


METHODS PAPER

# Data fusion of sparse, heterogeneous, and mobile sensor devices using adaptive distance attention

Jean-Marie Lepioufle<sup>1</sup> , Philipp Schneider<sup>1</sup>, Paul David Hamer<sup>1</sup>, Rune Åvar Ødegård<sup>1</sup>, Islen Vallejo<sup>1</sup>, Tuan Vu Cao<sup>1</sup>, Amir Taherkordi<sup>2</sup> and Marek Wojcikowski<sup>3</sup>

<sup>1</sup>The Climate and Environmental Research Institute NILU, Kjeller, Norway

<sup>2</sup>Department of Informatics, University of Oslo, Oslo, Norway

<sup>3</sup>Faculty of Electronics, Gdansk University of Technology, Gdansk, Poland

**Corresponding author:** Jean-Marie Lepioufle; Email: [jm@jeanmarie.eu](mailto:jm@jeanmarie.eu)

**Received:** 19 January 2023; **Revised:** 17 November 2023; **Accepted:** 21 May 2024

**Keywords:** attention; data fusion; environment; multi-sensors

## Abstract

In environmental science, where information from sensor devices are sparse, data fusion for mapping purposes is often based on geostatistical approaches. We propose a methodology called *adaptive distance attention* that enables us to fuse sparse, heterogeneous, and mobile sensor devices and predict values at locations with no previous measurement. The approach allows for automatically weighting the measurements according to *a priori* quality information about the sensor device without using complex and resource-demanding data assimilation techniques. Both ordinary kriging and the general regression neural network (GRNN) are integrated into this attention with their learnable parameters based on deep learning architectures. We evaluate this method using three static phenomena with different complexities: a case related to a simplistic phenomenon, topography over an area of 196 km<sup>2</sup> and to the annual hourly NO<sub>2</sub> concentration in 2019 over the Oslo metropolitan region (1026 km<sup>2</sup>). We simulate networks of 100 synthetic sensor devices with six characteristics related to measurement quality and measurement spatial resolution. Generally, outcomes are promising: we significantly improve the metrics from baseline geostatistical models. Besides, distance attention using the Nadaraya–Watson kernel provides as good metrics as the attention based on the kriging system enabling the possibility to alleviate the processing cost for fusion of sparse data. The encouraging results motivate us in keeping adapting distance attention to space-time phenomena evolving in complex and isolated areas.

## Impact Statement

Data fusion is commonly employed with the assumption of having access to a substantial volume of data. Unfortunately, measurement campaigns of complex phenomena in isolated areas often result in significantly reduced amount of information. We show that combining geostatistical tools and deep learning models into a distance attention overcomes this issue. We applied our method to static environmental phenomena using synthetic sensor devices of different measurement characteristics representative of real sensors. The encouraging results suggest that such methods can be applied to space-time environmental phenomena.

## 1. Introduction

Monitoring real-time environmental phenomena enables experts to detect unusual events such as abnormal air quality, greenhouse gas emission sources, and extreme meteorological events, among

others. For this purpose, numerical modeling tools are developed and networks of monitoring stations as well as satellite-based remote sensing provide data. Nonetheless, providing highly accurate information at high spatial and temporal resolution on vast areas requires heavy numerical processing, high-technology sensors, and an expensive maintenance that only national organizations and large companies can afford. To reduce these costs, high quality sensor devices are deployed more sporadically, and, given suitable data quality, low-cost sensors (Castell et al., 2017; Hassani et al., 2023; Schneider et al., 2023; Van Poppel et al., 2023) can complement these. In addition, numerical models are run from global to local scale with highly variable spatial resolution. Finally, a combination of these sources is processed to retrieve the target information, for instance, low-cost sensor calibration (De Vito et al., 2018, 2020; Ionascu et al., 2021) and land-use regression (Hong et al., 2019; Weichenthal et al., 2021) for air quality, atmospheric temperature downscaling (Chau et al., 2021), downscaling of satellite data for air quality (Stebel et al., 2021), and multi-sensor data fusion to estimate evapotranspiration (Semmens et al., 2016).

In environmental science, data fusion based on neural networks and machine learning is used to combine regularly spaced gridded datasets, such as satellite data (Schneider et al., 2021, Shetty et al., 2024) and images from unmanned aerial vehicles, thereby enabling resolution space-time enhancement, pansharpening, and classification (Ghamisi et al., 2019). For datasets with dense and irregular point cloud data, such as hyperspectral imaging and lidar, point fusion (Xu et al., 2018) enables classification, clustering, and point enrichment.

In the case of sparse point clouds, data fusion is often based on geostatistical techniques such as kriging (Wackernagel, 2003; Rue et al., 2017) (both with and without spatial auxiliary variables), for example, for spatiotemporal mapping of air quality (Schneider et al., 2017, 2018). In addition, data assimilation approaches, such as Kalman filter, Optimal Interpolation, 3D-Var, and 4D-Var (Miyoshi et al., 2010; Wattrelot et al., 2014; Lussana et al., 2019; Mijling, 2020; Hassani et al., 2023; Schneider et al., 2023), in which deep learning has been recently integrated (Arcucci et al., 2021; Peyron et al., 2021), use the uncertainty of each data source to determine their weight while fusing. While kriging requires solving the kriging equation system, other less computational processing demanding kernel regression approaches enable the prediction of space-time phenomena such as a graph convolution network (Appleby et al., 2020) and a GRNN (Specht, 1991; Robert et al., 2013).

This research work is carried out in the context of advancements in measurement campaigns, where heterogeneous, mobile, and autonomous devices (Jońca et al., 2022; Samad et al., 2022; Scheller et al., 2022) monitor local phenomena in isolated areas (Miner et al., 2022) for prediction purposes, for instance, spatial mapping (Hassani et al., 2023). We limit our paper to sensors being preprocessed at level 1, following (Schneider et al., 2019). We have thus observation devices providing sparse measurements at different spatial and temporal resolutions, with different measurement qualities, and possibly at non-regular sampling frequencies. In this context, data fusion of environmental sensor devices faces two challenges: i) fusing nonoverlapping multiple sources of information with heterogeneous characteristics and ii) predicting complex phenomena with sparse data. To overcome these challenges, we propose a methodology based on three axes: i) the use of *a priori* information about measurement characteristics and its quality to weight their influence in data fusion, ii) an inclusion of deep neural networks into ordinary kriging (OK) and GRNN, and iii) determining an attention framework as Vaswani et al. (2017) to enable inter-comparison between the prediction approaches.

Our paper is structured as follows. Section 2 describes the materials and methods used in this study. Section 2.1 describes the measurement characteristics in a network of sparse, heterogeneous, and mobile sensors devices. Section 2.2 describes the adaptive distance attention, Section 2.3 describes the GRNN and OK as adaptive distance attention, Section 2.4 describes the data fusion model architecture, Section 2.5 describes the three cases studies of this study, and Section 2.6 describes the experimentation plan. Section 3 presents the results and the discussion. Section 3.1 presents the metrics for the different data fusion models applied to the three case studies, Section 3.2 presents the effect of the data fusion model and the measurement campaign on the learnable parameters, and Section 3.3 presents a discussion of the results. Finally, the conclusion is presented in Section 4.

## 2. Materials and methods

### 2.1. Measurement characteristics in a network of sparse, heterogeneous, and mobile sensors devices

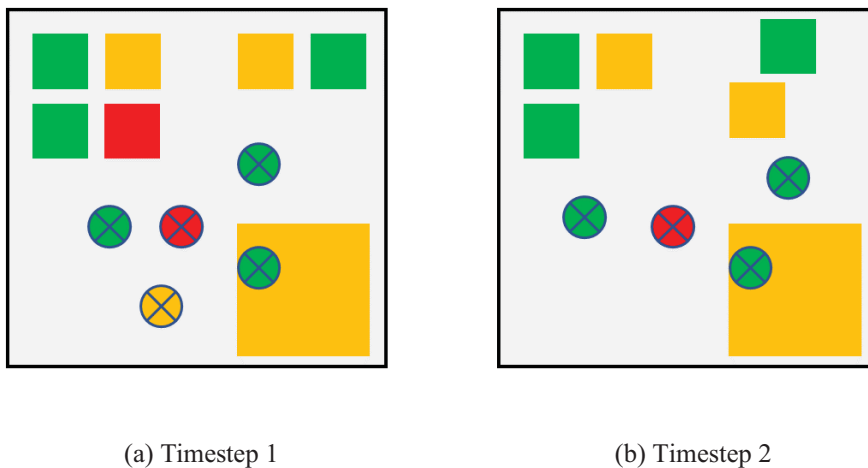
In this section, we describe the characteristics that are common to measurements from several heterogeneous sensor devices.

Let us assume a network of several sensor devices measuring the same physical phenomena. We assume all sensor devices to be at level 1 according to Schneider et al. (2019) and provide measurements with an identical unit. In this article, we use the term sensor device to describe any instrumentation that provides observations in space and time.

The measurements of these sensor devices are described by five characteristics: their spatial and temporal resolution, their location, their sampling frequency, and their quality. These characteristics are intrinsic to the device, for example, the quality of the sensor, the electronic hardware, the mechanical structure, the programmatic procedures, and the telecommunication method, to name a few. In our study, we assume sensor devices to be mobile, and being able to provide, for each sampling, a constant measurement over an area surrounding their location. The characteristics of the measurements used in this study are schematically presented in Figure 1.

#### 2.1.1. Measurement resolution

Without losing generality, we focus here on the spatial resolution of measurement characterized by a shape in Figure 1. A crossed circle represents a sensor device whose measure is representative at this point, and a square represents a sensor whose measure represents an average of the phenomena surrounding this area. The larger the size of the shape the larger the domain of the average and the lower is the spatial resolution. We assume each type of sensor device having an area representative of their measurement to be of any shape. Instead of using measurement resolution as key, and to increase the amount of information, we assume any points located under this shape to be constant. Each value within the shape is then characterized by the same measurement quality.

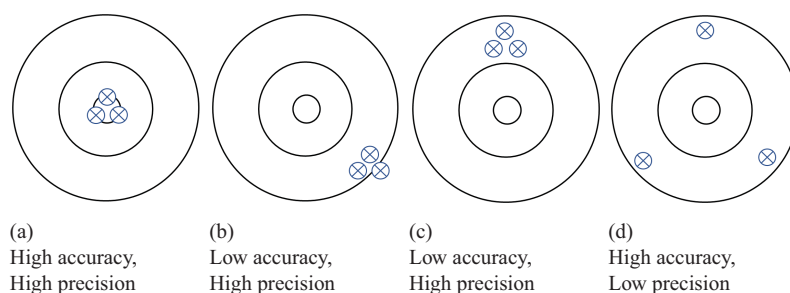


**Figure 1.** Sequence of two timesteps of a network of heterogeneous sensors measuring spatial phenomena. A crossed circle represents a sensor device whose measurement is representative at this point. A square represents a sensor whose measure represents an average of the phenomena surrounding this area. The quality of the measurement goes from high (color green), to medium (color orange) and to low (color red). Sensor device might be mobile and not providing measurement at every timestep.

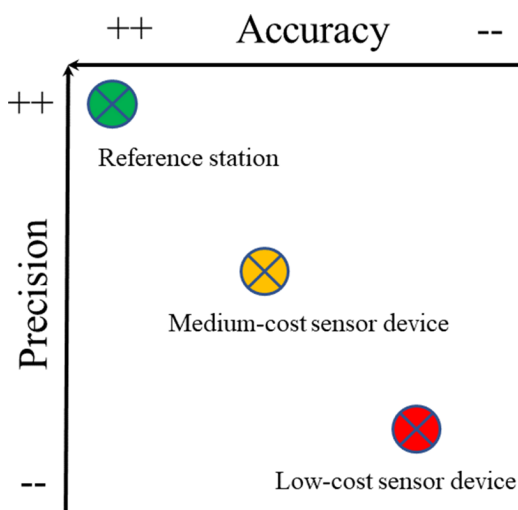
### 2.1.2. Measurement quality

As illustrated in Figure 1, the quality of the measurement ranges from high (color green), to medium (color orange), and finally to low (color red). For example, sensor devices designed as green crossed circle can be seen as reference monitoring stations and red crossed circle as low-cost sensor devices. Accuracy and precision are used in this study to quantify the quality of the measurements, and are schematically explained in Figure 2. A comparison of several sensor devices is possible by incorporating their specific accuracy and precision as depicted in Figure 3. The higher the accuracy and the precision, the closer is the measure to ground truth. Both accuracy and precision are considered as metrics processed over time. Because the ground truth is unknown, these metrics are determined with the measurements of the sensor device against the ones of a reference device whose measurements are of high quality. We assume that several items of a specific sensor device provided by one manufacturer get identical accuracy and precision. Realistically, we assume each type of sensor device to be tested beforehand either in laboratory conditions or by co-location with a reference sensor device (Castell et al., 2017; Schneider et al., 2018; Vogt et al., 2021). Accuracy and precision are thus *a priori* information about a sensor device.

In our study, we use root mean square error (RMSE) as accuracy and variance as precision. Large values of RMSE imply low accuracy and small values of RMSE implies high accuracy. In addition, large



**Figure 2.** Schematic description of accuracy and precision for three identical sensor devices measuring one phenomenon. The center of each target represents the phenomena to be measured. Case (a) sensor devices with high accuracy and high precision, cases (b) and (c) sensor devices with low accuracy and high precision, and case (d) sensor devices with high accuracy and low precision.



