# Augmenting Ground-Level PM2.5 Prediction via Kriging-Based Pseudo-Label Generation

**Lei Duan***
Department of Civil and Environmental Engineering
Duke University
Durham, NC 27705
lei.duan@duke.edu

**Ziyang Jiang***
Department of Civil and Environmental Engineering
Duke University
Durham, NC 27705
ziyang.jiang@duke.edu

**David Carlson**
Department of Civil and Environmental Engineering
Duke University
Durham, NC 27705
david.carlson@duke.edu

## Abstract

Fusing abundant satellite data with sparse ground measurements constitutes a major challenge in climate modeling. To address this, we propose a strategy to augment the training dataset by introducing unlabeled satellite images paired with pseudo-labels generated through a spatial interpolation technique known as ordinary kriging, thereby making full use of the available satellite data resources. We show that the proposed data augmentation strategy helps enhance the performance of the state-of-the-art convolutional neural network-random forest (CNN-RF) model by a reasonable amount, resulting in a noteworthy improvement in spatial correlation and a reduction in prediction error.

## 1 Introduction

Long-term exposure to fine particulate matter with an aerodynamic diameter of 2.5 $\mu m$ or smaller ($PM_{2.5}$) is is extensively recognized for increasing the risk of a variety of health problems such as stroke [1, 2], respiratory diseases [3], and ischemic heart disease [1, 4]. Accurate estimation and prediction of ground-level $PM_{2.5}$ concentrations are critical initial steps for subsequent epidemiological investigations into its health impacts. In recent years, advancements in satellite sensing techniques have enabled researchers to construct a high-resolution spatial mapping of $PM_{2.5}$ using machine learning models, in conjunction with ground measurements (e.g. $PM_{2.5}$, $PM_{10}$, temperature, humidity, etc.) acquired from air quality monitoring (AQM) stations [5–7]. However, in most cases, the satellite-based data is much more abundant compared to ground measurements, resulting in a large amount of unlabeled satellite data (i.e., satellite images at locations without a paired ground measurement). To elaborate, satellites have the capacity to cover entire urban areas and a majority of suburban regions of a city within a 24-hour window. In contrast, measurements from ground stations are only representative within their immediate vicinity, which we refer to as the “area of interest" (AOI). In most cases, the combined AOI of all ground stations is sparse compared to area covered by the satellite.

To address the aforementioned limitation, we come up with an approach to make full use of the unlabeled satellite data and effectively integrate them with ground measurements. Specifically, we propose to pair the unlabeled satellite data with *pseudo-labels* of ground measurements generated by a spatial interpolation method known as *ordinary kriging* and then augment the labeled dataset using these paired instances. In this study, we adopt the convolutional neural network-random forest (CNN-RF) model from Zheng et al. [6]. Our experimental results demonstrate the substantial advantages gained from this kriging-based augmentation, including decent reduction in errors and improvement in spatial correlation.

*Equal contribution

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2023.

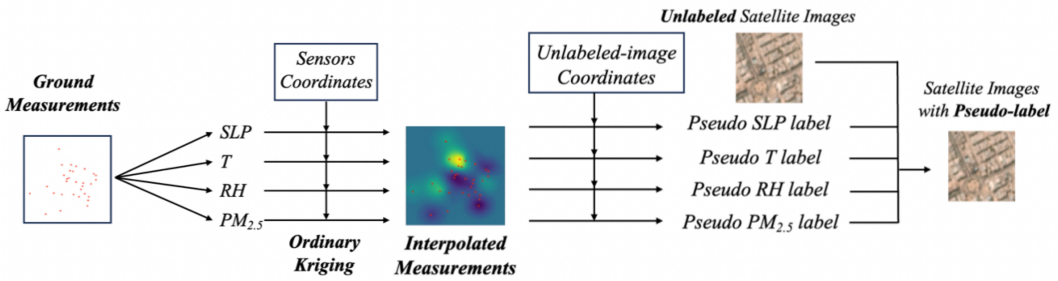


Figure 1: Illustration of pseudo-label generation process. The ground measurements consist of the ground-level $PM_{2.5}$ along with 3 meteorological attributes: sea level pressure (SLP), temperature (T), and relative humidity (RH). We first create a spatial mapping of interpolated measurements for the entire area of study using *ordinary kriging*. Then we pair unlabeled satellite images with corresponding interpolated measurements based on their geographical coordinates. Namely, each final data point with pseudo labels consists of 5 components: satellite image, interpolated SLP, T, RH, and $PM_{2.5}$ measurements.

