10^{3}$	$2.62 \cdot 10^{-1}$	$1.58 \cdot 10^{-3}$

Table 4.4 Model structures and cross-validation scores in the Nampyerong Bridge station

Model	Model structure	R^2_{CV}
Rating curve	$SSC = Q_{SL}/Q = 0.1203Q^{0.6942}$	0.687
Linear-1var	$\log_{10} SSC = 0.0405SCB - 2.5729$	0.765
SCB-msl	$\log_{10} SSC = -0.4237SCB - 4.3882h + 0.0309(h \cdot SCB) + 54.5258$	0.792
GRID-RFE-CV	SSC=SVR(SCB)	0.812
GO-RFE-CV	SSC=SVR(SCB)	0.852
MOSGO-SVR	SSC=SVR(SCB,Q)	0.892

and hyperparameters. Furthermore, Table 4.4 presents the model structures and cross-validation scores for each model. The one-variable linear model, denoted as linear-1var (formulated as Equation (1.1)), and SCB-msl (representing linear models employing two features, as per Equation (2.36)) are included.

Among these models, the rating curve presents the lowest R^2_{CV}, while the SCB-msl model outperforms the one-variable linear model. Similarly, incorporating additional variables improves model accuracy in SVR models. The three SVR models show higher R^2_{CV} compared to the rating curve and the linear models. Notably, the MOSGO-SVR model stands out as the top performer among all models. The result implies that considering additional hydraulic features and using SVR is favorable.

4.3 Application of MOSGO-SVR to various monitoring stations

4.3.1 SSC monitoring result

In this study, to assess the performance of utilizing the raw backscatter variable from H-ADCP, the effects of variable inputs in the MOSGO-SVR were investigated. In the first case, simultaneous calibrations were performed for all possible variables to identify the most important variables at the monitoring station. In Case 2, similar to Son (2021), additional variables were taken into consideration, such as water level, while SCB is forced to be included. In Case 3, only SCB was adopted, but SVR was employed to account for nonlinearity instead of using algebraic equations. The input variables combinations and SVR hyperparameters and corresponding R^2_{CV} are presented in Table 4.5.

The final model with minimum objective function is selected to enhance the practical applicability, and the SVR model was re-fitted with the entire dataset using the optimum solution as recommended by Hastie et al. (2009). Figure 4.4 shows the suspended load (Q_{SL}) estimations corresponding to flow rate for the eight monitoring stations with the models in Case 1. The sediment load prediction results of SCB-fixed models (Case 2) and the SCB-only models (Case 3) are depicted in Figures 4.5 and 4.6, respectively. In the figures, the H-ADCP-based sediment load prediction is denoted with black arrows, and the red arrows denote the measured suspended loads.

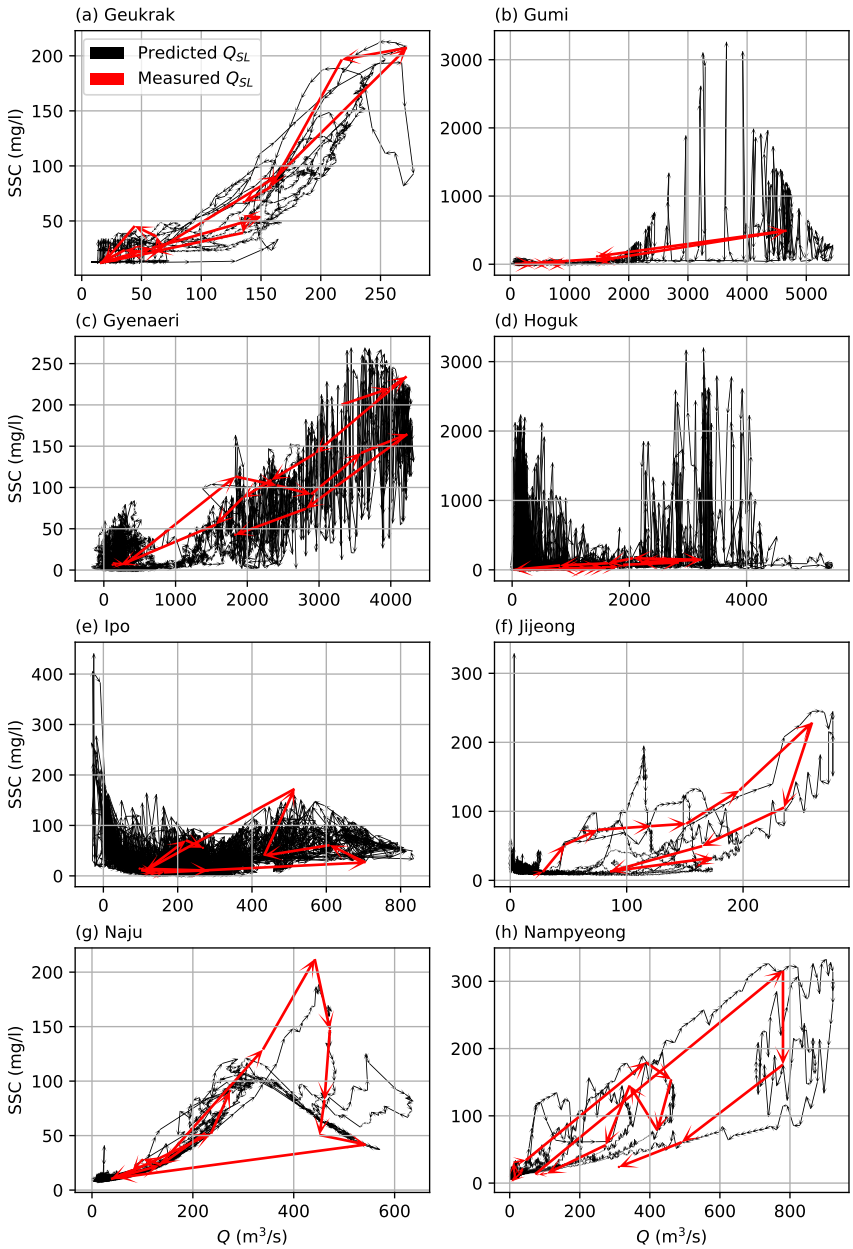


Figure 4.4 The graph of flow rate-suspended loads for Case 1, depicting temporal variations by using arrows during the monitoring periods.

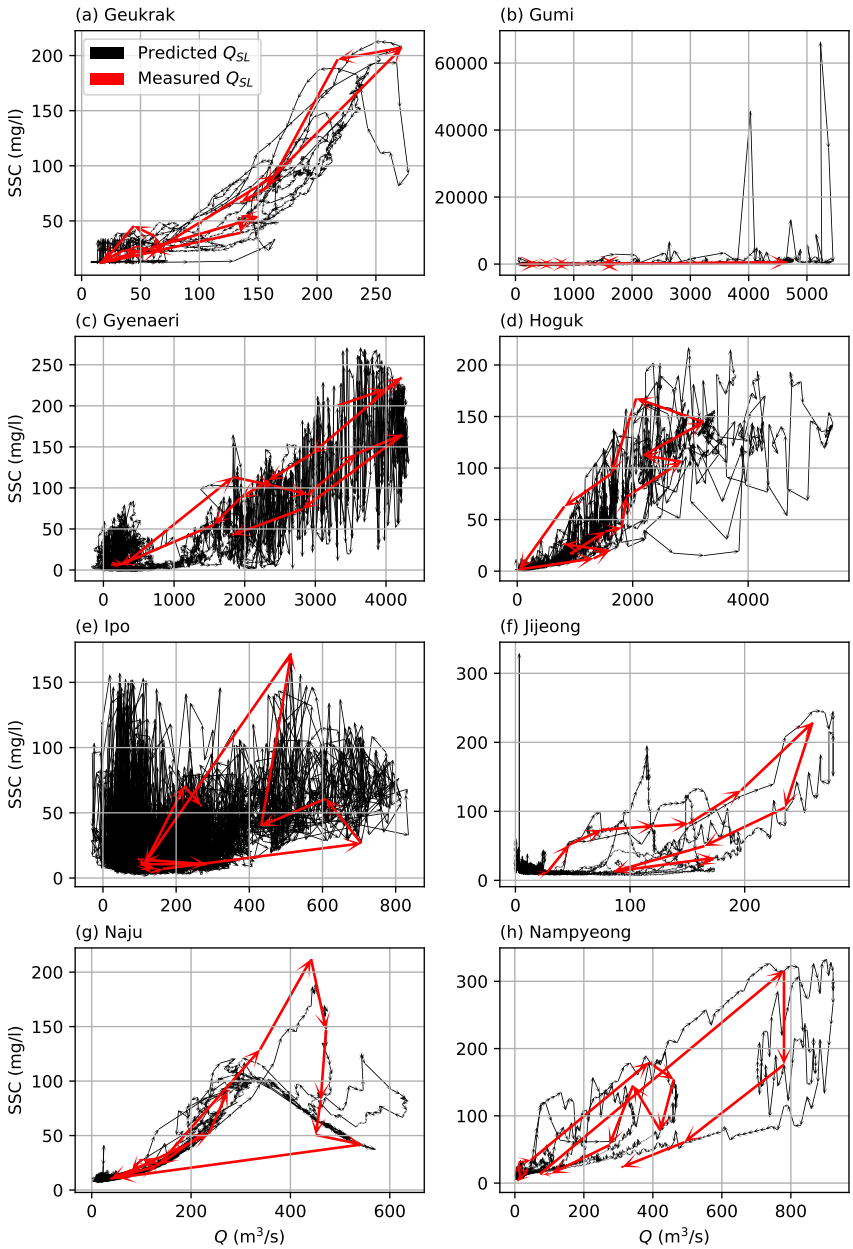


Figure 4.5 The graph of flow rate-suspended loads for Case 2, depicting temporal variations by using arrows during the monitoring periods.

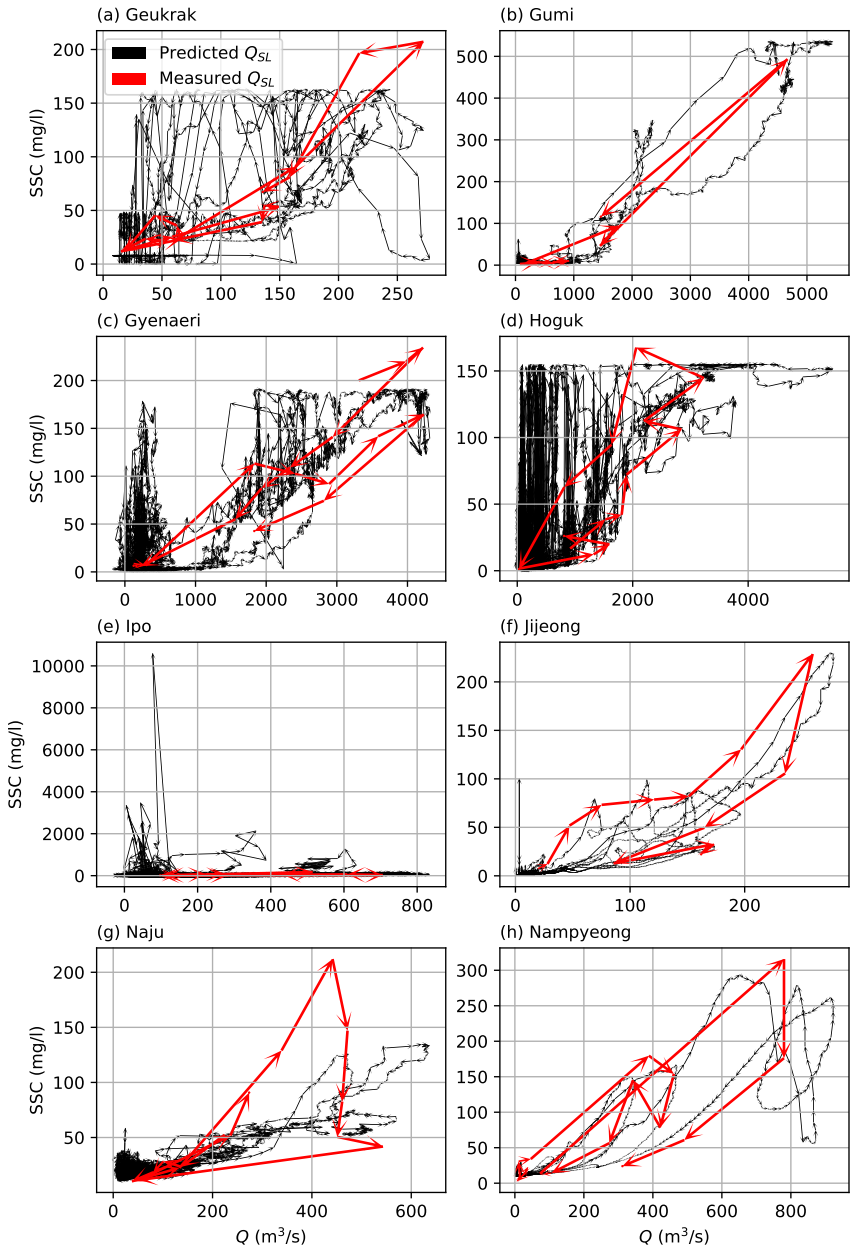


Figure 4.6 The graph of flow rate-suspended loads for Case 3, depicting temporal variations by using arrows during the monitoring periods.

Table 4.5 The MOSGO-SVR training results

Cases	Stations	Best fit variables	C_{SVR}	ϵ	γ	R^2_{CV}
Case 1 (No constraint)	Geukrak	SCB, y, Q	1.086E+00	1.417E-06	1.632E-01	0.958
	Gumi	Q, dQ/dt	1.337E+02	1.880E-02	3.407E-02	0.920
	Gyenaeri	SCB, y, dy/dt	9.670E+00	1.920E-02	1.033E-01	0.904
	Hoguk	y, dy/dt, dQ/dt	3.134E+01	3.722E-05	1.731E-01	0.713
	Ipo	y, Q, dQ/dt	7.294E+01	3.201E-02	3.127E-01	0.684
	Jijeong	SCB, Q, dy/dt	7.839E+00	2.903E-02	1.142E-01	0.928
	Naju	SCB, Q	4.297E+00	1.164E-08	9.955E-01	0.826
	Nampyeong	SCB, Q, dy/dt	1.276E+01	1.627E-06	2.517E-02	0.902
Case 2 (SCB forced)	Gumi	SCB, y, Q, dy/dt	7.581E+02	2.423E-06	1.008E-03	0.761
	Hoguk	SCB, Q, dy/dt	1.484E+01	3.814E-02	7.819E-02	0.711
	Ipo	SCB, y, Q, dQ/dt	6.277E+03	1.664E-02	2.065E-01	0.588
Case 3 (Only SCB)	Geukrak	SCB	9.765E+04	2.220E-01	8.195E-02	0.876
	Gumi	SCB	9.058E+00	4.484E-02	1.383E-01	0.230
	Gyenaeri	SCB	3.805E+01	2.253E-01	3.418E-01	0.860
	Hoguk	SCB	6.373E+00	7.540E-03	1.222E-01	0.461
	Ipo	SCB	2.197E+03	1.787E-03	8.941E-03	0.451
	Jijeong	SCB	3.806E+03	1.846E-01	1.005E-03	0.707
	Naju	SCB	1.188E+02	2.642E-01	2.280E-03	0.633
	Nampyeong	SCB	9.239E+04	2.060E-06	4.424E-03	0.833

In Case 1, all R^2_{CV} values exceeded 0.7. In particular, the H-ADCP backscattering was important in the five stations (Geukrak B., Gyenae-ri, Jijeong B., Naju B., and Nampyeong B.) where R^2_{CV} values were greater than 0.82. The flow rate was selected in most stations except for the Gyenae-ri station. In Case 1, dQ/dt, appeared as an important variable in the stations where SCB was not selected.

In case 2, MOSGO-SVR was performed, fixing SCB as the input variable for the Gumi B., Hoguk B., and Ipo weir upstream stations. Model fitness decreased in terms of R^2_{CV} when using SCB is forced. However, in this case, the Hoguk Bridge model's prediction turned out rather reasonable compared to that of Case 1. In Figure 4.4(d) SSC diverged to 10,000 kg/s, while the maximum SSC of Figure 4.5(d) is smaller than 1,000 kg/s. This result implies that SCB can work as a limiter when the

model is too sensitive to Q, y or their derivatives.

The results of Case 3 showed lower R^2_{CV} values compared to the other two cases. Particularly, at the Gumi Bridge station, the cross-validation score was approximately 0.69, lower than in Case 1. However, at Geukrak Bridge, Gyenae-ri, and Nampyeong Bridge, the decrease in R^2_{CV} was less than 0.1. The decrease in R^2_{CV} did not always negatively impact the model's predictive performance. According to Figure 4.6, smoother curves were observed for stations that previously showed significant fluctuations and divergence in Cases 1 and 2. This smoothing phenomenon was pronounced in stations with strong fluctuations and divergence, such as Gumi Bridge, Hoguk Bridge, and Gyenae-ri, located in the Nakdong catchment. However, in stations belonging to the Yeongsan catchment, such as Geukrak Bridge, Naju Bridge, and Nampyeong Bridge, the predictability of sediment concentration peak decreased. On the other hand, partially spiky estimations were observed at the Ipo Weir upstream station.

Ipo Weir upstream is located upstream of a weir gate, directly influenced by the flow from Ipo Weir's operation. This makes it difficult to establish a clear relationship between sediment loads and flow conditions. In poorly fitted stations, such as Gyenae-ri and Hoguk Bridge, the sediment sampling locations are situated at 0.44 km and 2.23 km downstream, respectively. These uncertainties lead to a low correlation between SCB and SSC.

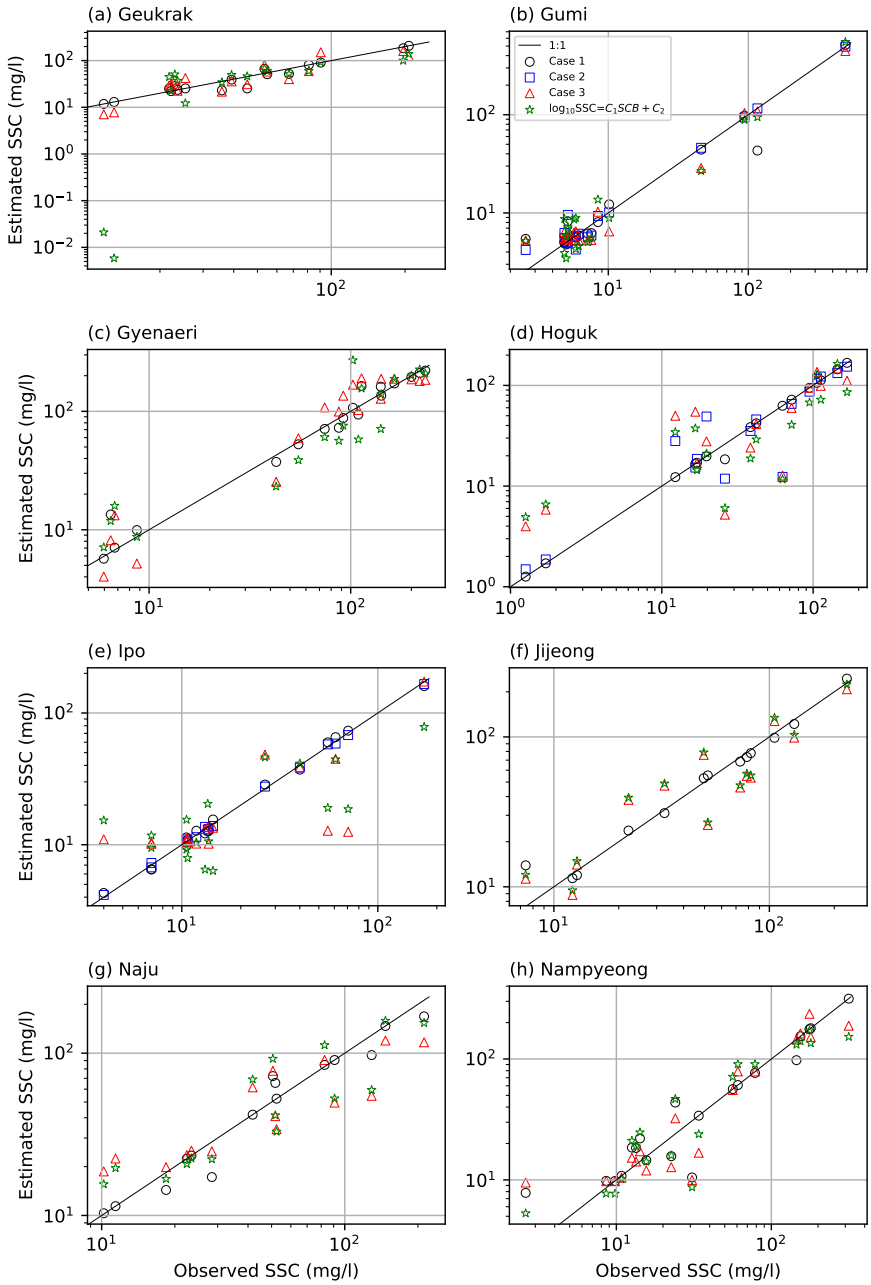


Figure 4.7 Scatter plots for Measured SSC versus estimated SSC using Cases 1–3

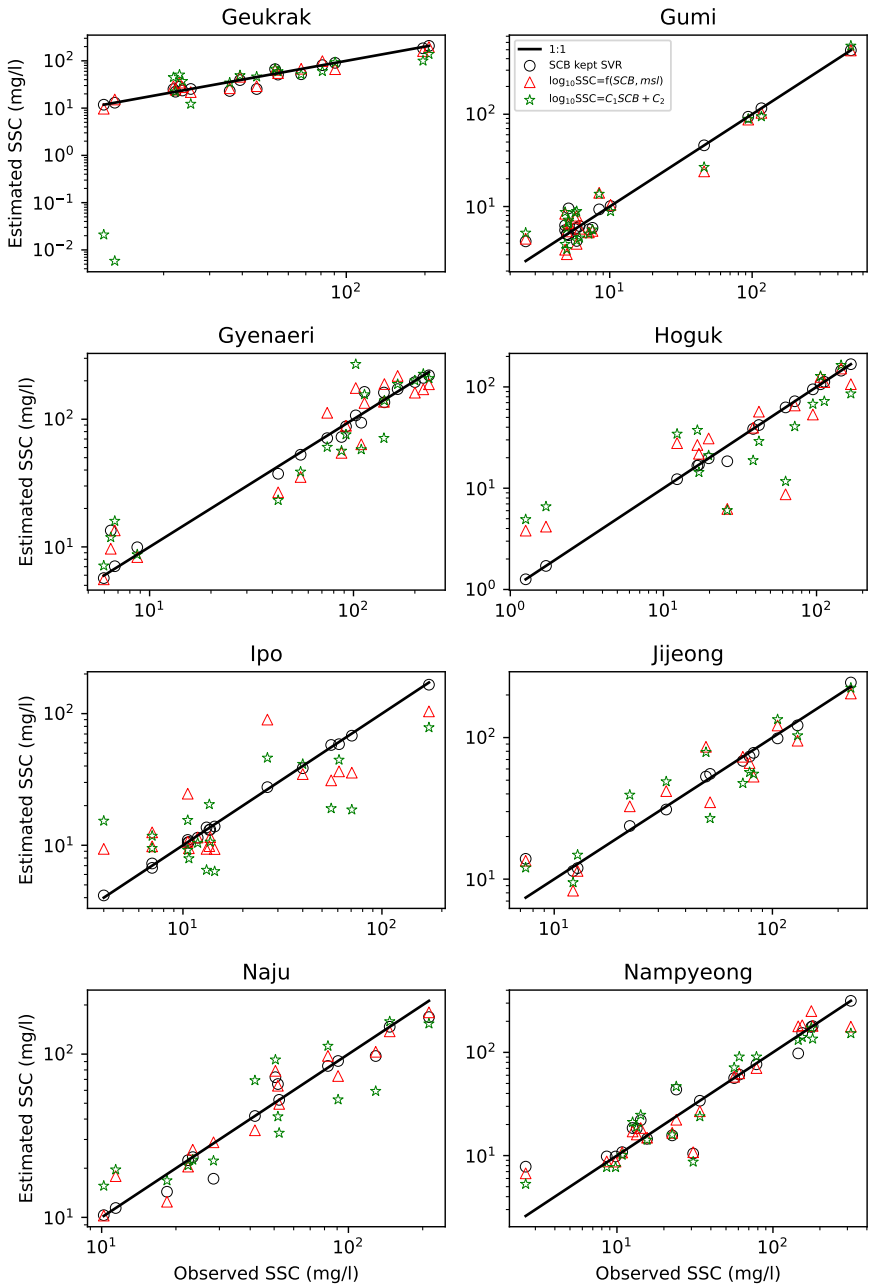


Figure 4.8 Scatter plots for Measured SSC versus estimated SSC using Case 2 and one- or two-variable linear models

Table 4.6 Regression coefficients of the linear models

	Linear-1var		SCB-msl			
Station	C_1	C_2	C_1	C_2	C_3	C_4
Geukrak	0.0599	-3.704	-0.1277	-1.7179	0.0234	10.6692
Gumi	0.0759	-6.1559	0.8555	3.4637	-0.0303	-95.2831
Gyenaeri	1.0771	-6.0976	0.4630	8.8243	-0.0801	-48.7645
Hoguk	0.0822	-6.6040	0.7692	4.5530	-0.0390	-87.7626
Ipo	0.0360	-1.6620	-3.2011	-7.6087	0.1143	214.1131
Jijeong	0.05376	-4.1514	0.7329	0.6730	-0.0106	-48.2121
Naju	0.0331	-1.6931	0.1982	5.7220	-0.0678	-15.9844
Nampyeong	0.0405	-2.5729	-0.4237	-4.3882	0.0359	54.5258

Figure 4.7 presents a pairwise comparison between the derived models and the observed Suspended Sediment Concentration (SSC). It is important to note that the SVR models in this comparison are re-fitted using the entire dataset to facilitate a fair comparison with linear models under the same conditions. Alongside the SVR models, the linear-1var models for each station are represented by star-shaped markers. The estimation results from Cases 1, 2, and 3 are indicated by circle, rectangle, and triangle markers, respectively. Figure 4.8 further illustrates the SCB-msl model with triangle markers. The coefficients of the linear models and the root mean squared errors (RMSEs) for all models are detailed in Tables 4.6 and 4.7, respectively, with the RMSEs of the best-performing models highlighted in bold.

Notably, circle markers, Case 1 models, are most closely aligned along the 1:1 estimation line among the compared models in Figure 4.7. However, at the Gumi Bridge station, an underestimation is observed around $SSC = 10^2$ mg/l. In many stations, the linear regression shows underestimation in high SSC and overestimation

Table 4.7 RMSEs of the refitted models for each station

Stations	RMSE (mg/l)				
	Case 1	Case 2	Case 3	Linear-1var	SCB-msl
Geukrak	**34.36**	-	122.29	130.37	54.93
Gumi	76.46	**5.75**	53.06	65.05	27.70
Gyenaeri	**62.81**	-	144.74	201.84	148.83
Hoguk	**7.68**	67.38	102.61	124.03	99.76
Ipo	14.82	**8.00**	77.59	117.68	107.12
Jijeong	**22.71**	-	76.67	74.08	70.35
Naju	**60.81**	-	137.61	117.62	58.18
Nampyeong	**57.03**	-	148.20	177.79	165.45

in low SSC. In Geukrak Bridge, SSC by the linear regression models are 10^3 times smaller than actual values.

In Figure 4.8, the SVR models retained through the SCB process demonstrate superior accuracy compared to other models. Notably, the SCB-msl model exhibits substantial enhancements over the one-variable linear models, underscoring the significance of incorporating hydraulic variables. However, at the Gyenari, Hoguk Bridge, and Ipo Weir upstream stations, certain data points display lower accuracy than the linear-1var model. Examining Table 4.5, MOSGO-SVR incorporates dy/dt in addition to SCB and y in those stations, a factor not considered in the SCB-msl model. This suggests the importance of accounting for the time derivative.

Comparing Case 3 and linear models, the Case 3 SVR models are superior to those in six stations. Compared to the linear models, the performance improvements of Case 1 and 2 models are more significant than those of Case 3. The models using

SCB and additional variables in most stations show better accuracy than the other cases, except for the Hoguk Bridge station. This result implies advantages in using SVR and employing hydraulic variables in addition to SCB. In the Hoguk Bridge station, the distance exceeding 2 km between the H-ADCP sensor and the sediment sampling point introduces a time lag between the SCB and SSC data. Consequently, the correlation between SCB and SSC diminishes, leading to lower accuracy.

The SCB-msl models exhibited superior performance compared to the linear-1var models. With the exception of the Gyenaeri, Ipo, and Nampyeong stations, RMSE values were smaller than those of Case 3. This suggests that manipulating nonlinearity through SVR is less impactful than including additional hydraulic variables. Nevertheless, the SCB-forced SVR models (Case 2) demonstrated better accuracy across all stations.