**Figure 3.** Accuracy-precision diagram representing the measurement quality of three types of sensor devices: a reference station, a medium-cost sensor device, and a low-cost sensor device.



values of variance imply low precision, and small values of variance imply high precision. The expression of RMSE reads:

$$RMSE(\hat{V}, V) = \sqrt{\frac{1}{N} \sum_i^N (\hat{V}_i - V_i)^2} \quad (1)$$

with  $\hat{V}_i$  the measurement of the sensor device at time  $i$  and  $V_i$ , the measurement of the reference sensor device, and  $N$  the amount of timesteps.

And the expression of variance reads:

$$variance(\hat{V}, V) = \frac{1}{N} \sum_i^N (|\hat{V}_i - V_i - bias|)^2 \quad (2.2)$$

where  $bias$  being determined as  $\frac{1}{N} \sum_i^N (\hat{V}_i - V_i)$ .

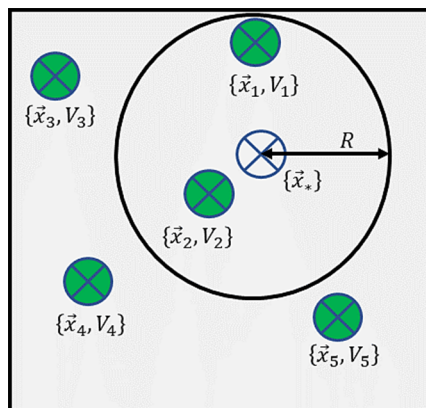
### 2.1.3. Measurement sampling and device mobility

As presented in Figure 1, any sensor device has a specific geographical location. Some sensor devices might have issues in providing a measurement at regular frequency; for instance, one red square and one orange crossed circle did not provide any measurement at timestep 2. Besides, some sensor devices might be mobile; for instance, two squares and two crossed circles moved from their original locations between timestep 1 and timestep 2. Each measurement is thus related to a location and a timestamp, used as keys. No measurement does not provide any information and will not be replaced by any fill-in methods.

## 2.2. Adaptive distance attention

This section aims at presenting adaptive distance attention as a framework for prediction that satisfies the measurement characteristics of a network of sparse and heterogeneous sensor devices.

Let us assume a network of  $k$  reference stations located in  $x_i$  with measurements  $V_i$  following a Normal distribution. We assume the mean and the standard deviation of the Normal distribution to be stationary. A general formulation of a spatial prediction that suits both OK (Wackernagel, 2003), and GRNN (Specht, 1991) is presented as follows: the prediction  $\hat{V}_*$  at location  $x_*$  is estimated from a) an ensemble of weights involving the Euclidean distance between the location of the target and the locations of the predictors  $\|x_*, x_i\|$ , that is, the distance of de-correlation of the phenomenon  $R$  and b) the value of the predictors  $V_i$ . Figure 4 illustrates this general formulation. The closer a station is to the prediction location, the higher is



**Figure 4.** Schematic illustration of a network where each station is identified as  $\{x_i, V_j\}$ , the location of the prediction is identified as  $x_*$ , and its area of similitude is characterized by a radius  $R$ .

its weight and thus the involvement of its value. For example, the prediction at location  $x_*$  will be characterized mostly by the value of the station located at  $x_2$ . The value located at  $x_1$  will get a lower impact. Besides, values from stations located at  $x_3, x_4$ , and  $x_5$  will get a minor weight due to their locations outside of the area of representativity delimited by the circle of radius  $R$ . Following this description, we write the prediction  $\hat{V}_*$  at location  $x_*$  as:

$$\hat{V}_* = \sum_{i=1}^k A\left(\frac{\|x_*, x_i\|}{R}\right) V_i \quad (2.3)$$

where  $A$  represents the attention weight based on a score involving the Euclidean distance between  $x_i$  and  $x_*$ .

We replace the expression  $1/R$  by  $W$  to avoid any issues of division by zero while training a model in Section 2.6. We call  $W$  the learnable parameter. Finally, following the notation related to attention in Vaswani et al. (2017), we call  $x_*$  the query  $Q$  and  $x_i$  the key  $K$ . Then, expression 2.3 writes:

$$\hat{V}_Q = \sum_{i=1}^k A(\|Q, K\|W) V_i \quad (2.4)$$

### 2.2.1. Multi-dimension

As done in Kyriakidis and Journal (1999) and (Li et al. (2020)), the query  $Q$  and the key  $K$  can represent both space and time. More generally, we let  $Q$  and  $K$  represent a  $d$ -dimensional space. Besides,  $Q$  and  $K$  can both represent multiple locations. Thus, we have  $Q$  a matrix  $\in \mathbb{R}^{q \times d}$ , and  $K$  a matrix  $\in \mathbb{R}^{k \times d}$ .

### 2.2.2. Adaptive parameter

In the  $d$ -dimension, processing the attention weight function faces an anisotropy effect. Thus, it requires the learnable parameter  $W$  to be adaptive (Robert et al., 2013) to each dimension of  $Q$  and  $K$ . More generally, we let the parameter be adaptive for any points of  $K$  and  $Q$ . Consequently, we have two learnable parameters depending, respectively, of  $Q$  and  $K$ , written, respectively,  $W_Q \in \mathbb{R}^{q \times d}$  and  $W_K \in \mathbb{R}^{k \times d}$ .

### 2.2.3. Multivariable

We let  $V$  and  $\hat{V}$  describe more than one variable  $v$ . Nonetheless, we keep the number of variables identical for both the value  $V$  and the prediction  $\hat{V}$ . Thus, we see  $V$  being a matrix  $\in \mathbb{R}^{k \times v}$  and  $\hat{V}$  a matrix  $\in \mathbb{R}^{q \times v}$ .

Our multivariable prediction with a multidimensional adaptive attention then reads:

$$\hat{V}_Q = \sum_{i=1}^k A(\|W_Q Q, W_K K_i\|) V_i \quad (2.5)$$

## 2.3. GRNN and OK as adaptive distance attention

This section is dedicated to the integration of two prediction methods, namely GRNN and OK, within the context of an adaptive distance attention framework.

GRNNs follow the Nadaraya–Watson kernel regression (Nadaraya, 1964; Watson, 1964):

$$\hat{V}_Q = \sum_{i=1}^k \frac{K_R(Q, K)}{\sum_{j=1}^k K_R(Q, K)} V_i \quad (2.6)$$

where  $K_R$  represents a kernel with bandwidth  $R$ .

GRNN is based on an isotropic radial basis function  $e^{-\|Q, K\|^2 / 2R^2}$  as parametric kernel. Finally, by using the softmax expression  $e^{u_i} / \sum_{j=1}^k e^{u_j}$ , expression 2.5 becomes:

$$\hat{V}_Q = \sum_{i=1}^k A_{S,G}(\|W_Q Q, W_K K\|) V_i \quad (2.7)$$

with subscript  $S$  as softmax function and superscript  $G$  as Gaussian kernel.

**OK** gets its attention weights by solving the kriging system  $\lambda_Q = \Lambda^{-1} \Lambda_Q$ , where  $\Lambda$  represents the semi-variogram matrix. In the case of  $Q$  representing a single location, we obtain:

$$\lambda_Q = \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_k \\ \mu \end{pmatrix}, \Lambda = \begin{pmatrix} \Lambda_{1,1} & \cdots & \Lambda_{1,k} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \Lambda_{k,1} & \cdots & \Lambda_{k,k} & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix}, \Lambda_Q = \begin{pmatrix} \Lambda_{1,Q} \\ \vdots \\ \Lambda_{k,Q} \\ 1 \end{pmatrix} \quad (2.8)$$

where  $\mu$  is a Lagrangian multiplier and  $\Lambda_{ij}$  represents the semi-variogram between location  $x_i$  and  $x_j$

Assuming an existing function  $\text{linalg}$  that solves the linear kriging system and using the case where the variogram follows an exponential function  $1 - e^{-\|x_* - x_i\|W}$  of variance 1 and range  $1/W$ , expression 2.5 becomes:

$$\hat{V}_Q = \sum_{i=1}^k A_{L,E}(\|W_Q Q, W_K K\|) V_i \quad (2.9)$$

with subscript  $L$  as  $\text{linalg}$  function and superscript  $E$  as an exponential kernel.

Although alternative kernel types and semi-variograms are applicable to GRNN and OK, our emphasis lies on the utilization of simpler variants. Specifically, we prioritize those that a) allow learnable parameters to impart richness to the structure and b) mitigate the risk of encountering issues related to infinite loss during model training.

#### 2.4. Data fusion model architecture

This section is dedicated to introducing our data fusion approach for predicting values, along with a detailed exploration of its underlying model architecture.

Our data fusion uses a similar approach as cross-kriging (Journel and Huijbregts, 1978) in the distance attention framework. An attention weight is processed using queries  $Q$  and keys  $K$  belonging to two different networks. During the training phase, a first network, called  $X$ , provides  $Q$  and their respective values as targets, and network  $Y$  provides  $K$  and values  $V$ . During the prediction phase, the network  $X$  only provides  $Q$ , and network  $Y$ , provides  $K$  and  $V$ . Given expressions 2.7 and 2.9, the data fusion expression reads:

$$\hat{V}_Q = W_O \sum_{i=1}^k A(\|W_Q Q, W_K K_i\|) V_i \quad (2.10)$$

where  $A$  is an adaptive distance attention and can be either  $A_{S,G}$  or  $A_{L,E}$ , and  $W_O$ , called the learnable parameter of the output, makes the expression able to adapt in case of trend in the measurement; it writes  $W_O \in \mathbb{R}^{q \times v}$ . A visualization of the model architecture is shown in Figure 5.

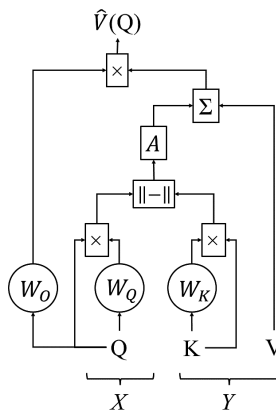
The learnable parameters  $W_O$ ,  $W_Q$ , and  $W_K$  are the outputs of three multilayer perceptrons. Each multilayer perceptron reads:

$$MLP(X) = \begin{cases} H^{(1)} = \sigma(W^{(1)}X + b^{(1)}) \\ H^{(2)} = \sigma(W^{(2)}H^{(1)} + b^{(2)}) \\ O = \sigma(W^{(3)}H^{(2)} + b^{(3)}) \end{cases} \quad (2.11)$$

where  $X$  is the input matrix,  $W^{(i)}$  are the hidden-layer weights matrices,  $b^{(i)}$  is the bias vectors, and  $\sigma(\cdot)$  is the ReLU activation function. We have thus:

$$W_O = MLP_O(Q) \quad (2.12)$$

with  $W_O^{(1)} \in \mathbb{R}^{d \times h}$ ,  $W_O^{(2)} \in \mathbb{R}^{h \times h}$ ,  $W_O^{(3)} \in \mathbb{R}^{h \times v}$ ,  $b_O^{(1)} \in \mathbb{R}^{1 \times h}$ ,  $b_O^{(2)} \in \mathbb{R}^{1 \times h}$ ,  $b_O^{(3)} \in \mathbb{R}^{1 \times v}$



**Figure 5.** Schematic illustration of the data fusion architecture. Network  $X$  provides input  $X$  and network  $Y$  provides inputs  $K$  and  $V$ . Learnable parameters are  $W_O$ ,  $W_Q$ , and  $W_K$ . Attention is represented by the symbol  $A$  without any subscript to ease the reading. The prediction at query  $Q$  is symbolized as  $\hat{V}$ .

$$W_K = \text{MLP}_K(K) \quad (2.13)$$

with  $W_K^{(1)} \in \mathbb{R}^{d \times h}$ ,  $W_K^{(2)} \in \mathbb{R}^{h \times h}$ ,  $W_K^{(3)} \in \mathbb{R}^{h \times d}$ ,  $b_K^{(1)} \in \mathbb{R}^{1 \times h}$ ,  $b_K^{(2)} \in \mathbb{R}^{1 \times h}$ ,  $b_K^{(3)} \in \mathbb{R}^{1 \times d}$

$$W_Q = \text{MLP}_Q(Q) \quad (2.14)$$

with  $W_Q^{(1)} \in \mathbb{R}^{d \times h}$ ,  $W_Q^{(2)} \in \mathbb{R}^{h \times h}$ ,  $W_Q^{(3)} \in \mathbb{R}^{h \times d}$ ,  $b_Q^{(1)} \in \mathbb{R}^{1 \times h}$ ,  $b_Q^{(2)} \in \mathbb{R}^{1 \times h}$ ,  $b_Q^{(3)} \in \mathbb{R}^{1 \times d}$ .

Ultimately, we address the challenge of overfitting by implementing dropout with a probability of  $p$  specifically applied to the attention mechanism.