## 2 Related work

**Kriging** The theoretical basis of kriging, alternatively known as Gaussian process (GP) regression, was developed by Georges Matheron [8]. Within geostatistics models, kriging serves as a spatial interpolation method where each data point is treated as a sample from a stochastic process [9–11]. It also includes several distinct variants (e.g., ordinary kriging [12, 13], universal kriging [14], co-kriging [14, 15], etc.) that are categorized based on the properties of the stochastic process. There are previous research efforts where kriging is applied to the context of remote sensing and climate modeling. To elaborate, Zhan et al. [16] developed a random forest-spatiotemporal kriging model to estimate the daily $NO_2$ concentrations in China from satellite data and geographical covariates. Wu and Li [17] employed residual kriging to interpolate average monthly temperature using latitude, longitude, and elevation as input variables. Besides these, researchers also used GPs for interpolation along with other sequence models [18] or classification models [19] in machine learning.

**Fusing Satellite and Ground Data** The large gap in data density between satellite observations and ground measurements is a common phenomenon in the fields of remote sensing and climate modeling, leading researchers to develop a variety of approaches to address this issue. For instance, Jiang et al. [20] introduced the spatiotemporal contrastive learning strategy, which involves pre-training deep learning models using unlabeled satellite images to improve the model performance on ground measurements. Verdin et al. [21] employed Bayesian kriging to model the ground observations of rainfall, with the mean parameterized as a linear function of satellite estimates and elevation. In light of these prior studies, our proposed method shares similarities with both of these approaches.

## 3 Methods

Our proposed method can be divided into 3 steps. First, we generate pseudo-labels of ground measurements using ordinary kriging and pair them with unlabeled satellite images. Then we integrate these satellite images with an existing training dataset containing true ground measurements. Lastly, we adopt the CNN-RF model from Zheng et al. [6] to train on this augmented training dataset and compare the performance to the case without introducing pseudo-labels. We will discuss these steps in detail in the following subsections.

### 3.1 Pseudo-label Generation

The process of generating pseudo-labels is illustrated in Figure 1, where we use ordinary kriging with an exponential semivariogram (analogous to the kernel function in GP) to interpolate the ground measurements over the entire area of study. We prefer kriging over other spatial interpolation methods because it serves as an *optimal linear predictor* under the modeling framework, accounting for factors such as spatial distance, continuity, and data redundancy. A more detailed discussion will be provided

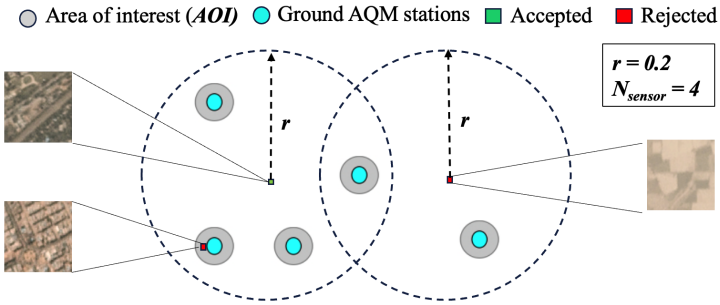


Figure 2: Filtering rules for unlabeled satellite images. To be included in the training dataset, an unlabeled image must not be located within the AOI of any AQM station and must have at least $N_{\text{sensor}}$ AQM stations within its vicinity (defined by a circular region with a radius of $r$). In our experiments, we set $N_{\text{sensor}} = 4$ and $r = 0.2$.