In summary, the fitted model reproduced a reasonable yet simple prediction curve throughout Cases 1 to 3, when SCB was used without incorporating other variables. However, this overly simplified approach underestimated SSC during high flows. Therefore, when flow-related variables were included in addition to SCB, the accuracy improved at higher flow rates. As more variables were applied, the model exhibited greater fluctuation and divergence. This demonstrates a trade-off between the model's complexity and accuracy when considering additional variables. Notably, it was observed that higher CV scores did not always lead to physically plausible

results. Therefore, in practical applications, the approach used in Case 2 is advisable. If the model based on estimation seems too erratic, then using a model derived solely from SCB, as in Case 3, would be more appropriate, applying the principle of Occam's razor.

4.3.2 Discussion on the optimized variable set

In the Naju Bridge station, during the initial rising phase of the flow, the peak suspended sediment concentration occurs, and surprisingly, the sediment concentration decreases when the flow reaches a gradual rise after the peak flow. Due to inconsistent behavior in suspended sediment concentration concerning the time derivative, it is not considered in the analysis. At the Geukrak Bridge, variations in SSC during both rising and falling flow phases are minimal, with fluctuations below 50 mg/l, leading to a limited consideration of the influence of time derivatives.

The Gumi Bridge station experiences large fluctuations in water level despite a modest 1.2 m amplitude of observed water level variations. On the other hand, especially with a flow amplitude exceeding 5,000 cms, significantly contributes to flow rate considerations in Case 1. In Case 2, where SCB is forcibly included, dy/dt is considered instead of dQ/dt.

Including SCB in all cases involves incorporating flow (Q), and simultaneous consideration of flow and SCB proves effective for estimating cross-sectional average suspended sediment concentration. In the case of Gyenaeri, Hoguk Bridge, and Ipo

Weir upstream, where water level fluctuations are frequent due to operational manipulations of the sluice gate, the consideration of water level, flow, or the time derivative of water level is relevant. In the Gyenaeri station, the flow exhibits significant fluctuations compared to other variables, leading to the exclusion of the flow rate from the analysis. For the Hoguk Bridge station, the correlation between SSC and SCB is low due to the difference in the location of sediment concentration measurements and H-ADCP, resulting in the initial exclusion of SCB. In the upstream region of Ipo Weir, frequent flow changes occur due to backwater effects, emphasizing the significant consideration of water level, flow, and the rate of flow change over SCB.

Chapter 5. Clustering of sediment characteristics in South Korean rivers and its expanded application strategy to H-ADCP based suspended sediment concentration monitoring technique

This chapter is partially reproduced from the following publication: Noh, Son, Kim & Park (2023) *J. Korea Water Resour. Assoc.*, **55**: 43-57. The content presented here has been adapted and expanded from the original publication to fit the context and objectives of this dissertation.

5.1 Linear model coefficient similarity

For convenience in the analysis, the coefficients of the derived H-ADCP-SSC relationship equation were plotted on a scatter plot with different markers for each river in Figure 5.1. Different marker shapes were used for cases where the frequency of the H-ADCP signal was different. The observation points represented by circles, where 300 kHz H-ADCPs were installed, showed that similar coefficient values were derived for the same river, supporting this study's assumption. However, for the tributary of the Namhan River, the coefficients were somewhat different from those of the two observation points in the mainstream of the Han River. A noteworthy point is that the derived equations have an accuracy of 0.98 in terms of the coefficient of determination,

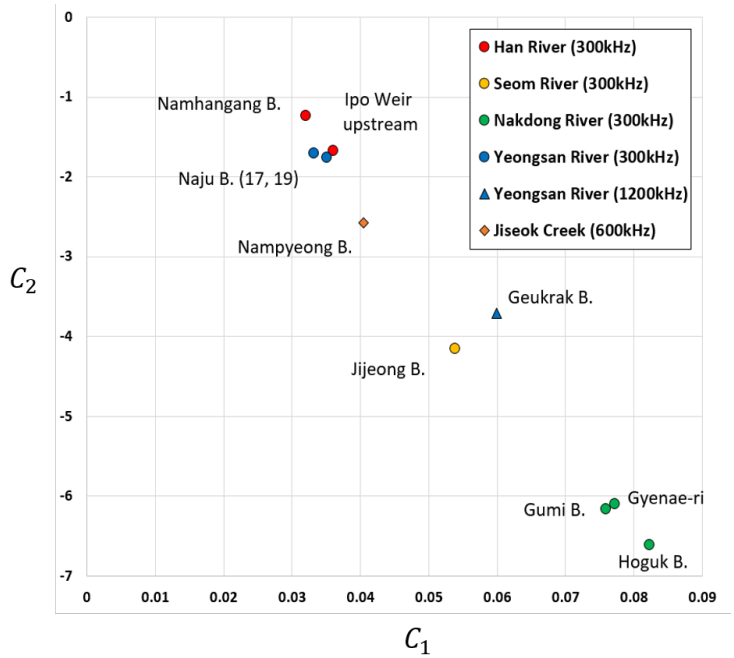


Figure 5.1 Scatter plot of H-ADCP-SSC equation coefficients corresponding to Table 4.6

which corresponds to the relationship in Eq 5.1.

$$C_2 = -103.678C_1 + 1.8862 \tag{5.1}$$

Using data from Naju Bridge located in the mainstream of Yeongsan River in 2017 and 2019, the coefficients of the H-ADCP-SSC equations derived for each year showed an error rate of less than 10%, despite the difference in timing. This result is less than the coefficient differences observed between adjacent stations in Han River or Nakdong River.

In stations located in the main and tributary streams of Yeongsan River, there

were stations that used H-ADCPs of different frequencies. In these cases, it was revealed that different coefficients were obtained even for the same main river. The Nampyeonggyo observation station, which used a 600 kHz H-ADCP and is in the same river basin as the Naju Bridge observation station, showed a value that was more than 0.006 larger than the mean value of the Naju Bridge station, and the value of C2 was about 0.850 smaller. The Geurak Bridge station in the Yeongsan River basin operates a 1,200 kHz H-ADCP, and in this case, a larger variation was observed with an increase of 0.026 and a decrease of 1.981 for C1 and C2, respectively. Considering that the Han River and Nakdong River observation stations have similar coefficient values, assuming that C1 and C2 for Geurak Bridge and Nampyeong Bridge can be approximated to the value of Naju Bridge using a 300 kHz H-ADCP, it is estimated that C1 and C2 will have positive and negative relationships, respectively, with the frequency of the H-ADCP.

5.2 Data description

In this study, the 2019 Annual Hydrological Report on Korea (MOE, 2019) was used as the clustering target. The report provides the coordinates and catchment areas of 44 sediment observation stations in Korea, as well as flow-suspended sediment discharge equations developed based on directly collected suspended sediment concentrations and measured suspended sediment discharge using the modified Einstein method

proposed by Colby and Hambree (1954). The flow-suspended sediment discharge equation is derived in the following form:

$$Q_{SL} = a_{SL} Q^{b_{SL}} \tag{5.2}$$

where, a_{SL} and b_{SL} are the regression coefficient. The total sediment load rating curve has the same form. To differentiate the notation, the total load and the regression coefficients are denoted as Q_{TL}, a_{TL}, and b_{TL}.

In addition to the suspended sediment discharge equation and suspended sediment characteristics, the report includes the grain size distribution of suspended sediment and bed sediment as similar features. The suspended sediment grain size distribution includes eight particle size categories ranging from 0.062 mm to 8 mm, each with corresponding weight distribution values. The bed sediment grain size distribution provides 20 particle size categories ranging from the 5% particle size d_5 to the 100% particle size d_{100} at 5% intervals, based on cumulative percentage values. The report also includes uniformity coefficient C_u, curvature coefficient C_g, and standard deviation σ_g calculated based on the parameterization of the grain size distribution characteristics in Eqs 5.3–5.5.

$$C_u = \frac{d_{60}}{d_{10}} \tag{5.3}$$

$$C_g = \frac{d_{30}^2}{d_{10}d_{60}} \tag{5.4}$$

$$\sigma_g = \sqrt{\frac{d_{84}}{d_{16}}} \tag{5.5}$$

Simons et al. (1981) emphasized the important of the gradation coefficient Gr and median grain size d_{50} in the total sediment load estimation. Therefore, as bed material characteristics, Gr and the dimensionless grain size d_* were computed as defined in Table 6.1. Table 5.1 presents the considered variables in 44 sediment monitoring stations.

5.3 Regional classification of the sediment monitoring stations

Clustering variables, including latitude, longitude, and catchment area, were additionally considered. In particular, 0.062 mm and 2 mm, thresholds for silt and sand were adopted as representative suspended sediment particle size distribution (SSPSD) variables, and d_{20}, d_{50}, d_{80} were adopted as bed material particle size distribution (BMPSD) characteristics. To investigate the combination of variables, 26 cases of clustering were independently conducted. The variables used in each case were highlighted in green and summarized in Table 3. The cases where all class values were used and the cases where only specific class values were used when applying the depth analysis data. In Table 5.2, the cases where all class values were used and the cases where only specific class values were used when applying the depth analysis data

Table 5.1 Variable summary of sediment monitoring stations

Location		Rating curve coeff.				SSPSD (mm)		BMPSD (mm or -)						
Station Name	Cat. Area (m^2)	a_{SL}	b_{SL}	a_{TL}	b_{TL}	0.062	2	d_{20}	d_{50}	d_{80}	C_u	C_g	σ_g	d_*
Jucheon B.	533.75	0.3536	1.3882	0.4729	1.5946	80.3	0.2	0.4	0.7	7.6	3.7	0.6	3.2	17.2
Jijeong B.	1,186.67	0.0537	1.9211	0.049	1.9781	59.7	0.2	0.5	1.1	6.2	4.7	1.1	4.7	27.07
Wonbu B.	519.53	0.7223	1.6889	1.0563	1.5802	82.9	0	0.3	1	3.7	6	0.7	4	24.54
Namhangang B.	10,947.38	0.0009	2.2668	0.0005	2.3932	70.7	4.8	0.5	0.9	7	3	1.2	3.1	21.75
Yeoju B.	11,114.18	0.0044	2.0392	0.0123	1.853	77.1	2.4	0.6	1.5	12.1	5.6	1.1	5.1	37.94
Yulgeuk B.	177.33	1.6772	1.7213	10.759	1.3184	89.8	0.2	0.4	0.9	3.8	3.8	0.8	3.4	22.77
Heungcheon B.	294.78	0.7853	1.7854	0.2471	2.0436	82.2	0.1	0.5	1.5	3.2	8.2	2	3.3	38.7
Ipo Weir upstream	11,774.88	0.0091	1.9562	0.0051	2.0853	40.5	14.6	0.4	1.2	13.7	8.9	0.8	6.7	30.61
Gyeongan B.	261.82	0.2261	2.0326	0.1218	2.2044	60.5	9.1	0.6	2.1	8.5	10.1	0.9	4.9	52.36
Hoeryong B.	1,514.28	0.1438	1.7813	0.4995	1.5273	75.2	2.5	0.7	1.3	1.9	3.2	1.1	2.1	32.13
Gimyong-ri	609.42	0.0612	1.6997	0.048	1.7594	54.3	3.1	1.6	9.9	27.7	21.3	0.8	5.4	250.94
Hwagye B.	177.23	2.3218	1.385	5.5478	1.2225	78.3	2.8	0.4	1	2.5	5.1	1.2	3.2	24.79
Bian B.	1,212.02	0.1461	1.7819	0.0891	2.1159	65.1	1.2	0.5	0.9	2.8	3.1	1	3.2	22.51
Museong-ri	472.69	0.0501	1.9671	0.1133	1.8683	73.6	2	0.8	4.4	21.5	21.9	0.5	6.8	110.29
Gimcheon B.	456.4	0.0794	1.9119	0.0745	2.0064	84.4	0	0.4	0.9	1.6	3.4	1.1	2.3	22.26
Seonju B.	987.52	0.3607	1.6316	0.7683	1.7826	68.8	2.7	0.5	1	1.6	3.7	1.3	2.2	24.28
Gumi B.	10,915.39	0.0108	1.7977	0.0001	2.5101	74.2	3.7	0.5	0.8	1.2	2.9	1.2	1.8	19.98
Hoguk B.	11,103.91	0.008	1.8303	0.0039	1.9219	71.5	3.6	0.4	1.4	11.3	11.1	0.6	6	34.15
Geumchang B.	926.93	0.0051	1.8773	0.3981	1.2967	48.3	9.9	3.9	12.8	23.3	9.6	2.5	3.3	323.03
Ansim B.	1,386.90	1.1932	1.1748	1.0177	1.2343	25.8	16	13.8	38.3	21.3	16	0.9	5.9	967.57
Gangchang B.	2,090.22	0.024	2.1127	0.0334	2.1284	51.4	9	0.5	0.8	1.3	2.8	1.1	1.8	19.73
Dojin B.	749.77	0.1396	1.7112	1.677	1.2872	75.3	0.5	0.5	0.8	1.2	2.7	1.2	1.7	20.74
Hwanggang B.	1,240.66	0.6958	1.3793	0.1611	1.6815	42.6	11.7	0.5	0.8	1.2	2.5	1.1	1.6	20.24
Jeokpo B.	16,433.12	0.0011	2.2325	0.0027	2.1234	69	0	0.3	0.4	0.6	2.2	1.1	1.5	9.11
Jeongam B.	2,990.66	0.2107	1.4763	0.0087	2.0131	64.5	7.3	0.3	0.4	0.9	2.4	1.1	2.1	9.87
Gyenae-ri	20,354.77	0.0044	1.9575	0.0027	1.9891	47.4	13.2	0.3	0.7	4.3	4.2	1.1	2.9	18.21
Singu B.	642.5	0.0751	1.9496	0.0118	2.556	68.8	0	1.9	12.5	28.4	26.1	0.3	6	315.95
Palgyeol B.	908	0.9223	1.635	1.1097	1.7519	80	1.1	0.9	1.9	7.6	4.5	0.8	3.8	47.81
Geumnam B.	6,946.30	0.0005	2.7478	0.0013	2.6217	75.6	0.3	0.4	1	3	4.2	0.9	3.3	25.04
Geumgang B.	7,213.30	0.0281	1.9442	0.0023	2.541	60.3	1.4	0.4	1.7	16.3	23	0.3	6.9	43.51
Jicheon B.	209	0.0411	2.1927	0.0053	2.1745	76.7	0.1	0.4	1	7.3	7.3	1.1	6.4	26.31
Baekjae B.	8,328.80	0.09	1.5481	0.0522	1.6805	79.5	0.9	0.2	0.5	1.1	4.5	1.1	2.9	12.65
Yongsan B.	442.58	0.0979	1.8422	0.0562	2.0517	84.1	0	0.5	1.9	13	13.6	0.9	7.3	49.07
Yuchon B.	103.47	3.7123	1.4493	3.1277	1.5521	30.4	12.4	0.6	3.8	19.1	19.2	0.6	7.3	95.62
Geukrak B.	683.5	0.073	1.8954	0.9305	1.8424	74.8	1.6	0.6	1.4	7.4	5.7	1	4.9	35.67
Jangrok B.	555.08	0.0876	1.8602	0.2719	1.7	77.6	0.2	0.3	3.2	17.2	41.5	0.7	9.9	80.19
Nampyeong B.	585.05	0.0602	1.8845	0.0997	1.7579	78.4	0	1	4.1	17.6	11.3	0.4	5.8	103.71
Naju B.	2,055.78	0.0308	1.9557	0.0207	2.0263	63.1	4.7	0.4	4.4	12	24.7	0.2	9.5	111.05
Donggang B.	2,599.85	0.5893	1.3265	0.4066	1.4097	89.5	0.2	0.5	2	5.4	13.3	1.3	4.6	49.58
Nakdan B.	9,399.97	-	-	-	-	78.09	0	0.66	2.22	13.61	9.99	0.69	5.83	56.16
Ilseon B.	9,532.76	0.003	2.0493	-	-	81.12	0	0.46	1.16	5.73	6.52	1.13	5.13	29.34
Oin B.	109.21	-	-	-	-	66.7	0.92	0.26	0.79	2.21	7.43	1.15	4.3	19.98
Gukjae B.	257.4	0.1793	1.8529	-	-	66.57	0	0.72	1.39	3.74	3.25	0.91	2.83	35.16
Pungyeongjeongcheon 2 B.	66.85	-	-	-	-	79.97	0	1.02	5.14	15.37	19.66	1.33	5.3	130.02

were distinguished, and the variables used in each case were highlighted in green.

In this study, cases with excessive clustering that resulted in similar spatial classification or did not converge in AIC+BIC values were excluded from the cluster analysis for 44 observation sites. As a result, six representative cases were identified and organized in Figure 5.2. SSPSD and BMPSD in Figure 5.2. are cases that use all class values from floating and bed sediment distributions, respectively, as input

Table 5.2 Viarable combinations of the clustering cases

									SSPSD (%)			BMPSD (%)						
Case	K	Lon	Lat	Cat. Area (m^2)	a_{SL}	b_{SL}	a_{TL}a	b_{TL}	all	0.062	0.062,2	all	d_{20}, d_{50}, d_{80}	C_u	C_g	σ_g	Gr	d_*
1	4																	
2	4																	
3	9																	
4	9																	
5	4																	
6	4																	
7	4																	
8	7																	
9	6																	
10	5																	
11	5																	
12	5																	
13	5																	
14	6																	
15	5																	
16	5																	
17	2																	
18	6																	
19	6																	
20	11																	
21	12																	
22	6																	
23	12																	
24	4																	
25	4																	
26	4																	

variables. Other cases that use only the representative size of floating and bed sediments are separately indicated. The Han River, Nakdong River, Geum River, Yeongsan River, and Seomjin River basins are distinguished in grayscale on the map, and the clustering results of the observation sites are shown in color on the map.

Figure 5.2-(a), -(b), and -(c) are the clustering results of observation sites with different floating and bed sediment distributions based on their geographic locations (latitude and longitude). These correspond to cases 3, 5, and 1, respectively, in Table 3. The case that uses only the location and bed sediment distribution shows similar results to the case that considers both floating and bed sediment distributions, which is similar

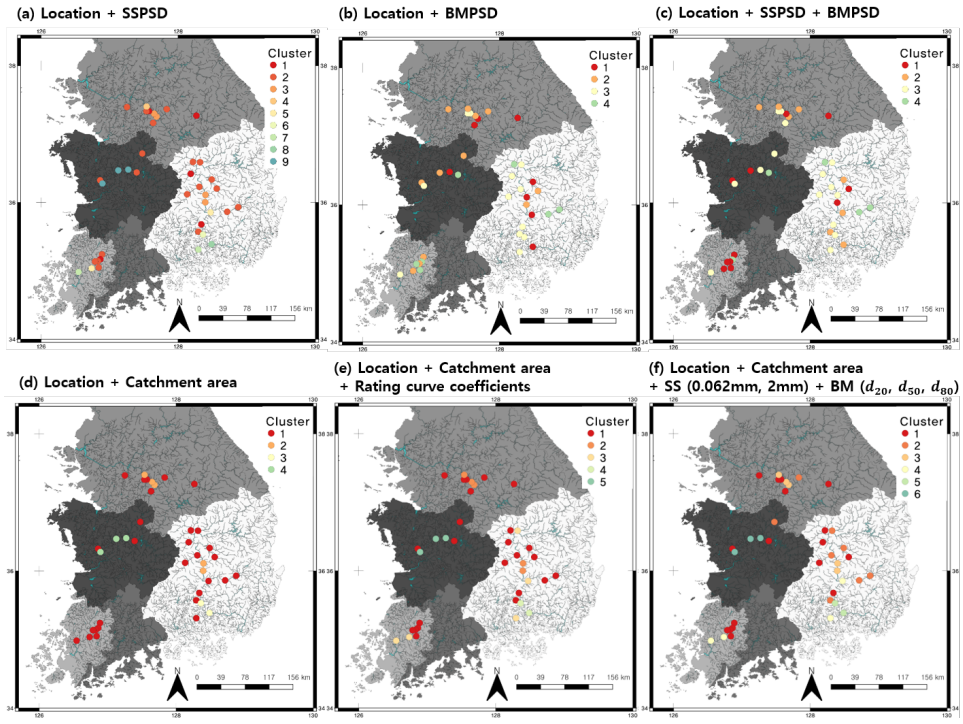


Figure 5.2 Representative clustering cases for sediment measurement stations

to the case that considers only the location and floating sediment distribution. This indicates that the bed sediment distribution has a more significant impact on clustering sediment monitoring sites among the 44 sites, clearly dividing the boundaries of the clusters compared to the floating sediment distribution.

Cases 7, 15, and 19 were compared to investigate the influence of the basin area in the remaining three cases (Figure 5.2-(d) (f)). All three cases classified The upstream tributaries into cluster 1 (red). When the basin area was considered, the three observation sites located in the Geum River mainstream were classified into independent clusters regardless of their location variables. These sites were also

classified into similar clusters in Figure 5.2-(a), indicating that the floating sediment characteristics of the Geum River basin have a similar pattern to the basin area.

Figures 5.2 (e) and (f) show the results of the analysis for the Yeongsan River and Nakdong River watersheds using two different approaches. The first approach involves incorporating a coefficient into the equation based on the catchment area, while the second approach considers the sediment size distribution of both suspended load and bed material. Both of these approaches provide more detailed results compared to the analysis based solely on the catchment area and location coordinates in Figure 5.2 (d), especially for the main channels of the Yeongsan River and Nakdong River watersheds. The second approach, which considers the sediment size distribution, results in a more pronounced differentiation in smaller tributaries than the first approach, which incorporates a coefficient based on the catchment area. For example, in the Han River watershed, the Seomgang Bridge monitoring station, which is located in a tributary of the Namhan River, and the Namhan River Bridge monitoring station, which is located in the main channel, are differentiated when considering both the catchment area and sediment size characteristics. In the Yeongsan River watershed, the observation station located near the Naju Bridge is classified into cluster 1, while the observation stations located upstream are differentiated from those located downstream.

Table 5.3 summarizes the statistical values of the monitoring stations assigned to each cluster in Figure 5.2(f) (Case 19). The results show that 17 monitoring stations

are assigned to cluster 1, with the smallest mean catchment area, while only 5 or fewer monitoring stations are assigned to clusters 3, 4, 5, and 6. The average catchment area of cluster 2 is intermediate between clusters 1 and 4, but the average d_{50} of cluster 2 is larger than that of cluster 1. The Seomgang Bridge monitoring station, which is assigned to cluster 2 in the Han River watershed, is located in a tributary composed of cobblestones, as reported by Lee et al. (2010). On the other hand, the Yulgeuk Bridge monitoring station, which is assigned to cluster 1 in the Han River watershed, is located in Yanghwa Creek, a smaller tributary with a smaller d_{50} composed of gravel and sand.

Table 5.3 Statistics summary of each cluster

Cluster		Stats.	1	2	3	4	5	6
Count			17	8	5	4	2	3
Location	Catch. Area	mean	439.641	1,170.37	11,171.15	2,434.13	18,393.95	7,496.13
		std	198.933	218.317	349.193	446.696	2,773.03	733.364
		min	103.47	908	10,915.39	2,055.78	16,433.12	6,946.30
		max	749.77	1,514.28	11,774.88	2,990.66	20,354.77	8,328.80
Rating curve Coeff.	a_{SL}	mean	0.6214	0.4401	0.0066	0.2137	0.0028	0.0395
		std	1.0247	0.4447	0.004	0.2649	0.0023	0.0458
		min	0.0411	0.0051	0.0009	0.024	0.0011	0.0005
		max	3.7123	1.1932	0.0108	0.5893	0.0044	0.09
	b_{SL}	mean	1.7862	1.6478	1.978	1.7178	2.095	2.08
		std	0.2224	0.2562	0.1884	0.3759	0.1945	0.6113
		min	1.385	1.1748	1.7977	1.3265	1.9575	1.5481
		max	2.1927	1.9211	2.2668	2.1127	2.2325	2.7478

Table 5.3 (continued)

Cluster		Stats.	1	2	3	4	5	6
	a_{TL}	mean	1.4483	0.5116	0.0044	0.1174	0.0027	0.0186
		std	2.8047	0.4154	0.0049	0.1931	0	0.0291
		min	0.0053	0.049	0.0001	0.0087	0.0027	0.0013
		max	10.759	1.1097	0.0123	0.4066	0.0027	0.0522
	b_{TL}	mean	1.7953	1.671	2.1527	1.8944	2.0563	2.2811
		std	0.3571	0.3077	0.2886	0.3272	0.095	0.5217
		min	1.2225	1.2343	1.853	1.4097	1.9891	1.6805
		max	2.556	2.1159	2.5101	2.1284	2.1234	2.6217
SSPSD (mm)	0.062	mean	73.67	58.18	66.81	67.15	58.19	71.77
		std	14.14	18.19	14.94	16.02	15.22	10.17
		min	30.41	25.79	40.46	51.44	47.43	60.25
		max	89.77	80	77.1	89.51	68.95	79.48
	2	mean	1.89	5.64	5.81	5.29	6.6	0.85
		std	3.52	5.99	4.97	3.81	9.29	0.57
		min	0	0.17	2.41	0.24	0.03	0.28
		max	12.37	15.99	14.57	9.01	13.17	1.41
	d_{20}	mean	0.65	2.65	0.49	0.4	0.28	0.34
		std	0.44	4.67	0.09	0.1	0.04	0.11
		min	0.32	0.45	0.43	0.25	0.25	0.22
		max	1.86	13.84	0.64	0.48	0.31	0.41
	d_{50}	mean	3	7.24	1.14	1.88	0.54	1.07
		std	3.35	13.18	0.31	1.8	0.25	0.61
		min	0.68	0.8	0.79	0.39	0.36	0.5
		max	12.49	38.25	1.5	4.39	0.72	1.72

Table 5.3 (continued)

Cluster		Stats.	1	2	3	4	5	6
	d_{80}	mean	11.25	8.24	9.03	4.9	2.45	6.79
		std	9.05	9	5.04	5.12	2.67	8.29
		min	1.15	1.15	1.17	0.94	0.56	1.09
		max	28.43	23.32	13.65	11.97	4.33	16.3
	C_u	mean	12.4	5.91	6.3	10.79	3.21	10.58
		std	10.5	4.63	3.61	10.55	1.4	10.72
		min	2.66	2.53	2.94	2.39	2.22	4.24
		max	41.5	15.98	11.07	24.67	4.2	22.96
	C_g	mean	0.85	1.2	0.94	0.93	1.07	0.76
		std	0.4	0.55	0.26	0.47	0.03	0.39
		min	0.26	0.75	0.56	0.23	1.05	0.34
		max	1.96	2.5	1.16	1.29	1.09	1.1
	σ_g	mean	5.03	3.33	4.53	4.48	2.2	4.39
		std	2.14	1.43	2.02	3.56	0.93	2.2
		min	1.73	1.64	1.81	1.75	1.54	2.94
		max	9.92	5.86	6.66	9.48	2.86	6.92
	d_*	mean	75.95	183.08	28.89	47.56	13.66	27.07
		std	84.7	333.37	7.79	45.57	6.44	15.53
		min	17.2	20.24	19.98	9.87	9.11	12.65
		max	315.95	967.57	37.94	111.05	18.21	43.51

Two downstream observation stations in Geum River were classified into cluster 6. According to Table 5.3, these stations have a moderate catchment area compared to other clusters. However, they were classified as independent clusters in

multiple cases due to their smallest sediment sizes compared to other clusters.

The Nakdong River basin is the most subdivided cluster, with clusters 4 and 5 only appearing in the Nakdong River basin. Cluster 5 consists of locations in the lower reaches of the Nakdong River mainstream, and Cluster 4 consists of upstream observation stations relatively located in the lower reaches. a_{SL} is a measure of how sensitive sediment load is to changes in flow rate and is used as a variable in this case. Cluster 5 had the lowest a_{SL} value, indicating a region with strong resistance to changes in sediment load. Although a_{SL} was not used as an analytical variable in this case, the regional characteristics of a_{SL} were revealed. This result was considered obvious considering that sediment load and grain size of bed material and catchment area are widely used variables in determining sediment transport characteristics. On the other hand, cluster 4 was distinguished from other clusters in that the bed material grain size was relatively smaller despite having a slightly larger average sediment size of 2 mm.

This result shows that observation stations with similar sediment loads and bed material characteristics are classified into different clusters depending on the river. Considering this, it is considered desirable to consider the overall similarity when selecting the clustering results. However, as small watershed observation stations with a catchment area of less than 749.77 m^2 accounted for the highest proportion in the clustering process, Figure 5.2-(f), which was the most subdivided case among the

cases considered the catchment area, was selected as the representative case.