#### 2.4.1. Models overview

We highlight 12 data fusion models in the adaptive distance attention framework following expression 2.10. They differ both in terms of attention  $A_{LE}$  or  $A_{SG}$ , as well as with different assumptions simplifying expressions 2.12, 2.13, and 2.14: i) the learnable parameters of each dimension of  $Q$  and  $K$  are either constant or not ii) the presence or absence of the learning parameter  $W_O$ , iii) both networks  $X$  and  $Y$  measure a physical phenomenon with identical or different spatial structures. For example, a model with attention  $A_{LE}$ , in absence of  $W_O$ , and where networks  $X$  and  $Y$  are measuring a physical phenomenon with identical spatial structure is a data fusion approach based on OK. An overview of each model with their name, their attention, and the characteristics of their learnable parameters is given in Table 1. For readability, we designate models incorporating a kriging system like OK as “krig,” and models involving the Nadaraya–Watson kernel, such as GRNN, as “NW.” The addition of the “NN” suffix to the name signals the involvement of learnable parameters through multilayer perceptrons.

#### 2.5. Cases studies

This section aims at describing the three case studies of this article. It describes first the phenomena that synthetic sensor devices will measure, the construction of heterogeneous networks of mobile sensor devices, and finally the presence of several networks used for the experimentation.

For each case study, we assume a phenomenon representing ground truth to be constant in time and provided by a model or a dataset  $M$ . We chose three case studies spanning a spectrum of complexities, ranging from simple to intricate. The complexity is related to the ground truth to be measured, and the spatial resolution of the measurements of the sensor devices.

**Table 1.** Description of the 12 data fusion models, with their name, their attention, and the characteristics of their learnable parameters

Learnable parameters				
Name	Attention	Constant	Conditions	Number of parameters
krig	$A_{L,E}$	yes	$W_O = 1, W_K = W_Q$	$d$
NW	$A_{S,G}$	yes	$W_O = 1, W_K = W_Q,$	$d$
krig2	$A_{L,E}$	yes	$W_O \neq 1, W_K = W_Q$	$v + d$
NW2	$A_{S,G}$	yes	$W_O \neq 1, W_K = W_Q$	$v + d$
krig3	$A_{L,E}$	yes	$W_O \neq 1, W_K \neq W_Q$	$v + 2d$
NW3	$A_{S,G}$	yes	$W_O \neq 1, W_K \neq W_Q$	$v + 2d$
krigNN	$A_{L,E}$	no	$W_O = 1, MLP_K = MLP_Q$	$h(2d + h + 2) + d$
NWNN	$A_{S,G}$	no	$W_O = 1, MLP_K = MLP_Q$	$h(2d + h + 2) + d$
krigNN2	$A_{L,E}$	no	$W_O = MLP_O, MLP_K = MLP_Q$	$h(3d + 2h + v + 4) + v + d$
NWNN2	$A_{S,G}$	no	$W_O = MLP_O, MLP_K = MLP_Q$	$h(3d + 2h + v + 4) + v + d$
krigNN3	$A_{L,E}$	no	$W_O = MLP_O, MLP_K \neq MLP_Q$	$h(5d + 3h + v + 6) + v + 2d$
NWNN3	$A_{S,G}$	no	$W_O = MLP_O, MLP_K \neq MLP_Q$	$h(5d + 3h + v + 6) + v + 2d$

2.5.1. Simplistic

This simplistic ground truth evolves over an area with dimensions  $x \in [0; 1]$  and  $y \in [0; 1]$ . Its spatial area is 1 unit, and its spatial resolution is  $2.5 \cdot 10^{-5}$  unit. Its values follow the expression  $V(x, y) = \cos^2(2\pi x) + \sin^2(2\pi x)$  (Figure 6).

2.5.2. Topography

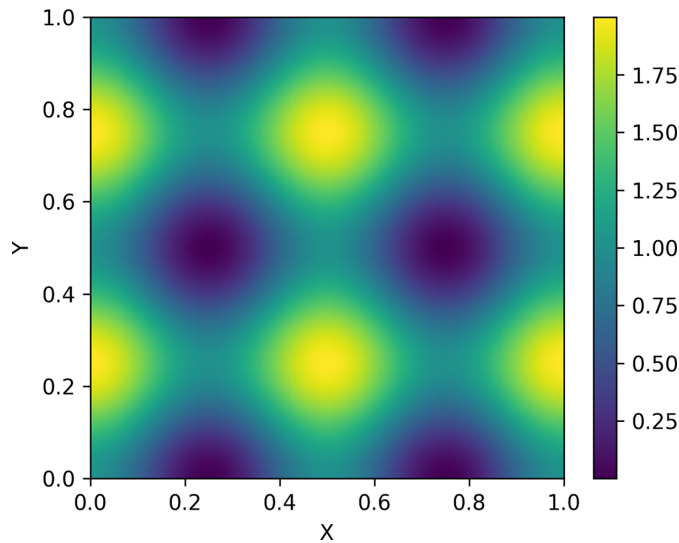
This ground truth is a subset of the 25-m spatial resolution Digital Elevation Model EU-DEM v1.1 (Copernicus, 2016) over an area of  $196 \text{ km}^2$  with x-coordinates between 4342031 m and 4356031 m and y-coordinates between 4085001 m and 4099001 m in the reference-system EPSG:3035 (Figure 7).

2.5.3. Annual hourly nitrogen dioxide concentration

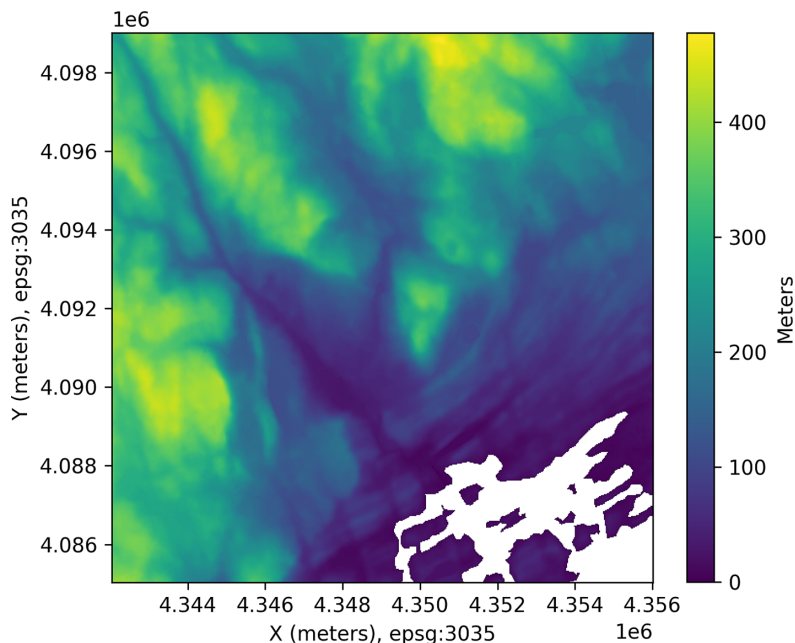
This ground truth is the result of an annual average of hourly nitrogen dioxide concentration (AH  $\text{NO}_2$ ) in 2019 over an area of  $1026 \text{ km}^2$  over the Oslo metropolitan region resulting from the simulation using the EPISODE dispersion model (Hamer et al., 2020). EPISODE is a two-step model: first, a 3D Eulerian model provides a  $1\text{-km}^2$  spatial resolution grid, and then a sub-grid model using preprocessed point and line source emissions provides  $\text{NO}_2$  concentrations at, in this case, 21209 point locations (also called “receptor points”). The spatial density of these locations is irregularly distributed: most of the information is located over the urban areas and on main roads with large sources of traffic-related  $\text{NO}_2$  emissions and strong spatial gradients in pollution patterns. Outside of these areas, the spatial density of the output is lower, with receptor points distributed at every 1 km. Instead of processing a spatial interpolation between each location to get a grid of 100-m spatial resolution over the whole area of interest, we directly exploit the receptor point data (Figure 8).

2.5.4. Heterogeneous networks of mobile sensor devices

This section presents the characteristics used in creating synthetic mobile sensor devices in order for them to be as representative as possible compared to environmental sensors as described in Section 2.1 and illustrated in Figure 1. We use six types of sensor devices, whose type is characterized by their spatial resolution and their measurement quality. A network is composed of  $k$  sensor devices moving over  $N$



**Figure 6.** Illustration of the static two-dimensional ground truth used in case study “Simplistic.”



**Figure 7.** Illustration of the static two-dimensional ground truth used in case study “Topography.” It represents a subset of the 25-m spatial resolution Digital Elevation Model EU-DEM v1.1 over an area of 196 km<sup>2</sup> in Norway.

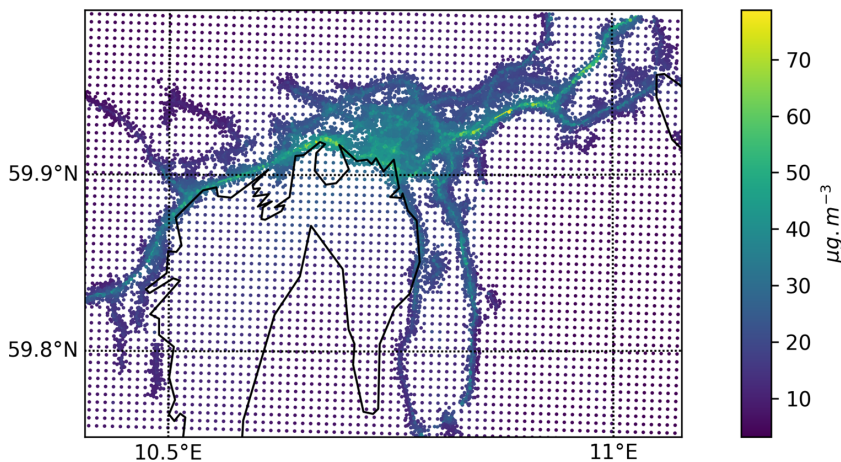
predefined locations uniformly distributed over the domain for each case study. At each timestep,  $k$  sensor devices randomly chosen among the six types of sensor devices provide one measurement each.

The sensor devices are characterized by two types of measurement spatial resolution. The first type is chosen identical to the spatial resolution of the dataset. The second type is at a coarser spatial resolution where a measurement is a spatial average over a square area. None of the sensor devices provides measurement as multi-pixel but only as a single pixel. For the case studies of the simplistic phenomena



**Table 2.** Information about measurement spatial resolution for the three cases studies

Spatial resolution type	Case studies		
	Simplistic	Topography	AHNO <sub>2</sub>
High	$2.5 \cdot 10^{-5}$	25 m	Point <sup>a</sup>
Low	$3.44 \cdot 10^{-2}$	482.75 m	1000 m <sup>b</sup>

<sup>a</sup>Output from the sub-grid model.<sup>b</sup>Output from the 3d-Eulerian model.**Figure 8.** Illustration of the static two-dimensional ground truth used in case study “AH NO<sub>2</sub>.” It represents the annual average of hourly nitrogen dioxide concentrations in 2019 over the Oslo metropolitan region simulated with the EPISODE dispersion model.

and the topography, a spatial average is processed on the ground truth. For the case study AH NO<sub>2</sub>, the data with a coarser spatial resolution comes from the 3D-Eulerian model and data with the higher resolution comes from the sub-grid model. The characteristics of the measurement spatial resolution for the three case studies are presented in Table 2.

The sensor devices are characterized by three types of measurement quality: high, medium, and low. We produce measurements  $V_t^D$  for a sensor device at a specific time and location by adding uncertainty to the ground truth  $M_t^D$  with spatial resolution  $D$ , following a Normal distribution in a total error framework (Working Group on Guidance for the Demonstration of Equivalence, 2010; Lepioufle et al., 2021). Thus, the measurements of a sensor device are given by:

$$V_t^D \sim \mathcal{N}(\beta_0 \cdot M_t^D + \beta_1, \sigma_{g,t}^2 + \sigma_{h,t}^2 + \sigma_r^2) \quad (2.15)$$

Given ground truth to be perfect, its structural error is nil, and so is its standard deviation  $\sigma_{h,t}$ . The standard deviation  $\sigma_{g,t}$  is the parameter of the sensor device error. It is chosen proportionally to the measurements as used in Ref. (Translation of the Report on the Suitability Test of the Ambient Air Measuring System, 2007), that is,  $\sigma_{g,t} = M_t^D \sigma_g$ . It is usually determined as a percentage. The choice of the parameter  $\sigma_g$  follows Refs. (Translation of the Report on the Suitability Test of the Ambient Air Measuring System, 2007; Directive 2008/50/EC, 2008). For instance, a reference monitoring station has  $\sigma_g \leq 5\%$ , and a low-cost sensor device has  $\sigma_g \geq 30\%$ . In addition, we make the simplifying assumption that the sensor of the device does not exhibit any aging effect, that is,  $\beta_0$  is equal to one and  $\sigma_g$  is constant over time. In our case, errors due to both external effects (e.g., meteorology, environment) and internal



**Table 3.** Parameters related to the three measurement quality types

Measurement quality type	$\sigma_g$ (in %)	$\beta_1$	$\sigma_r$
High	2	1	1
Medium	10	2	2
Low	30	5	5

Note. Case study “Simplistic” sees its parameters  $\beta_1$  and  $\sigma_r$  multiplied by  $1.10^2$ .

effects (e.g., mechanical and electronic components) are represented in a remnant error characterized by the parameters  $\beta_1$  and  $\sigma_r$ . We assume the sensor devices of the first type to run on an internal system characterized as robust, the second type on a medium one, and the third on a weak one. Besides, we assume the external environment of the three case studies to be different: from gentle for the “Simplistic” case study to difficult for “AH  $NO_2$ ” case study. As a consequence, the external environment affects the signal of the sensor devices by amplifying the error related to the internal system. We chose the parameters of the remnant error in a heuristic manner. Parameters describing the three types of measurement quality are given in Table 3. Remnant error parameters remain identical for the case studies “Topography” and “AH  $NO_2$ .” Nonetheless, given the lower values for case study “AH  $NO_2$ ,” these parameters will get a higher impact on the signal of the sensor device. In addition, based on empirical testing, the remnant error parameters are multiplied by  $1.10^2$  for case “Simplistic” to keep the effect of the external environment gentle.