in Appendix A.1. The parametric form for the exponential semivariogram is presented as follows:

$$\gamma(h) = \begin{cases} \tau^2 + \sigma^2(1 - \exp(-\phi h)) & \text{if} \quad h > 0, \\ 0 & \text{otherwise}, \end{cases} \tag{1}$$

where $h$ stands for the distance between two ground measurements, while $\tau^2$, $\tau^2 + \sigma^2$, and $1/\phi$ represent the nugget, sill, and range parameters, respectively. Concepts and details of semivariogram will be discussed in Appendix A.2. We choose this particular form due to its superior fit with the empirical semivariogram compared to other isotropic semivariograms (e.g., Gaussian, spherical, etc.), as demonstrated in Appendix A.3. Following the generation of pseudo-labels, we pair them with unlabeled satellite images based on their geographical coordinates. We then blend these pseudo-labeled images with existing satellite images with true ground measurements to form an *augmented* training dataset. It is also possible to include uncertainty during imputation, but we only use the mean here as prior work shows that high-quality imputations make the biggest impact on performance [19].

### 3.2 Training of CNN-RF Joint Model

With the augmented dataset, we proceed to the phase of model training. We adopt the convolutional neural network-random forest (CNN-RF) model from a previous study conducted by Zheng et al. [6] as it demonstrates the state-of-the-art performance on the $PM_{2.5}$ prediction task. Specifically, the RF learns a preliminary estimation of $PM_{2.5}$ levels $\hat{y}_{\text{RF}}$ from the meterological attributes (i.e., SLP, T, RH) and the CNN learns a residual $\epsilon = y - \hat{y}_{\text{RF}}$ between the true $PM_{2.5}$ label $y$ and the RF-predicted $PM_{2.5}$ $\hat{y}_{\text{RF}}$. The model training phase is divided into two key components: Random Forest (RF) and Convolutional Neural Network (CNN). The detailed model architecture and hyperparameters can be found in Figure 2 in Zheng et al. [6].

### 3.3 Model evaluation

To ensure a direct comparative analysis, we test the effectiveness of our method using the same evaluation metrics adopted by Zheng et al. [6], which include the root-mean-square error (RMSE), mean absolute error (MAE), Pearson correlation coefficient (Pearson R), and spatial Pearson correlation coefficient (spatial Pearson R) as defined as follows:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y - \hat{y})^2}, \qquad \text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y - \hat{y}| \tag{2}$$

$$\text{Pearson R} = \frac{\sum(y-\bar{y})(\hat{y}-\bar{\hat{y}})}{\sqrt{\sum(y-\bar{y})^2\sum(\hat{y}-\bar{\hat{y}})^2}}, \qquad \text{Spatial Pearson R} = \frac{\sum(y'-\bar{y'})(\hat{y}-\bar{\hat{y}})}{\sqrt{\sum(y-\bar{y})^2\sum(\hat{y}-\bar{\hat{y}})^2}} \tag{3}$$

where $y$, $\hat{y}$ are the true $PM_{2.5}$ values and predicted $PM_{2.5}$ values, and $\bar{y}$,$\bar{\hat{y}}$ are the average true $PM_{2.5}$ values and the average predicted $PM_{2.5}$ values, $N$ is number of data points in the test set.

## 4 Experiments

### 4.1 Data

We download the three-band (red-blue-green, RGB) scene visual products developed by Planet Labs [22] with a spatial resolution of 3 m/pixel, covering the period from January 1, 2018, to June 28,