5.4 Extended application strategy of H-ADCP-to-SSC models using the clustering result

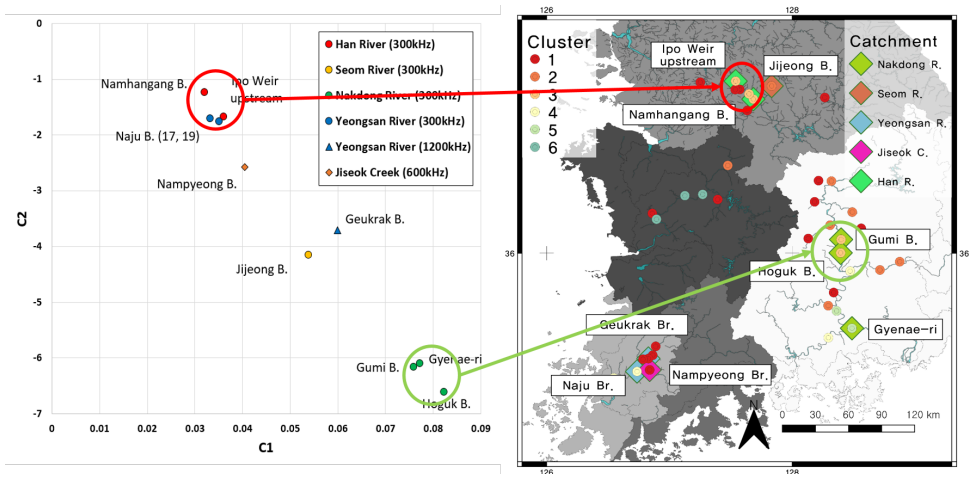


Figure 5.3 Spatial overlapping of the clustering result with the stations where the H-ADCP-SSC equations exist (left-hand side figure is originally from Figure 5.1)

Figure 5.3 shows the cluster results of Korean suspended sediment monitoring stations based on their location, watershed area, suspended sediment concentration, and bed sediment particle size distribution. The figure also includes the H-ADCP-SSC regression equations for each station, with a diamond marker indicating the location of each station. To aid in the analysis, Figure 5.1 is also included.

In the Han River basin, the Ipo Weir and Namhangang Bridge are clustered together in Cluster 3 due to their similar regression coefficients. On the other hand, the two stations located in the mainstream of the Han River and the Jijeong Bridge

station on the Seom River, which have slightly different coefficients, were successfully separated into Cluster 2.

As seen in Figure 5.3, when comparing the derived coefficients, the Naju Bridge in the Yeongsan River basin showed similar coefficients to the Ipo Weir and Namhangang Bridge in the Han River basin. While the Han River mainstream stations were classified into Cluster 3, Naju Bridge in the Yeongsan River basin was classified into Cluster 4. Although the coefficient values were similar, it was determined that they were classified into different clusters because they were geographically far apart.

The Gumi Bridge and Hoguk Bridge monitoring stations in the Nakdong River basin were clustered into Cluster 3, the same as the Ipo Weir upstream station and Namhangang Bridge station in the Han River basin. However, the Haman Gyenaeri station, which has a very similar H-ADCP-SSC regression equation to the Gumi Bridge station, was assigned to Cluster 5.

In the case of the Han River basin, it can be said that the clustering by location is effective within the same basin. However, it is inferred from the fact that even though they have similar H-ADCP-SSC regression equations, they were classified into different clusters in the Han River and Yeongsan River basins that the clustering results are relatively less significant between different basins. Therefore, as an alternative, it was proposed to use the coefficients of monitoring stations classified into the same cluster within the same basin in areas where the H-ADCP-SSC regression equation

has not been developed.

Based on the analysis described above, an alternative protocol for applying H-ADCP-based SSC monitoring methods at monitoring stations, where the H-ADCP-SSC regression equation has not been developed, was presented. This protocol can be applied by distinguishing between the 44 suspended sediment monitoring stations used in this study for clustering and those that were not. To explain the proposed protocol, Figure 5.4 shows the clustering results with Voronoi polygons indicating each cluster's influence range, and two examples are added. QGIS's Voronoi polygon function was used in this process.

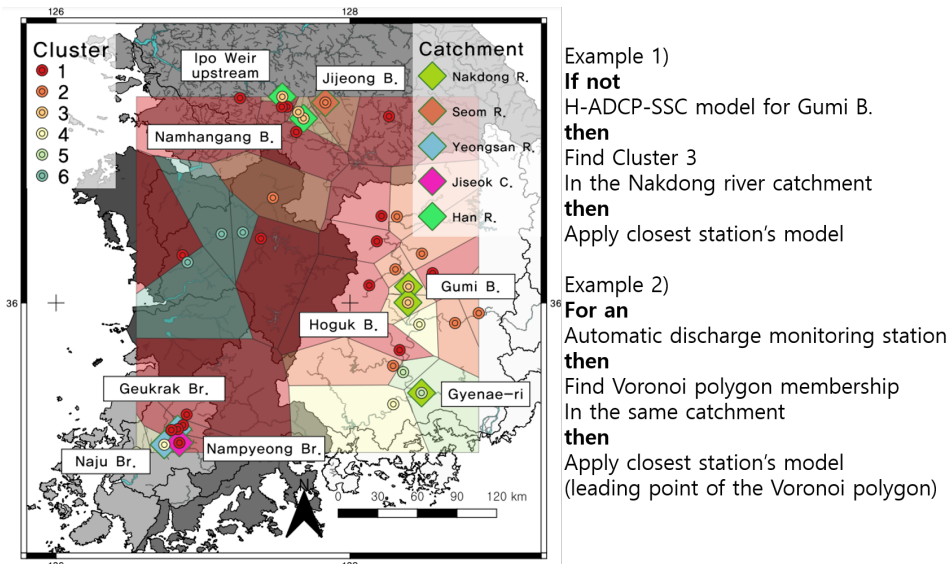


Figure 5.4 Example of the H-ADCP-SSC equation determination protocol

The first case is when applying the protocol to a similar quantity observation station where the H-ADCP-SSC relationship equation has not been developed. First,

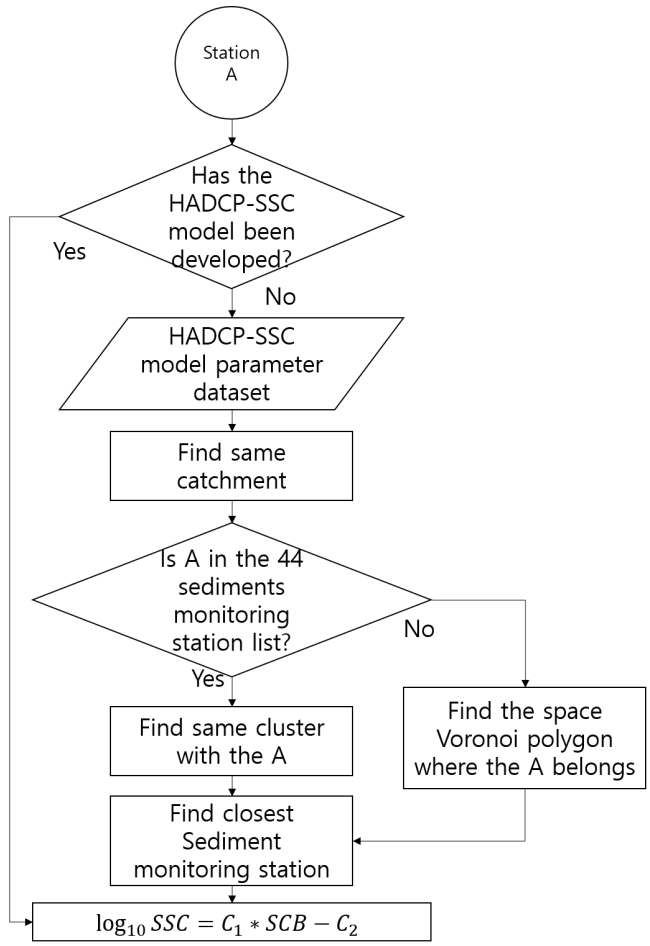


Figure 5.5 Flowchart of the H-ADCP-SSC equation determination protocol

select equations developed in the same region from the H-ADCP-SSC relationship equation dataset. Then, select the same clustering results in the same region and use the H-ADCP-SSC relationship equation of the observation station that is geographically closest among the observation stations classified into the same cluster in the same region. Example 1 in Figure 5.4 is when the Gumi Bridge observation station is used as an example of an undeveloped equation area. For the Gumi Bridge, selecting point

3 of the cluster in the Nakdong River basin selects the Hoguk Bridge and Gumi Bridge. Therefore, the floating bed measurement model developed at the Hoguk Bridge floating bed observation station can be applied.

The second case is when an alternative protocol is applied to an automatic flow monitoring station that is not a floating bed observation station targeted for clustering in this study. In this case, the process of selecting observation stations in the same region is the same as in the previous case. However, in this case, knowing which cluster the observation station belongs to is impossible. Instead of directly referring to the cluster number, if it is determined which Voronoi polygon area the target observation station is included in Figure 5.4, it can be determined which cluster the area belongs to. By selecting the same cluster using this method, the model of the closest observation station can be used instead, as in the first case. A flowchart is provided in Figure 5.5 to understand better the protocol described above.

To test the applicability of the extended relationship equation for H-ADCP-SSC, which includes a range of concentrations from 2.57 mg/l to 493.29 mg/l observed in 2019, the range of SCB was set between 80 dB and 118 dB to cover the observed range, and the H-ADCP SSC relationship equation for each of the three observation points, Hoguk Bridge, Jijeong Bridge, and Namhangang Bridge, was applied and the results were shown in Figure 5.6. As a result, the R^2 the Gumi Bridge observation point was 0.14, -0.39, and -0.44 for Hoguk Bridge, Jijeong Bridge, and Namhangang

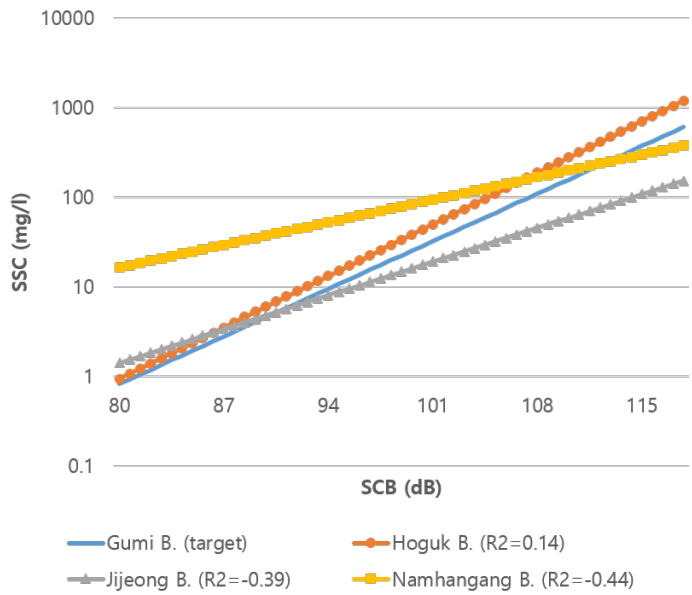


Figure 5.6 Estimated SSC graphs of the four tested H-ADCP-SSC models for a given SCB range.

Bridge, respectively. The relationship equation for all three observation points showed a low correlation coefficient of 0.2 or less with the Gumi Bridge equation. In addition, the RMSE and PBIAS were evaluated, and models with lower RMSE and PBIAS were considered to have lower errors. The RMSE values for the Hoguk Bridge, Jijeong Bridge, and Namhangang Bridge equations were 156.8 mg/l, 124.3 mg/l, and 62.13 mg/l, respectively, showing results opposite to the $R2$ values. On the other hand, the PBIAS for Hoguk Bridge was 56%, which was more than 5.6 times lower than the PBIAS for Namhangang Bridge with the lowest RMSE of 315%. In Figure 5.6, the suspended sediment concentrations at the junctions increase in the order of Hoguk Bridge, Jijeong Bridge, and Namhangang Bridge, and the root mean squared

error of the error increases significantly at higher suspended sediment concentrations. Therefore, although the RMSE for the Hoguk Bridge equation was the largest and the RMSE for the Namhangang Bridge observation point was the smallest, the overall applicability of the Namhangang Bridge equation is considered the poorest with an error ratio of 1,166% at low suspended sediment concentrations. Considering the $R2$, RMSE, and PBIAS, it is concluded that the Hoguk Bridge observation point is the most suitable alternative model for the Gumi Bridge observation point among the three observation points.

To explore the interrelations between not only the linear-1var model but also the SVR models, Table 5.4 is presented. This table compiles the R^2 when cross-applying the stations denoted in the second column on the left to the stations on the second rows. A greener shade indicates a higher performance. For example, row 1, column 1 means R^2 when measured SSC at the Geungnak Bridge station was estimated using the Geuknak Bridge model, and row 1, column 2 means R^2 calculated at the Geuknak Bridge station using the Gumi Bridge model. The diagonal components represent the accuracy of the refitted model, which may differ from Table 4.5, presenting the cross-validation score.

In most stations, the diagonal components in the stations where the model is derived show the fittest results. However, in the Gyenaeri station, the R^2 of the model derived from the Hoguk Bridge station was greater than the Gyenaeri station model.

The accuracy of the Gyenaeri model was insignificant in the Hoguk Bridge station, highlighting its inaccuracy due to the low correlation between SCB and SSC in that station. Also, in Naju Nampyeong Bridge station, observation points located in the same catchment exhibited high applicability to each other. Therefore, it is expected that incorporating additional catchment information in model determination will yield more robust results.

In the previously mentioned example monitoring station, Gumi Bridge station, the Nampyeong Bridge station located in a different watershed showed an R^2 of more than 0.5. In contrast, the Hoguk Bridge station revealed a score of 0.3 in the clustering approach. Looking solely at alternative application scores, the Naju Bridge and Nampyeong Bridge stations located in the Yeongsan River basin showed high cross-application scores compared to the Jijeong Bridge station located in the Seom River.

The Ipo Weir model shows the worst score, except when applied to the Ipo Weir upstream station. In that station, the Geukrak Bridge model was the second most accurate, even though the Geukrak Bridge is in a different basin.

Table 5.5 presents the RMSE values obtained by applying the Hoguk Bridge and Nampyeong Bridge models to the Gumi Bridge station for four rainfall events and a pre-rainfall period. The RMSE of each model was calculated using the Gumi Bridge time series estimation as the reference. In the pre-rainfall period, the Hoguk Bridge

Table 5.4 Coefficient of determination (R^2) by cross-application of SVR models using only SCB

		Alternative application							
		Geukrak	Gumi	Gyenaeri	Hoguk	Ipo	Jijeong	Naju	Nampyeong
Target	Geukrak	0.725	-0.856	-0.651	-0.553	-689273356.628	-0.858	-0.276	-3545.016
	Gumi	-0.475	0.988	0.157	0.304	-2093.399	0.398	0.347	0.538
	Gyenaeri	-2.221	0.544	0.770	0.786	-354.452	-0.668	-0.112	-0.558
	Hoguk	-2.371	0.455	0.626	0.745	-67.756	-0.407	0.200	-0.285
	Ipo	0.451	-0.270	-0.019	-0.028	0.783	-0.394	0.104	-0.188
	Jijeong	-1.942	-5.104	-2.094	-0.967	-107387.122	0.870	0.690	0.485
	Naju	-1.142	-3.234	-0.459	0.118	-24778.965	0.468	0.601	0.582
	Nampyeong	-1.013	-1.528	-0.109	0.241	-20134.557	0.699	0.599	0.834

model demonstrates an RMSE of 39.19 mg/l, representing a 73% improvement over the Nampyeong Bridge model. For Events 1 to 4, the Hoguk Bridge model consistently outperforms the Nampyeong Bridge model, showing RMSE values that are 88%, 72%, 95%, and 101% smaller, respectively.

Despite the higher R^2 values for the Nampyeong Bridge model shown in Table 5.5, it is essential to note that these results are influenced by the relatively higher accuracy of Event 4's high concentration. When considering the RMSE calculated using estimated SSC with 10-minute intervals, the difference is minimal at only 0.5%. This underscores the robustness of the Hoguk Bridge model, which remains superior. The real-time SSC estimation results affirm the effectiveness of the sediment ungauged station application strategy, favoring the nearest station in the same cluster, particularly the Hoguk Bridge model in this case.

Rivers undergo constant changes due to a variety of factors, including natural occurrences like floods and human-induced activities such as development. Recognizing the dynamic nature of natural rivers, it is essential to update the data used for

Table 5.5 RMSE values of applications of Hoguk Bridge and Nampyeong Bridge models to the Gumi Bridge station

	RMSE (mg/l)				
	Pre-rainfall	Event 1	Event 2	Event 3	Event 4
Hoguk to Gumi	39.19	144.68	112.62	3,582.57	90,939.86
Nampyeong to Gumi	147.52	164.13	156.31	3,759.03	90,428.03
RMSE ratio (-)	0.27	0.88	0.72	0.95	1.005

SCB-SSC model determination annually. The sediment load-flow rate relationship at the sediment monitoring site is revised yearly in the Annual Hydrological Report on Korea. It is crucial to note that the results presented in this study are based solely on data from 2019, introducing the possibility of errors in practical applications. However, given the minimal temporal variation observed in the coefficients of the H-ADCP-SSC relationship developed at the Naju Bridge in 2017 and 2019, the model is considered valid at least every two years.

Moreover, this study focused exclusively on the 44 sediment monitoring stations designated for suspended sediment observations in 2019. Applying these results to rivers in the Seomjin River Basin or along the East Coast, which are not part of the target data, may entail a high level of uncertainty.

Figure 5.5 suggests using the values from the nearest observed station within the cluster, which includes the target sediment monitoring station and demonstrates similarity in estimation when using a linear model at the Gumi Bridge station in Figure 5.6. However, when estimating SSC with SVR models at the Gumi Bridge station, Table 5.4 shows that the Nampyeong Bridge model has the highest R^2 for measured

SSC estimation. Nevertheless, in Table 5.5, the time series estimation similarity to the Gumi Bridge model for the Hoguk Bridge model, located in the same watershed, surpasses that of the Nampyeong Bridge model. Additionally, sediment transport exhibits strong locality, emphasizing the need to consider locational information. Therefore, when determining the model for sediment ungagged stations, it is advisable to follow the instructions in Figure 5.5 and prioritize a model from a monitoring station located in the same watershed.

Chapter 6. A novel efficient method of estimating suspended-to-total sediment load fraction in natural rivers

This chapter is partially reproduced from the following publication: Noh, Park & Seo (2023) *Water Resour. Res*. **59**: e2022WR034401. The material presented in this chapter has been modified and expanded from the original publication to align with the context and objectives outlined in this dissertation.

6.1 Dimensional analysis

First, dimensionless numbers were deduced based on Buckingham's Pi theorem to obtain reasonable dimensionless numbers for total sediment transport estimations. The dimensionless variables examined in a previous study (Tayfur et al., 2013) were additionally referred to and rearranged to avoid duplications. Table 6.1 compiles the dimensionless variables presented in this study, where ρ_s is the sediment density, respectively; W is the channel width; d_{84}, and d_{16} are the sediment particle sizes of the 84%, and 16% of the material by weight, respectively; and τ is the shear stress.

The selection of appropriate input variables requires extensive sediment transport observations and analyses. Table 6.2 lists the published empirical equations for estimating the total loads and the dimensionless parameters of the equations. In the table, C_w and C_{ppm} denote the total sediment concentration by the sediment weight

Table 6.1 Dimensionless variables related to sediment transport

Variables	Definitions	Variables	Definitions
$G_s = \frac{g\rho_s}{g\rho_w} = \frac{\gamma_s}{\gamma_w}$	Specific gravity	$\frac{W}{h}$	Channel width depth ratio
$\frac{U}{u_*} \approx \frac{U}{\sqrt{gR_hS_0}} \approx \frac{U}{\sqrt{ghS_0}}$	Friction factor	$\frac{US_0}{w_s}$	Dimensionless stream power
$Gr = \frac{1}{2}(\frac{d_{84}}{d_{50}} + \frac{d_{50}}{d_{16}})$	Gradation coefficient	$\sigma_g = (\frac{d_{84}}{d_{16}})^{1/2}$	The gradation of the sediment mixture
$d_* = d_{50}[\frac{g(G_s-1)}{\nu^2}]^{1/3}$	Dimensionless particle size	$\frac{R_h}{d_{50}} \approx \frac{h}{d_{50}}$	Dimensionless hydraulic radius
$Re_{d50} = \frac{Ud_{50}}{\nu}$	Particle Reynolds number	$Re_h = \frac{Uh}{\nu}$	Flow Reynolds number
$Re_* = \frac{u_*h}{\nu}$	Shear Reynolds number	$Re_{d*} = \frac{u_*d_{50}}{\nu}$	Particle shear Reynolds number
$Re_w = \frac{w_sd_{50}}{\nu}$	Falling particle Reynolds number	$Fr = \frac{U}{\sqrt{gh}}$	Froude number
$Fr_d = \frac{U}{\sqrt{g(G_s-1)d_{50}}}$	Densimetric Froude number	$Ro = \frac{w_s}{\beta\kappa u_*}$	Rouse number
$\tau_* = \frac{\tau}{g\rho_w(G_s-1)d_{50}} = \frac{u_*^2}{g(G_s-1)d_{50}}$	Shields number	$F_{sus} = \frac{Q_{SL}}{Q_{SL}+Q_{BL}}$	suspended-to-total sediment load fraction

per total weight and parts per million units, respectively. Harun et al. (2021) developed six equations by setting two variable sets with three machine learning models. The two multi-gene genetic-programming (MGGP) equations are presented as the representative models of the two variable sets.

In improvements of the modified Einstein procedure (Colby and Hembree, 1954; Shah-Fairbank et al., 2011; Shah-Fairbank and Julien, 2015; Yang and Julien, 2019), u_*/w_s and h/d_{50} were considered governing factors related to the suspended and total loads. For example, Shah-Fairbank et al. (2011) demonstrated that u_*/w_s and h/d_{50} are the major factors determining the ratio of suspended to total sediment discharge and that u_*/w_s is more influential than h/d_{50}.

Although a few variables in Table 6.1 do not appear in Table 6.2, the following analyses embrace all possible dimensionless variables. For example, W/h significantly influences the suspended to total load ratio (Edwards et al., 1999). W/h is a morphologically important factor resulting from stream bank stability, along with

Table 6.2 Empirical equations for total loads with dimensionless variables

References	Formulae	Dim.less parameters
References	Formulae	Dim.less parameters
Bagnold (1966)	$\frac{Q_{TL}}{W} = q_t = q_b + q_s = \frac{\tau_0 U}{G_s - 1}(e_B + \frac{0.01U}{w_s})$, where $0.2 < e_b < 0.3$	$C = f(\frac{U}{w_s})$
Engelund and Hansen (1967)	$C_w = 0.05(\frac{G_s}{G_s - 1})\frac{US_0}{\sqrt{(G_s-1)gd_{50}}}\frac{R_h S_0}{d_{50}(G_s-1)}$	$C = f(\frac{U}{\sqrt{g(G_s-1)d_{50}}}, S_0, \frac{R_h}{d_{50}})$
Shen and Hung (1972)	$\log C_{ppm} = [-107,404.459 + 324,214.747Sh$ $-326,309.589Sh^2 + 109,503.872Sh^3]$ where, $Sh = (\frac{US_0^{0.57159}}{w_s^{0.31988}})^{0.00750189}$	$C = f(\frac{US_0}{w_s})$
Ackers and White (1973)	$C_w = c_{AW2}G_s(\frac{d_{50}}{R_h})(\frac{U}{u_*})^{c_{AW1}}(\frac{c_{AW5}}{c_{AW3}} - 1)^{c_{AW4}}$ $c_{AW5} = \frac{u_*^{c_{AW1}}}{\sqrt{(G_s-1)gd_{50}}}(\frac{U}{\sqrt{32}\log(10h/d_{50})})^{1-c_{AW1}}$ for $1.0 < d_* \leq 60.0$ $c_{AW1} = 1.0 - 0.56\log d_*$ $c_{AW2} = 2.86\log d_* - (\log d_*)^2 - 3.53$ $c_{AW3} = \frac{0.23}{\sqrt{d_*}} + 0.14$ $c_{AW4} = \frac{9.66}{d_*} + 1.34$ for $d_* > 60.0$, $c_{AW1} = 0$, $c_{AW2} = 0.025$, $c_{AW3} = 0.17$,$c_{AW4} = 1.50$	$C = f(\frac{U}{u_*}, \frac{R_h}{d_{50}}, \frac{u_*}{\sqrt{(G_s-1)gd_{50}}}, d_*)$
Yang (1979)	for sand, $C_{ppm} = 5.435 - 0.286\log\frac{w_s d_{50}}{\nu} - 0.457\log\frac{u_*}{w_s}$ $+(1.799 - 0.409\log\frac{w_s d_{50}}{\nu} - 0.314\log\frac{u_*}{w_s})\log(\frac{US_0}{w_s} - \frac{U_{cr}S_0}{w_s})$ for $1.2 < \frac{u_* d_{50}}{\nu} < 70.0$, $\frac{U_{cr}}{w_s} = \frac{2.5}{\log(\frac{u_* d_{50}}{\nu}) - 0.06} + 0.66$ for $70 \leq \frac{u_* d_{50}}{\nu}$, $\frac{U_{cr}}{w_s} = 2.05$	$C = f(\frac{US_0}{w_s}, \frac{u_*}{w_s}, \frac{w_s d_{50}}{\nu}, \frac{u_* d_{50}}{\nu}, S_0)$
Karim (1998)	$\frac{q_t}{\sqrt{(G_s-1)d_{50}^3}} = 0.00139(\frac{U}{g\sqrt{(G_s-1)d_{50}}})^{2.97}(\frac{u_*}{w_s})^{1.47}$	$C = f(\frac{U}{g\sqrt{(G_s-1)d_{50}}}, \frac{u_*}{w_s})$
Molinas and Wu (2001)	$C_{ppm} = \frac{1430(0.86+\sqrt{\Psi})\Psi^{1.5}}{0.016+\Psi}$ where, $\Psi = \frac{U^3}{(G_s-1)ghw_s(\log(h/d_{50}))^2}$	$C = f(\frac{U}{u_*}, \frac{U}{w_s}, \frac{h}{d_{50}})$
Tayfur et al. (2013)	$C_{ppm} = [0.00075(\frac{u_* d_{50}}{\nu})^{2.5047}(\frac{1}{d_*^3})^{0.2117}(\frac{R_h}{d_{50}})^{1.2405}$ $(\frac{q_t}{\sqrt{(G_s-1)gd_{50}^3}})^{-0.3637}(\frac{u_*^2}{gd_{50}})^{0.7975}(\frac{U}{\sqrt{g(G_s-1)d_{50}}})^{0.9561}]$	$C = f(\frac{u_* d_{50}}{\nu}, d_*, \frac{R_h}{d_{50}}, \frac{q}{\sqrt{g(G_s-1)d_{50}^3}}, \frac{u_*^2}{gd_{50}}, \frac{U}{\sqrt{g(G_s-1)d_{50}}})$
Okcu et al. (2016)	$C_{ppm} = 34.45\frac{P^{3.239}J^{0.005}}{L^{0.066}R^{0.146}}$ where, $P = \frac{U}{\sqrt{(G_s-1)gd_{50}}}$ $J = \exp[(\ln S_0)^3]$ $L = \exp[(\ln(h/d_{50}))^2]$ $R = \frac{u_* d_{50}}{\nu}$	$C = f(\frac{U}{g\sqrt{(G_s-1)d_{50}}}, S_0, \frac{h}{d_{50}}, \frac{u_* d_{50}}{\nu})$
Harun et al. (2021)	$Q_{TL} = 25.8\frac{u_*}{U}(e^{\frac{u_*}{U}} - 0.869\frac{U^2}{gh}log(\frac{U^2}{gh})) - 200\frac{u_*}{U}$ $-4.2log(Q) - 6.15log(\frac{U^2}{gh})) - 0.787Q\frac{U^2}{gh})$ $-0.135Q + 1311Q(\frac{u_*}{U})^2\frac{U^2}{gh}) - 1.96$	$C = f(\frac{u_*}{U}, \frac{U^2}{gh})$
Harun et al. (2021)	$Q_{TL} = 0.953\frac{\sqrt{g(G_s-1)d_{50}}}{UR_h}(Q(Q + \frac{R_h}{d_{50}}) + Q^{\frac{\sqrt{g(G_s-1)d_{50}}}{UR_h}})$ $-10.7\frac{US_0}{w_s} - 0.0724Qlog(\frac{\sqrt{g(G_s-1)d_{50}}}{UR_h}) - 0.00157Q^2$ $+1.16Q^2log((\frac{\sqrt{g(G_s-1)d_{50}}}{UR_h})^{\frac{\sqrt{g(G_s-1)d_{50}}}{UR_h}})$ $-2000Q^2\frac{US_0}{w_s}\frac{\sqrt{g(G_s-1)d_{50}}}{UR_h}log(\frac{\sqrt{g(G_s-1)d_{50}}}{UR_h})$	$C = f(\frac{U}{g\sqrt{(G_s-1)d_{50}}}, \frac{R_h}{d_{50}}, \frac{US_0}{w_s})$

sinuosity and S_0 (Rosgen, 1994). Gr is also considered a particle size distribution indicator because of its apparent contributions (e.g., entrained suspended particle size; Van Rijn 1993).