Finally, we describe a sensor device by its two characteristics: its spatial resolution ( $R$ ) and its measurement quality ( $Q$ ), to each of which we add a subscript to describe the type of characteristics: high ( $H$ ), medium ( $M$ ), and low ( $L$ ). Thus, the six types of sensor device read  $R_HQ_H$ ,  $R_HQ_M$ ,  $R_HQ_L$ ,  $R_LQ_H$ ,  $R_LQ_M$ , and  $R_LQ_L$ .

2.6. Experimentation plan

In this section, we describe our experimentation. It consists of testing our model architecture, described in Section 2.4 on the three case studies described in Section 2.5.

2.6.1. Seven heterogeneous networks

For each case study, the experiment is based on seven heterogeneous networks of sensor devices, all distinct from each other. Our data fusion model requires two networks  $X$  and  $Y$  for the three phases: the simultaneous training and validation phases, and the evaluation phase of the prediction model. In addition, one last network is used to assess the measurement quality as it is carried out for real measurement campaigns, either with co-location or in a laboratory with a climatic chamber: every measurement of one type of sensor device is compared to a reference instrument representing a high quality point sensor device. The resulting metrics are then used as *a priori* information about the measurement quality of the sensor device.

2.6.2. Networks  $X$  and  $Y$

During the training, validation, and evaluation phases, both networks  $X$  and  $Y$  are built-up in the same manner. They consist of 600 sensor devices moving across 1000 fixed locations. For each network  $X$  and  $Y$ , several sensor devices might occupy the same location. However, the 1000 locations of the sensor devices in  $X$  will be different from the 1000 locations in  $Y$ . At each timestep, 100 sensor devices randomly chosen within network  $X$  and 100 sensor devices randomly chosen within network  $Y$  provide measurements of one variable. We have, thus,  $q = k = 100$  and  $v = 1$ . The network used for the calibration consists of 600 sensors representing the six measurement characteristics co-located with sensor devices with high-quality point measurements. Ground truth being constant over time, we do not use time as a dimension

describing the values. Therefore, the keys  $K$  and the queries  $Q$  are represented in a four-dimensional space, that is, the x-coordinate (shortened as  $x$ ), the y-coordinate (shortened as  $y$ ), the *a priori* accuracy (shortened as  $acc$ ), and the *a priori* precision (shortened as  $prec$ ), thus  $d = 4$ .

### 2.6.3. Model architecture

For every model architecture, we use hidden layers of size  $h = 32$ . According to Table 1, we have the models “krig” and “NW” described by 4 parameters, the models “krig2” and “NW2” with 5 parameters, the models “krig3” and “NW3” with 9 parameters, the models “krigNN” and “NWNN” with 1860 parameters, the models “krigNN2” and “NWNN2” with 2597 parameters, and the models “krigNN3” and “NWNN3” with 3945 parameters. We use a dropout of  $p = 0.1$ . In addition, for models using  $A_{LE}$ , solving the kriging system with sensor devices of network  $Y$  with potentially identical locations will result in nonuniqueness of the solution. We overcome this issue by using a function *linalg* that computes a solution to the least squares problem of the kriging system. Finally, we write the learnable parameters  $W_K$  and  $W_Q$  related to each dimension as  $W_{\cdot,x}$ ,  $W_{\cdot,y}$ ,  $W_{\cdot,acc}$ , and  $W_{\cdot,prec}$  where the dot determines either  $K$  or  $Q$ .

### 2.6.4. Training, validation, and evaluation

The training of the models is done using the optimization algorithm Adam (Kingma and Ba, 2014) with a learning rate of  $1 \cdot 10^{-3}$ . We use mean square error (MSE) as loss while training and validating the models. We use 200 epochs with an early exit stopping the training phase if the loss does not improve during 20 consecutive epochs with the validation dataset. During training, validation, and evaluation, the prediction is established with standardized  $V$ ,  $Q$ , and  $K$ . During training and validation, the losses are processed by keeping standardized outputs while this is not the case during the evaluation phase. Metrics such as RMSE, variance and coefficient of determination ( $R^2$ ) are used to evaluate the prediction of the models.

### 2.6.5. Experimentation

The first part of the experiment consists of evaluating the 12 models of Section 2.4 on the three case studies described in Section 2.5 with heterogeneous sensor devices as input and with high quality point measurement data as target. The six types of sensor devices, as described in Section 2.5, are equally represented. Besides, each location of network  $X$  can be predicted using several sets of 100 sensor devices of network  $Y$  as input. We thus produce an ensemble of predictions for each location of network  $X$  and evaluate the median of the ensemble during the evaluation phase. Hereafter, to enhance clarity, we use the terms single prediction and ensemble median, respectively.

The second part of the experiment focuses on highlighting and visualizing the effect of the model architecture, the sequence over time of the mobile sensor device locations, and their characteristics on the learnable parameters and the predictions. We focus on the “Topography” case study, and krigNN2 and NWNN2 as model architectures. These models provide good metrics with a reasonable amount of parameters for non-simplistic phenomena. For this experiment, the 100 sensor devices belonging to network  $X$  can move on 1000 predetermined locations. The same applies to the 100 sensor devices belonging to network  $Y$ . Only the sequence over time of the location of the sensor devices and their characteristics change. We train four models (two krigNN2 and two NWNN2) using four distinct sequences of mobile sensors. For each of the four trained models, we produce an ensemble of predictions using network  $X$ . Finally, for each model architecture, we highlight the difference by comparing i) the maps of the learnable parameters and ii) the maps of the dispersion of the members of the ensemble. Quantifying the dispersion of the ensemble is done by producing two maps; a first one by subtracting the 5-percentile of the ensemble to the median on every point of prediction, and a second-one by subtracting the median to its 95-percentile. The first map is called the lower dispersion and the second is called the upper dispersion. Finally, iii) the maps of the metrics (RMSE and variance) between the members of the ensemble and the observation at each location. To ease the visualization, the prediction is inferred over 6400 locations uniformly distributed over the area of the case study.

2.6.6. Implementation

We developed the Python package *Steams* based on Pytorch (Paszke et al., 2019), and ran the experiment on a machine equipped with an Intel Core i5-9500 CPU @ 3.00GHz x 6.

3. Results and discussion

3.1. Metrics for the different data fusion models applied to the three case studies

We present the metrics of the 12 models for the 3 case studies in Tables 4–6 for the case studies “Simplistic,” “Topography,” and “AH NO<sub>2</sub>,” respectively. Generally, involving deep neural networks in the learnable parameters  $W_K$ ,  $W_Q$ , and  $W_O$  has a positive impact on the metrics of single prediction. Nonetheless, each case study gets its proper metrics pattern: a light impact on the metrics for case study “Simplistic” with coefficient of determination going from 0.65 to 0.89, a strong impact on the metrics for case study “Topography” with coefficient of determination going from −1.09 to 0.91, and an average impact on the metrics for case study “AH NO<sub>2</sub>” with a coefficient of determination going from −6.43 to 0.69. Furthermore, increasing the amount of parameters of a model architecture does not automatically increase the metrics. In addition, using the ensemble median increases the metrics RMSE and variance. Nonetheless, regarding metric  $R^2$ , it tends to increase this metric for positive values and worsen it for negative ones. Finally, for single prediction, the model NWNN3 provides better metrics for case study “Simplistic” and “Topography.” For the ensemble median, model NWNN3 provides better metrics for case study “Simplistic,” and both models NWNN2 and NWNN3 provide close metrics for case study “Topography.” For case study “AH NO<sub>2</sub>” and single prediction, the model krigNN2 provides better RMSE and variance and the model NWNN2 provides a better coefficient of determination. For the ensemble median, the model krigNN provides better RMSE and variance and NWNN2 provides a better coefficient of determination.

As an illustration, we present, for each case study and based on the model with the best RMSE and variance metrics, a prediction on 6400 locations made by 100 heterogeneous sensor devices chosen randomly over the area. Case study “Simplistic” has its prediction based on model NWNN3 and is shown in Figure 9. Case study “Topography” has its prediction based on model NWNN3 and is shown in Figure 10. Finally, case study “AH NO<sub>2</sub>” has its prediction based on model krigNN2 and is shown in Figure 11. Generally, the prediction well reproduces the phenomena presented as ground truth and keep prediction values in the same range of values of the ground truth.

**Table 4.** Metrics of the data fusion models for case study “Simplistic”. Bold values represent the best metrics.

Name	Single prediction			Ensemble median		
	RMSE	variance	R <sup>2</sup>	RMSE	variance	R <sup>2</sup>
krig	0.23	0.03	0.65	0.17	0.02	0.77
NW	0.26	0.04	0.54	0.18	0.02	0.72
krig2	0.22	0.02	0.76	0.13	0.01	0.89
NW2	0.25	0.04	0.63	0.16	0.01	0.82
krig3	0.22	0.03	0.75	0.15	0.01	0.87
NW3	0.27	0.04	0.60	0.18	0.02	0.77
krigNN	0.22	0.03	0.71	0.19	0.02	0.77
NWNN	0.16	<b>0.01</b>	0.85	0.13	0.01	0.89
krigNN2	0.16	<b>0.01</b>	0.84	0.13	0.01	0.88
NWNN2	0.15	<b>0.01</b>	0.85	0.13	0.01	0.89
krigNN3	0.16	<b>0.01</b>	0.82	0.15	0.01	0.85
NWNN3	<b>0.13</b>	<b>0.01</b>	<b>0.89</b>	<b>0.11</b>	<b>0.00</b>	<b>0.93</b>

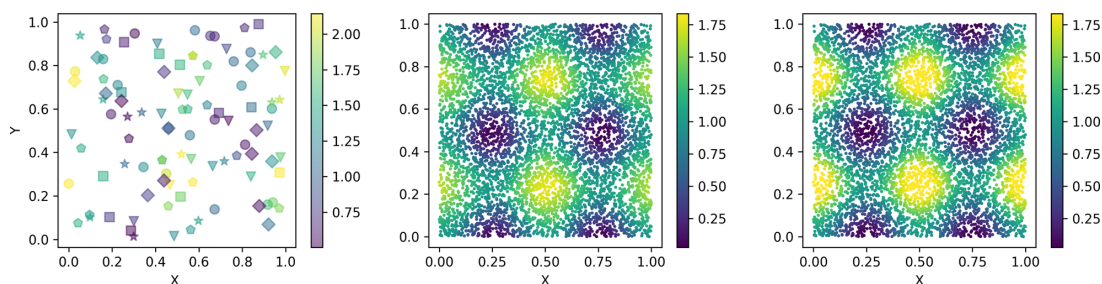
**Table 5.** Metrics of the data fusion models for case study “Topography”. Bold values represent the best metrics.

Name	Single prediction			Ensemble median		
	RMSE	variance	R <sup>2</sup>	RMSE	variance	R <sup>2</sup>
krig	84.14	4543.58	−0.62	75.29	3858.17	−0.69
NW	88.47	5031.97	−1.09	80.12	4436.96	−1.44
krig2	81.77	4207.19	0.09	66.21	2935.13	0.22
NW2	83.74	4400.30	0.01	67.93	3055.72	0.14
krig3	79.97	3980.34	0.15	64.28	2834.48	0.29
NW3	80.67	4190.81	0.13	66.17	2958.29	0.28
krigNN	36.04	727.87	0.87	32.13	589.05	0.90
NWNN	39.17	874.31	0.84	34.95	689.88	0.87
krigNN2	31.88	583.51	0.90	30.06	505.17	0.91
NWNN2	31.13	541.95	<b>0.91</b>	<b>29.25</b>	460.75	<b>0.92</b>
krigNN3	31.93	583.69	0.90	31.27	550.28	0.91
NWNN3	<b>30.96</b>	<b>495.90</b>	<b>0.91</b>	30.17	<b>454.74</b>	<b>0.92</b>

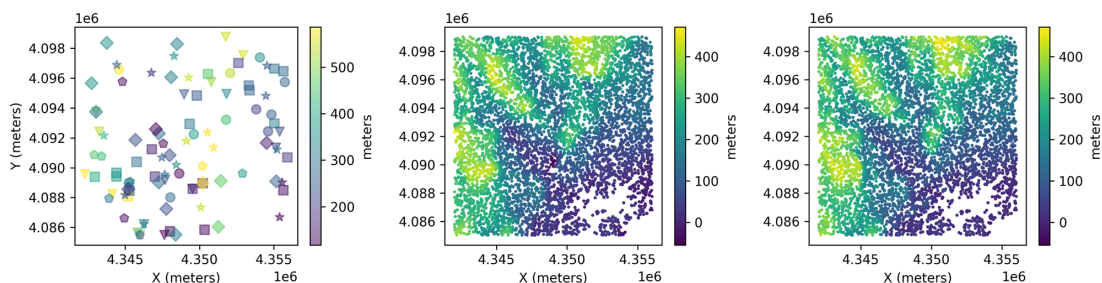
**Table 6.** Metrics of the data fusion models for case study “AH NO<sub>2</sub>”. Bold values represent the best metrics.