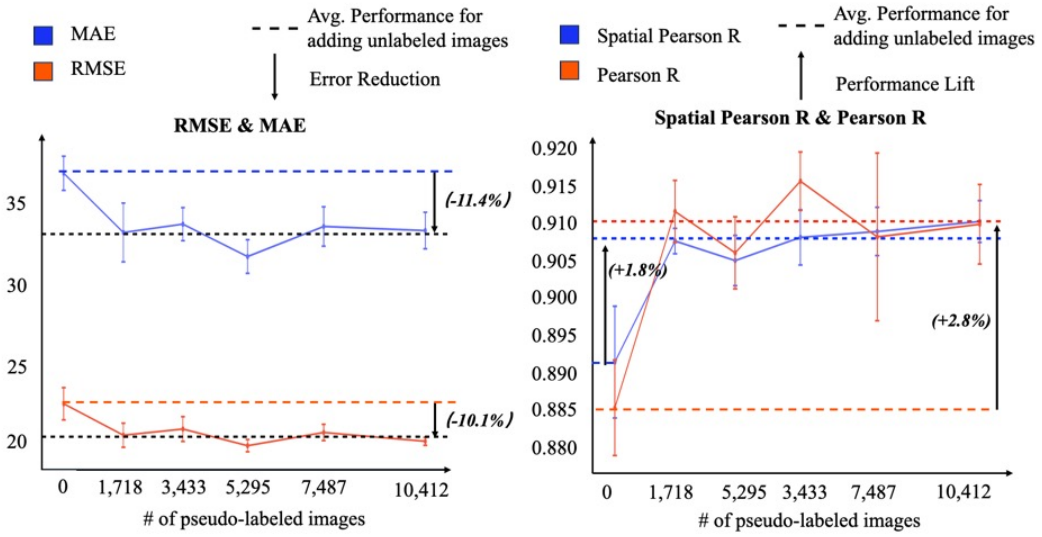


Figure 3: Performance of CNN-RF model on the test data with different number of pseudo-labeled images. (a) Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), (b) Pearson R and Spatial Pearson R.

2020. The existing training dataset contains 31,568 satellite images with true ground-level $PM_{2.5}$ labels and meteorological attributes obtained from 51 ground AQM stations spanning the National Capital Territory (NCT) of Delhi and its satellite cities such as Gurgaon, Faridabad, Noida, etc. We filter the unlabeled satellite images (i.e., images outside the AOI of AQM stations) as illustrated in Figure 2. This procedure ensures that each image can be effectively paired with pseudo-labels that are spatially interpolated using a sufficient amount of surrounding ground measurement data.

### 4.2 Results

We test the CNN-RF model on datasets containing different number of pseudo-labeled satellite images and compare them with the baseline case with only true-labeled images (i.e., case with 0 pseudo-labeled images). For each scenario, we repeat the experiment by 10 times and calculate the average and standrd deviation of the evaluation metrics. As illustrated in Figure 3, our results show substantial improvements in both prediction accuracy and spatial correlation. Specifically, we observe a 10.1% reduction in average root mean square error (RMSE), decreasing from 37.64 to 32.56, and a 11.4% decrease in mean absolute error (MAE), dropping from 24.39 to 20.31. Moreover, there is a 2.8% increase in Pearson's correlation coefficient (Pearson R) from 0.880 to 0.912, and a 1.8% increase in spatial Pearson's correlation coefficient (spatial Pearson R) from 0.878 to 0.907. These results highlight the substantial benefits gained from the integration of pseudo-labeled images within the training process.

## 5 Conclusion

In this research, we introduce a data augmentation approach that involves the incorporation of pseudo-labeled images into the model training procedure, allowing us to fully leverage the abundancy of unlabeled satellite images. Our approach involves the generation of pseudo-labeled images through a two-step process: firstly, applying ordinary kriging on ground measurements to produce pseudo-labels, and subsequently, pairing these labels with unlabeled images that are carefully selected through a heuristic criterion, ensuring that each unlabeled image has sufficient spatial information for spatial interpolation. The results demonstrate that including pseudo-labeled images successfully improve the state-of-the-art model performance in terms of prediction error and spatial correlation on the test data, showing the effectiveness of our proposed data augmentation strategy. We expect this strategy to be useful for various other scenarios in the field of remote sensing and geostatistics.

## References


[1] Stacey E Alexeeff, Noelle S Liao, Xi Liu, Stephen K Van Den Eeden, and Stephen Sidney. Long-term pm2. 5 exposure and risks of ischemic heart disease and stroke events: review and meta-analysis. *Journal of the American Heart Association*, 10(1):e016890, 2021.

[2] Sheng Yuan, Jiaxin Wang, Qingqing Jiang, Ziyu He, Yuchai Huang, Zhengyang Li, Luyao Cai, and Shiyi Cao. Long-term exposure to pm2. 5 and stroke: a systematic review and meta-analysis of cohort studies. *Environmental research*, 177:108587, 2019.

