Water Research 229 (2023) 119429

Contents lists available at ScienceDirect

Water Research

journal homepage: www.elsevier.com/locate/watres

Prediction of attachment efficiency using machine learning on a comprehensive database and its validation

Allan Gomez-Flores [a], Scott A. Bradford [b], Li Cai [c], Martin Urík [d], Hyunjung Kim [a,*]

[a] Department of Earth Resources and Environmental Engineering, Hanyang University, 222 Wangsimni-ro, Seongdong-gu, Seoul 04763, Republic of Korea
[b] USDA, ARS, Sustainable Agricultural Water Systems Unit, 239 Hopkins Road, Davis, CA 95616, United States
[c] College of Environmental Science and Engineering, Donghua University, Shanghai 201620, PR China
[d] Institute of Laboratory Research on Geomaterials, Faculty of Natural Sciences, Comenius University in Bratislava, Ilkovičova 6, 84215 Bratislava, Slovakia

ARTICLE INFO

Keywords:
Attachment efficiency
Machine learning
Missing data
Colloid deposition

ABSTRACT

Colloidal particles can attach to surfaces during transport, but the attachment depends on particle size, hydrodynamics, solid and water chemistry, and particulate matter. The attachment is quantified in filtration theory by measuring attachment or sticking efficiency (Alpha). A comprehensive Alpha database (2538 records) was built from experiments in the literature and used to develop a machine learning (ML) model to predict Alpha. The training (r-squared: 0.86) was performed using two random forests capable of handling missing data. A holdout dataset was used to validate the training (r-squared: 0.98), and the variable importance was explored for training and validation. Finally, an additional validation dataset was built from quartz crystal microbalance experiments using surface-modified polystyrene, poly (methyl methacrylate), and polyethylene. The experiments were performed in the absence or presence of humic acid. Full database regression (r-squared: 0.90) predicted Alpha for the additional validation with an r-squared of 0.23. Nevertheless, when the original database and the additional validation dataset were combined into a new database, both the training (r-squared: 0.95) and validation (r-squared: 0.70) increased. The developed ML model provides a data-driven prediction of Alpha over a big database and evaluates the significance of 22 input variables.

1. Introduction

Particle transport occurs between air, terrestrial, and aquatic compartments because they are interconnected (Guo et al., 2020; Horton and Dixon, 2018; Koutnik et al., 2021). Attachment of particles to soil grains and sediments during transport is important in the soil subsurface and aquatic compartments (Kim and Walker, 2009; Koutnik et al., 2021; Li et al., 2020). The attachment efficiency (Alpha) expresses the probability of particles sticking to surfaces after collision (Bradford et al., 2014; Petosa et al., 2010). Environmental fate modeling (EFM) has been developed to simulate the distribution of particles in compartments and uses Alpha as a critical parameter to predict the attachment of particles to soil and sediments (Krol et al., 2013; Meesters et al., 2014). EFM is commonly used for engineered nanoparticles (Meesters et al., 2014; Suhendra et al., 2020) and some microplastics (Besseling et al., 2017). For convenience, various EFMs have been assumed Alpha values, but the use of experimentally calculated Alpha leads to a more realistic EFM.

Alpha is a value obtained from experimental data and its advantages and disadvantages have been discussed elsewhere (Cornelis, 2015; Landkamer et al., 2013). However, the use of Alpha in EFM is limited to its availability for various conditions of particle surface chemistry, water chemistry, and hydrodynamics. Thus, prediction tools for estimating Alpha can be developed to address those limitations. Moreover, these tools can provide valuable information on the effects of input variables on Alpha, helping to interpret attachment behaviors. To our knowledge, there is only one pure empirical effort to predict Alpha in the literature (Park et al., 2012), but limitations of experimental conditions remain. In detail, Park et al. (2012) used data from six articles to predict Alpha for *C. parvum Oocysts* in packed columns using linear regression of the inverse Debye length and pH. The predictive ability of this empirically determined linear model was tested using Alpha from 3 other articles. The r-squared of predictions ranged from 0.71 to 0.76, indicating that ionic strength (IS) and pH, including a few records, cannot fully capture Alpha's behavior. It was suggested that including variables such as natural organic matter (NOM) coating content on sand grains, grain size distribution, porosity, pore water velocity, and *Oocyst* properties can

* Corresponding author.
E-mail address: kshjkim@hanyang.ac.kr (H. Kim).

https://doi.org/10.1016/j.watres.2022.119429
Received 20 September 2022; Received in revised form 17 November 2022; Accepted 20 November 2022
Available online 25 November 2022