Name	Single prediction			Ensemble median		
	RMSE	variance	R <sup>2</sup>	RMSE	variance	R <sup>2</sup>
krig	10.93	65.94	−2.04	9.87	54.18	−9.11
NW	10.14	60.55	−3.41	9.52	54.28	−6.32
krig2	10.56	62.03	−6.43	10.27	60.33	−18.03
NW2	10.15	58.75	−2.38	9.34	51.29	−4.17
krig3	10.46	60.84	−5.84	10.20	59.68	−13.08
NW3	10.09	57.54	−2.10	9.32	50.49	−3.37
krigNN	6.12	18.97	0.61	<b>5.70</b>	<b>16.50</b>	0.65
NWNN	6.47	22.35	0.45	6.12	19.71	0.49
krigNN2	<b>5.90</b>	<b>18.63</b>	0.63	5.75	17.02	0.64
NWNN2	6.14	19.83	<b>0.69</b>	5.96	17.84	<b>0.70</b>
krigNN3	6.06	18.70	0.63	5.96	16.95	0.64
NWNN3	6.05	18.84	0.66	5.82	16.53	0.68

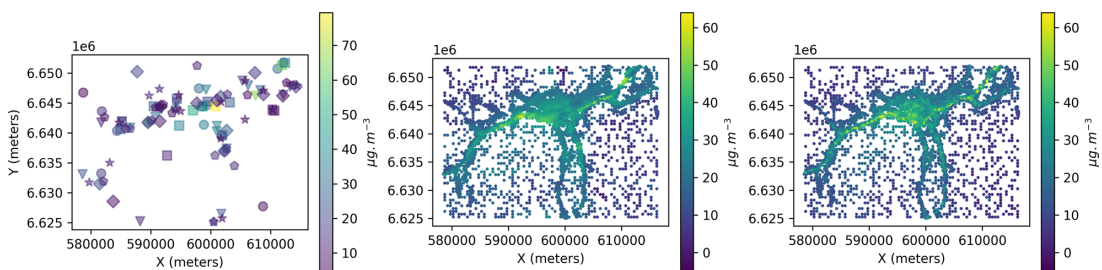
We show accuracy-precision diagrams to compare the metrics of the 12 models used for single prediction with the six types of sensor devices used as *a priori* information in Figures 12–Figure 14 for the case studies “Simplistic,” “Topography,” and “AH NO<sub>2</sub>,” respectively. Sensor devices of high spatial resolution and high measurement quality ( $R_H Q_H$ ) are chosen as reference. First, we observe that given the different case studies, the “quality” order of the different types of sensors differ. For example, a sensor device of low spatial resolution and of high measurement quality is the second best sensor device for case study “Simplistic,” and among the last ones for the case study “AH NO<sub>2</sub>.” In the case study “AH NO<sub>2</sub>,” a sensor device of low spatial resolution and of low measurement quality can get a better accuracy than a sensor device of low spatial resolution and of high measurement quality. Finally, we observe that models involving deep neural networks in the learnable parameters  $W_K$ ,  $W_Q$ , and  $W_O$  achieve



**Figure 9.** Illustration of one prediction for the case study “Simplistic” based on the measurements of 100 sensor devices (left). Each type of sensor device is described as a symbol: circle:  $R_H Q_H$ , triangle down:  $R_H Q_M$ , square:  $R_H Q_L$ , pentagon:  $R_L Q_H$ , star:  $R_L Q_M$ , and diamond:  $R_L Q_L$ . The prediction is based on the model NWN3 and is carried out at 6400 locations (middle). Ground truth on these 6400 locations is presented on the panel on the right.

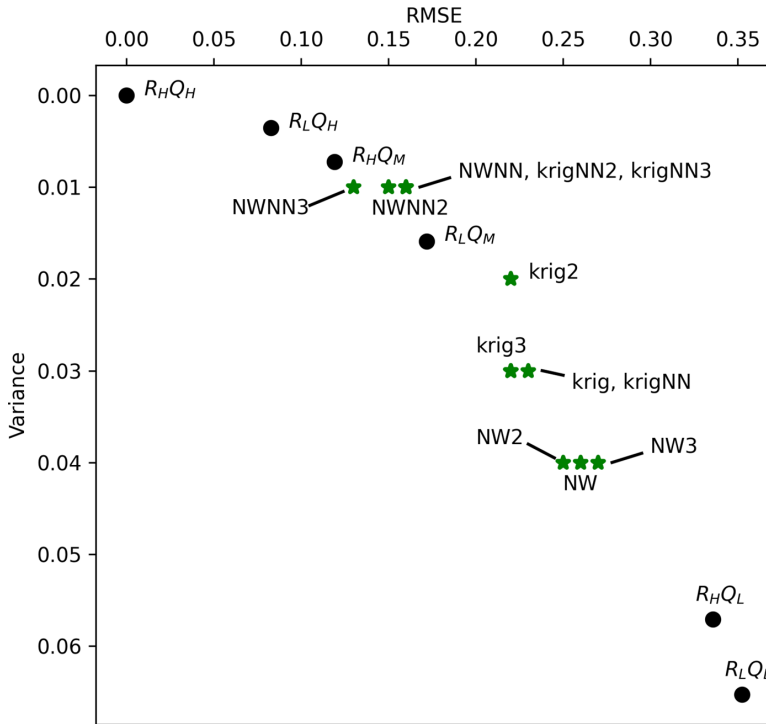


**Figure 10.** Illustration of one prediction for the case study “Topography” based on the measurements of 100 sensor devices (left). Each type of sensor device is described as a symbol: circle:  $R_H Q_H$ , triangle down:  $R_H Q_M$ , square:  $R_H Q_L$ , pentagon:  $R_L Q_H$ , star:  $R_L Q_M$ , and diamond:  $R_L Q_L$ . The prediction is based on the model NWN3 and is carried out at 6400 locations (middle). Ground truth on these 6400 locations is presented on the panel on the right.



**Figure 11.** Illustration of one prediction for the case study “AH  $\text{NO}_2$ ” based on the measurements of 100 sensor devices (left). Each type of sensor device is described as a symbol: circle:  $R_H Q_H$ , triangle down:  $R_H Q_M$ , square:  $R_H Q_L$ , pentagon:  $R_L Q_H$ , star:  $R_L Q_M$ , and diamond:  $R_L Q_L$ . The prediction is based on the model krigNN2 and is carried out at 6400 locations (middle). Ground truth on these 6400 locations is presented on the panel on the right.

metrics close to those from sensor devices of high spatial resolution and of medium measurement quality. Models not involving deep neural networks achieve metrics similar to a sensor device of high spatial resolution and low measurement quality for the case studies “Topography” and “AH  $\text{NO}_2$ ,” and between sensor devices of low spatial resolution and high measurement quality and sensor devices of high spatial resolution and low measurement quality for the case study “Simplistic.”



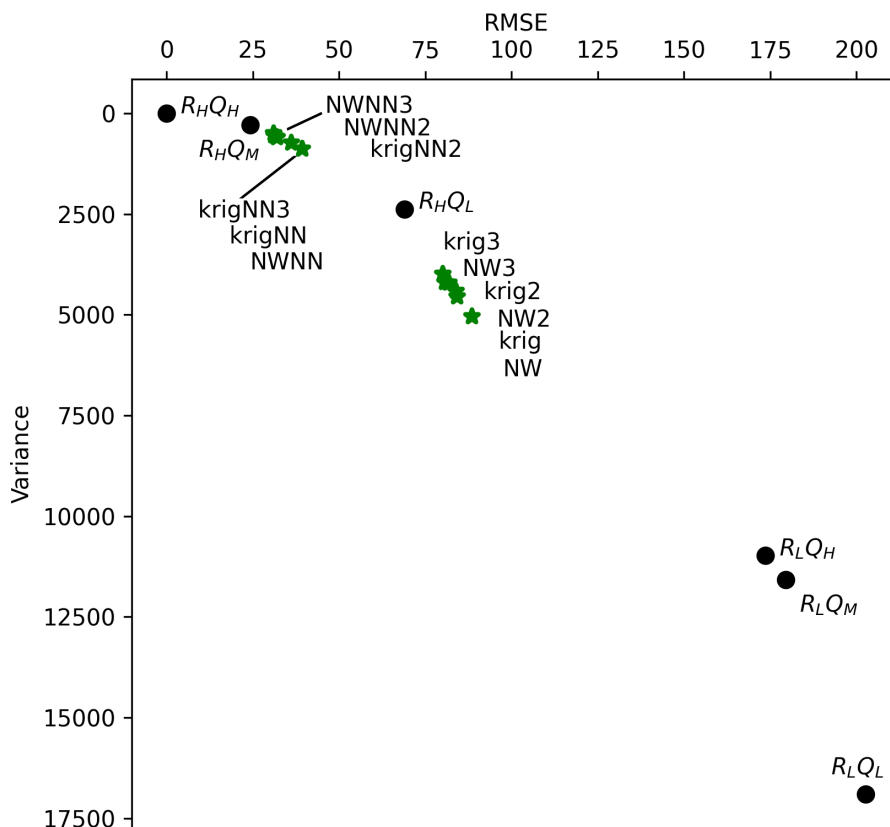
**Figure 12.** Accuracy-precision diagram for case study “Simplistic” with both a priori measurement quality related to the 6 types of sensor devices (black circles) and the metrics of the 12 data fusion models (green stars). Some green stars represent several models with identical metrics (e.g., NWNN, krigNN2, and krigNN3).

### 3.2. Variability of the learnable parameters from architecture krigNN2 and NWNN2 for case study “Topography”

We show the 2D representations of the learnable parameters  $W_K$  and  $W_O$  with sensor devices of high spatial resolution and high measurement quality as input in Figures 15 and 16, respectively. The figures represent the learnable parameters obtained using the two trained krigNN2 models and two trained NWNN2 models. Our attention models being based on the Euclidean distance, we ease the visualization, without changing any meaning, by plotting the absolute values of the learnable parameters related to their dimension. The higher the value of a learnable parameter, the higher the weight of its dimension in the distance attention. The maps of  $W_{K,x}$ ,  $W_{K,y}$ ,  $W_{K,acc}$ , and  $W_{K,prec}$  each have their own color scale to highlight the patterns and order of magnitude. For two trained models of identical architecture, we see their learnable parameters to have different patterns. Nonetheless, they all keep a coherent pattern distributed in space. In addition, their values respect an identical order of magnitude for each dimension. Finally, the learnable parameters corresponding to model architecture krigNN2 are constrained by its kriging system and are thus characterized by lower values than the ones corresponding to model architecture NWNN2.

We then show the 2D-maps of the dispersion of the ensemble of prediction obtained from the two trained krigNN2 models in Figure 17 and the two trained NWNN2 models in Figure 18. Each ensemble is composed of around 600 members. The upper figures show the lower dispersion, and the lower figures show the upper dispersion. From one model architecture to another, we see two distinct dispersion patterns. Nevertheless, when employing an identical model architecture, we observe a comparable dispersion pattern, despite a slight variance in magnitude. Furthermore, for each trained model, the upper dispersion and the lower one have almost symmetric patterns. Only the isolated patches with higher values alter the symmetry.





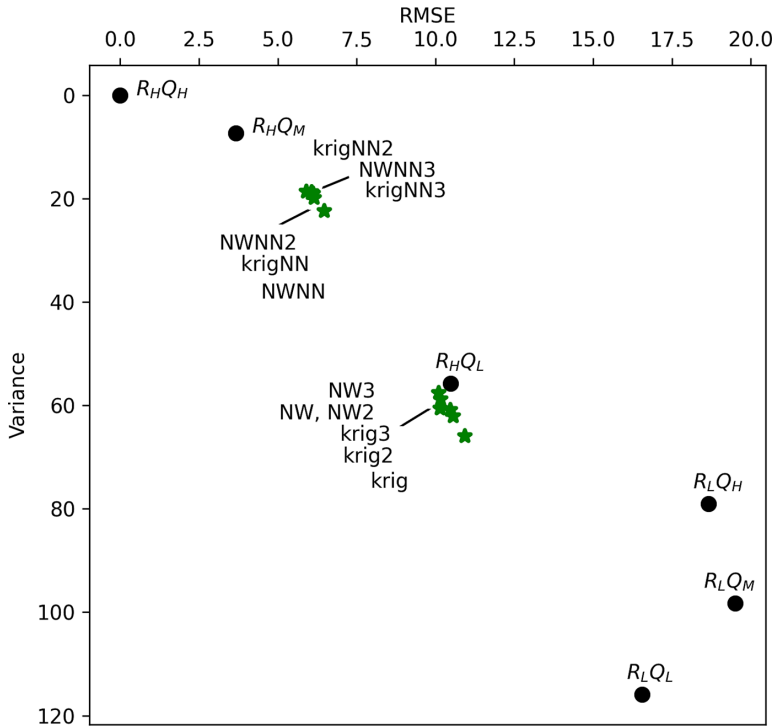
**Figure 13.** Accuracy-precision diagram for case study “Topography” with both *a priori* measurement quality related to the 6 types of sensor devices (black circles) and the metrics of the 12 data fusion models (green stars).