6.2 Data

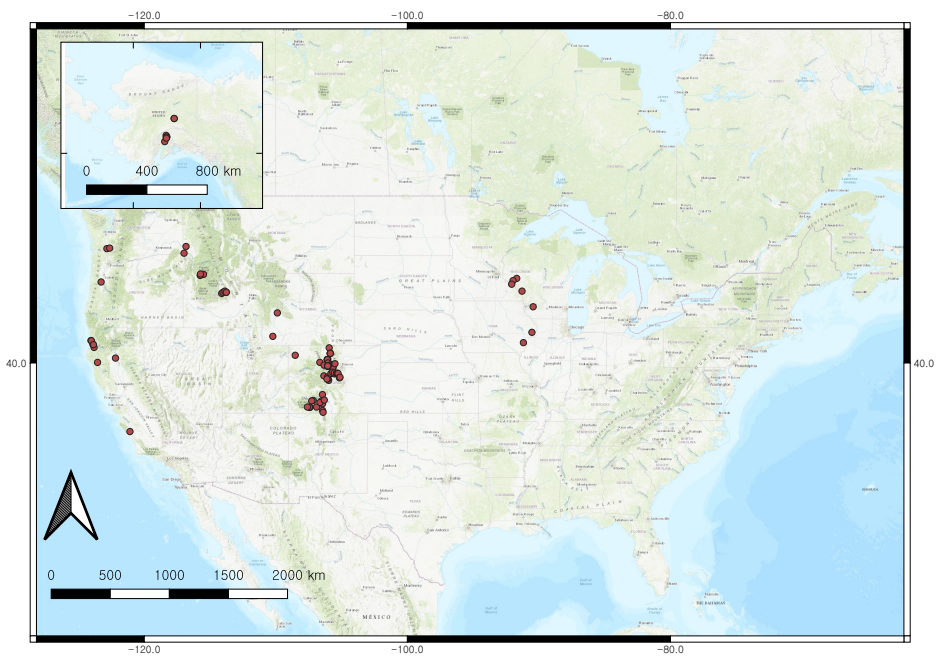


Figure 6.1 The sediment load measurement sites in (Williams and Rosgen, 1989). The measurement sites are marked with red dots.

The analyses in this study require not only the integrated total sediment loads but also the suspended and bed loads with hydraulic variables. However, in South Korea, direct total load measurement data in South Korea do not exist. Therfore, the target dataset includes data from the United States Geological Survey (USGS) report

on the measurement of suspended and bed loads in 93 natural rivers (Williams and Rosgen, 1989). Figure 6.1 displays the 93 measurement sites in Williams and Rosgen (1989): Colorado (54 sites), Alaska (9 sites), Idaho (9 sites), California (8 sites), Wisconsin (5 sites), Washington (3 sites), Iowa (2 sites), Wyoming (2 sites), and Oregon (1 site). All the locations were obtained from the USGS database, and descriptions were included in the paper. As shown in the figure, most sites are distributed in the western US.

The targeted dataset is a natural river sediment load monitoring dataset based on field sampling that includes sample analysis of both suspended and bed loads with hydraulic variable measurements. The input variables and calculated dimensionless numbers are summarized in Table 6.3.

The kinematic viscosity of water, $\nu = \mu/g$, was obtained based on the Vogel equation (Vogel, 1921), which is calculated as follows:

$$\mu = g\nu = \exp[-3.7188 + \frac{578.919}{-137.546 + T_K}], \tag{6.1}$$

where μ is the dynamic viscosity of water and T_K is the temperature in Kelvin. The coefficients from the above equation were obtained from the website of Dortmund Data Bank Software and Separation Technology (GmbH, nd).

The National Institute of Standards and Technology (Maryland, USA) adopts the model from Wagner and Pruß (2002) for density calculation, but it is known to

Table 6.3 Summary of the dataset (Nan rows excluded)

	Count	Mean	Std.	Min.	Max.
Q (cms)	1,957	2.26×10^{2}	5.15×10^{2}	7.00×10^{-3}	3.77×10^{3}
U (m/s)	1,721	1.05	6.41×10^{-1}	4.70×10^{-2}	3.40
W (m)	1,894	5.70×10^{1}	8.95×10^{1}	6.40×10^{-1}	5.18×10^{2}
h (m)	1,764	1.01	1.18	4.00×10^{-2}	5.80
S_0	650	7.39×10^{-3}	2.14×10^{-2}	9.30×10^{-5}	1.88×10^{-1}
u_* (m/s)	632	1.48×10^{-1}	8.51×10^{-2}	3.02×10^{-2}	6.37×10^{-1}
Temp. (°C)	1,026	9.92	5.19	5.00×10^{-1}	3.00×10^{1}
C_w (mg/l)	1,957	3.31×10^{2}	1.39×10^{3}	1.00	2.91×10^{4}
Q_{SL} (kg/s)	1,957	1.81×10^{2}	7.68×10^{2}	2.50×10^{-5}	1.41×10^{4}
Q_{BL} (kg/s)	1,928	7.75	2.32×10^{1}	3.20×10^{-7}	3.38×10^{2}
d_{16} (mm)	1,487	9.95	1.39×10^{1}	1.06×10^{-1}	9.04×10^{1}
d_{50} (mm)	1,530	3.77×10^{1}	4.07×10^{1}	2.78×10^{-1}	2.16×10^{2}
d_{65} (mm)	1,530	5.58×10^{1}	5.78×10^{1}	3.26×10^{-1}	2.89×10^{2}
d_{84} (mm)	1,530	9.85×10^{1}	1.02×10^{1}	4.25×10^{-1}	4.46×10^{2}
ν (m^2/s)	1,957	1.17×10^{-6}	2.00×10^{-7}	8.04×10^{-7}	1.71×10^{-6}
σ_g	1,487	5.23	4.66	1.46	2.37×10^{1}
Gr	1,487	8.09	1.12×10^{1}	1.46	5.99×10^{1}
F_{sus}	1,928	7.49×10^{-1}	2.69×10^{-1}	1.82×10^{-3}	1.00
W/h	1,755	4.74×10^{1}	5.63×10^{1}	3.03	6.32×10^{2}
H/d_{50}	1,409	3.59×10^{2}	1.10×10^{3}	5.10×10^{-1}	1.19×10^{4}
d_*	1,530	8.65×10^{2}	9.20×10^{2}	5.54	4.35×10^{3}
w_s	1,530	6.27×10^{-1}	3.86×10^{-1}	3.43×10^{-2}	1.76
US_0/w_s	389	1.03×10^{-2}	1.36×10^{-2}	9.20×10^{-5}	7.61×10^{-2}
U/u_*	589	9.58	4.57	2.06×10^{-1}	2.04×10^{1}
Re_h	1,720	1.35×10^{6}	2.21×10^{6}	6.16×10^{3}	1.60×10^{7}
Re_{d50}	1,366	2.96×10^{4}	3.12×10^{4}	1.33×10^{2}	2.05×10^{5}
Re_{d*}	431	5.66×10^{3}	1.02×10^{4}	1.05×10^{1}	6.07×10^{4}
Re_*	632	1.95×10^{5}	2.46×10^{5}	4.65×10^{3}	1.29×10^{6}
Re_w	1,530	3.31×10^{4}	5.13×10^{4}	6.69	2.70×10^{5}
Fr	1,720	3.97×10^{-1}	1.48×10^{-1}	3.00×10^{-2}	1.24
Fr_d	1,366	2.64	2.90	2.90×10^{-2}	2.39×10^{1}
U/w_s	1,366	3.05	3.85	3.08×10^{-2}	4.66×10^{1}
Ro	431	8.57	4.70	8.98×10^{-1}	2.33×10^{1}
τ_*	431	2.25×10^{-1}	4.35×10^{-1}	9.74×10^{-3}	4.07

be extremely complicated. Thus, all density-related variables were calculated using Equation (6.2), which was improved for both brevity and correctness (Civan, 2007).

$$\ln(1-\frac{\rho_w}{1065}) = 1.2538 - \frac{-1.4496\cdot 10^3}{T_C+175} + \frac{-1.2971\cdot 10^5}{(T_C+175)^2}(kg/m^3), \qquad (6.2)$$

where T_C is the temperature in Celsius.

When the falling velocity w_s and Rouse number Ro are estimated, the median suspended grain size d_{50ss} is considered the characteristic grain size, particularly in the MEP. To ensure the applicability of the proposed models, d_{50} was used, which can be readily obtained from databases such as Abeshu et al. (2022), instead of d_{50ss}. For example, in remote sensing using aerial images for SSC, obtaining d_{50ss} for every monitoring event may not be reasonable, and thus measuring d_{50}. In the characteristic size percentile, the median bed material size d_{50} is used if the particle size percentile for a dimensionless variable is not explicitly expressed. Similarly, the falling velocity w_s was calculated using the following equation:

$$w_s = \frac{8\nu}{d_{50}}[(1+0.0139d_*^3)^{1/2} - 1] \qquad (6.3)$$

The shear velocity u_* was calculated using the water surface slope by approximating $u_* \sim \sqrt{ghS_0}$.

Equation 6.3 indicates that the falling velocity of the suspended particles is

influenced by temperature because d_* depends on both the viscosity and density of water. If the temperature is greater than approximately 4 $°C$, both the density and viscosity decrease as the temperature increases. This results in an increase in ρ_s/ρ_w and a decrease in the viscous drag, which increases the falling velocity. Figure 6.2 shows the falling velocity changes owing to temperature and grain size variations. The y-axes in Figures 6.2(a) and (b) represent the dimensionless number $w_{s*} =$

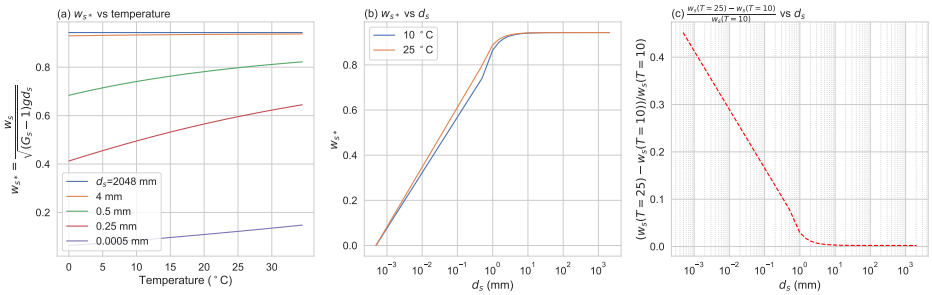


Figure 6.2 The temperature and grain size effects on the falling velocity: (a) w_s vs T; (b) w_s vs d_s; (c) $\frac{w_s(T=25)-w_s(T=10)}{w_s(T=25)}$ vs d_s.

$w_s/\sqrt{(G_s-1)gd_s}$, which is the ratio of the falling velocity computed by Equation 6.3 to the terminal velocity under buoyancy force. Figure 6.2(c) shows the acceleration rate of the falling velocity by changing the temperature from 10 $°C$ to 25 $°C$. It must be noted that the falling velocity of the figure may differ from that of a real-world phenomenon because the silt or clay particles are likely to flocculate (Julien, 2010).

As shown in Figures 6.2(a) and (b), the effect of increasing falling velocity is negligible when the grain size is larger than 4 mm. For larger particles (d_s >>4 mm), w_{s*} converges to 0.94. For particles smaller than 4 mm (fine gravel, sand, silt,

and clay), the viscous drag is discernible, accompanying the temperature effect. The temperature effect is apparent in the range $10^{-3} < d_s < 4mm$. The gap between the orange and blue lines is maximized for sand-sized particles. As shown in Figure 6.2(c), the actual falling velocity of particles larger than fine gravel is insensitive to temperature variations. By contrast, $\frac{w_s(T=25)-w_s(T=10)}{w_s(T=25)}$ continues to increase as d_s decreases. Although the ratio of the gravity force to w_s appears to be insensitive to the temperature variation for small particles, the viscosity change due to temperature affects the actual falling velocity. For extremely fine sand, $d_s \approx 10^{-2}$ mm, the falling velocity changes by approximately 30%.

Overall, the analysis implied that the temperature effect should be considered for sand, silt, and clay particles. The average value of d_{50} of the dataset is 3.76 mm, and the inflection point is observed in Figure 6.2. Therefore, the dimensionless variables related to ρ_w and ν, such as w_s, are computed using Equations 6.1 and 6.2, respectively, considering the temperature effect.

The flow rate Q was recorded on the entire dataset. However, other hydraulic variables have some missing data including the width and depth. In particular, S_0 is only given at 650 points, and thus the available number of shear velocity $u_* \sim \sqrt{ghS_0}$ reduces to 632. Results of the bed material sample analysis are recorded for each river. The data include the distance of each bed material sampling location from the sediment discharge monitoring location. The characteristic grain size of each river, d_{50}, was

obtained by analyzing the closest sample. Subsequently, the number of available data differs depending on the combination of input variables. For example, 371 out of 1,354 data contain u_*. Therefore, depending on whether u_* is included as a variable, the number of available data changes substantially.

6.3 Results

6.3.1 GRID-RFE-SVR

For SVR parameter determination, the kernels and other parameters were tuned, such as C_{SVR}, γ_{RBF}, and ϵ. Because the field sediment measurement data are accompanied by noise owing to various sources of uncertainties, it is important to reasonably determine noise regulation parameters (C_{SVR} and ϵ) for an acceptable prediction of F_{sus}. Considering noise and overfitting, the parameters were tuned by grid searching using a cross-validation (grid-CV) approach. Table 6.4 lists the hyperparameter nominee grid points.

The ϵ-insensitive SVR does not impose a fitting penalty on the data points within ϵ. Accordingly, the grid range of ϵ is $[2^{-6}, 2^3]$ that includes the possible maximum value of $10^{F_{sus}} = 10$. Additionally, 0.001 was added.

Table 6.4 Tested hyperparameter grid for the GRID-RFE-CV

Hyperparameters	Values
ϵ	$10^{-3}, \{2^i \vert i = [-6, 3] \text{ and } i \in \mathbf{I}\}$
C_{SVR}	$\{2^i \vert i = [-6, 10] \text{ and } i \in \mathbf{I}\}$
γ_{RBF}	$\{2^i \vert i = [-6, 10] \text{ and } i \in \mathbf{I}\}$

In each hyperparameter combination of the grid-CV sequence, RFE-SVR was additionally performed, hereafter referred to as GRID-RFE-CV. In this GRID-RFE-CV system, the user can determine the hyperparameter values and input variables of the model with a generalized capability supported by the cross-validation score.

All the dimensionless variables discussed in Section 6.1 were nominated to GRID-RFE-CV. To check the variable scaling effect of SVR fitting, the target variable F_{sus} and dimensionless input variables were scaled. In addition to F_{sus} without scaling, the scaling cases included the logarithmic scaling ($log(F_{sus})$) and power scaling ($10^{F_{sus}}$).

Table 6.5 presents the GRID-RFE-CV results for all the cases. The first and second numbers of the case names are distinguished by the input variables and F_{sus}, respectively. To compare the model performances, three criteria were evaluated, namely, the mean squared error (MSE), percent bias (PBIAS), and coefficient of determination R^2. The performance criteria in Table 6.5 can be defined as follows:

$$MSE = \frac{\sum_{i=1}^{n}(Y_{i,(obs)} - Y_{i,(est)})^2}{n}, \tag{6.4}$$

$$PBIAS = \frac{100}{n}\sum_{i=1}^{n}\frac{Y_{i,(est)} - Y_{i,(obs)}}{Y_{i,(obs)}}, \tag{6.5}$$

$$R^2 = \frac{\sum_{i=1}^{n}(Y_{i,(obs)} - Y_{i,(est)})^2}{\sum_{i=1}^{n}(Y_{i,(obs)} - \overline{Y_{(obs)}})^2}, \tag{6.6}$$

where $Y_{i,(obs)}$ and $Y_{i,(est)}$ are the observed and estimated values, respectively, and

$\overline{Y_{(obs)}}$ is the mean observed value. Both MSE and R^2 describe the erraticism of the model. The former reflects the scale of the error, whereas the latter focuses on model predictability compared to lumped mean prediction. PBIAS is a useful indicator of over or underestimation of signs (+ or -). In addition, PBIAS measures errors corresponding to each data, whereas MSE and R^2 provide data-lumped error information.

The performance criteria values define the best variable model from GRID-RFE-CV. Table 6.5 shows the average test score matrices in the 5-fold cross-validation step. For C12, C13, C22, and C23, the matrices were computed after transforming scaled variables back to F_{sus}, with $0 \leq F_{sus} \leq 1$.

Table 6.5 The condition of each case and cross-validation scores of the best model results from GRID-RFE-CV

Case	F_{sus}	Inputs	MSE	PBIAS	R^2	Selected variables
C11	F_{sus}	X	0.037	47.8	0.538	W/h,d_*,Re_h,Fr_d,Re_w
C12	log(F_{sus})	X	0.039	33.3	0.505	W/h,d_*,Re_h,Fr_d,Re_w
C13	$10^{F_{sus}}$	X	0.042	38.7	0.472	US_0/w_s,U/u_*,Re_h,Re_w,Gr
C21	F_{sus}	log(X)	0.045	62.0	0.428	Re_h,Fr,Fr_d
C22	log(F_{sus})	log(X)	0.046	70.8	0.425	Re_h,Fr,Fr_d
C23	$10^{F_{sus}}$	log(X)	0.042	63.8	0.464	H/d_{50},Re_h,Fr_d

In the cases where the input variables are not scaled, all the performance criteria support C11. In particular, the R^2 of C11 is 0.538, which is the best in all the cases. Although the R^2 score of C11 is superior to C12 and C13, the PBIAS of C12 and C1 are better than that of C12. Thus, C11 is taken to be the best case among the cases without input-variable scaling.

C21 in F_{sus} exhibits the lowest PBIAS for no scaling, and the $log(F_{sus})$ scaling case shows poor PBIAS and R^2 score. R^2 of C23 is slightly larger than that of the other cases.

Considering the four performance measures, deriving the SVR models without F_{sus} scaling is preferable. The surviving input variables differ depending on whether the input variables are scaled. but they are independent of the F_{sus} scaling. The effective input variables are revealed from the frequencies of the surviving variables, as presented in Table 6.5. W/h, d_*, Re_h, Fr_d, and Re_w survived when the input variables were not scaled, whereas Re_h, Fr, and Fr_d survived for C21, C22, and C23. Notably, Re_h and Fr_d were the two most frequent features. Re_h survived in all of the cases, and Fr_d was excluded in C13.

Two different SVR models were selected based on GRID-RFE-CV analysis. The two SVR models use five and three surviving variables in C11 and C21, respectively. The names of the models are distinguished by the number of input variables, namely, SVR5 and SVR3. The optimal hyperparameter settings for the SVR models are set as follows: SVR3 [kernel: RBF, $C_{SVR} = 0.25$, $\gamma_{RBF} = 256$, $\epsilon = 0.0625$], and SVR5 [kernel: RBF, $C_{SVR} = 1$, $\gamma_{RBF} = 128$, $\epsilon = 0.0625$].

6.3.2 Explicit equations

Although crucial features for F_{sus} were identified by RFE-SVR with acceptable accuracy, the functional relationship remained hidden. The following subsection presents

how the input variables interact with the help of explicit expressions aided by symbolic regression. Cutting-edge machine-learning methods, MGGP and Operon, were used to identify the underlying sediment transport physics in F_{sus}. The analysis continues with clustering and sensitivity analyses. Note that all the corresponding input variables, such as Re_h and Fr_d, in the following explicit equations are post-processed values by the MinMaxScaler.

MGGP

Formulation using MGGP requires certain parameter settings. The parameters that can be tuned in MGGP consist of formula shape and genetic algorithm parameters. Determining the functional form depends on the mathematical operator used in MGGP. In addition to the arithmetic operations, exponential operators (power, $\tanh$, $\log$, and $\exp$) were included. A formula can be generated under the function set and formula size parameter (maximum gene number and tree depth)using the genetic algorithm parameters. Thus, the population size and generations must be sufficiently large to appropriately examine the functional structure to obtain reasonable results. However, increasing the population size and generation is not a solution. Essentially, genetic algorithms lose solution diversity, converging individual solutions to a certain form for one sequence. Therefore, in this step, the population using the number of runs was reset to 200. However, an increase in shuffling within the genetic algorithm operators (crossover, mutation, and replacement) results in a trade-off between population

diversity and the dismantling of the population. The determined MGGP parameter settings are presented in Table 6.6.

Table 6.6 MGGP parameter settings

Parameter	Settings
Mathematical operators	$+, -, \times, \div, \sqrt{}$, square, cube, exp, tanh, log, power
Population size	500
Number of generations	500
Runs	200
Maximum number of genes	4
Maximum tree depth	6
Tournament size	15
Elitism	0.15 of population
Crossover events	0.84
High-/low-level crossover	0.2 / 0.8
Mutation events	0.14
Sub-tree mutation	0.9
Replacing input terminal with another random terminal	0.05

MGGP provides Pareto optimal equations; thus, several optional equations can be selected as the final product. In this study, the best models with respect to the test set scores were chosen and compared. For the perceptibility of the explicit models, a few terms such as A_{M3} were included as separate expressions. The replaced symbols use A, B, C, D, and E with the subscripts denoting the symbolic regression method. For

example, $M3$ is the three-variable MGGP model and $O5$ is the five-variable Operon model.

The three-variable MGGP model (MGGP3) is as shown in Equations (6.7) – (6.8).

$$F_{sus} = 0.406\, e^{A_{M3}} - 1.97\, e^{-Re_h} - 0.779\, e^{{Fr_d}^2} + 0.779\, e^{-{Re_h}^3} + 1.45\, {Fr_d}^2 + 1.77 \tag{6.7}$$

$$A_{M3} = e^{-6\, Fr_d - 3\, Re_h} - Fr^2\, {Re_h}^3 \tag{6.8}$$

Fr appears in only once in Equation (6.8), with the accompanying Re_h. For Fr, F_{sus} decreases with an increase in Fr. In addition, Re_h with Fr appears to affect the scaling of Fr in the last term of Equation (6.11).

The MGGP5 model has a more complicated structure than MGGP3. Equations (6.9) – (6.11) are mathematical expressions for MGGP5.

$$F_{sus} = 0.365\, e^{A_{M5}} - 0.549\, d_* - 0.0521 (e^{B_{M5}} + Re_h + \sqrt{(\frac{W}{h})^{d_*}}) + 0.222\, \frac{W}{h}\, d_* + 0.708 \tag{6.9}$$

$$A_{M5} = \frac{e^{-\frac{\tanh(Re_h)}{Re_h + d_*}}}{\tanh\left((e^{-Re_w})^{Re_h\, d_*}\right)} \tag{6.10}$$

$$B_{M5} = 3\, e^{-Re_h} \tag{6.11}$$

In the above formulation, MGGP considers all five surviving variables (W/h, d_*,

Re_h, Fr_d, and Re_w). However, the resultant equation does not contain Fr_d, which is related to the grain size-flow interaction. Instead, d_* and Re_w are included. Notably, composite effects of W/h and d_* are observed.

Operon

The low computational cost and accuracy of Operon enable heuristic input parameter tuning with less effort compared to MGGP. Hence, in this study, the input parameters of Operon were determined by a grid search with cross-validation using multiple Operon runs. The test parameter grid was identical to that in a previous study (La Cava et al., 2021).

Operon3 (Equations 6.12 – 6.17) requires three variables but is the most complicated among the explicit formulations proposed in this study.

$$F_{sus} = \frac{1.012\ (2.616\,Re_h - 11.552\,Fr + A_{O3} - B_{O3} + C_{O3})}{\sqrt{(0.711\,Re_h - 11.392\,Fr + D_{O3})^2 + 1}} - 0.009 \quad (6.12)$$

$$A_{O3} = \frac{20.192\,Fr - 1.331}{\sqrt{(7.505\,Re_h - 0.567\,Fr + E_{O3} - 0.04)^2 + 1}} \quad (6.13)$$

$$E_{O3} = \frac{45.229\,Fr_d}{\sqrt{\frac{11.916304\,Fr^2}{387.893025\,{Re_h}^2 + 1} + 1}} \quad (6.14)$$

$$B_{O3} = \frac{(3.364\,Fr - 1.587)}{\sqrt{8330.395441\,{Re_h}^2 + 1}} \quad (6.15)$$

$$C_{O3} = (3421.821\,Fr_d + 0.005)\ (0.075\,Re_h + 0.004\,Fr + 0.005) \quad (6.16)$$

$$D_{O3} = (0.057\, Re_h + 0.015)\ (9.269\, Re_h + 3739.117\, Fr_d + 31.422) \tag{6.17}$$

The five-variable Operon model was produced using the following equations:

$$F_{sus} = 0.499\, \frac{W}{h} - A_{O5} - B_{O5} + 2.622 \tag{6.18}$$

$$A_{O5} = \frac{\left(2.878\, \frac{W}{h} + 1.345\, d_* + 2.235\, Fr_d\right)}{\sqrt{5670.843025\, {Re_h}^2 + 1}} \tag{6.19}$$

$$B_{O5} = \frac{\left(27.784\, Re_h - 0.657\, d_* - 2.446\, Fr_d + \frac{0.563}{\sqrt{38808.212\, {Re_w}^2 + 1}} + 1.331\right)}{\sqrt{288.388324\, {Re_h}^2 + 1}} \tag{6.20}$$

Operon5 uses five complete variable sets, including Fr_d, which are not included in MGGP5.

The formulations of MGGP3 and MGGP5 show dependence on $exp[Re_h]$, resulting in the potential for computational overhead. However, the equations derived using Operon consist of multi-fractional expressions.

Nonlinear least-squares local optimization coefficient tuning distinguishes Operon from the MGGP models. For example, some terms in MGGP models share coefficients (the third and fourth terms in Equation (6.7)). Each term in the Operon model has a particular fine-tuned coefficient value. This coefficient tuning increases the predictability but lengthens the equation. The above Operon models were additionally rearranged, and the coefficient values were truncated to the sixth decimal

place for simplicity.

6.3.3 Model performances

Table 6.7 5-fold cross-validation score of the empirical equations in estimation of F_{sus}.

	MSE	PBIAS	R^2
SVR3	0.045	62.0	0.428
SVR5	0.037	47.8	0.538
MGGP3	0.059	101.9	0.264
MGGP5	0.055	98.2	0.310
Operon3	0.045	56.8	0.441
Operon5	0.046	50.8	0.427

Table 6.7 shows the F_{sus} estimation performance of the derived models. Every proposed model may estimate a value outside of the range [0,1]. Because values with $F_{sus} > 1$ or negative values are physically incorrect, all estimated values over one are corrected to 1. The negative values are adjusted to 10^{-4} to prevent infinite total load values when $Q_{TL} = Q_{SL}/0 = \infty$. These physical limitations must be applied to practical applications of these models.

In terms of MSE, the two SVR-driven models were superior to other symbolic regression models. Operon3 and Operon5 were next in terms of performance.

A distinct result of PBIAS is the poor performance of SVR3, which has a smaller PBIAS than those of the Operon models. MGGP3 yielded the lowest absolute value of PBIAS, and MGGP 5 follows. All models overestimated F_{sus} more than 47%.

SVR5 showed excellent accuracy in terms of R^2 (0.538). R^2 values of Operon3 ranked second. Operon3 was superior in MSE, and R^2 to Operon5. The two MGGP-driven models showed low R^2 values for all the performance criteria compared to the other methods. MGGP5 showed better accuracy than MGGP3 in all criteria.

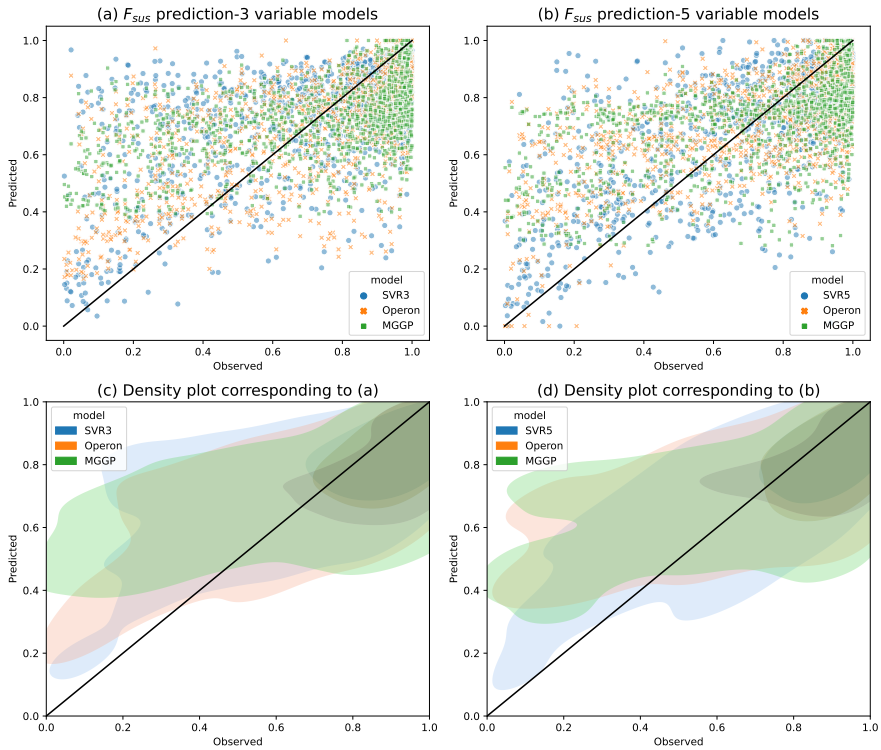


Figure 6.3 Scatter plots for F_{sus} estimation using all available data. (a) scatter plot of the three variable models; (b) scatter plot of the three variable models. (c–d) are the kernel density plots corresponding to (a–b).