[3] C. Arden Pope and Douglas W. Dockery. Health Effects of Fine Particulate Air Pollution: Lines that Connect. *Journal of the Air & Waste Management Association*, 56(6): 709–742, June 2006. ISSN 1096-2247. doi: 10.1080/10473289.2006.10464485. URL https://doi.org/10.1080/10473289.2006.10464485. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10473289.2006.10464485.

[4] George D Thurston, Richard T Burnett, Michelle C Turner, Yuanli Shi, Daniel Krewski, Ramona Lall, Kazuhiko Ito, Michael Jerrett, Susan M Gapstur, W Ryan Diver, et al. Ischemic heart disease mortality and long-term exposure to source-related components of us fine particle air pollution. *Environmental health perspectives*, 124(6):785–794, 2016.

[5] Tongshu Zheng, Michael H Bergin, Shijia Hu, Joshua Miller, and David E Carlson. Estimating ground-level pm2. 5 using micro-satellite images by a convolutional neural network and random forest approach. *Atmospheric Environment*, 230:117451, 2020.

[6] Tongshu Zheng, Michael Bergin, Guoyin Wang, and David Carlson. Local PM2.5 Hotspot Detector at 300 m Resolution: A Random Forest–Convolutional Neural Network Joint Model Jointly Trained on Satellite Images and Meteorology. *Remote Sensing*, 13(7):1356, April 2021. ISSN 2072-4292. doi: 10.3390/rs13071356. URL https://www.mdpi.com/2072-4292/13/7/1356.

[7] Ziyang Jiang, Tongshu Zheng, Yiling Liu, and David Carlson. Incorporating prior knowledge into neural networks through an implicit composite kernel. *arXiv preprint arXiv:2205.07384*, 2022.

[8] Georges Matheron. Krigeage d'un panneau rectangulaire par sa périphérie. *Note géostatistique*, 28, 1960.

[9] M. A. Oliver and R. Webster. Kriging: a method of interpolation for geographical information systems. *International Journal of Geographical Information Systems*, 4(3):313–332, July 1990. ISSN 0269-3798. doi: 10.1080/02693799008941549. URL https://www.tandfonline.com/doi/ref/10.1080/02693799008941549. Publisher: Taylor & Francis.

[10] Sudipto Banerjee, Bradley P Carlin, and Alan E Gelfand. *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC, 2003.

[11] Michael L Stein. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 1999.

[12] Noel Cressie. Spatial prediction and ordinary kriging. *Mathematical geology*, 20:405–421, 1988.

[13] Hans Wackernagel and Hans Wackernagel. Ordinary kriging. *Multivariate geostatistics: an introduction with applications*, pages 79–88, 2003.

[14] A Stein and LCA Corsten. Universal kriging and cokriging as a regression procedure. *Biometrics*, pages 575–587, 1991.

[15] Jeffrey D Helterbrand and Noel Cressie. Universal cokriging under intrinsic coregionalization. *Mathematical Geology*, 26:205–226, 1994.

[16] Yu Zhan, Yuzhou Luo, Xunfei Deng, Kaishan Zhang, Minghua Zhang, Michael L Grieneisen, and Baofeng Di. Satellite-based estimates of daily no2 exposure in china using hybrid random forest and spatiotemporal kriging model. *Environmental science & technology*, 52(7):4180–4189, 2018.

[17] Tingting Wu and Yingru Li. Spatial interpolation of temperature in the united states using residual kriging. *Applied Geography*, 44:112–120, 2013.

[18] Joseph Futoma, Sanjay Hariharan, and Katherine Heller. Learning to detect sepsis with a multitask gaussian process rnn classifier. In *International conference on machine learning*, pages 1174–1182. PMLR, 2017.

[19] Steven Cheng-Xian Li and Benjamin M Marlin. A scalable end-to-end gaussian process adapter for irregularly sampled time series classification. *Advances in neural information processing systems*, 29, 2016.