Finally, we present the 2D-maps of the metrics between the members of the ensemble obtained from the two trained krigNN2 models and the two trained NWNN2 models against the observations, respectively, in Figures 19 and 20. The upper panels show the RMSE and the lower panels show the variance. Identically to the dispersion 2D-maps, we see two distinct patterns from one model architecture to another. In addition, the trained models with identical architecture provide similar patterns with a small difference in the order of magnitude. Furthermore, for each trained model, contrarily to 2D-maps of metric RMSE described by local variability, 2D-maps of metric variance have larger patterns that match the asymmetry between the dispersion 2D-maps. Finally, given the global metrics in Table 5, where the RMSE is around 32 m and variance around 550 m<sup>2</sup> for both krigNN2 and NWNN2, these 2D-maps of the metrics highlight local but large errors in the prediction; for instance, the 2D-map visualizing the RMSE of the krigNN2 model can reach 200 m and 120 m for NWNN2, and the 2D-map variance of the krigNN2 and NWNN2 models can reach 1400 m<sup>2</sup>.

### 3.3. Discussion

Adaptive distance attention allows the fusion of the measurements collected by sparse, heterogeneous and mobile sensor devices and the prediction of values at locations with no measurements. We tested this method on three static phenomena over time with different complexities. For each case study, a first network of 100 moving and heterogeneous sensor devices were deployed and trained using a second network of 100 moving high quality sensor devices. In general, the results are positive. By including deep learning



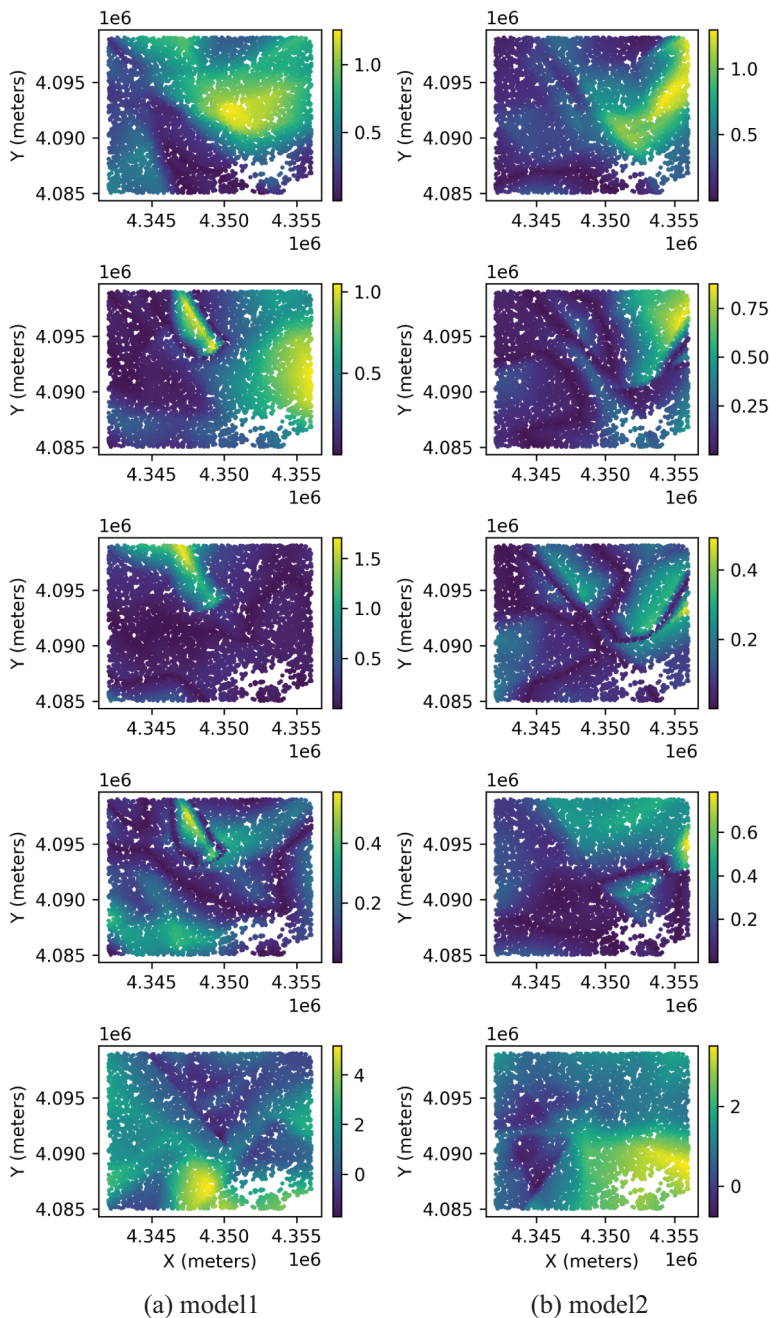


**Figure 14.** Accuracy-precision diagram for case study “AHNO<sub>2</sub>” with both *a priori* measurement quality related to the 6 types of sensor devices (black circles) and the metrics of the 12 data fusion models (green stars).

models into learnable parameters, we improved the metrics from the baseline models OK and GRNN, called “krig” and “NW” in this study. For the three case studies, accuracy-precision diagrams highlight the capability of adaptive distance attention to provide predictions at arbitrary locations with a quality close to sensor devices of medium quality, that is, with an uncertainty of 10% of the signal. Furthermore, the method allows for automatically incorporating the way measurements are weighted according to their *a priori* quality without using any methods such as Kalman filter or data assimilation. Distance attention using the Nadaraya–Watson kernel provides metrics in the same order of magnitude as the attention based on the kriging system; while solving the kriging system involves a matrix inversion, the Nadaraya–Watson kernel is a good alternative to alleviate processing cost for data fusion of sparse data.

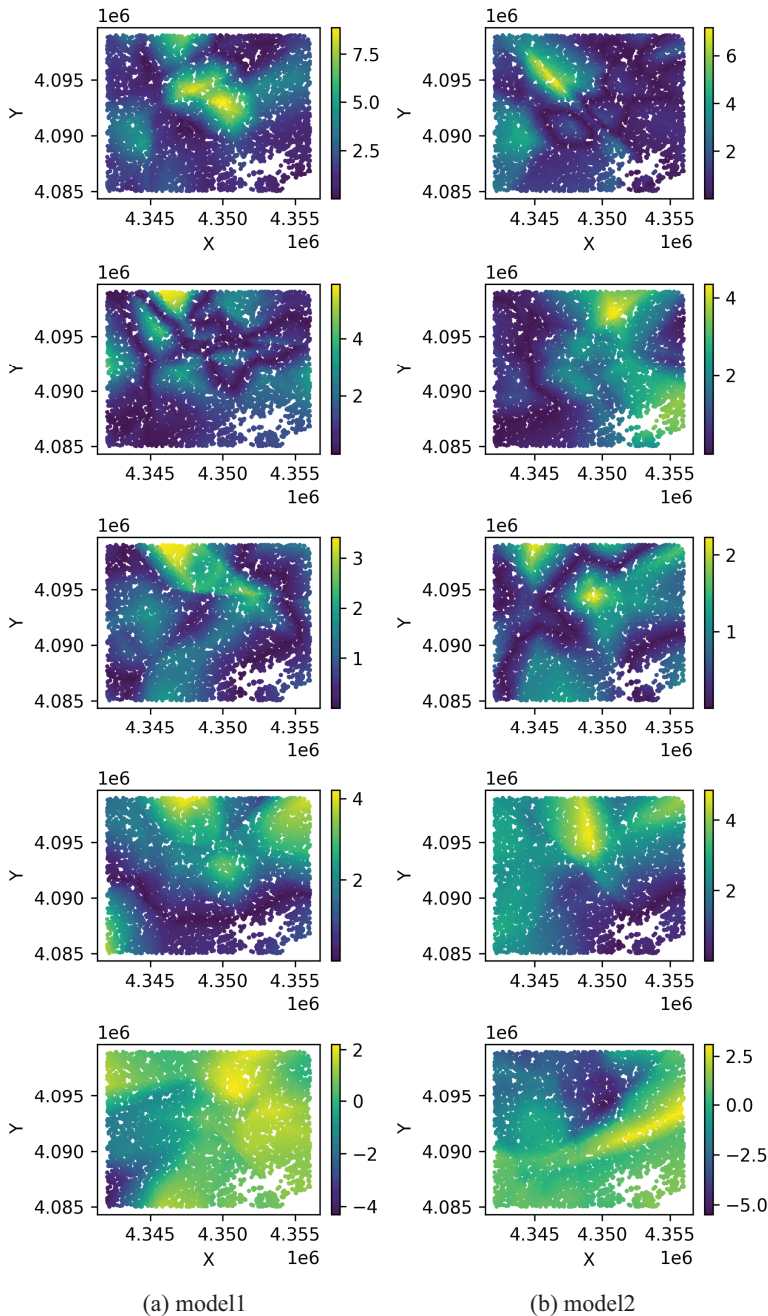
In this study, we assume the existence of 100 sensor devices of high quality at high spatial resolution used as targets to train the data fusion model. This choice is useful to test our data fusion model architecture; it represents nonetheless a high instrumental cost for a measurement campaign. Reducing the instrumentation cost might be done by training a model with sensor devices of different qualities and different spatial resolutions as targets. For this purpose, future work will focus on connecting the raw signal output of the sensors described as level 0, following Ref. (Schneider et al., 2019), to both their external environment and their internal system and bring these variables as keys into the data fusion model. This approach will allow the modeling of the ageing effect of the sensor and the hardware.

Our study focuses on evaluating a trained model with observations belonging to the same bounding area as the training and testing datasets. To enable the use of the trained model with heterogeneous sensors in areas outside of this domain we will test other keys  $K$  connected to the phenomena of interest. In the case of “AHNO<sub>2</sub>,” and in addition to the coordinates, such auxiliary datasets could include information on the underlying emissions (Grythe et al., 2022), the characteristics of the cities from OpenStreetMap as in Steininger et al. (2020) and meteorological information. Further investigation will be required to test the potential of our method from extrapolation to transfer learning for cities with difference ranging from



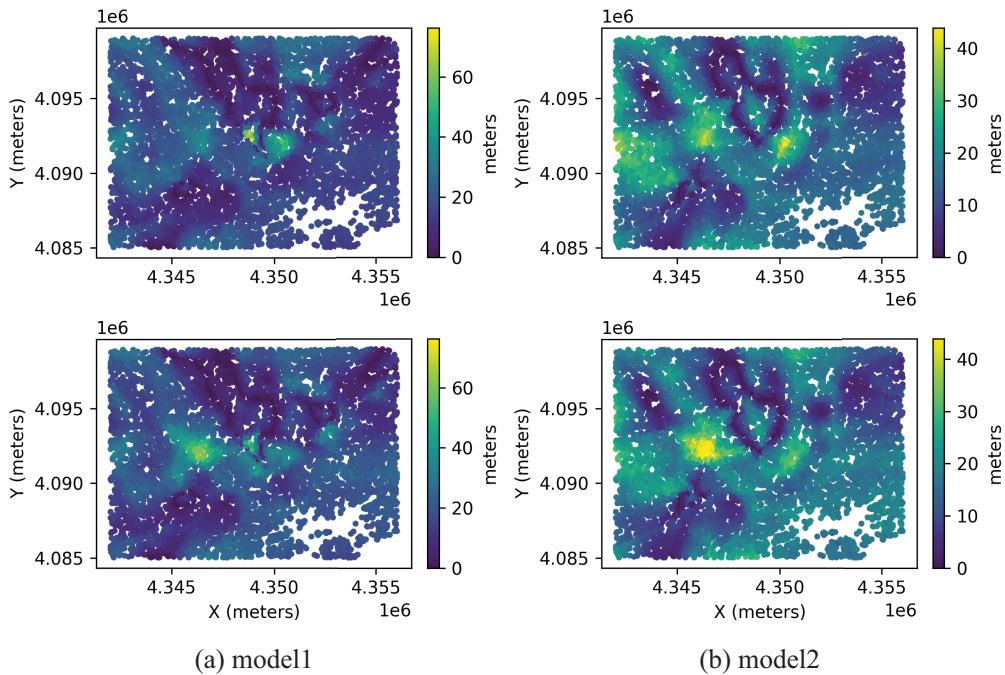
**Figure 15.** 2D-maps of the learnable parameters  $W_K$  and  $W_O$  for a sensor device of high spatial resolution and high measurement quality. Two sets of learnable are presented corresponding to two krigNN2 models trained using two different sequences of sensor devices evolving on a same network  $X$  and  $Y$ . From top to bottom:  $W_{K,x}$ ,  $W_{K,y}$ ,  $W_{K,acc}$ ,  $W_{K,prec}$ , and  $W_O$ .

subtle to significant; for example, it would be interesting to start testing the trained model to predict “AH  $NO_2$ ” on highways connected to the Oslo metropolitan area but outside this area, then to use the trained model to predict “AH  $NO_2$ ” over other cities within Norway, and finally to test the trained model to predict “AH  $NO_2$ ” in cities worldwide.