Figure 6.3 shows the estimation results of the six models as scatter and density plots. The figures on the left-hand side are for the three-variable models, and those on the right-hand side are for the five-variable models; the symbols represent the

derivation methods. The black lines are the 1:1 lines of perfect estimations.

In the scatter plots, almost all markers are under the 1:1 line when F_{sus} is close to 1, while for low values, the markers are over the 1:1 line. All models appear to fit, centering approximately on the average of F_{sus}, 0.749. In addition, the overestimation of the lower values establishes the lower limit barriers in cases of Operon3, MGGP3, and MGGP5.

Additionally, two density plots were drawn for perceptibility. The two circles indicate the two density levels for each color, which are the same as those in the scatter plots. The closer to the 1:1 line and thinner, the more accurate the model is. Most F_{sus} observations are distributed in the range from 0.75 to 1, and the inner circles cover the range. Using the two distinguished circles, the performance at large and low values can be resolved.

As proven above, SVR5 exhibits the best performance among the proposed models, with the densest distribution around the 1:1 prediction line. In Figure 6.3 (c), Operon3 appears at a comparable level to SVR3, which is the best-performing three-variable model. Although SVR3 is the most accurate model for $F_{sus} < 0.75$ among the three-variable models, it presented underestimation for the larger F_{sus} range, as evidenced by the inner circle of the density plot. Contrary to the high predictability of Operon3, Operon5 does not perform well, covering a range similar to that of MGGP5.

In the performance and applicability evaluation of this work, all results of SVR-

based models are from test set estimation of 5-fold cross-validation. In Table 6.7, the Operon and MGGP models present full evaluation results rather than cross-validation because the symbolic regression model was derived without cross-validation. Because refitting models with optimal hyperparameters using the entire dataset is practically recommended (Hastie et al., 2009), the practical applicability of the SVR models in the following performance assessments can be underestimated. For example, R^2 of the refitted SVR3 and SVR5 models based on the full dataset are 0.648 and 0.742, respectively.

6.4 Discussion

6.4.1 Regional applicability of the models

Table 6.8 F_{sus} estimation performance and mean F_{sus} for each geographical location of the entire data in Williams and Rosgen (1989). The cells were green colored for high scores and red colored for low scores.

State	Data size	Mean F_{sus}	R^2					
			SVR3	SVR5	MGGP3	MGGP5	Operon3	Operon5
Alaska	265	0.919	-0.266	-0.111	-0.548	-0.366	0.108	0.134
California	136	0.476	0.699	0.793	0.591	0.653	0.805	0.819
Colorado	982	0.747	0.320	0.514	0.189	0.238	0.383	0.350
Idaho	261	0.788	0.640	0.647	0.630	0.649	0.653	0.642
Iowa	20	0.909	-4.998	-1.472	-4.668	-6.638	-3.846	-2.675
Oregon	43	0.926	-0.442	-1.779	-3.398	-0.793	-0.620	-0.308
Washington	86	0.885	-0.435	-0.063	-0.981	-1.317	-0.780	-0.820
Wisconsin	89	0.497	0.079	0.384	0.312	0.451	0.342	0.349
Wyoming	46	0.377	0.095	0.170	-0.603	-0.513	-0.188	-0.013

To assess the regional applicability of the model, the R^2 scores of six models presented in this study were individually calculated based on the geographical location

included in the dataset in Table 6.8. In this table, to provide macroscopic information with a large dataset, geographical locations were divided by state. In Oregon streams, U is not available, so instead, U was estimated with $\frac{Q}{Wh}$ for estimating F_{sus}. The table includes the number of data in each region and region averaged F_{sus}.

The overall prediction performance of the SVR5 model was the best, and in the symbolic regression method, the two Operon models showed similar performance. MGGP3 was the least accurate, followed by MGGP5. With respect to states, the prediction performance of Iowa was the worst among all regions, followed by Oregon. However, in the case of SVR5, it was found that Oregon was the worst predicted state, and Iowa had a relatively high R^2, indicating a better fit. In all models, streams in California showed the best prediction performance. In particular, Operon5 showed the most accurate result with an R^2 of over 0.819 for California among all regions. Additionally, the streams in Wisconsin were the second-best fit for SVR5 with an R^2 of 0.9086, but R^2 of Wisconsin was smaller than that of Colorado and Idaho in the Operon models. The mean F_{sus} values of California and Wisconsin being close to 0.5, SVR5 well-predicts the cases in a range of $0.45 < F_{sus} < 0.75$. Although Colorado streams comprise more than 50% of the dataset, streams in Idaho, Wisconsin, and California, which have a relatively small dataset proportion, were better predicted. In the case of Washington, it was found to have a higher accuracy than in Alaska, only in the SVR5 model.

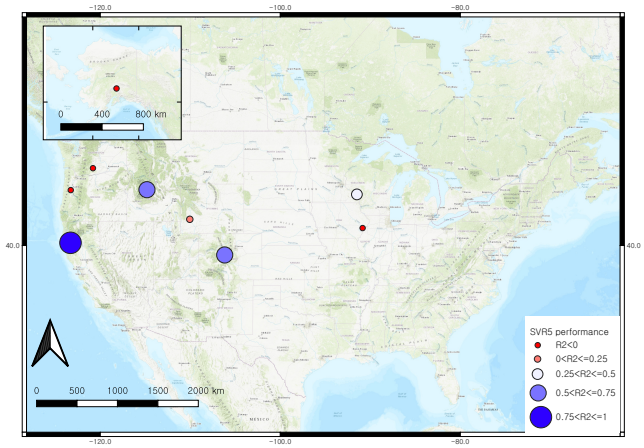


Figure 6.4 Gerographical projection of SVR5 model performance corresponding Table 6.8. The marker size increases in a order of $R^2 < 0, 0 < R^2 \leq 0.25, 0.25 < R^2 \leq 0.5$, $0.5 < R^2 \leq 0.75$, and $0.75 < R^2 \leq 1$, turning colors from red to blue.

Additionally, the geographical performance mapping of SVR5 was illustrated in Figure 6.4. The figure shows that the regional specificity of the SVR5 model is not clearly clustered with geographic locations. California and Colorado appeared to have R^2 values of 0.5 or greater, whereas $R^2 < 0$ for the adjacent region, Oregon. Iowa was the second-lowest region, and Wisconsin was the fourth-highest region whilst they are both in the Mid-west. Wyoming and Colorado showed an R^2 of 0.170 and 0.514, respectively.

Practically, it is common to refit the model with determined hyperparameters during the cross-validation step (Hastie et al., 2009). Consequently, SVR3 and SVR5 were refitted to the entire dataset using the optimal parameter and variable configurations. The geographical performance mapping results of the refitted models (SVR3 and SVR5) are presented in Table 6.9. The overall R^2 scores notably increased follow-

Table 6.9 F_{sus} estimation performance (R^2) of the refitted SVR3 and SVR5 model for entire data and each geographical location. The cells were green colored for high scores and red colored for low scores.

	SVR3-refitted	SVR5-refitted
Entire data	0.5296	0.7710
Alaska	-0.1670	0.6417
California	0.8461	0.9546
Colorado	0.4491	0.7189
Idaho	0.7862	0.8217
Iowa	-2.9354	0.2476
Oregon	-0.9163	-0.2302
Washington	0.0874	0.8301
Wisconsin	0.5783	0.9086
Wyoming	0.3599	0.4789

ing refitting for both models. Particularly noteworthy is the substantial improvement in estimation using the refitted SVR5 model in California and Wisconsin, demonstrating significant agreements with R^2 exceeding 0.9. The lowest recorded R^2 value is -0.2302, indicating challenges in estimating F_{sus} in Oregon state streams using the refitted models. Conversely, notable improvement is observed in Washington with an R^2 value of 0.8301. Colorado, Washington, and Idaho exhibit R^2 values of 0.7 or higher, while the adjacent region, Oregon, has an R^2 below 0. Iowa ranks as the second-worst region, and Wisconsin as the second-best, despite both being located in the Midwest. Wyoming and Colorado show R^2 values of 0.4789 and 0.7189, respectively.

One practical approach involves considering site-specific information, given the variability in local sediment transport and hydrologic characteristics. Incorporating site-specific information is anticipated to enhance predictability. For instance, models

can be derived for each category after classifying characteristic regions.

Several studies have clustered hydrological factors in the United States, including (Shinker, 2010; Berghuijs et al., 2014; Ho et al., 2017). They grouped Oregon and Washington as hydrologically homogeneous regions and Idaho, Wyoming, and Colorado as another cluster. However, Agarwal et al. (2016) combined some streams in Idaho, Oregon, Colorado, and Wyoming as one cluster. According to (Agarwal et al., 2016; Dettinger et al., 2011), precipitation in this cluster is concentrated during winter storms. In contrast, Shinker (2010); Agarwal et al. (2016) revealed that Wyoming and Colorado, which were better predicted than Oregon and Idaho, are relatively diverse regions with various combinations of clusters compared to Idaho. This is consistent with our results showing that the locality of Colorado and Wyoming is diverse, resulting in low accuracy. With regard to the Midwestern areas, the climate characteristics in Iowa and Wisconsin can be classified into different clusters (Ho et al., 2017; Berghuijs et al., 2014). In particular, Berghuijs et al. (2014) reported that the correlation between their past rainfall records was only 0.07. Notice that local hydrologic and sediment transport characteristics can vary. It is advised to employ site-specific information in building models whenever possible.

Despite achieving R^2 scores above 0.6 in Table 6.9, further accuracy improvement is still necessary. Consideration of sediment transport locality is crucial for practical application. For instance, the model performs well in streams with scores

higher than 0.8, while its applicability is limited in streams with scores below 0.5.

It is important to note that all the data used for model training were obtained from US streams. The applicability of the model to South Korean streams is not explicitly validated. For practical application in South Korea, the optimal approach would be to develop an estimation model using a dataset collected from South Korean streams.

Alternatively, by comparing the spatial estimation scores of US streams with the clustering results of South Korean streams, it is possible to identify clusters that are similar to those with high scores, facilitating a more appropriate application. It should be acknowledged that the conditions of upstream urbanization in South Korean streams may differ from those in US streams.

6.4.2 Clustering analysis

A clustering analysis was performed to investigate the relationships between the dimensionless variables by grouping data points of similar distributions. Before applying the clustering algorithm, the correlations between the derived dimensionless variables were inspected. Figure 6.5 presents a correlation heat map for the dimensionless variables. For F_{sus}, which is the key parameter of this study, six variables were filtered based on the condition that the absolute values of the Pearson correlation coefficient were greater than 0.5. The six selected variables that significantly correlate with F_{sus} are W/h, US_0/w_s, U/u_*, H/d_{50}, Re_h, and Fr_d, which are also marked

in the correlation map. The variables with a maximum-to-minimum ratio higher than 10^4 were analyzed on a logarithmic scale.

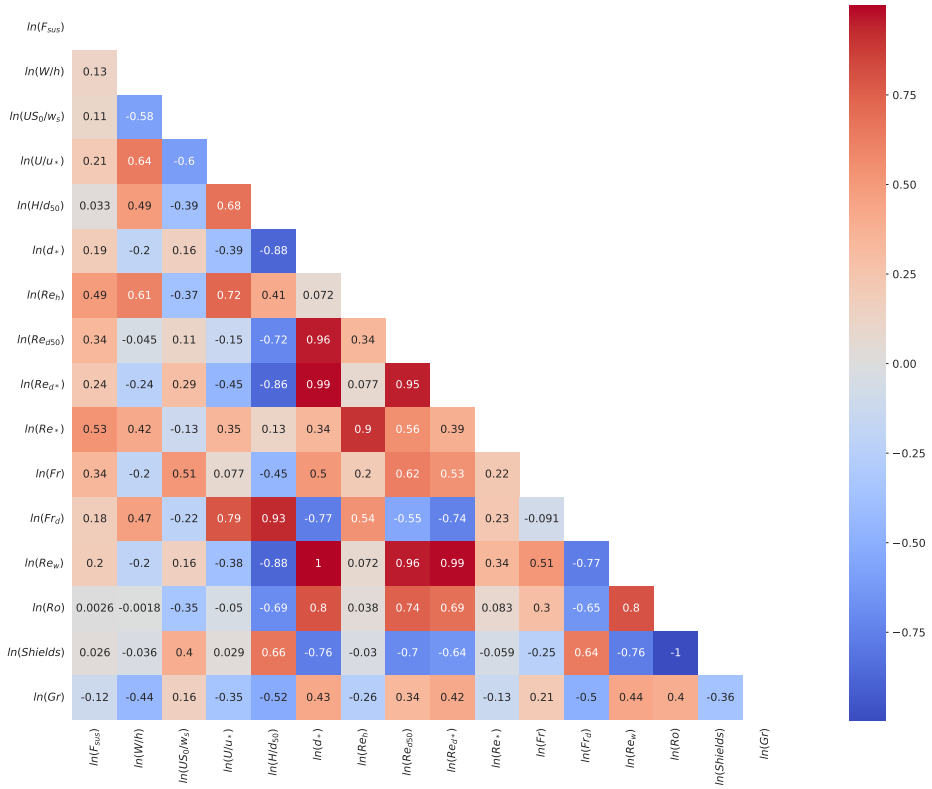


Figure 6.5 Correlation heat map for all dimensionless variables. The correlation coefficient values are written in the box, and colored with the corresponding color bar.

For the SOM analysis, the grid size was determined according to the relationship $p \times q = 5\sqrt{n}$ (Vesanto et al., 2000). The data length was 1,346, and the corresponding optimal SOM map size was calculated as $5\sqrt{1346} = 183.5$. Thus, the grid size of the SOM was set as $14 \times 13 = 182$.

The trained SOM map was additionally partitioned by GMM partitioned,

and the clustering case with the smallest $AIC + BIC$ score was selected as the final clustering result using an iterative method. that was similar to a method used previously in (Noh et al., 2021). The test range of the epochs of the SOM and the number of GMM clusters K were [0,1000] and [2, 10], respectively.

To optimize the SOM training, the training epoch was optimized, minimizing both QE and TE (Equations (2.55) and (2.56)). The QE-TE test results are shown in Figure 6.6. Both QE and TE rebounded after 300 epochs of the SOM update. GMM was performed after fixing the SOM to 250 epochs to ensure the lowest QE and TE.

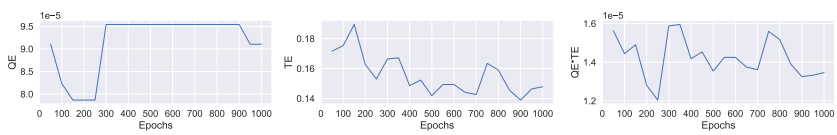


Figure 6.6 QE and TE epochs for the seven dimensionless variables [F_{sus}, W/h, d_*, Re_h, Fr, Fr_d, and Re_w]

The iterative GMM procedure is illustrated in Figure 6.7. The figure shows the minimum scores for each cluster. The minimal AIC+BIC value was 5. However, $K = 4$ was selected because the BIC increased when $K > 4$.

Two cluster plots were drawn to analyze the SOM-GMM results. Figure 6.9 shows a pair of scatter plots, and Figure 6.8 shows the corresponding SOM component planes.

Based on the frequency of the dimensionless variables, it is evident that Re_h and Fr_d are sufficiently informative to explain F_{sus} through the following inferences. Cheng et al. (2020) demonstrated that Fr_d is preferable to explain sediment en-

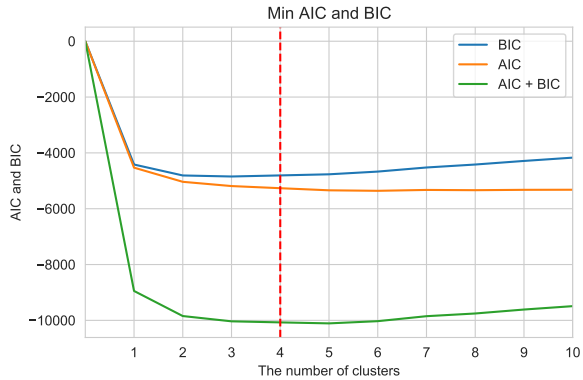


Figure 6.7 Minimum AIC+BIC values for each cluster number for the seven dimensionless variables [F_{sus}, W/h, d_*, Re_h, Fr, Fr_d, and Re_w]

trainment, and considering turbulent kinetic energy can enhance estimation accuracy. Furthermore, all of the dimensionless numbers, excluding the slope-related numbers u_* and S_0 with high uncertainties, can be approximated by combining Re_h and Fr_d. For example, $Re_h Fr_d = f(h/\sqrt{d_{50}})$, such that h/d_{50} can be expressed in a scaled manner.

With respect to physical inference, these two variables are related to suspended and bed loads. Fr_d is identical to the drag-bed friction balance, which can be expressed using Equation 6.21.

$$\frac{\text{Drag force}}{\text{Friction force}} = \frac{C_d \pi r_p^2 u^2}{\lambda_f N} = \frac{C_d \pi r_p^2 u^2}{\lambda_f g(G_s - 1)\pi \frac{4}{3} r_p^3} = f(\frac{u^2}{g(G_s - 1)r_p}) = f(Fr_d^2), \tag{6.21}$$

where C_d denotes the drag coefficient, r_p denotes the particle radius, u_p denotes the effective velocity of the particle, λ_f denotes the friction coefficient on the bed, and N

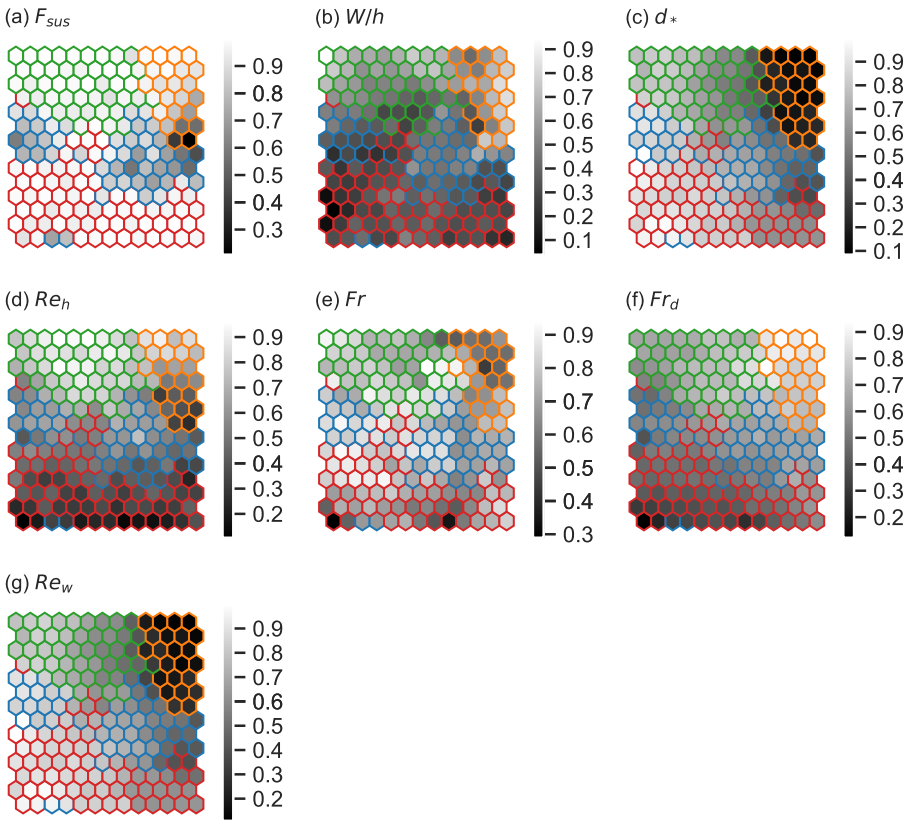


Figure 6.8 Component planes of the trained SOM grid: (a) F_{sus}; (b) W/h; (c) d_*; (d) Re_h; (e) Fr; (f) Fr_d; (g) Re_w. The grey scale face color denotes the values of variables. The determined clusters are differentiated by the edge colors of hexagons.

is the normal force. This interpretation of the initiation of particle motion aligns with the observations made by Aguirre-Pe et al. (2003). In another aspect, with respect to coastal or ocean environments, similar interpretations have been conveyed by Fischer et al. (2002) regarding the denominator of Equation (6.21) as a representation of the buoyancy force. In this respect, as shown in Table 6.2, Fr_d is considered as the main input variable in total load estimation formulas, especially in recent studies (Tayfur

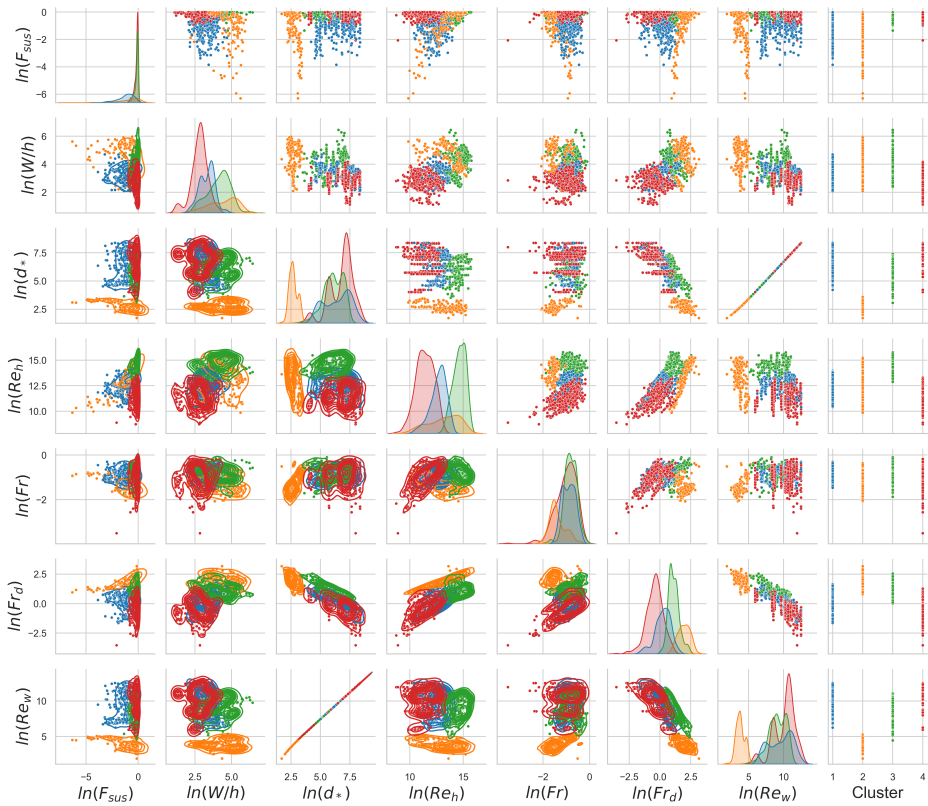


Figure 6.9 Pair scatter plots with kernel density plots for the seven dimensionless variables [F_{sus}, W/h, d_*, Re_h, Fr, Fr_d, and Re_w]. The colors of clusters were mapped into the dot and density contours with the same colors in Figure 6.8.

et al., 2013; Okcu et al., 2016). Fr_d has been highlighted as the main parameter along with d_{50}/h, the main parameter of MEP, in the bed load transport mechanism Hager (2018), sewer deposition problem (Safari and Mehr, 2018).

On the other hand, the Shields number (τ_*) is the most commonly considered parameter for sediment transport, describing the incipient motion of particles. However, it was not included in the optimal input variable sets in this study. This omission stems from the observation that incorporating τ_* did not significantly enhance the

estimation of F_{sus} compared to the contributions of W/h, d_*, Re_h, Fr, Fr_d, and Re_w.

Notably, the role of Fr_d, which considers flow force instead of shear stress (as in τ_*), was found to represent the sediment transport dynamics effectively. Since the work of Hager and Oliveto (2002), there has been a growing focus on analyzing bedload transport and sediment entrainment by incorporating Fr_d, surpassing the importance of τ_*. Several studies (Cheng and Emadzadeh, 2016; Sulaiman et al., 2017; Cheng et al., 2020; Wahl, 2023) have underscored the stronger correlation of Fr_d with sediment transport regimes. Particularly, Sulaiman et al. (2017) emphasized the continuous and robust relationship exhibited by Fr_d for both highland and lowland streams, whereas the Shields parameter and Einstein's exponential formula (Einstein, 1950) are applicable only to lowland streams dominated by suspended load.

Attempts to explain the higher correlation of Fr_d compared to the Shields number highlight that, fundamentally, the Shields number considers excessive total shear stress over the critical shear stress. In situations of weak sediment transport, where stress is very close to or lower than the critical shear stress, the explanation becomes challenging with the Shields number (Cheng, 2002). Another reason is that, in highland streams, the primary factor of total shear stress, a key component of the Shields stress, is significantly influenced by form drag (Pitlick et al., 2008), making it challenging to explain particle movement (Cheng, 2009).

The Reynolds number is known as the turbulence criterion. Thus, Re_h may contribute to increasing the turbulent diffusion, causing particles to remain in suspension. The imbalance of the drag force on a single particle and the friction between the particle and bed materials initiate incipient motions (e.g., sliding, saltating, etc.).

In the high Re_h region, F_{sus} approaches 1. In cases of sufficiently strong turbulence dispersion forces, bed loads in unmeasured areas of suspended samplers become suspended and disperse to the measurable area, corresponding to the suspended sediment region. Consequently, intense suspension allows suspended sediment loads to be approximated to the total sediment loads (as shown in Figure 6.9). Previous studies (Shen and Lemmin, 1999; Best, 2005; Hardy et al., 2009) have also reported that an increase in Re_h strengthens the coherent turbulent structures near the bottom (e.g., hairpin vortices and larger wakes), leading to particle movement. In a Rousean profile, the increase in shear stress associated with Re_h leads to a decrease in the Rouse number, resulting in a stronger contribution from suspended sediment.

From a different perspective, the drag coefficient C_d is commonly considered a function of Re_h (Van Nierop et al., 2007; Cheng, 2009; Wallwork et al., 2022). Upon revisiting the drag force equation in Equation (6.21), an interrelationship between Fr_d and Re_h becomes evident. Additionally, as highlighted by Brown and Lawler (2003); Cheng (2009), d_* can also contribute to C_d.

As observed from the structures of MGGP3 and Operon3, Fr, which Re_h

always accompanies, plays a role in scaling h. Furthermore, $Fr^2 = U^2/(gh)$ is the ratio of the flow energy head to the suspended sediment region. For $h = h_s + h_b$, where h_s and h_b represent the suspended sediment and bed load regions, respectively, h_b is constant owing to the sampler size, and thus, a variation in h indicates a variation in h_s. If the flow velocity is fixed, a decrease in Fr implies an increase in h_s, which in turn increases Q_{SL}. In terms of fixing the water depth h, laboratory experiments demonstrated that the suspended load contribution increases for larger Fr in dune migration dominated by bed loads (Naqshband et al., 2014). In Figures 6.8 and 6.9, the cover range of a low Fr decreases in the order of red, blue, and orange clusters for $12 < ln(Re_h) < 14$. For the same Re_h value, F_{sus} increases in the same order, thus supporting the above inference. Camenen et al. (2006) reported that roughness height can be predicted using Fr and dimensionless falling velocity. Shen et al. (1990) observed that the combined effect of skin resistance and form resistance constitutes the overall resistance to flow for modeled alluvial bed forms in situations of open channel flow with a Froude number less than 0.4. Fr controls the wavenumber and stability of bedforms, altering the form drag induced by bedforms in the equilibrium state and suppressing suspension (Fourriere et al., 2010). Subsequently, Fr can be interpreted as scaling of Re_h and bed roughness predictor.

d_* is a fundamental characteristic for sediment particles, defining the falling velocity of particles as indicated in Equation 6.3. Additionally, the well-known Shields

diagram illustrates incipient motion concerning τ_* and d_* for critical Shields stress. Several studies have classified bedforms based on d_* (for example, Julien and Raslan 1998; van Rijn 1984a,b). Ackers and White (1973) adopted d_* to consider sediment transport regime shift.

In both MGGP5 and Operon5 formulations, W/h accompanies d_*. Stewart (1983) reported that the fluvial channel, predominantly composed of suspended sediment, possessed features, such as silt/clay and steep bench/point bar, owing to a low W/h. In morphological transitions, streams with low W/h are likely to be eroded, and excessive deposition occurs in streams with high W/h (Rosgen, 1994, 2019). Another report (Edwards et al., 1999) describes the influence of W/h on F_{sus} and its temporal change. For fine bed materials, W/h can be reciprocal to C_w. According to a previous study (Xu, 2002), W/h can have a positive relation with C_w for low C_w, with the assumption that for a coarser grain, the flow is prone to be related to bed load. The low W/h coverage is smaller in the order of red, blue, orange, and green clusters for $ln(Re_h) < 12.5$. F_{sus} decreases in the order of the red, blue, and orange clusters. However, F_{sus} for the green cluster is the largest, despite the high W/h and d_*. As shown in the upper two rows of Figures 6.8 (b) and (c), the green cluster is characterized by a high Re_h. For large total loads, the Q_{TL} fraction becomes dominant, as depicted by the linearly increasing lower bound in the 1×4 plot in Figure 6.9. This suspended sediment-dominant flow of the green cluster was due to the excessively

large Re_h. The nonlinear relation between W/h and d_* in MGGP5 and Operon5 is valid for the calibration of the regime shift. The same interpretation can be applied to Re_w because its correlation to d_* is 1 and curved for low Re_w (the orange cluster).