[20] Ziyang Jiang, Tongshu Zheng, Mike Bergin, and David Carlson. Improving spatial variation of ground-level pm2. 5 prediction with contrastive learning from satellite imagery. *Science of Remote Sensing*, 5:100052, 2022.

[21] Andrew Verdin, Balaji Rajagopalan, William Kleiber, and Chris Funk. A bayesian kriging approach for blending satellite and ground precipitation observations. *Water Resources Research*, 51(2):908–921, 2015.

[22] Planet Labs PBC. Planet application program interface: In space for life on earth, 2018–. URL `https://api.planet.com`.

# A Details of Kriging

## A.1 Advantages of Kriging

There are several spatial interpolation methods in the existing geostatistics literature, such as bilinear interpolation, nearest neighbor, inverse distance weighting (IDW), modified Shepard interpolation etc. However, all these interpolation methods have certain limitations. Specifically, nearest neighbor approach gives very poor estimation when the ground stations are sparse. Both IDW and modified Shepard compute the estimated measurement at a target location as a weighted linear combination of its neighboring stations, with the weights being a nonlinear function of the distance $h$ between the target location and the neighboring station. A closer station (i.e., with smaller $h$) typically receives larger weights. Two commonly used weight functions are the inverse power function $w(h) = 1/h^\beta$ and the exponential function $w(h) = \exp\left(-(h/h_0)^\beta\right)$ where $\beta$ and $h_0$ are predetermined parameters.

However, a major drawback of IDW and modified Shepard lies in their failure to account for data redundancy. For instance, consider a scenario where we have measurements from ground stations denoted as $y_1, y_2, ..., y_m, y_{m+1}$, with the measurement we seek to estimate labeled as $y_0$. Let's assume that $y_1, y_2, ..., y_m$ are closely clustered together, while $y_{m+1}$ stands alone and further assume all of the stations, namely $y_1, y_2, ..., y_m, y_{m+1}$, are equidistant from $y_0$. With IDW or modified Shepard, $\{y_1, ..., y_m\}$ will receive much larger weights *as a cluster* compared to $y_{m+1}$. Nevertheless, in reality, the cluster $y_1, ..., y_m$ contributes roughly the same amount of information to the estimation of $y_0$ as $y_{m+1}$ does. Therefore, it is reasonable to expect that $\{y_1, ..., y_m\}$ should receive similar weighting as $y_{m+1}$.

Kriging is a linear estimation method that minimizes the expected squared error, and therefore is also referred to as the *best linear unbiased estimator*. To be specific, say we again have measurements from ground stations denoted as $y_1, y_2, ..., y_m$, with the measurement we want to estimate labeled as $y_0$. With Kriging, our estimation is formulated as a linear combination, denoted as $\hat{y}_0 = \sum_{i=1}^m \alpha_i y_i$, subject to the constraint $\sum_{i=1}^m \alpha_i = 1$. The determination of the coefficients $\alpha_i$ is closely tied to a function known as the *variogram*, which will be elaborated in the following section. Compared to the aforementioned methods, Kriging takes into account spatial distance, spatial continuity, and data redundancy simultaneously. Consequently, it is the preferred method for our research, as it offers a more comprehensive approach to spatial estimation.

## A.2 Variogram

The variogram can be viewed as a function that describes the correlation between two ground measurements as a function of the distance $h$ between them. Mathematically, the semivariogram, which is half the value of the variogram, can be expressed by the general formulation as given below:

$$\gamma(h) = \frac{1}{2}\,\mathbb{E}[y(s+h) - y(s)]^2, \tag{4}$$

where $y(s)$ and $y(s+h)$ represent the ground measurements at location $s$ and $s+h$, respectively. Some candidates of variogram include spherical, Gaussian, and exponential, whose functional forms are given as:

$$\gamma_{\text{spherical}}(h) = \begin{cases} \tau^2 + \sigma^2 & \text{if} \quad h \geq 1/\phi \\ \tau^2 + \sigma^2 \left[\frac{3\phi h}{2} - \frac{1}{2}(\phi h)^3\right] & \text{if} \quad 0 < h \leq 1/\phi \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$$\gamma_{\text{Gaussian}}(h) = \begin{cases} \tau^2 + \sigma^2\left(1 - \exp(-\phi^2 h^2)\right) & \text{if} \quad h > 0 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