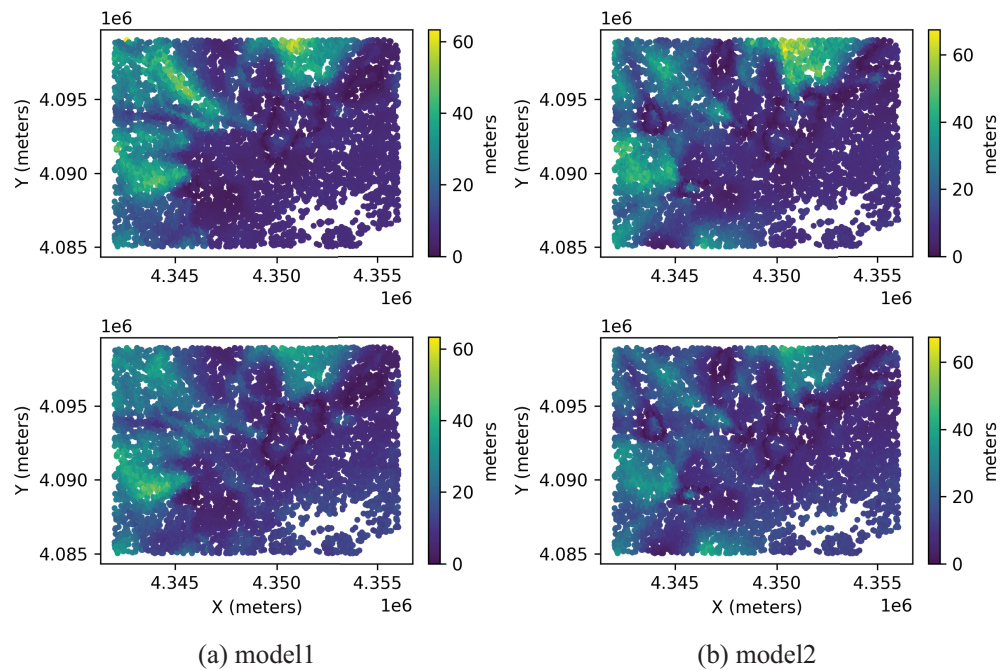


**Figure 16.** 2D-maps of the learnable parameters  $W_K$  and  $W_O$  for a sensor device of high spatial resolution and high measurement quality. Two sets of learnable parameters are presented corresponding to two NWNN2 models trained using two different sequences of sensor devices evolving on a same network  $X$  and  $Y$ . From top to bottom:  $W_{K,x}$ ,  $W_{K,y}$ ,  $W_{K,acc}$ ,  $W_{K,prec}$ , and  $W_O$ .

Our case studies assume phenomena constant in time. This choice is useful to test our data fusion model architecture. Adapting our method for the prediction of time-dependent phenomena will require adding variables related to time into the keys  $K$ . Keeping in mind our interest in predicting hourly urban air quality, we will first follow the work of Stojanović et al. (2023) by using B-splines to encode periodic time-related

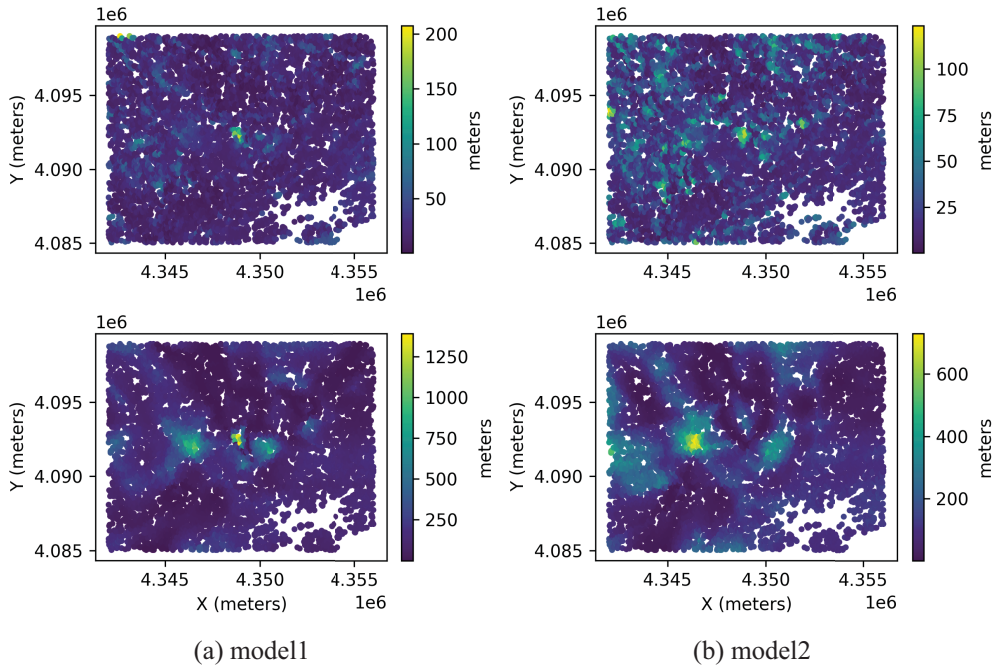


**Figure 17.** 2D-maps of the lower dispersion (top) and upper dispersion (bottom). Two sets of dispersion are presented corresponding to two krigNN2 models trained using two different sequences of sensor devices evolving on a same network  $X$  and  $Y$ .

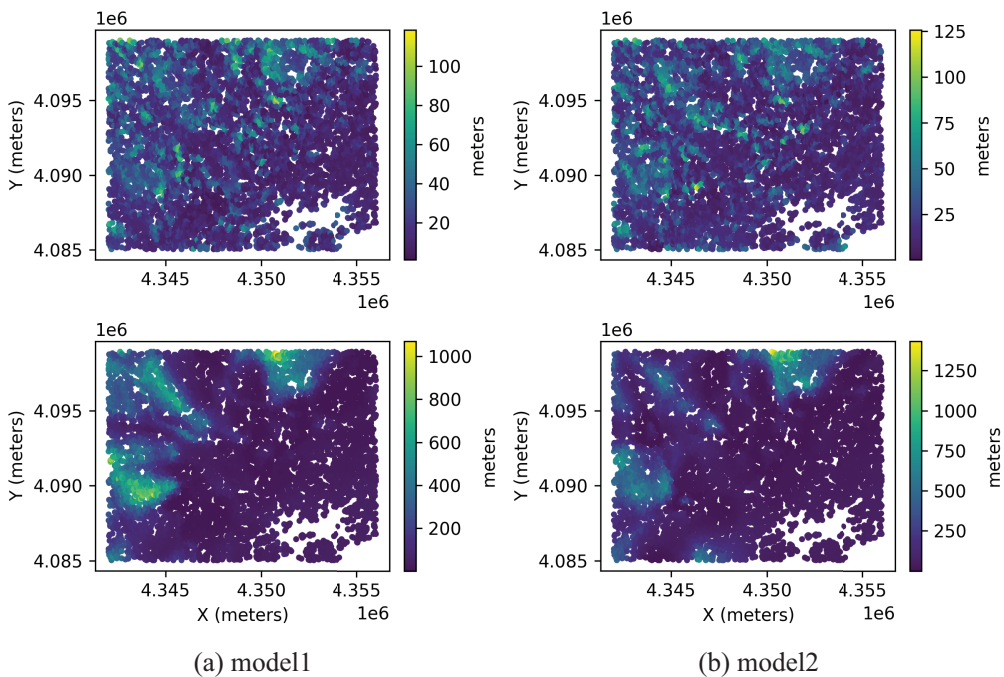


**Figure 18.** 2D-maps of the lower dispersion (top) and upper dispersion (bottom). Two sets of dispersion are presented corresponding to two NWNN2 models trained using two different sequences of sensor devices evolving on a same network  $X$  and  $Y$ .





**Figure 19.** 2D-maps of metrics RMSE (top) and variance (bottom). Two sets of dispersion are presented corresponding to two krigNN2 models trained using two different sequences of sensor devices evolving on a same network  $X$  and  $Y$ .



**Figure 20.** 2D-maps of metrics RMSE (top) and variance (bottom). Two sets of dispersion are presented corresponding to two NWNN2 models trained using two different sequences of sensor devices evolving on a same network  $X$  and  $Y$ .

features related to the human behavior. Then, we will test time-embedding methods, such as Kazemi et al. (2019), on meteorological variable influencing air quality, such as temperature, wind, and relative humidity. Finally, for forecast purposes, we will adapt the transformer architecture (Lin et al., 2022) to our method.

Quantifying the uncertainty automatically is crucial for optimizing measurement campaigns and sensor selection. In our study, we employ ensemble prediction to create error maps. However, this approach has a significant computational cost and provides post-measurement insights. Our future work will focus on refining the model architecture for real-time and cost-effective error prediction. We will follow the work of Tagasovska and Lopez-Paz (2019), where the uncertainty in deep neural networks is estimated using a single model and simultaneous quantile regression as a loss function. This method effectively captures all conditional quantiles, enabling well-calibrated prediction intervals with complex characteristics such as asymmetry, multimodality, and heteroscedasticity.

We assume that measurement campaigns are random sequences of sensor devices deployed at different locations, with different measurement qualities, and different spatial resolutions following predefined characteristics. Our results show that the patterns of learnable parameters differ from one measurement campaign to another; contrary to the constant pattern from feature extraction (Steininger et al., 2020), adaptive distance attention extracts representative information of the phenomena that are ad hoc to one measurement campaign. Nonetheless, even though metrics of the same model architecture are of the same order of magnitude, some local errors characterized as spikes can occur. In a measurement campaign, localizing the areas with potentially significant errors is useful to plan further campaigns and minimize these errors. Highlighting, *a posteriori*, the locations of these errors with ensemble prediction is possible but has a processing cost. Avoiding the local errors while keeping a reasonable processing cost might be possible by planning the measurement campaign to catch relevant information while minimizing local metrics. For this purpose, future studies should focus on designing the measurement campaign workflow (Vasiljević et al., 2020) of the sensor device while letting them adapt to any external or internal constraints using reinforcement learning (Zhou et al., 2020). In doing so, it is important to limit the computational requirements reasonable while keeping models that allow accurate predictions. Finally, we will take the direction of combining this approach with intelligent instrumentation design (Ballard et al., 2021) to help designing new sensor devices to reach better metrics, for instance, in the case study of “AH NO<sub>2</sub>.”

#### 4. Conclusion

We describe the methodology and demonstrate the potential of an adaptive distance attention technique that allows for i) the fusion of observations made by sparse, heterogeneous, and mobile sensor devices; ii) the prediction of values at locations with no measurements; and iii) the automatic weighting of the measurements according to *a priori* quality information about the sensor device without using any methods of data assimilation.

We integrate both OK and a GRNN into this attention with their learnable parameters based on deep learning architectures. We evaluate this method using three static phenomena with different complexities: a case related to a simplistic phenomenon, topography over an area of 196 km<sup>2</sup> and to the annual hourly NO<sub>2</sub> concentration in 2019 over the Oslo metropolitan region (1026 km<sup>2</sup>).

We simulate networks of 100 synthetic sensor devices with six characteristics related to measurement quality and measurement spatial resolution. This approach allows us to generate a set of sensor devices describing reference monitoring stations, low-cost sensor devices, and pixels of satellites.

Outcomes are promising: we significantly improve the metrics from baseline geostatistical models without using any methods of data assimilation.

For the three case studies, accuracy-precision diagrams highlight the capability of adaptive distance attention to provide predictions at arbitrary locations with a quality close to sensor devices of medium quality, that is, with an uncertainty of 10% of the signal of ground truth.

In addition, distance attention using the Nadaraya–Watson kernel provides as good metrics as the attention based on the kriging system enabling the possibility to alleviate the processing cost for fusion of sparse data.

Finally, fusing heterogeneous sensor devices with adaptive distance attention can be used for measurement campaigns of local phenomena in isolated areas. The results are encouraging, and we are planning to continue adapting this approach to space-time phenomena evolving in complex areas.

**Author contribution.** Conceptualization, methodology, and software: J.M.L. Data curation: P.D.H., P.S., R.Ø., and I.V. Writing—original draft: J.M.L. Writing—review and editing: J.M.L. and P.S. Funding acquisition: A.T., T.V.C., J.M.L., P.S., and M.W. All authors approved the final submitted draft.

**Competing interest.** The authors declare none

**Data availability statement.** The 25-m spatial resolution Digital Elevation Model EU-DEM v1.1 is available at <https://www.eea.europa.eu/data-and-maps/data/copernicus-land-monitoring-service-eu-dem>. The Python package Steams is available at <https://pypi.org/project/steams/>.

**Ethics statement.** The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

**Funding statement.** This research was supported by grants from the Research Council of Norway (project number 322473) and the National Centre for Research and Development of Poland (Grant No. NOR/POLNOR/HAPADS/0049/2019-00). Additional partial funding provided by the European Space Agency within the framework of the CitySatAir project (4000131513/20/I-DT), by the Norwegian Research Council in the URBANITY project (321118), and the European Union in the CitiObs project (101086421), is gratefully acknowledged.