6.4.3 Sensitivity analysis

This section presents the sensitivity of the models developed in this study obtained by changing the input variables. The sensitivity analysis was conducted on Operon3 and SVR5, the best explicit and implicit models, respectively. In addition, a sensitivity analysis was conducted on SVR3 to inspect the effect of a nonlinear complexity increase.

Figure 6.10 presents the one-at-a-time (OAT) sensitivity analysis results. The upper plots are spyder plots indicating the change in F_{sus} owing to a 50% variation in the input variables. The sensitivity index (SI) defined by Equation 6.22 is computed for quantitative comparison.

$$SI = \frac{max(F_{sus}) - min(F_{sus})}{max(F_{sus})} \tag{6.22}$$

For perceptibility, three-dimensional surface plots were drawn using the two influential variables Fr_d and Re_h.

To properly apply the Optimal Attribute Transform (OAT) method, multi-collinearity should be examined. This can be checked by evaluating the Variance

Inflation Factor (VIF), which is defined as the R^2 score when a specific variable is taken as the dependent variable, and the remaining variables are used to regress the model. In three-variable models, there is no multicollinearity between the features when the VIF is less than 2. However, in five-variable models, it is revealed that Re_w and d_* are correlated. Therefore, this analysis focuses on the three-variable models.

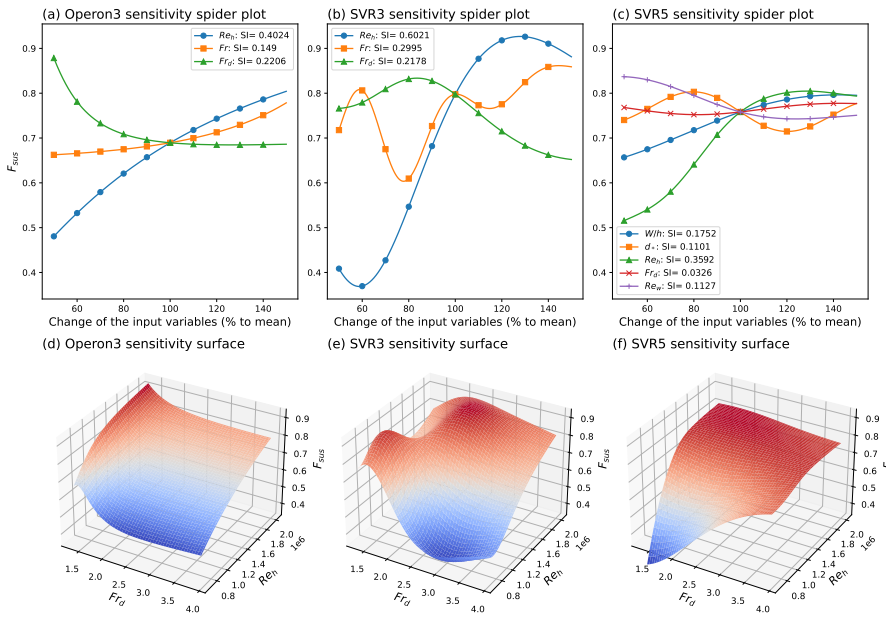


Figure 6.10 Spyder and three-dimensional surface plots for the three proposed algebraic equations: (a,d) tanh-type; (b,e) MGGP1; (c,f) MGGP2. The figure shows changes in the F_{sus} value as a function of specific variables. Different colors and markers are used to denote these changes in the spyder plots in (a, b, c). The surface grids in (d, e, f) represent the F_{sus} values obtained by combining Re_h and Fr_d.

The most sensitive variable in the case of Operon3 is Re_h (SI = 0.4024) in a positive relationship. Fr_d is reciprocal to F_{sus} and only half as influential as Re_h. Fr is the most insensitive variable with an SI value of 0.149 and an exponential-like

increment.

The effect of Re_h is prominent (SI = 0.5306). F_{sus} diminishes after a change of 120%. The increasing and decreasing behavior was observed for both Fr_d and Fr, but the fluctuation in Fr was exceptional. The fluctuation observed in Operon3 indicates a nonlinear relationship between the three variables.

In SVR5, the curve of Re_h that in SVR3. The SI associated with Re_h was the largest at 0.359. However, it was 1.67 times smaller than the maximum SI values obtained in the spyder plots of Operon3 and SVR3. This indicates the tuning effect of the two additional variables. d_* and Re_w demonstrated similar trends when increasing. For a negative change in d_*, F_{sus} drastically decreased with the local maximum point. Re_w, which represents the falling velocity, was negatively related to F_{sus}.

The proportionality of Re_h is clearly illustrated in the bottom row of Figure 6.10. For Operon3 and SVR3, the sensitivity of Fr_d is as high as Re_h is small. The surfaces of SVR3 and SVR5 have local maximum points. However, F_{sus} increases corresponding to Fr_d, as shown in Figure 6.10(f). This growth may be because SVR5 expresses the grain-size effect using not only Fr_d but also d_* and Re_w.

Chapter 7. Integrated sediment load assessment framework

7.1 Discussions on sediment load assessment

7.1.1 Simultaneous monitoring of total sediment load using MOSGO-SVR

Directly measuring bed loads in addition to the suspended sediment is the best way to obtain the total sediment load, Q_{TL}. However, bed load sampling shares similar, if not more, difficulties as in suspended sediment sampling. Instead, a common practice is to apply empirical models to estimate total loads. The most intuitive approach is applying a total sediment load model, such as Ackers and White (1973) and Yang (1979), but these direct estimation models do not contain information from suspended loads. Another method is to use the relationship between Q_{TL} with Q_{SL}. One popular method is the modified Einstein procedure (MEP) (Colby and Hembree, 1954; Son, 2021). However, the MEP is limited to sandy streams, and it also yields anomalous results, i.e., $Q_{TL} < Q_{SL}$, in some cases (Shah-Fairbank et al., 2011).

One may also utilize the suspended-to-total loads ratio, Q_{SL}/Q_{TL} (Turowski et al., 2010; Noh et al., 2023a). This method requires fewer parameters compared to MEP. Since Q_{SL}/Q_{TL} is in the range of [0,1], the resulting total loads are always equal to or larger than the suspended loads.

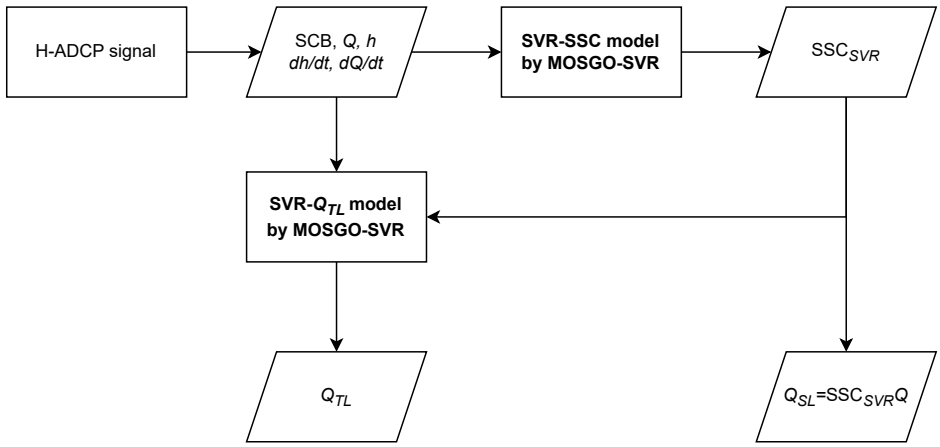


Figure 7.1 Examplar flowchart of real-time total sediment load monitoring system

If the total sediment load is measured or estimated at the monitoring station, the same approach can be applied to develop a model for estimating total loads as described in the previous section. Total loads are defined as the sum of suspended and bed loads. Therefore, the model developed using MOSGO-SVR to estimate SSC can be adapted for Q_{TL}. By nominating SSC as the input variable and using the variables obtained by H-ADCP, a framework can be established to simultaneously measure Q_{SL} and Q_{TL} in real-time. Figure 7.1 shows the real-time total sediment load monitoring system protocol when the suspended sediment sampling data is available. The protocol derives two SVR models, for SSC and Q_{TL}, utilizing MOSGO-SVR. Firstly, the SVR-SSC model is trained with information from H-ADCP. Next, the SVR model for total loads, Q_{TL}, is trained using the H-ADCP deduced variables and SSC$_{SVR}$. The flowchart can be applied not only in model derivation but also in estimation. However, the estimated sediment loads may lead to physically impossible

situations where the prediction model gives $Q_{TL} < Q_{SL}$. In such cases, it is necessary to adjust the total loads to be equal to the suspended loads.

In South Korea, using measured SSC, MEP is applied to estimate total loads. Then, the derived total loads are applied to develop rating curves for practical applications. In this study, using the MEP-estimated total loads as reference values, the framework in Figure 7.1 was applied in order to derive a real-time total load estimation model based on the results of Case 2. Figure 7.2 illustrates the total load estimation with cross-validation score for each station. The optimal variables combination frequently included that SSC estimated by SVR. The modeling accuracy showed that the CV score of SSC variables ranged from as low as 0.7, while that of Q_{TL} was above 0.9 for all models. This result indicates that the efficiency of sediment monitoring can be significantly enhanced by employing this real-time total loads monitoring framework.

7.1.2 Total sediment concentration estimation using F_{sus}

Overall, the analysis showed that SVR5 was the best model for estimating accuracy. In practical use, Operon3 shows promise, considering its explicit expression. However, the underestimation of PBIAS amplifies Q_{TL} in Operon3. By contrast, SVR5 is likely to underrate Q_{TL}. Based on these characteristics, SVR5 is considered suitable for users who want to determine F_{sus} correctly. Operon3 can be appropriately used for conservative river channel designs.

The practical use of F_{sus} involves the estimation of the total load Q_{TL} using

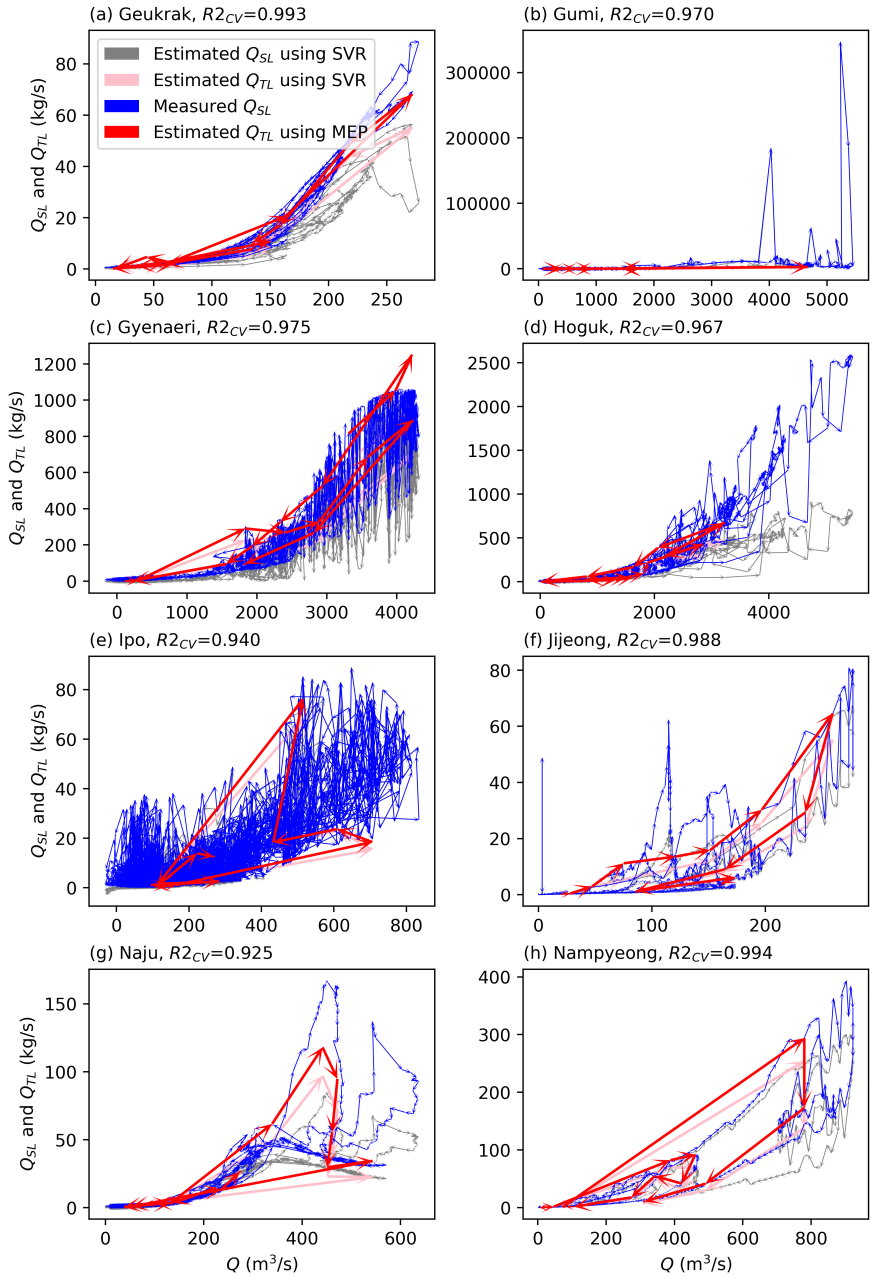


Figure 7.2 Application of MOSGO-SVR to the total load estimation procedure and cross-validation scores using MEP estimations.

the following relationship:

$$Q_{TL} = Q_{SL} + Q_{BL} = \frac{Q_{SL}}{F_{sus}} \tag{7.1}$$

in which Q_{SL} can be approximated to $Q \cdot C_w$.

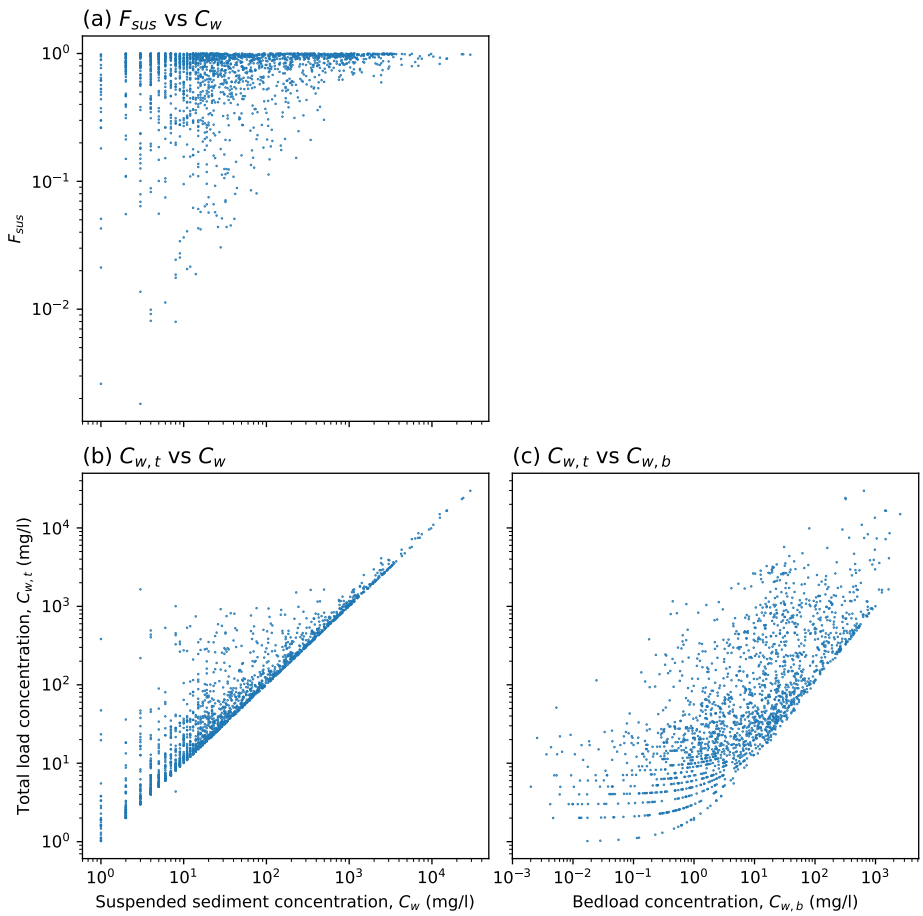


Figure 7.3 Scatter plots between sediment load concentrations: (a) F_{sus} vs suspended sediment concentration; (b) total load concentration vs suspended load concentration; and (c) bedload concentration vs total load concentration.

Figure 7.3 shows the relationships between F_{sus}, C_w, total load concentration ($C_{w,t}$), and bedload concentration ($C_{w,b}$). $C_{w,t}$ and $C_{w,b}$ were computed by dividing Q_{TL} and Q_{BL} with flow rate. Figure 7.3(b) shows that C_w is distributed along the 1:1 line. In the physical sense, C_w should be smaller or equal to $C_{w,t}$. For a highly tractive flow, water sweeps the bed material, resulting in rapid bed load transport. If the flow is sufficiently rapid to convey bed materials, there is also a high possibility of suspended sediment-governed flows that develop suspension. Thus, C_w can be approximated as $C_{w,t}$ even though a large amount of $C_{w,b}$ is transported. However, $C_{w,b}$ contributes more to a low C_w, as shown in the relationship between F_{sus} and C_w.

On the other hand, there is a point below the 1:1 line ($C_{w,t}<C_w$) in Figure 7.3(b). The case was Cross Creek near South Fork, Colorado State, measured on June 30, 1983. In that case, the reported C_w was 8 mg/l, but $Q_{SL}/Q = 4$ mg/l by recalculation, and thus an error is suspected.

Because C_w dominates over $C_{w,t}$, R^2 is equal to 0.985, where the R^2 value of $C_{w,b}$ is 0.053. However, estimating F_{sus} using only C_w is not recommended because the R^2 evaluation yields a value of -2.879·10^7. Despite the high R^2, estimating $C_{w,t}$ using F_{sus} is advantageous over using only Q_{SL} in a conservative design because an estimation using F_{sus} always yields $C_w, t \geq C_w$ with R^2 over 0.999.

The approximation values for Fsus approaching 0 were not evident in Figures 6.9 and 7.3. However, as shown in Figure 6.9, when Re_h is large, F_{sus} converges to

1. For $Re_h > 1.34 \cdot 10^7$, F_{sus} exceeded 0.87, and when $Re_h > 1.34 \cdot 10^7$, 75% of the data had $F_{sus} > 0.9$. In Figure 7.3, F_{sus} can be approximated to 1 as the suspended sediment concentration increases. When C_w is greater than 12,500 mg/l, the minimum value of F_{sus} was greater than 0.9. When C_w >500 mg/l, 75% of the data had F_{sus} exceeding 0.9. Subsequently, this study proposes a threshold of C_w >12,500 mg/l and $Re_h > 1.34 \cdot 10^7$ for suspended sediment dominant regime ($F_{sus} \approx 1$).

Other thresholds can be referred to in previous studies. For instance, in cases of large F_{sus}, Dade and Friend (1998) proposed a regime where u/w_s is less than 0.3 for suspended dominant flow and greater than 3 for bedload dominant flow. Yang and Julien (2019) used SEMEP to classify the suspended sediment dominant condition considering both C_w and u/w_s with the categorization of riverbed material by sand, gravel, and cobble.

MEP interprets that the nonlinear relationship between the Rouse number Ro and d_{50} governs F_{sus}. The Einstein integral contains the velocity profile information from the turbulent velocity profile, causing the ratio of suspended load to total load to vary with d_s, h, and Ro (Yang and Julien, 2019). u_* in Ro alternatively depends on g, h, and S_0. An issue arises when our equations do not contain u_* and d_{ss}, which are key factors for Ro. In contrast, Lara (1966) proved that Ro could be estimated using $Ro = Aw_{ss}^B$. Ro can be implicitly applied as a nonlinear expression of the explicit equations obtained in this study. Moreover, excluding u_* is beneficial for minimizing

uncertainty. In other words, the strict measurement of the slopes for u_* is challenging because natural streams have various bedforms and platforms.

Essentially, MEPs assume sand-bed streams. In this context, Shah-Fairbank et al. (2011) observed that applying different schemes for Ro regimes was favorable because of the applicability of MEP. The suggested empirical models are widely applicable using a previously published dataset (Williams and Rosgen, 1989), which covers bed material sizes ranging from sand (0.28 mm) to cobbles (216 mm).

Recently, river-monitoring techniques have been developed. The empirical models designed in this study can be implemented in recently developed flow-suspended sediment-monitoring techniques to estimate Q_{TL} because the required input variables can be obtained by these techniques. For example, at the river scale, drone-based remote-sensing techniques have been applied to SSCs (Kwon et al., 2022b,a; Gwon et al., 2023), bathymetry, and flows (Legleiter and Harrison, 2019; Legleiter and Kinzel, 2021; Eltner et al., 2020). ADCPs can be utilized to simultaneously measure flow and suspended sediment (Son et al., 2021; Noh et al., 2022). For bed grain-size estimation, one method is to use image-processing software packages, such as pyDGS (Buscombe, 2013) and Basegrain (Detert and Weitbrecht, 2012); however, sieving is the only reliable method that can be used for sand or finer grains (Harvey et al., 2022). If sieving is the only option, it is advantageous to create a dictionary of the median size of bed material on the probable areas before applying the

above methods. Safety and cost minimization can be achieved if the aforementioned monitoring technologies can be combined and applied appropriately.

7.1.3 Noisy behaviors of the MOSGO-SVR models

In Chapter 4, demonstrating the estimates of SSC for MOSGO-SVR models, noisy SSC curves were observed. This phenomenon may be attributed to the natural oscillations of flow variables (Q and h) or the step-shaped flow rate phenomenon, which arises due to the measurement resolution of H-ADCP being too long to capture temporal flow variations. This issue could amplify errors in the overall behavior. Therefore, to address errors arising from these two causes, both preprocessing and postprocessing steps can be considered.

In the preprocessing step, there are two approaches. The first is smoothing the time series of flow rate and water level. This can be implemented using techniques such as moving averages or mode decomposition. By doing so, when calculating instantaneous time derivatives, the step-like patterns due to the resolution of flow rate changes can be mitigated, retaining only large-scale variations. The second approach involves setting a window size when computing the time series flow rate from H-ADCP data, calculating the window-averaged time derivative instead of the instantaneous slope. While this method can reduce errors due to water level and flow rate resolution, choosing an appropriate window size is challenging for effectively controlling short-time-scale fluctuations. Postprocessing is similar to the first preprocessing method but

is applied to the estimated SSC time series. This approach is based on the assumption that the fluctuations in the input time derivatives are strictly trustworthy.

Investigating such preprocessing and postprocessing techniques can significantly contribute to analyzing physical data using time-series information. Nevertheless, as the primary objective of this study is the optimization of the estimation model rather than analysis through filtering, this aspect will not be further explored here and is left for future work. Instead, the influence of postprocessing on total load estimation will be briefly reviewed in Section 7.3.

7.2 The integrated sediment load assessment framework using hydro-acoustic backscatter

In the context of applying the SSC model derived from backscatters, limitations arise when considering the regional variations in flow and water level parameters. This prompts the need for a cautious approach, favoring models that align with the SCB's regional characteristics, especially when directly applying the MOSGO-SVR model to unmeasured observation points. Additionally, for estimating the total sediment load, it is proposed to leverage physics-based models when the observed sediment load is available.

For estimating the total sediment load, the framework suggests applying the MOSGO-SVR method when sediment load observations are available. In instances

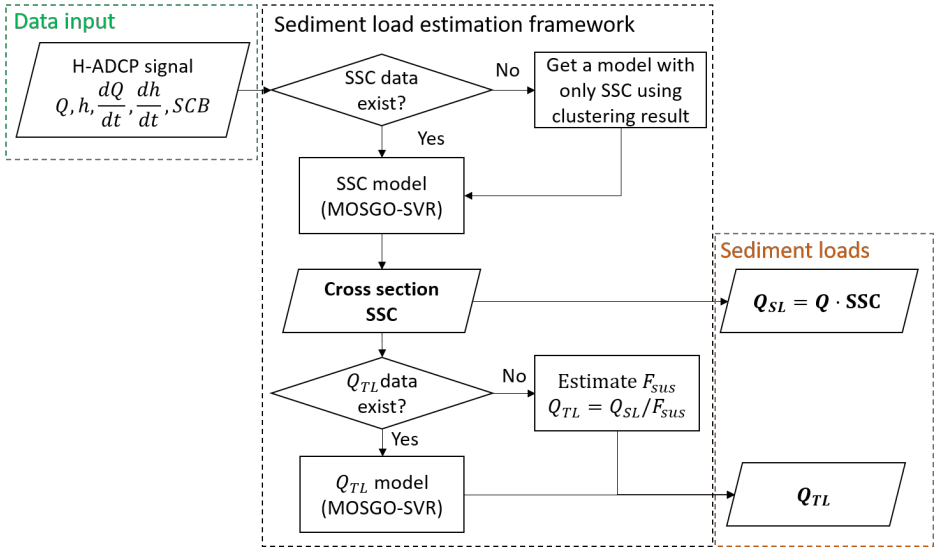


Figure 7.4 Flowchart of the integrated sediment load assessment framework

where only estimated total sediment load data is accessible, MEP for example, one can train a MOSGO-SVR model with estimated total load values.

In summary, the final sediment load assessment framework involves a systematic approach. Starting with H-ADCP input data, the framework assesses the availability of SSC observational data. If such data is available, the MOSGO-SVR method is deployed for SSC estimation. In cases where observational data is absent, cluster analysis outcomes are utilized to form models. For estimating total sediment load, models can also be derived if observational data exists, while the Fsus relationship formula is applied for estimations without total sediment load data. Figure 7.4 illustrates the integrated sediment load assessment framework.

Table 7.1 Integrated sediment load assessment framework test cases with different model combinations

Cases	SSC	Q_{TL}
Control	Measured SSC	MEP estimation
MM	MOSGO-SVR	MOSGO-SVR
MF	MOSGO-SVR	F_{sus} (refitted SVR5)
UF	USAS*	F_{sus} (refitted SVR5)

7.3 Instream applications

In this chapter, to demonstrate the effectiveness of the framework, the framework was applied to the estimation of SSC and Q_{TL} in various cases. The benchmark for comparing SSC values is the field-measured SSC. As for Q_{TL}, field-measured values are not available, and therefore, MEP estimations from the *Annual Hydrological Report of Korea* (MoE 2018; 2019) are used. For validation purposes, three cases were established and are organized in Table 7.1, being denoted as MM, MF, and UF, indicating the models employed for SSC and total load estimation. The MM case uses MOSGO-SVR for both SSC and total load estimation. In the second case, the SVR5 F_{sus} model is adopted for total load estimation instead of using MOSGO-SVR. In the UF Case, the strategy for ungauged station application and F_{sus} model are employed for SSC and total load estimation, respectively.

For all F_{sus} estimation, the SVR5 model refitted to the entire F_{sus} dataset was employed. The dimensionless hydraulic variables were calculated using the channel width and mean depth obtained from the integration of the cross-section below water

levels. The flow velocity was estimated by $U = Q/A$.

Similarly to the approach used for the Gumi Bridge station in Chapter 5, the framework was applied to this station. For the USAS strategy, the primary model was based on the Hoguk Bridge station, with an additional application of the Nampyeong Bridge station model. The coefficient of determination was utilized as the accuracy criterion. Table 7.2 presents the accuracies of SSC and Q_{TL} estimations for each case.

The MM Case demonstrates the most noteworthy accuracies in both SSC and Q_{TL}, surpassing 0.98. The MM Case, being a model directly trained on the MEP of the Gumi Bridge station, may exhibit higher accuracy in Q_{TL} than in SSC. However, uncertainties may be propagated for Q_{TL} estimation in MF and UF cases, where models are derived using SSC as input and extrapolated to other data. Despite the existence of error propagation and F_{sus} being derived from US streams, it reasonably estimated Q_{SL} in the Nakdong River with $R^2 = 0.891$. In UF Cases, the R^2 values for Q_{TL} estimation were 0.241 and 0.434 when using the Hoguk Bridge and Nampyeong Bridge models, respectively. In other words, the R^2 deterioration when using F_{sus} for Q_{TL} estimation is on the order of 0.063 to 0.1, suggesting that the performance of SSC estimation strongly influences the accuracy of total load estimation.

On the other hand, it is not accurate to claim that a model is more applicable simply because it exhibits higher accuracy. Note that the SVR models utilized for USAS employ only SCB as input variables. This is particularly true due to the lower

Table 7.2 SSC and Q_{TL} estimation accuracy (R^2) on the Gumi Bridge station for each case.