$$\gamma_{\text{exponential}}(h) = \begin{cases} \tau^2 + \sigma^2(1 - \exp(-\phi h)) & \text{if} \quad h > 0, \\ 0 & \text{otherwise}, \end{cases} \tag{7}$$

where $\tau^2$, $\tau^2 + \sigma^2$, and $1/\phi$ represent the nugget, sill, and range parameters, respectively. Here the nugget $\tau^2$ represents the semivariogram value at an infinitesimally small separation distance, i.e.,

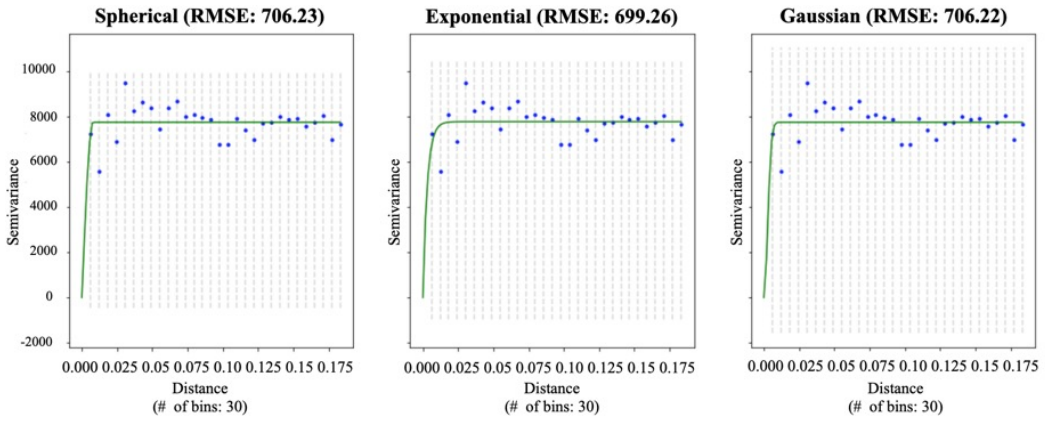


Figure A1: Empirical semivariograms (plotted as blue dots) with different theoretical fits (plotted as green lines), i.e., spherical, exponential, and Gaussian.

$\tau^2 = \gamma(0^+) \equiv \lim_{h \to 0^+} \gamma(h)$. The range $1/\phi$ is the distance at which $\gamma(h)$ first reaches its ultimate level, i.e., the sill $\tau^2 + \sigma^2$.

### A.3 Empirical semivariogram

As we elaborated in Section 3.1, we plot the *empirical semivariogram*:

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{(s_i, s_j) \in N(h_k)} \left[ y(s_i) - y(s_j) \right]^2, \tag{8}$$

where $N(h_k)$ is the set of pairs of points $(s_i, s_j)$ such that $\|s_i - s_j\| \in I_k$. Here $I_k$ represents the interval $(h_{k-1}, h_k)$ such that $0 < h_1 < h_2 < \cdots < h_K$ and $k \in \{1, 2, ..., K\}$. Figure A1 provides a visual comparison between the empirical semivariograms computed from ground measurements and the theoretical forms of spherical, exponential, and Gaussian semivariograms. Among these models, the exponential semivariogram exhibits the lowest RMSE when compared to the other two models. The RMSE between the empirical semivariogram and the theoretical model is calculated as:

$$\text{RMSE} = \sqrt{\frac{1}{N_{\text{bins}}} \sum_{i=0}^{N_{\text{bins}}} (\hat{\gamma}_i - \gamma_i)^2}, \tag{9}$$

where $N_{\text{bins}}$ is the total number of bins used for plotting the empirical semivariogram, $\hat{\gamma}_i$ is the empirical semivariogram value for the $i^{th}$ bin, and $\gamma_i$ is the theoretical value for the $i^{th}$ bin.