## References

- Appleby G, Liu L and Liu LP (2020) Kriging convolutional networks. *Proceedings of the AAAI Conference on Artificial Intelligence* 34(4), 3187–3194.
- Arcucci R, Zhu J, Hu S and Guo YK (2021) Deep data assimilation: integrating deep learning with data assimilation. *Applied Sciences* 11(3), 1114.
- Ballard Z, Brown C, Madni AM and Ozcan A (2021) Machine learning and computation-enabled intelligent sensor design. *Nature Machine Intelligence* 3(7), 556–565.
- Castell N, Dauge FR, Schneider P, Vogt M, Lerner U, Fishbain B, Broday D and Bartonova A (2017) Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? *Environment International* 99, 293–302.
- Chau SL, Bouabid S and Sejdinovic D (2021) Deconditional downscaling with gaussian processes. *Advances in Neural Information Processing Systems* 34, 17813–17825.
- Copernicus, L.M.S. (2016) European Digital Elevation Model (EU-DEM), version 1.1. Available at <https://land.copernicus.eu/imagery-in-situ/eu-dem/eu-dem-v1.1>.
- De Vito S, Esposito E, Castell N, Schneider P and Bartonova A (2020) On the robustness of field calibration for smart air quality monitors. *Sensors and Actuators B: Chemical* 310, 127869.
- De Vito S, Esposito E, Salvato M, Popoola O, Formisano F, Jones R and Di Francia G (2018) Calibrating chemical multisensory devices for real world applications: an in-depth comparison of quantitative machine learning approaches. *Sensors and Actuators B: Chemical* 255, 1191–1210.
- Directive 2008/50/EC (2008) Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe. *Official Journal of the European Union* 152, 1–44.
- Ghamisi P, Rasti B, Yokoya N, Wang Q, Hoffer B, Bruzzone L, Bovolo F, Chi M, Anders K, Gloaguen R and Atkinson PM (2019) Multisource and multitemporal data fusion in remote sensing: a comprehensive review of the state of the art. *IEEE Geoscience and Remote Sensing Magazine* 7(1), 6–39.
- Grythe H, Lopez-Aparicio S, Høyem H and Weydahl T (2022) Decoupling emission reductions and trade-offs of policies in Norway based on a bottom-up traffic emission model. *Atmosphere* 13(8), 1284.
- Hamer PD, Walker SE, Sousa-Santos G, Vogt M, Vo-Thanh D, Lopez-Aparicio S, Ramacher MO, Karl M (2020) The urban dispersion model EPISODE. Part 1: A Eulerian and subgrid-scale air quality model and its application in Nordic winter conditions. *Geoscientific Model Development* 13(9), 4323–4353.
- Hassani A, Bykuć S, Schneider P, Zawadzki P, Chaja P and Castell N (2023) Low-cost sensors and Machine Learning aid in identifying environmental factors affecting particulate matter emitted by household heating. *Atmospheric Environment* 314, 120108.
- Hassani A, Castell N, Watne ÅK and Schneider P (2023) Citizen-operated mobile low-cost sensors for urban PM<sub>2.5</sub> monitoring: field calibration, uncertainty estimation, and application. *Sustainable Cities and Society* 95, 104607.
- Hassani A, Schneider P, Vogt M and Castell N (2023) Low-cost particulate matter sensors for monitoring residential wood burning. *Environmental Science and Technology* 57(40), 15162–15172.
- Hong KY, Pinheiro PO, Minet L, Hatzopoulou M and Weichenthal S (2019) Extending the spatial scale of land use regression models for ambient ultrafine particles using satellite images and deep convolutional neural networks. *Environmental Research* 176, 108513.



- Ionascu ME, Castell N, Boncalo O, Schneider P, Darie M and Marcu M (2021) Calibration of CO, NO<sub>2</sub>, and O<sub>3</sub> using Airify: a low-cost sensor cluster for air quality monitoring. *Sensors* 21(23), 7977.
- Jońca J, Pawnuk M, Bezyk Y, Arsen A and Sówka I (2022) Drone-assisted monitoring of atmospheric pollution—A comprehensive review. *Sustainability* 14(18), 11516.
- Journel AG and Huijbregts CJ (1978) *Mining geostatistics*. 600 pp, Academic Press London, UK.
- Kazemi SM, Goel R, Eghbali S, Ramanan J, Sahota J, Thakur S, Wu S, Smyth C, Poupart P and Brubaker M (2019) *Time2vec: learning a vector representation of time*. Preprint, [arXiv:1907.05321](https://arxiv.org/abs/1907.05321).
- Kingma DP and Ba J (2014) *Adam: a method for stochastic optimization*. Preprint, [arXiv:1412.6980](https://arxiv.org/abs/1412.6980).
- Kyriakidis PC and Journel AG (1999) Geostatistical space–time models: a review. *Mathematical Geology* 31(6), 651–684.
- Lepioufle JM, Marsteen L and Johnsrud M (2021) Error prediction of air quality at monitoring stations using random forest in a total error framework. *Sensors* 21(6), 2160.
- Li T, Wang Y and Yuan Q (2020) Remote sensing estimation of regional NO<sub>2</sub> via space-time neural networks. *Remote Sensing* 12(16), 2514.
- Lin T, Wang Y, Liu X and Qiu X (2022) A Survey of Transformers. *AI Open*. (3), 111–132.
- Lussana C, Tveito OE, Dobler A and Tunheim K (2019) seNorge 2018, daily precipitation, and temperature datasets over Norway. *Earth System Science Data* 11(4), 1531–1551.
- Mijling B (2020) High-resolution mapping of urban air quality with heterogeneous observations: a new methodology and its application to Amsterdam. *Atmospheric Measurement Techniques* 13(8), 4601–4617.
- Miner KR, Turetsky MR, Malina E, Bartsch A, Tamminen J, McGuire AD, Fix A, Sweeney C, Elder CD and Miller CE (2022) Permafrost carbon emissions in a changing Arctic. *Nature Reviews Earth and Environment* 3(1), 55–67.
- Miyoshi T, Sato Y and Kadowaki T (2010) Ensemble Kalman filter and 4D-Var intercomparison with the Japanese operational global analysis and prediction system. *Monthly Weather Review* 138(7), 2846–2866.
- Nadaraya EA (1964) On estimating regression. *Theory of Probability and its Applications* 9(1), 141–142. <https://doi.org/10.1137/1109020>.
- Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L and Desmaison A (2019) Pytorch: an imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems* 32, 8026–8037.
- Peyron M, Fillion A, Gürol S, Marchais V, Gratton S, Boudier P and Goret G (2021) Latent space data assimilation by using deep learning. *Quarterly Journal of the Royal Meteorological Society* 147(740), 3759–3777.
- Robert S, Foresti L and Kanevski M (2013) Spatial prediction of monthly wind speeds in complex terrain with adaptive general regression neural networks. *International Journal of Climatology* 33(7), 1793–1804.
- Rue H, Riebler A, Sørbye SH, Illian JB, Simpson DP and Lindgren FK (2017) Bayesian computing with INLA: a review. *Annual Review of Statistics and Its Applications* 4, 395–421.
- Samad A, Alvarez Florez D, Chourdakis I and Vogt U (2022) Concept of using an unmanned aerial vehicle (UAV) for 3D investigation of air quality in the atmosphere—example of measurements near a roadside. *Atmosphere* 13(5), 663.
- Scheller JH, Mastepanov M and Christensen TR (2022) Toward UAV-based methane emission mapping of Arctic terrestrial ecosystems. *Science of the Total Environment* 819, 153161.
- Schneider P, Bartonova A, Castell N, Dauge FR, Gerboles M, Hagler GS, Huglin C, Jones RL, Khan S, Lewis AC, Mijling B, Müller M, Penza M, Spinelle L, Stacey B, Vogt M, Wesseling J and Williams RW (2019) Toward a unified terminology of processing levels for low-cost air-quality sensors. *Environmental Science & Technology* 53(15), 8485–8487.
- Schneider P, Castell N, Dauge FR, Vogt M, Lahoz WA and Bartonova A (2018) A network of low-cost air quality sensors and its use for mapping urban air quality. In: Bordogna G, Carrara P(eds) *Mobile Information Systems Leveraging Volunteered Geographic Information for Earth Observation. Earth Systems Data and Models*, 4. Springer, Cham. pp. 93–110. [https://doi.org/10.1007/978-3-319-70878-2\\_5](https://doi.org/10.1007/978-3-319-70878-2_5).
- Schneider P, Castell N, Vogt M, Dauge FR, Lahoz WA and Bartonova A (2017) Mapping urban air quality in near real-time using observations from low-cost sensors and model information. *Environment International* 106, 234–247.
- Schneider P, Hamer PD, Kylling A, Shetty S and Stebel K (2021) Spatiotemporal patterns in data availability of the sentinel-5P NO<sub>2</sub> product over urban areas in Norway. *Remote Sensing* 13(11), 2095.
- Schneider P, Vogt M, Haugen R, Hassani A, Castell N, Dauge FR and Bartonova A (2023) Deployment and evaluation of a network of open low-cost air quality sensor systems. *Atmosphere* 14(3), 540.
- Semmens KA, Anderson MC, Kustas WP, Gao F, Alfieri JG, McKee L, Prueger JH, Hain CR, Cammalleri C, Yang Y and Xia T (2016) Monitoring daily evapotranspiration over two California vineyards using Landsat 8 in a multi-sensor data fusion approach. *Remote Sensing of Environment* 185, 155–170.
- Shetty, S., Schneider, P., Stebel, K., Hamer, P. D., Kylling, A., & Bernsten, T. K. (2024). Estimating surface NO<sub>2</sub> concentrations over Europe using Sentinel-5P TROPOMI observations and Machine Learning. *Remote Sensing of Environment*, 312, 114321.
- Specht DF (1991) A general regression neural network. *IEEE Transactions on Neural Networks* 2, 568–576.
- Stebel K, Stachlewska IS, Nemuc A, Horálek J, Schneider P, Ajtai N, ... Zehner C (2021) SAMIRA - satellite based monitoring initiative for regional air quality. *Remote Sensing* 13(11), 2219.
- Steininger M, Kobs K, Zehe A, Lautenschlager F, Becker M and Hotho A (2020) Maplur: exploring a new paradigm for estimating air pollution using deep learning on map images. *ACM Transactions on Spatial Algorithms and Systems* 6(3), 1–24.

- Stojanović DB, Kleut DN, Davidović MD, Jovašević-Stojanović MV, Bartonova A and Lepioufle J-M** (2023) Low-processing data enrichment and calibration for PM<sub>2.5</sub> low-cost sensors. *Thermal Science* 27(3b), 2229–2240.
- Tagasovska N and Lopez-Paz D** (2019) Single-model uncertainties for deep learning. *Advances in Neural Information Processing Systems* 32, 6414–6425.
- Translation of the Report on the Suitability Test of the Ambient Air Measuring System** (2007) Translation of the Report on the Suitability Test of the Ambient Air Measuring System M200E of the Company Teledyne Advanced Pollution Instrumentation for the Measurement of NO, NO<sub>2</sub> and NO<sub>x</sub>. Technical Report 936/21205926/A2, TÜV, Cologne, Germany.
- Van Poppel M, Schneider P, Peters J, Yarkin S, Gerboles M, Matheeußen C, Bartonova A, Davila S, Signorini M, Vogt M, Dauge FR, Skaar JS and Haugen R** (2023) SensEURCity: a multi-city air quality dataset collected for 2020/2021 using open low-cost sensor systems. *Scientific Data* 10(1), 322.
- Vasiljević N, Vignaroli A, Bechmann A and Wagner R** (2020) Digitalization of scanning lidar measurement campaign planning. *Wind Energy Science* 5(1), 73–87.
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł and Polosukhin I** (2017) Attention is all you need. *Advances in Neural Information Processing Systems* 30, 5998–6008.
- Vogt M, Schneider P, Castell N and Hamer P** (2021) Assessment of low-cost particulate matter sensor systems against optical and gravimetric methods in a field co-location in Norway. *Atmosphere* 12(8), 961.
- Wackernagel H** (2003) *Multivariate Geostatistics: An Introduction with Applications*. Springer-Verlag, Berlin.
- Watson GS** (1964) Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*. 26(4), 359–372.
- Wattrelot E, Caumont O and Mahfouf JF** (2014) Operational implementation of the 1D+ 3D-Var assimilation method of radar reflectivity data in the AROME model. *Monthly Weather Review* 142(5), 1852–1873.
- Weichenthal S, Dons E, Hong KY, Pinheiro PO and Meysman FJ** (2021) Combining citizen science and deep learning for large-scale estimation of outdoor nitrogen dioxide concentrations. *Environmental Research* 196, 110389.
- Working Group on Guidance for the Demonstration of Equivalence** (2010) Guide to the Demonstration of Equivalence of Ambient Air Monitoring Methods Technical Report, European Commission, Brussels, Belgium.
- Xu D, Anguelov D and Jain A** (2018) Pointfusion: deep sensor fusion for 3d bounding box estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 244–253.
- Zhou T, Chen M and Zou J** (2020) Reinforcement learning based data fusion method for multi-sensors. *IEEE/CAA Journal of Automatica Sinica* 7(6), 1489–1497.