	R^2	
Case	SSC	Q_{TL}
MM	0.988	0.9996
MF	0.988	0.891
UF-Hoguk	0.304	0.241
UF-Nampyeong	0.538	0.434

correlation between SCB and SSC at the Hoguk Bridge station. Therefore, when implementing USAS, a more advanced approach, such as employing ensemble models with fuzzy clustering inference, could be considered to mitigate uncertainties arising from the SSC model itself.

In this section, comparisons were made with MEP estimation data. Since there are no actual measured true values, MEP was utilized for model validation, but it is applicable only to sandy rivers and remains an estimated value indefinitely. Therefore, it should be acknowledged that all the total load estimation results presented in this section may not represent true values.

The F_{sus} model was derived based on data that encompasses not only sand but also a diverse distribution of gravel, the following dataset by Williams (1989). Hence, in rivers where MEP does not perform well, especially those classified as gravel rivers with a d50 of more than 2 mm among the eight target observation stations in this study (Gyenaeri Bridge, Hoguk Bridge, Naju Bridge, Nampyeong Bridge), the applicability of F_{sus} may be relatively closer to the true values than MEP.

7.4 A brief guideline for the integrated sediment load assessment framework using hydro-acoustic backscatter

7.4.1 Scope of application

The monitoring of sediment transport is vital for effective river management. However, the costs and manpower requirements of conventional sediment monitoring methods impose limitations on expanding observation points and frequency. Therefore, the objective is to economically enhance sediment monitoring efficiency by utilizing fixed hydro-acoustic sensors, especially horizontal acoustic Doppler current profilers (H-ADCP), installed in automated flow monitoring stations for real-time sediment monitoring (e.g., Figure 7.6). Additionally, this approach encompasses the determination protocol for both suspended sediment concentration (SSC) and total sediment load (Q_{TL}).

The flowchart of the framework is presented in Figure 7.4. The following sections will provide a brief description of how to apply the framework.

7.4.2 Data and model sources

7.4.2.1 Sediment loads

Sediment loads encompass both suspended sediment load and bedload sediment load (Figure 7.5). In this framework, the suspended load is considered as the measured suspended load obtained using a suspended sediment sampler. The suspended load can

Figure 7.5 Sediment sampling photographs. (a) wading-type suspended sediment monitoring; (b) D-74 suspended sediment sampler; (c) bedload sampling.

be calculated by multiplying the flow rate and the suspended sediment concentration, which is estimated through sample analysis. The weight of bedload samples and the recorded time can be directly converted to bedload discharge.

Sediment sampling and sample analysis are advised to adhere to the guidelines outlined in Edwards et al. (1999) for obtaining cross-section averaged sediment load. Alternatively, in South Korea, the sediment dataset provided in the *Annual Hydrological Report of Korea* (MoE 2018; 2019) can be used for data collection.

7.4.2.2 Sediment corrected backscatter

SCB stands for sediment-corrected backscatter. Through the analysis of water-corrected backscatter derived from raw acoustic signals, SCB can be assessed, as explained in Chapter 2.2.1. As emphasized in Aleixo et al. (2020), SCB analysis should be carried

Figure 7.6 H-ADCP installation at a bank with sediment cloud passing

out using effective cells that fall within the sensible range without noise from the water surface or bottom.

7.4.2.3 Flow-Related Variables

H-ADCP can be installed at artificial structures or stream banks (Figure 7.6). Cell-wise flow velocity and water level are outputs from H-ADCP monitoring. As H-ADCP covers a limited area without mobility, the flow rate needs to be calculated using the index velocity method. In South Korea, official water level and flow rate data are provided by the government through the websites of flood control offices at the following URLs: https://www.hrfco.go.kr/ (Han River), https://www.nakdongriver.go.kr (Nakdong River), https://www.yeongsanriver.go.kr/ (Yeongsan River).

Additionally, cross-sectional shapes are available in the *Annual Hydrological Report of Korea* or on the flood control office websites. The integration of the cross-sectional shape, utilizing water level information, allows the evaluation of the cross-sectional area and channel width. Dividing the flow rate by the area yields the mean

hydraulic depth.

7.4.2.4 Suspended sediment concentration estimation

This sediment load assessment framework utilizes the MOdel Selection with Global Optimization for SVR (MOSGO-SVR), which optimizes SVR model hyperparameters and input variables in SSC estimation. The method is implemented in the Python library for Global Optimization and SHallow machine learning (pyGOSH). You can download the pyGOSH library from the following URL: https://github.com/hyoddubi1/pyGOSH

7.4.2.5 Total sediment load estimation

For total load estimation, the default setting is to use the MOSGO-SVR model if measured total load data are available. If not, the F_{sus} estimation models developed in this study can be used as an alternative. All the F_{sus} estimation models developed in this study are stored in the following GitHub repository: https://github.com/hyoddubi1/Fsus-sediment-fraction-models. Alternatively, explicit models can directly calculate F_{sus}. Then, the total load can be estimated by dividing suspended load with F_{sus}.

7.4.3 Practical Implementation

The list of required data for deriving the suspended sediment concentration (SSC) estimation model is as follows:

- H-ADCP raw signal
- Flow rate and water level
- SSC data

The list of required data for deriving the total sediment load estimation model is as follows:

- H-ADCP raw signal
- Flow rate and water level
- Estimated SSC data
- Total load data

7.4.3.1 Model derivation

Initially, the model derivation process is conducted utilizing the MOSGO-SVR method, which is implemented within the pyGOSH library. Below is an example of MOSGO-SVR model derivation in Python.

```
from pyGOSH.Utils import CheckMakeFolder
from pyGOSH.RFECVSVR import MOSGOSVR
from pyGOSH import GlobalOptimization as go
from sklearn.pipeline import Pipeline
from sklearn.svm import SVR
from sklearn.model_selection import KFold
```

```
from sklearn.preprocessing import StandardScaler

k_fold =  KFold(n_splits=4, random_state=42,  shuffle=True)
svr = SVR(kernel='rbf')
scaler = StandardScaler()
estimator = Pipeline([('scaler',scaler), ('regressor',svr)])
optimizer = go.Optimizer(lb = [10**-9,10**-9,10**-3],
        ub = [100000,10,1],
        algorithm = 'MCCE',
        stop_step = 200,
        stop_span = 10**-6,
        stop_obj_percent = 0.1,
        stop_fcal = 40000,
        dimension_restore = True,
        n_complex = 15,
        n_complex_size = 8,
        iseed = int(np.random.random()* 100),
        pop_init_method = 'LHS',
        init_pop = None,
        verbose = True,
        n_jobs = 1,
        pre_dispatch = '2*n_jobs',
        obj_eval = 'serial' )

mosgo = MOSGOSVR(X=None,
```

```
            estimator=estimator,
            cv = k_fold,
            optimizer=optimizer,
            verbose = True,
            X_log_flag = [True,True,True],
            yScale = 'log10',
            n_jobs= 4,
            fix_variables = [False, False, True,False, False ])
mosgo.fit(X,y)
```

Listing 7.1 Python example for five variable MOSGO-SVR execution with the variable keeping operation

The hyperparameter search space can be set as lb and ub variables. It is recommended that the MOSGO-SVR procedure be executed repeatedly for the robust model derivation. To apply the model and estimate the total sediment load, if there is total sediment load measurement data available for training, the user can also derive a total sediment load model using the code provided above.

7.4.3.2 Model application

To estimate SSC or total load using the models derived in Section 7.4.3.1, the user can input 'mosgo.predict(X)'. For the case where the total load needs to be estimated using the F_{sus} model, Python code is provided below.

```
from FsusModels.FsusModels import main, CsustSL, FsustoTL
```

```
#Manual input
    ============================================================
U = [0.5, 2]    # m/s
h = 0.5         # m
W = 2           # m
d50 = 1         # mm


Cunit = 'mg/L' # Unit of Csus
SLunit = 'kg/s'#

model = 'SVR'  # Empirical model derivation method ('SVR','Operon','
    MGGP','GPGOMEA','BPGP')-str
nv = 5          # Number of variables used to model derivation (3,5)-
    int
#============================================================
    Processing
U = np.array(U),h = np.array(h),W = np.array(W),d50 = np.array(d50),
    S0 = np.array(S0)

Q = U * W * h # cms

Csus = mosgo.predict(X)
Fsus = main(U,h,W = W, d50=d50,nu=10**-6, T=20, ustar = None, rhos =
     None,
```

```
      nv = nv,model = model)
SL = CsustoSL(Csus,Q, Cunit=Cunit, SLunit=SLunit)
TL = FsustoTL(Fsus,SL)
```

Listing 7.2 Python script for total load estimation using MOSGO-SVR and F_{sus} model (SVR5)

Here, the example hydraulic variables are entered manually, but the user can also load pre-measured and calculated hydraulic variables for input. Additionally, within the code, MOSGO-SVR may have loaded models for stations determined through the ungauged station application strategy instead of the models derived directly.

7.4.4 Limitations and recommendation

The F_{sus} estimation models were developed based on steady flow data and assume a bed material load. Similarly, empirical sediment transport equations rely on three key assumptions (Gomez and Church, 1989): 1) steady flow; 2) an algebraic relationship between hydraulic variables and similarity variables; 3) the maximum sediment load being transported under specific hydraulic conditions. Despite the disparities in assumptions, empirical bedload formulas are applied for unsteady time-series sediment yield assessments in practical applications (Li et al., 2021; Cohen et al., 2022). Moreover, while empirical bedload estimation models can be practically used following H-ADCP-driven SSC estimations, which encompass wash load and unsteady flow, it should be noted that unsteady flow bedload transport may deviate from estimates

provided by these equations.

If there is a hydraulic structure, the pattern of sediment transport and the velocity profile may change, potentially posing limitations in the models proposed in this work, i.e., F_{sus} models. However, since the scattering of sound waves follows the physical relationship between sediment concentration and the sound waves themselves, it is deemed applicable, irrespective of hydraulic conditions. Nevertheless, H-ADCP may introduce structural errors when estimating cross-sectional average sediment concentrations, leading to reduced accuracy in cases where flow patterns vary significantly.

The advantage of the ADCP method lies in its ability to observe both flow and sediment transport after installation simultaneously. While this makes it suitable for long-term, real-time observations, the inherent costliness of the equipment remains a challenge. In developing countries, where cost considerations weigh heavily, direct manual measurements of sediment transport might be a more cost-effective approach than employing ADCP. For situations where increased temporal resolution in measurements is required, especially if cost is a concern, one can use RGB cameras as an index of SSC rather than ADCP backscatter.

While this study contributes significantly to sediment monitoring, several areas for further research improvement have been identified in each chapter. It is noteworthy that these suggestions aim to guide future research toward achieving more efficient

and accurate sediment monitoring. The following paragraphs will briefly overview the identified areas for improvement and propose directions for future research to address current limitations.

As demonstrated in prior studies (Guerrero et al., 2016; Guerrero and Di Federico, 2018; Aleixo et al., 2020), the median particle size and the standard deviation of particle size distribution can characterize the backscattering features. In Chapter 4, although the importance of particle size distribution was highlighted, the particle size features were not considered in deriving the SSC model. While the real-time acquisition of particle size distribution is challenging in the field, incorporating particle size information along with flow features in the H-ADCP-based model derivation will contribute to the advancement of SSC monitoring theory. Once the relationship is explicitly established, particle sizes can be inversely modeled during H-ADCP monitoring.

This study initially classified sediment monitoring stations based on on-station values. However, sediment transport mechanisms in unsteady flows are influenced by upstream catchment land use or cover (Gellis, 2013). Additionally, considering factors like tributaries and stream order may be essential for understanding the hysteresis in SSC. These additional variables need careful consideration, especially given the theoretically dominant role of particle size distributions and the temporally varying nature of rivers.

As highlighted in Noh et al. (2023b), the H-ADCP-based sediment monitoring technique accurately analyzes fluvial sediment input. Furthermore, acoustic devices, such as an acoustic wave and current profiler (AWAC), are deployed for monitoring coastal currents. Real-time SSC monitoring in coastal environments can be effectively achieved by applying a similar method to the one presented in this study using moored acoustic devices.

In Chapter 7, the applicability in South Korean streams is not fully investigated since the training dataset is collected from the U.S. streams. A comparable approach can be employed if a substantial dataset of South Korean total load measurements is available for model derivation. While focusing solely on bed material size in F_{sus} estimation may offer practical advantages, it is advisable from a physical accuracy standpoint to consider suspended sediment particle size, as it holds significant information regarding the interaction between suspended and total loads.

Chapter 8. Summary and concluding remarks

This paper introduces methodologies to improve sediment transport's predictability and measurability using H-ADCPs and innovative machine learning techniques, specifically MOSGO-SVR and iterative SOM-GMM. Mainly, three contributions were made to enhance hydro-acoustic-based sediment transport estimation accuracy and applicability by optimized parameters: (1) enhancing the accuracy of H-ADCP-based sediment monitoring with high temporal resolution (10 minutes); (2) addressing the limitation of spatial applicability; and (3) extending real-time total load monitoring while providing insights into the suspended-to-total sediment load transport mechanism. The proposed procedures enable sediment data collection at higher spatiotemporal resolutions compared to current conventional techniques, including total sediment load estimation. Consequently, advancements in sediment transport management, both qualitatively and quantitatively, are anticipated. The contributions of this study are summarized in Table 8.1.

Chapter 3 presents machine learning model selection methods, employing iterative approaches and global optimization, particularly for SVR and clustering algorithms. It discusses the rational advantages of the newly presented machine learning methods. Note that all the methods were implemented in the pyGOSH library and shared in public.

In Chapter 4, the investigation into the accuracy improvement of the H-ADCP-based SSC monitoring technique was conducted by incorporating hydraulic variables alongside SCB. The results demonstrated that considering additional variables while maintaining SCB shows good accuracy. However, comparing estimation results with the SCB-only case is advisable. Moreover, the two SVR model selection methods, GRID-RFE-CV and MOSGO-SVR, were applied, with MOSGO-SVR yielding better accuracy than GRID-RFE-CV. Furthermore, a systematic connection of H-ADCP-based SSC monitoring to the total load, Modified Einstein Procedure estimations, was presented with reasonable predictability.

In Chapter 5, clustering analysis, using the iterative GMM, of sediment characteristics, including particle size distributions and flowrate-sediment load rating curve coefficients, in South Korean sediment monitoring stations was conducted. The clusters were primarily classified based on the catchment area and particle sizes. Given that the SSC-backscattering relationship depends on suspended sediment particle size distribution, a protocol for determining the SSC-backscattering relationship was introduced. The protocol establishes a surrogate model with the model in the closest station of the cluster where the target station is assigned.

Chapter 6's objective of this study is to present a globally applicable total estimation method using suspended load and hydraulic variables, whereas the total load model in Chapter 3 is locally applicable. As the target variable, the suspended-to-total

load fraction, F_{sus}, six machine learning models were developed. The importance of the feature was assessed through the GRID-RFE-CV step with SVR fitting. The flow Reynolds number (Re_h), densimetric Froude number (Fr_d), and Froude number (Fr) were identified as the three dominant parameters. In addition to SVR, explicit equations were derived using the two symbolic regression algorithms, MGGP and Operon. The iterative SOM-GMM protocol revealed the influences of hydraulic variables on F_{sus}.

Subsequently, Chapter 7 systematically incorporates the major results in Chapters 3–6 proposing a total sediment load estimation framework. Possible model combinations of the framework were tested and showed commendable predictability. From a more practical point of view, a brief guideline and related discussions were provided.

Table 8.1 Summary of this study

	Chapter 4	**Chapter 5**	**Chapter 6**
Contribution	-Predictability enhancement of SSC monitoring using H-ADCP & MOSGO-SVR -SSC monitoring using H-ADCP & SVR	-Extended application of TL monitoring -Catchment-wise temporal transport pattern	-Fsus model development -Suspended-to-total load transport physics
Predictive models	H-ADCP to SSC, TL models (SVR)	H-ADCP-SSC model determination protocol	Fsus (=SL/TL) model (SVR, Operon, MGGP)
Machine learning models	GRID-RFE-CV and MOSGO-SVR	Itarative GMM	GRID-RFE-CV and iterative SOM-GMM
Related publications	-Noh, Son, Kim & Park (2023) *KSCE J. Civ. Environ. Eng. Res.*, **43**: 321-335. -Noh, Son, Kim & Park (2023) *Proc. Coastal Sediment* **5**: 1801–1808 -Noh, Son, Kim & Park (accepted) *Adv. Water Resour.*	-Noh, Son, Kim & Park (2023) *J. Korea Water Resour. Assoc.*, **55**: 43-57	-Noh, Park & Seo (2023) *Water Resour. Res.* **59**: e2022WR034401

References

Abeshu, G. W., Li, H.-Y., Zhu, Z., Tan, Z., and Leung, L. R. (2022). "Median bed-material sediment particle size across rivers in the contiguous us." *Earth System Science Data*, 14(2), 929–942.

Ackers, P. and White, W. R. (1973). "Sediment transport: new approach and analysis." *Journal of the Hydraulics Division*, 99(11), 2041–2060.

Agarwal, A., Maheswaran, R., Kurths, J., and Khosa, R. (2016). "Wavelet spectrum and self-organizing maps-based approach for hydrologic regionalization-a case study in the western united states." *Water resources management*, 30, 4399–4413.

Agrawal, Y. C. and Pottsmith, H. C. (2000). "Instruments for particle size and settling velocity observations in sediment transport." *Marine Geology*, 168(1-4), 89–114.

Aguirre-Pe, J., Olivero, M. ı. a. L., and Moncada, A. T. (2003). "Particle densimetric froude number for estimating sediment transport." *Journal of Hydraulic Engineering*, 129(6), 428–437.

Akaike, H. (1974). "A new look at the statistical model identification." *IEEE transactions on automatic control*, 19(6), 716–723.

Aleixo, R., Guerrero, M., Nones, M., and Ruther, N. (2020). "Applying adcps for long-term monitoring of ssc in rivers." *Water Resources Research*, 56(1), e2019WR026087.

Alibrahim, H. and Ludwig, S. A. (2021). "Hyperparameter optimization: Comparing genetic algorithm against grid search and bayesian optimization." *2021 IEEE Congress on Evolutionary Computation (CEC)*, IEEE, 1551–1559.

Alvarez-Guerra, M., González-Piñuela, C., Andrés, A., Galán, B., and Viguri, J. R. (2008). "Assessment of self-organizing map artificial neural networks for the classification of sediment quality." *Environment International*, 34(6), 782–790.

Awad, M., Khanna, R., Awad, M., and Khanna, R. (2015). "Support vector regression." *Efficient learning machines: Theories, concepts, and applications for engineers and system designers*, 67–80.

Bagnold, R. A. (1966). *An approach to the sediment transport problem from general physics*. US government printing office.

Berghuijs, W. R., Sivapalan, M., Woods, R. A., and Savenije, H. H. (2014). "Patterns of similarity of seasonal water balances: A window into streamflow variability over a range of time scales." *Water Resources Research*, 50(7), 5638–5661.

Berrar, D. (2019). "Cross-validation." *Encyclopedia of Bioinformatics and Compu-*

tational Biology, S. Ranganathan, M. Gribskov, K. Nakai, and C. Schönbach, eds., Academic Press, Oxford, 542–545.

Best, J. (2005). "The fluid dynamics of river dunes: A review and some future research directions." *Journal of Geophysical Research: Earth Surface*, 110(F4).

Bezdek, J. C., Ehrlich, R., and Full, W. (1984). "Fcm: The fuzzy c-means clustering algorithm." *Computers & geosciences*, 10(2-3), 191–203.

Bishop, C. M. (2006). "Pattern recognition." *Machine learning*, 128(9).

Boateng, I., Bray, M., and Hooke, J. (2012). "Estimating the fluvial sediment input to the coastal sediment budget: A case study of ghana." *Geomorphology*, 138(1), 100–110.

Brown, P. P. and Lawler, D. F. (2003). "Sphere drag and settling velocity revisited." *Journal of environmental engineering*, 129(3), 222–231.

Burkham, D. and Dawdy, D. R. (1980). *General study of the modified Einstein method of computing total sediment discharge*, Vol. 2066. US Government Printing Office.

Burlacu, B., Kronberger, G., and Kommenda, M. (2020). "Operon c++ an efficient genetic programming framework for symbolic regression." *Proceedings of the 2020 Genetic and Evolutionary Computation Conference Companion*, 1562–1570.

Buscombe, D. (2013). "Transferable wavelet method for grain-size distribution from

images of sediment surfaces and thin sections, and other natural granular patterns." *Sedimentology*, 60(7), 1709–1732.

Camenen, B., Bayram, A., and Larson, M. (2006). "Equivalent roughness height for plane bed under steady flow." *Journal of Hydraulic Engineering*, 132(11), 1146–1158.

Chalov, S., Moreido, V., Ivanov, V., and Chalova, A. (2022). "Assessing suspended sediment fluxes with acoustic doppler current profilers: Case study from large rivers in russia." *Big Earth Data*, 6(4), 504–526.

Chaudry, M. H. (2008). *Open-channel flow*. Springer, NY, U.S.

Cheng, N.-S. (2002). "Exponential formula for bedload transport." *Journal of Hydraulic Engineering*, 128(10), 942–946.

Cheng, N.-S. (2009). "Comparison of formulas for drag coefficient and settling velocity of spherical particles." *Powder Technology*, 189(3), 395–398.

Cheng, N.-S. and Emadzadeh, A. (2016). "Estimate of sediment pickup rate with the densimetric froude number." *Journal of Hydraulic Engineering*, 142(3), 06015024.

Cheng, N. S., Wei, M. X., Chiew, Y.-M., Lu, Y., and Emadzadeh, A. (2020). "Combined effects of mean flow and turbulence on sediment pickup rate." *Water Resources Research*, 56(2), e2019WR026181.

Choubin, B., Darabi, H., Rahmati, O., Sajedi-Hosseini, F., and Kløve, B. (2018). "River suspended sediment modelling using the cart model: A comparative study of machine learning techniques." *Science of the Total Environment*, 615, 272–281.

Chu, W., Gao, X., and Sorooshian, S. (2010). "Improving the shuffled complex evolution scheme for optimization of complex nonlinear hydrological systems: Application to the calibration of the sacramento soil-moisture accounting model." *Water Resources Research*, 46(9).

Chu, W., Gao, X., and Sorooshian, S. (2011). "A new evolutionary search strategy for global optimization of high-dimensional problems." *Information Sciences*, 181(22), 4909–4927.

Civan, F. (2007). "Critical modification to the vogel- tammann- fulcher equation for temperature effect on the density of water." *Industrial & engineering chemistry research*, 46(17), 5810–5814.

Cohen, S., Syvitski, J., Ashley, T., Lammers, R., Fekete, B., and Li, H.-Y. (2022). "Spatial trends and drivers of bedload and suspended sediment fluxes in global rivers." *Water Resources Research*, 58(6), e2021WR031583.

Colby, B. R. and Hembree, C. H. (1954). "Computations of total sediment discharge, niobrara river near cody, nebraska." *Science*, 119(3097), 657–658.

Colby, B. R. and Hubbell, D. W. (1961). "Simplified methods for computing total

sediment discharge with the modified einstein procedure." *Report No. 1593*, G.P.O., U.S. Report.

Dade, W. B. and Friend, P. F. (1998). "Grain-size, sediment-transport regime, and channel slope in alluvial rivers." *The Journal of Geology*, 106(6), 661–676.

Davies, D. L. and Bouldin, D. W. (1979). "A cluster separation measure." *IEEE transactions on pattern analysis and machine intelligence*, PAMI-1(2), 224–227.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). "Maximum likelihood from incomplete data via the em algorithm." *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1), 1–22.

Detert, M. and Weitbrecht, V. (2012). "Automatic object detection to analyze the geometry of gravel grains–a free stand-alone tool." *River flow*, Taylor & Francis Group London, 595–600.

Dettinger, M. D., Ralph, F. M., Das, T., Neiman, P. J., and Cayan, D. R. (2011). "Atmospheric rivers, floods and the water resources of california." *Water*, 3(2), 445–478.

Downing, A., Thorne, P. D., and Vincent, C. E. (1995). "Backscattering from a suspension in the near field of a piston transducer." *The Journal of the Acoustical Society of America*, 97(3), 1614–1620.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A., and Vapnik, V. (1996). "Support vector regression machines." *Advances in neural information processing systems*, 9.

Duan, Q., Gupta, V. K., and Sorooshian, S. (1993). "Shuffled complex evolution approach for effective and efficient global minimization." *Journal of optimization theory and applications*, 76, 501–521.

Duan, Q., Sorooshian, S., and Gupta, V. (1992). "Effective and efficient global optimization for conceptual rainfall-runoff models." *Water resources research*, 28(4), 1015–1031.

Eder, A., Strauss, P., Krueger, T., and Quinton, J. (2010). "Comparative calculation of suspended sediment loads with respect to hysteresis effects (in the petzenkirchen catchment, austria)." *Journal of hydrology*, 389(1-2), 168–176.

Edwards, T. K., Glysson, G. D., Guy, H. P., and Norman, V. W. (1999). *Field methods for measurement of fluvial sediment*. US Geological Survey Denver, CO.

Eghbal, M., Saha, T. K., and Hasan, K. N. (2011). "Transmission expansion planning by meta-heuristic techniques: a comparison of shuffled frog leaping algorithm, pso and ga." *2011 IEEE power and energy society general meeting*, IEEE, 1–8.

Einstein, H. A. (1950). "The bed-load function for sediment transportation in open channel flows." *Report No. 1026*, US Department of Agriculture.

Eltner, A., Sardemann, H., and Grundmann, J. (2020). "Flow velocity and discharge measurement in rivers using terrestrial and unmanned-aerial-vehicle imagery." *Hydrology and Earth System Sciences*, 24(3), 1429–1445.

Engelund, F. and Hansen, E. (1967). "A monograph on sediment transport in alluvial streams." *Technical University of Denmark 0stervoldgade 10, Copenhagen K.*

Eusuff, M. M. and Lansey, K. E. (2003). "Optimization of water distribution network design using the shuffled frog leaping algorithm." *Journal of Water Resources planning and management*, 129(3), 210–225.

Fischer, P. F., Leaf, G. K., and Restrepo, J. M. (2002). "Forces on particles in oscillatory boundary layers." *Journal of Fluid Mechanics*, 468, 327–347.

Flammer, G. H. (1962). "Ultrasonic measurement of suspended sediment." *Bulletin 1141A*, US Government Printing Office, Washington, D.C.

Fourriere, A., Claudin, P., and Andreotti, B. (2010). "Bedforms in a turbulent stream: formation of ripples by primary linear instability and of dunes by nonlinear pattern coarsening." *Journal of Fluid Mechanics*, 649, 287–328.

Garrido-Merchán, E. C. and Hernández-Lobato, D. (2020). "Dealing with categorical and integer-valued variables in bayesian optimization with gaussian processes." *Neurocomputing*, 380, 20–35.

Gellis, A. C. (2013). "Factors influencing storm-generated suspended-sediment concentrations and loads in four basins of contrasting land use, humid-tropical puerto rico." *Catena*, 104, 39–57.

GmbH, D. (n.d.). "Liquid dynamic viscosity. Accessed: 2022-11-09.

Gomez, B. and Church, M. (1989). "An assessment of bed load sediment transport formulae for gravel bed rivers." *Water Resources Research*, 25(6), 1161–1186.

Guerrero, M. and Di Federico, V. (2018). "Suspended sediment assessment by combining sound attenuation and backscatter measurements–analytical method and experimental validation." *Advances in water resources*, 113, 167–179.

Guerrero, M., Rüther, N., Szupiany, R., Haun, S., Baranya, S., and Latosinski, F. (2016). "The acoustic properties of suspended sediment in large rivers: consequences on adcp methods applicability." *Water*, 8(1), 13.

Guerrero, M., Szupiany, R. N., and Latosinski, F. (2013). "Multi-frequency acoustics for suspended sediment studies: an application in the parana river." *Journal of Hydraulic Research*, 51(6), 696–707.

Guo, J. and Julien, P. Y. (2004). "Efficient algorithm for computing einstein integrals." *Journal of hydraulic engineering*, 130(12), 1198–1201.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). "Gene selection for cancer classification using support vector machines." *Machine learning*, 46(1), 389–422.

Gwon, Y., Kwon, S., Kim, D., Seo, I. W., and You, H. (2023). "Estimation of shallow stream bathymetry under varying suspended sediment concentrations and compositions using hyperspectral imagery." *Geomorphology*, 433, 108722.

Hager, W. H. (2018). "Bed-load transport: advances up to 1945 and outlook into the future." *Journal of Hydraulic Research*, 56(5), 596–607.

Hager, W. H. and Oliveto, G. (2002). "Shields' entrainment criterion in bridge hydraulics." *Journal of hydraulic engineering*, 128(5), 538–542.

Hardy, R. J., Best, J. L., Lane, S. N., and Carbonneau, P. E. (2009). "Coherent flow structures in a depth-limited flow over a gravel surface: The role of near-bed turbulence and influence of reynolds number." *Journal of geophysical research: earth surface*, 114(F1).

Harun, M. and Ab. Ghani, A. (2020). "Revised equations of total bed material load for rivers in malaysia." *ICDSME 2019: Proceedings of the 1st International Conference on Dam Safety Management and Engineering*, Springer, 332–340.

Harun, M. A., Safari, M. J. S., Gul, E., and Ab Ghani, A. (2021). "Regression models for sediment transport in tropical rivers." *Environmental Science and Pollution Research*, 28(38), 53097–53115.

Harvey, E. L., Hales, T. C., Hobley, D. E., Liu, J., and Fan, X. (2022). "Measuring

the grain-size distributions of mass movement deposits." *Earth Surface Processes and Landforms*.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, Vol. 2. Springer.

Haught, D., Venditti, J. G., and Wright, S. A. (2017). "Calculation of in situ acoustic sediment attenuation using off-the-shelf horizontal a dcps in low concentration settings." *Water Resources Research*, 53(6), 5017–5037.

Heil, J., Häring, V., Marschner, B., and Stumpe, B. (2019). "Advantages of fuzzy k-means over k-means clustering in the classification of diffuse reflectance soil spectra: A case study with west african soils." *Geoderma*, 337, 11–21.

Ho, M., Lall, U., Sun, X., and Cook, E. R. (2017). "Multiscale temporal variability and regional patterns in 555 years of conterminous us streamflow." *Water Resources Research*, 53(4), 3047–3066.

Hoffmann, T. O., Baulig, Y., Fischer, H., and Blöthe, J. (2020). "Scale breaks of suspended sediment rating in large rivers in germany induced by organic matter." *Earth Surface Dynamics*, 8(3), 661–678.

Holmquist-johnson, C. L. (2006). "Bureau of reclamation automated modified einstein procedure (boramep) program for computing total sediment load.

Jain, A. K. (2010). "Data clustering: 50 years beyond k-means." *Pattern recognition letters*, 31(8), 651–666.

Julien, P. and Raslan, Y. (1998). "Upper-regime plane bed." *Journal of Hydraulic Engineering*, 124(11), 1086–1096.

Julien, P. Y. (2010). *Erosion and sedimentation*. Cambridge university press.

Julien, P. Y. (2018). *River mechanics*. Cambridge University Press.

Jung, W. S., Kim, S. E., and Kim, Y. D. (2021). "Prediction of surface water quality by artificial neural network model using probabilistic weather forecasting." *Water*, 13(17), 2392.

Karim, F. (1998). "Bed material discharge prediction for nonuniform bed sediments." *Journal of Hydraulic Engineering*, 124(6), 597–604.

Kazemi, M. S., Banihabib, M. E., and Soltani, J. (2021). "A hybrid svr-pso model to predict concentration of sediment in typical and debris floods." *Earth Science Informatics*, 14, 365–376.

Keulegan, G. H. (1938). *Laws of turbulent flow in open channels*, Vol. 21. National Bureau of Standards Gaithersburg, MD.

Kim, K.-H., Yun, S.-T., Yu, S., Choi, B.-Y., Kim, M.-J., and Lee, K.-J. (2020). "Geochemical pattern recognitions of deep thermal groundwater in south korea using

self-organizing map: Identified pathways of geochemical reaction and mixing." *Journal of Hydrology*, 589, 125202.

Kim, S. E. and Seo, I. W. (2015). "Artificial neural network ensemble modeling with conjunctive data clustering for water quality prediction in rivers." *Journal of Hydro-Environment Research*, 9(3), 325–339.

Kim, S. E., Seo, I. W., and Choi, S. Y. (2017a). "Assessment of water quality variation of a monitoring network using exploratory factor analysis and empirical orthogonal function." *Environmental Modelling & Software*, 94, 21–35.

Kim, S. E., Shin, J., Seo, I. W., and Lyu, S. (2016). "Development of stage-discharge rating curve using hydraulic performance graph model." *Procedia Engineering*, 154, 334–339.

Kim, Y. D., Kim, J. M., and Kang, B. (2017b). "Projection of runoff and sediment yield under coordinated climate change and urbanization scenarios in doam dam watershed, korea." *Journal of Water and Climate Change*, 8(2), 235–253.

Kiviluoto, K. (1996). "Topology preservation in self-organizing maps." *Proceedings of International Conference on Neural Networks (ICNN'96)*, Vol. 1, IEEE, 294–299.

Kohonen, T. (1990). "The self-organizing map." *Proceedings of the IEEE*, 78(9), 1464–1480.

Kohonen, T. (2012). *Self-organizing maps*, Vol. 30. Springer Science & Business Media.

Kohonen, T. (2013). "Essentials of the self-organizing map." *Neural networks*, 37, 52–65.

Koza, J. R. (1992). *Genetic programming: on the programming of computers by means of natural selection*, Vol. 1. MIT press.

Kwon, S., Seo, I. W., Noh, H., and Kim, B. (2022a). "Hyperspectral retrievals of suspended sediment using cluster-based machine learning regression in shallow waters." *Science of The Total Environment*, 833, 155168.

Kwon, S., Shin, J., Seo, I. W., Noh, H., Jung, S. H., and You, H. (2022b). "Measurement of suspended sediment concentration in open channel flows based on hyperspectral imagery from uavs." *Advances in Water Resources*, 159, 104076.

La Cava, W., Orzechowski, P., Burlacu, B., de França, F. O., Virgolin, M., Jin, Y., Kommenda, M., and Moore, J. H. (2021). "Contemporary symbolic regression methods and their relative performance." *arXiv preprint arXiv:2107.14351*.

Landers, M. N. (2012). "Fluvial suspended sediment characteristics by high-resolution, surrogate metrics of turbidity, laser-diffraction, acoustic backscatter, and acoustic attenuation." Ph.D. thesis, Dep. of Civil and Environ. Eng., Georgia Inst. of Technol., Atlanta, GA.

Landers, M. N., Straub, T. D., Wood, M. S., and Domanski, M. M. (2016). "Sediment acoustic index method for computing continuous suspended-sediment concentrations." *Report no.*, U.S. Geological Survey.

Landers, M. N. and Sturm, T. W. (2013). "Hysteresis in suspended sediment to turbidity relations due to changing particle size distributions." *Water Resources Research*, 49(9), 5487–5500.

Lane, E. and Borland, W. (1951). "Estimating bed load." *Eos, Transactions American Geophysical Union*, 32(1), 121–123.

Lara, J. M. (1966). *Computation of "Z's" for Use in the Modified Einstein Procedure*. Department of the Interior, Bureau of Reclamation, Office of Chief Engineer, Denver.

Laursen, E. M. (1958). "The total sediment load of streams." *Journal of the Hydraulics Division*, 84(1), 1–36.

Lee, C.-J., Kim, J.-S., Kim, Y.-J., and Kim, W. (2010). "Method for estimation of roughness coefficient by field measurement and its application." *Proceedings of the Korea Water Resources Association Conference*, Korea Water Resources Association, 504–508. (in Korean).

Legleiter, C. J. and Harrison, L. R. (2019). "Remote sensing of river bathymetry:

Evaluating a range of sensors, platforms, and algorithms on the upper sacramento river, california, usa." *Water Resources Research*, 55(3), 2142–2169.

Legleiter, C. J. and Kinzel, P. J. (2021). "Improving remotely sensed river bathymetry by image-averaging." *Water Resources Research*, 57(3), e2020WR028795.

Li, H.-Y., Tan, Z., Ma, H., Zhu, Z., Abeshu, G., Zhu, S., Cohen, S., Zhou, T., Xu, D., and Leung, L.-Y. R. (2021). "A new large-scale suspended sediment model and its application over the united states." *Hydrology and Earth System Sciences Discussions*, 2021, 1–44.

Li, T., Sun, G., Yang, C., Liang, K., Ma, S., and Huang, L. (2018). "Using self-organizing map for coastal water quality classification: Towards a better understanding of patterns and processes." *Science of the Total Environment*, 628, 1446–1459.

Lin, J.-Y., Cheng, C.-T., and Chau, K.-W. (2006). "Using support vector machines for long-term discharge prediction." *Hydrological sciences journal*, 51(4), 599–612.

Liu, H.-H., Chang, L.-C., Li, C.-W., and Yang, C.-H. (2018). "Particle swarm optimization-based support vector regression for tourist arrivals forecasting." *Computational Intelligence and Neuroscience*, 2018.

Lloyd, S. (1982). "Least squares quantization in pcm." *IEEE transactions on information theory*, 28(2), 129–137.

Maddock, T. and Borland, W. M. (1950). *Sedimentation Studies for the Planning of Reservoirs by the Bureau of Reclamation*. Bureau of Reclamation, Hydrology Division.

Malik, A., Tikhamarine, Y., Souag-Gamane, D., Kisi, O., and Pham, Q. B. (2020). "Support vector regression optimized by meta-heuristic algorithms for daily streamflow prediction." *Stochastic Environmental Research and Risk Assessment*, 34, 1755–1773.

Melesse, A., Ahmad, S., McClain, M., Wang, X., and Lim, Y. (2011). "Suspended sediment load prediction of river systems: An artificial neural network approach." *Agricultural Water Management*, 98(5), 855–866.

Meshram, S. G., Singh, V. P., Kisi, O., Karimi, V., and Meshram, C. (2020). "Application of artificial neural networks, support vector machine and multiple model-ann to sediment yield prediction." *Water Resources Management*, 34(15), 4561–4575.

Ministry of Environment (2018). "Annual hydrological report on korea (sediment, soil moisture, and evapotranspiration)." *Report no.*, Ministry of Environment. (in Korean).

Ministry of Environment (2019). "Annual hydrological report on korea (sediment, soil moisture, and evapotranspiration)." *Report no.*, Ministry of Environment. (in Korean).

Mirjalili, S., Mirjalili, S. M., and Lewis, A. (2014). "Grey wolf optimizer." *Advances in engineering software*, 69, 46–61.

Molinas, A. and Wu, B. (2001). "Transport of sediment in large sand-bed rivers." *Journal of Hydraulic Research*, 39(2), 135–146.

Moore, S., Le Coz, J., Hurther, D., and Paquier, A. (2012). "On the application of horizontal adcps to suspended sediment transport surveys in rivers." *Continental Shelf Research*, 46, 50–63.

Muste, M., Kim, D., and Kim, K. (2022a). "A flood-crest forecast prototype for river floods using only in-stream measurements." *Communications Earth & Environment*, 3(1), 78.

Muste, M., Kim, D., and Kim, K. (2022b). "Insights into flood wave propagation in natural streams as captured with acoustic profilers at an index-velocity gaging station." *Water*, 14(9), 1380.

Muste, M., Lee, K., Kim, D., Bacotiu, C., Oliveros, M. R., Cheng, Z., and Quintero, F. (2020). "Revisiting hysteresis of flow variables in monitoring unsteady streamflows." *Journal of hydraulic research*, 58(6), 867–887.

Naeini, M. R., Analui, B., Gupta, H., Duan, Q., and Sorooshian, S. (2019). "Three decades of the shuffled complex evolution (sce-ua) optimization algorithm: Review and applications." *Scientia Iranica*, 26(4), 2015–2031.

Naeini, M. R., Yang, T., Sadegh, M., AghaKouchak, A., Hsu, K.-l., Sorooshian, S., Duan, Q., and Lei, X. (2018). "Shuffled complex-self adaptive hybrid evolution (sc-sahel) optimization framework." *Environmental Modelling & Software*, 104, 215–235.

Nagy, H., Watanabe, K., and Hirano, M. (2002). "Prediction of sediment load concentration in rivers using artificial neural network model." *Journal of Hydraulic Engineering*, 128(6), 588–595.

Naqshband, S., Ribberink, J. S., Hurther, D., and Hulscher, S. J. (2014). "Bed load and suspended load contributions to migrating sand dunes in equilibrium." *Journal of Geophysical Research: Earth Surface*, 119(5), 1043–1063.

Nelder, J. A. and Mead, R. (1965). "A simplex method for function minimization." *The computer journal*, 7(4), 308–313.

Noh, H., Kwon, S., Seo, I. W., Baek, D., and Jung, S. H. (2020). "Multi-gene genetic programming regression model for prediction of transient storage model parameters in natural rivers." *Water*, 13(1), 76.

Noh, H., Park, Y. S., and Lee, M. (2021). "Regional classification of total suspended matter in coastal areas of south korea." *Estuarine, Coastal and Shelf Science*, 254, 107339.

Noh, H., Park, Y. S., and Seo, I. W. (2023a). "A novel efficient method

of estimating suspended total sediment load fraction in natural rivers, <https://doi.org/10.5281/zenodo.7707130> (March).

Noh, H., Son, G., Kim, D., and Park, Y. S. (2022). "Clustering of sediment characteristics in south korean rivers and its expanded application strategy to h-adcp based suspended sediment concentration monitoring technique." *Journal of Korea Water Resources Association*, 55(1), 43–57.

Noh, H., Son, G., Kim, D., and Park, Y. S. (2023b). "Importance of bedload sediment supply in the riverine sediment supply revealed from a real-time total load monitoring using horizontal-adcp and the support vector regression." *Coastal Sediments 2023*, 1801–1808.

Noh, H., Son, G., Kim, D., and Park, Y. S. (2023c). "A SVR based-pseudo modified einstein procedure incorporating h-adcp model for real-time total sediment discharge monitoring." *KSCE Journal of Civil and Environmental Engineering Research*, 43(3), 321–335.

Okcu, D., Pektas, A. O., and Uyumaz, A. (2016). "Creating a non-linear total sediment load formula using polynomial best subset regression model." *Journal of Hydrology*, 539, 662–673.

Oyelade, J., Isewon, I., Oladipupo, F., Aromolaran, O., Uwoghiren, E., Ameh, F.,

Achas, M., and Adebiyi, E. (2016). "Clustering algorithms: their application to gene expression data." *Bioinformatics and Biology insights*, 10, BBI–S38316.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). "Scikit-learn: Machine learning in python." *Journal of machine learning research*, 12(Oct), 2825–2830.

Pektaş, A. O. (2015). "Determining the essential parameters of bed load and suspended sediment load." *International Journal of Global Warming*, 8(3), 335–359.

Pektaş, A. O. and Doğan, E. (2015). "Prediction of bed load via suspended sediment load using soft computing methods." *G eofizika*, 32(1), 27–46.

Pitlick, J., Mueller, E. R., Segura, C., Cress, R., and Torizzo, M. (2008). "Relation between flow, surface-layer armoring and sediment transport in gravel-bed rivers." *Earth Surface Processes and Landforms: The Journal of the British Geomorphological Research Group*, 33(8), 1192–1209.

Pomázi, F. and Baranya, S. (2022). "Acoustic based assessment of cross-sectional concentration inhomogeneity at a suspended sediment monitoring station in a large river." *Acta Geophysica*, 70(5), 2361–2377.

Raghavendra. N, S. and Deka, P. C. (2014). "Support vector machine applications in the field of hydrology: A review." *Applied Soft Computing*, 19, 372–386.

Rajaee, T., Nourani, V., Zounemat-Kermani, M., and Kisi, O. (2011). "River suspended sediment load prediction: application of ann and wavelet conjunction model." *Journal of Hydrologic Engineering*, 16(8), 613–627.

Riahi-Madvar, H. and Seifi, A. (2018). "Uncertainty analysis in bed load transport prediction of gravel bed rivers by ann and anfis." *Arabian Journal of Geosciences*, 11, 1–20.

Rosgen, D. (2019). "The rosgen stream classification system." *Encyclopedia of Water: Science, Technology, and Society*, 1–15.

Rosgen, D. L. (1994). "A classification of natural rivers." *Catena*, 22(3), 169–199.

Rouse, H. (1937). "Modern conceptions of the mechanics of fluid turbulence." *Transactions of the American Society of Civil Engineers*, 102(1), 463–505.

Safari, M. J. S. and Mehr, A. D. (2018). "Multigene genetic programming for sediment transport modeling in sewers for conditions of non-deposition with a bed deposit." *International Journal of Sediment Research*, 33(3), 262–270.

Schulkin, M. and March, H. W. (1962). "Sound absorption in sea water." *The Journal of the Acoustical Society of America*, 34(6), 864–865.

Schwarz, G. (1978). "Estimating the dimension of a model." *The annals of statistics*, 461–464.

Searson, D. P. (2015). "Gptips 2: an open-source software platform for symbolic data mining." *Handbook of genetic programming applications*, Springer, 551–573.

Seeger, M., Errea, M.-P., Beguerıa, S., Arnáez, J., Martı, C., and Garcıa-Ruiz, J. (2004). "Catchment soil moisture and rainfall characteristics as determinant factors for discharge/suspended sediment hysteretic loops in a small headwater catchment in the spanish pyrenees." *Journal of Hydrology*, 288(3-4), 299–311.

Shah-Fairbank, S. C. (2009). "Series expansion of the modified einstein procedure." Ph.D. thesis, Colorado State University, Fort Collins, CO.

Shah-Fairbank, S. C. and Julien, P. Y. (2015). "Sediment load calculations from point measurements in sand-bed rivers." *International Journal of Sediment Research*, 30(1), 1–12.

Shah-Fairbank, S. C., Julien, P. Y., and Baird, D. C. (2011). "Total sediment load from semep using depth-integrated concentration measurements." *Journal of Hydraulic Engineering*, 137(12), 1606–1614.

Shen, C. and Lemmin, U. (1999). "Application of an acoustic particle flux profiler in particleladen open-channel flow." *Journal of Hydraulic Research*, 37(3), 407–419.

Shen, H. and Hung, C. (1972). "An engineering approach to total bed-material load by regression analysis.." *Proc. of Sedimentation Symposium, Berkeley, California*.

Shen, H. W., Fehlman, H. M., and Mendoza, C. (1990). "Bed form resistances in open channel flows." *Journal of Hydraulic Engineering*, 116(6), 799–815.

Shen, H. W. and Hung, C. S. (1983). "Remodified einstein procedure for sediment load." *Journal of Hydraulic Engineering*, 109(4), 565–578.

Sheng, J. and Hay, A. E. (1988). "An examination of the spherical scatterer approximation in aqueous suspensions of sand." *The Journal of the Acoustical Society of America*, 83(2), 598–610.

Shinker, J. J. (2010). "Visualizing spatial heterogeneity of western us climate variability." *Earth Interactions*, 14(10), 1–15.

Sidle, R. C. and Campbell, A. J. (1985). "Patterns of suspended sediment transport in a coastal alaska stream 1." *JAWRA Journal of the American Water Resources Association*, 21(6), 909–917.

Simons, D., Li, R., and Fullerton, W. (1981). "Theoretically derived sediment transport equations for pima county, arizona." *Prepared for Pima County DOT and Flood Control District, Tucson, AZ, Fort Collins, CO: Simons, Li and Associates*.

Smola, A. J. and Schölkopf, B. (2004). "A tutorial on support vector regression." *Statistics and computing*, 14(3), 199–222.

Son, G. (2021). "Riverine sediment load measurement technique using acoustic

doppler current profilers considering hysteresis." Ph.D. thesis, Dankook University, Yongin-Si, Gyeonggi-Do, South Korea. (in Korean).

Son, G., Kim, D., Kwak, S., Kim, Y. D., and Lyu, S. (2021). "Characterizing three-dimensional mixing process in river confluence using acoustical backscatter as surrogate of suspended sediment." *Journal of Korea Water Resources Association*, 54(3), 167–179 (in Korean).

Stenger, D. and Abel, D. (2022). "Benchmark of bayesian optimization and meta-heuristics for control engineering tuning problems with crash constraints." *arXiv preprint arXiv:2211.02571*.

Stewart, D. (1983). "Possible suspended-load channel deposits from the wealden group (lower cretaceous) of southern england." *Modern and ancient fluvial systems*, 369–384.

Storn, R. and Price, K. (1997). "Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces." *Journal of global optimization*, 11, 341–359.

Sulaiman, M. S., Sinnakaudan, S. K., Azhari, N. N., and Abidin, R. Z. (2017). "Behavioral of sediment transport at lowland and mountainous rivers: a special reference to selected malaysian rivers." *Environmental Earth Sciences*, 76, 1–15.

Szupiany, R. N., Lopez Weibel, C., Guerrero, M., Latosinski, F., Wood, M.,

Dominguez Ruben, L., and Oberg, K. (2019). "Estimating sand concentrations using adcp-based acoustic inversion in a large fluvial system characterized by bi-modal suspended-sediment distributions." *Earth Surface Processes and Landforms*, 44(6), 1295–1308.

Tayfur, G., Karimi, Y., and Singh, V. P. (2013). "Principle component analysis in conjunction with data driven methods for sediment load prediction." *Water resources management*, 27(7), 2541–2554.

Tikhamarine, Y., Souag-Gamane, D., and Kisi, O. (2019). "A new intelligent method for monthly streamflow prediction: hybrid wavelet support vector regression based on grey wolf optimizer (wsvr–gwo)." *Arabian Journal of Geosciences*, 12, 1–20.

Topping, D. J., Wright, S. A., Melis, T. S., and Rubin, D. M. (2006). "High-resolution monitoring of suspended-sediment concentration and grain size in the colorado river using laser-diffraction instruments and a three-frequency acoustic system." *in Proceedings of the Eighth Federal Interagency Sedimentation Conference (8thFISC)*, 539–546.

Topping, D. J., Wright, S. A., Melis, T. S., and Rubin, D. M. (2007). "High-resolution measurements of suspended-sediment concentration and grain size in the colorado river in grand canyon using a multi-frequency acoustic system.." *in Proceedings, 10th International Symposium on River Sedimentation*, 330–339.

Torabi, H. and Dehghani, R. (2018). "Comparison and evaluation of intelligent models for river suspended sediment estimation (case study: Kakareza river, iran)." *Environmental Resources Research*, 6(2), 139–148.

Turowski, J. M., Rickenmann, D., and Dadson, S. J. (2010). "The partitioning of the total sediment load of a river into suspended load and bedload: a review of empirical data." *Sedimentology*, 57(4), 1126–1146.

Urick, R. (1948). "The absorption of sound in suspensions of irregular particles." *The Journal of the acoustical society of America*, 20(3), 283–289.

Urick, R. J. (1975). *Principles of underwater sound for engieneers*. McGraw Hill, NY, U.S.

Van Nierop, E. A., Luther, S., Bluemink, J. J., Magnaudet, J., Prosperetti, A., and Lohse, D. (2007). "Drag and lift forces on bubbles in a rotating flow." *Journal of fluid mechanics*, 571, 439–454.

van Rijn, L. C. (1984a). "Sediment transport, part ii: suspended load transport." *Journal of hydraulic engineering*, 110(11), 1613–1641.

van Rijn, L. C. (1984b). "Sediment transport, part iii: bed forms and alluvial roughness." *Journal of hydraulic engineering*, 110(12), 1733–1754.

Van Rijn, L. C. (1993). *Principles of sediment transport in rivers, estuaries and coastal seas*, Vol. 1006. Aqua publications Amsterdam.

Vesanto, J., Himberg, J., Alhoniemi, E., and Parhankangas (2000). "Som toolbox for matlab 5." *Report No. A57*, Helsinki University of Technology, Neural Networks Research Centre, Espoo, Finland.

Vogel, H. (1921). "Das temperaturabhängigkeitsgesetz der viskosität von flüssigkeitenn." *Phys. Z.*, 22, 645–646.

Wagner, W. and Pruß, A. (2002). "The iapws formulation 1995 for the thermodynamic properties of ordinary water substance for general and scientific use." *Journal of physical and chemical reference data*, 31(2), 387–535.

Wahl, T. L. (2023). "History and physical significance of the roughness froude number." *Journal of Hydraulic Research*, 61(2), 173–182.

Wallwork, J. T., Pu, J. H., Kundu, S., Hanmaiahgari, P. R., Pandey, M., Satyanaga, A., Khan, M. A., and Wood, A. (2022). "Review of suspended sediment transport mathematical modelling studies." *Fluids*, 7(1), 23.

Williams, G. P. (1989). "Sediment concentration versus water discharge during single hydrologic events in rivers." *Journal of Hydrology*, 111(1-4), 89–106.

Williams, G. P. and Rosgen, D. L. (1989). *Measured total sediment loads (suspended loads and bedloads) for 93 United States streams*. US Geological Survey Washington, DC.

Willis, C. M. and Griggs, G. B. (2003). "Reductions in fluvial sediment discharge by coastal dams in california and implications for beach sustainability." *The Journal of Geology*, 111(2), 167–182.

Xu, J. (2002). "Complex behaviour of natural sediment-carrying streamflows and the geomorphological implications." *Earth Surface Processes and Landforms: The Journal of the British Geomorphological Research Group*, 27(7), 749–758.

Yadav, A., Chatterjee, S., and Equeenuddin, S. M. (2018). "Suspended sediment yield estimation using genetic algorithm-based artificial intelligence models: case study of mahanadi river, india." *Hydrological Sciences Journal*, 63(8), 1162–1182.

Yan, K. and Zhang, D. (2015). "Feature selection and analysis on correlated gas sensor data with recursive feature elimination." *Sensors and Actuators B: Chemical*, 212, 353–363.

Yang, C. T. (1979). "Unit stream power equations for total load." *Journal of Hydrology*, 40(1-2), 123–138.

Yang, C. T., Marsooli, R., and Aalami, M. T. (2009). "Evaluation of total load sediment transport formulas using ann." *International Journal of Sediment Research*, 24(3), 274–286.

Yang, C. T. and Molinas, A. (1982). "Sediment transport and unit stream power function." *Journal of the Hydraulics Division*, 108(6), 774–793.

Yang, C.-Y. and Julien, P. Y. (2019). "The ratio of measured to total sediment discharge." *International Journal of Sediment Research*, 34(3), 262–269.

Zounemat-Kermani, M., Mahdavi-Meymand, A., Alizamir, M., Adarsh, S., and Yaseen, Z. M. (2020). "On the complexities of sediment load modeling using integrative machine learning: Application of the great river of loíza in puerto rico." *Journal of Hydrology*, 585, 124759.

국문초록

자연하천의 유사 이송 현상은 공학적으로, 환경적으로 중요한 의의를 가진다. 그러나 시료 채취에 의존하는 노동집약적인 전통적 유사량 계측 방법으로 인해 유사량 조사의 시공간 해상도를 높이는 데에 한계가 있다. 최근에는 횡방향 도플러 유속계(H-ADCP, Horizontal Acoustic Doppler Current Profiler)의 후방산란 강도와 부유사 농도 관계식을 유도해 부유사 모니터링의 시간 해상도를 높이고자 하는 노력이 지속되고 있다. H-ADCP 기반 모니터링은 데이터 취득 효율을 크게 높이지만 후방산란 신호 발생 기작의 비선형성과 데이터 취득의 어려움으로 인해 예측 성능과 적용성이 제한적이다. 본 연구는 대한민국의 자동 유량 관측소에서 얻은 H-ADCP 후방산란 신호를 기반으로 부유사 농도 모델을 결정하는 파이프라인을 기반으로 예측성과 계측 가능 지점과 항목을 향상시키는 것을 목적으로 한다. 단면 평균 부유사 농도 산정의 비선형성을 고려하기 위해 후방산란 신호 외에도 유량과 유량의 시간 변화율과 같은 수리 변수를 추가로 사용하였다. 이를 위해 본 연구에서는 입력변수의 조합과 서포트벡터회귀 모형의 초매개변수를 동시에 결정하는 방법론인 전역최적화 기반 서포트벡터회귀 모형 결정법(MOSGO-SVR, MOdel Selection by Global Optimization for support vector regression)와 최적 군집수 및 군집화 결과를 결정하는 반복 군집화 기법을 새로 제시하고 자동유량관측소에 H-ADCP 자료에 적용해 H-ADCP 기반 유사량 모니터링 기법의 산정 정확도

및 계측성을 높일 수 있는 방법을 제시한다. 첫번째로, H-ADCP로 취득할 수 있는 후방산란강도와 더불어 수위, 유량 그리고 수위와 유량의 시간 변화율 자료에 MOSGO-SVR를 적용함으로써 입력변수 조합을 포함한 최적의 부유사 농도 산정 모형을 결정해 정확도를 높인다. 이 때 총유사량 산정값을 추가로 학습해 H-ADCP 신호로부터 부유사 농도와 총유사량을 동시에 체계적으로 산정할 수 있는 파이프라인을 추가로 제시한다. 다음으로는 반복적 군집화 기법으로 국내 유사량 관측소의 군집 분석을 수행한 뒤 그 결과를 바탕으로 유사량 미계측 관측소에 H-ADCP 기반 부유사 농도 모니터링 방법을 적용할 수 있는 방안에 대해 논의한다. 세번째로, 특정 관측소의 유사량 산정자료를 학습시키는 대신 보다 수리적인 관계를 바탕으로 총유사량을 산정하기 위해 부유사농도로부터 총유사량을 산정하는 수리모형을 SVR과 기호회귀법을 이용해 유도한다. 최종적으로 이 결과를 종합해 H-ADCP를 이용한 부유사 농도 및 총유사량 산정 프레임워크를 제안함으로써 기존 유사량 모니터링의 정확도 향상 및 시공간 격차 해소에 기여한다. 본 연구의 결과를 적용함으로써 하천 유사량 관리 및 유사이송 기작에 대한 이해를 발전시킬 수 있을 것으로 기대된다.

주요어: 총유사량, 유사이송, 유사량 모니터링, H-ADCP, 후방산란, 최적화, 기계학습회귀, 군집화

학번: 2019-38726