Towards Fine-Grained Sensor-Based Probabilistic Individual Air Pollution Exposure Prediction using Wind Information

Paul Tremper[1], Till Riedel[2]

Abstract:

The estimation of pollutant exposure is highly dependent on the spatial and temporal resolution of the underlying model. This work presents a street-level Gaussian Process Regression model for urban air quality that uses a novel covariance kernel based on physical considerations to process wind information. This model can be driven by information from observations from low-cost sensor networks. We present the model, including the construction of the wind angle kernel, and discuss the inconclusive evaluation to date, the current challenges, and the way forward.

Keywords: Gaussian Process Regression; kriging; air quality; interpolation; wind direction; prediction

1 Introduction

Air quality plays a crucial role for our health [Or23]. Accurately assessing air quality is challenging. Existing air quality dispersion models[3], are computationally intensive. These models further require often incomplete emission inventories to start with. Immission modelling takes a different approach by using observations to predict the temporal and/or spatial distribution of pollutants. The emergence of low-cost air quality sensors deployed by citizen science [Oy22] or research initiatives [Bu17] has made immission models increasingly attractive. However, those models currently lack the resolution of simulations. Sokhi et al. recently conducted an extensive review on air quality research [So22]. They emphasised the importance of data from dense air quality networks which makes interpolation naturally come into focus. We construct a Bayesian inference-based interpolation model that uses wind information to improve predictions and push the applicability of such models to the street scale. To achieve this, we use Gaussian Process Regression to construct the wind information processing from physical considerations. After discussing existing work, we describe the structure of the used kernel in section 3 and particularly the construction of the wind angle kernel and demonstrate its behaviour in section 4. We use these kernels to make a spatial prediction of NO_2 values from simulated measurements and compare it with purely spatial kernels to showcase the effect of the wind field information. We discuss our current challenges and propose approaches to overcome them.

[1] KIT, TM-PCS/TECO, Vincenz-Prießnitz-Str. 1, 76131 Karlsruhe, Germany paul.tremper@kit.edu
[2] KIT, TM-PCS/TECO, Vincenz-Prießnitz-Str. 1, 76131 Karlsruhe, Germany till.riedel@kit.edu
[3] e.g. GRAMM/GRAL, `https://github.com/GralDispersionModel` or PALM `https://gitlab.palm-model.org`

2 Background and Related Work

Multiple methods are being used for spatial interpolation of air quality data. *Land Use Regression* (LUR) links air pollution measurements with land use to predict pollution at unmeasured locations using statistics, e.g. [Su16], [No20]. Supervised *Machine Learning* (ML) uses historical pollution data and features such as weather, traffic and land use for real-time or future air quality prediction, e.g. Random Forest, [Yu16]. *Kriging* (Gaussian processes) estimates pollution using observed data correlations [DMŽ17; Ki14; Pa16].

Wind information has been incorporated into spatial interpolation methods to improve accuracy. For example, [CF16] use a modified inverse distance weighting (IDW) model that favours windward points in a 30° sector, reporting positive results for traffic emissions such as NO, NO_2 and SO_2. [Zh21a] use kriging with upwind points, reporting a more stable RMSE than ordinary kriging. In the context of land use regression (LUR), [Ar07] incorporate wind fields, emphasising their influence on the intra-urban distribution of NO_2. [LGZ14] and [Zh21b] adapt IDW and deep learning, respectively, using wind paths for better predictions. [Zh18] integrate wind directions as labels in a multiple kernel learning ML model.

All of these models operate at much larger scales (10km to 100km): To the best of our knowledge, no study has addressed the interpolation of air pollutants at street level without relying on dispersion models for predictions. Furthermore, we have not found any study that has attempted to model local wind angle dependence at a granular level into the covariance function of a kriging model.

Various approaches have attempted to integrate data from dispersion modelling and sensor data information: Data fusion methods (for example, [Sc17]), data assimilation techniques (such as, [Jo22]), and machine learning methods applied to improve dispersion models ([Ka23]) have been employed to enhance predictions. We employ a dispersion model in our study as synthetic ground truth to assess our interpolation model. The method does not need such model when trained or applied on real data.

When previously comparing several interpolation methods [TRB21], one of which was Gaussian Process Regression with a Gaussian (RBF) kernel, to determine their effectiveness in interpolating sensors of the SmartAQnet deployment [4], we found that the interpolation was able to capture the general trends of the time series at a distance of approximately 500m. In this paper, we discuss improvements to the Gaussian Process's kernel function by incorporating wind information to enhance its prediction accuracy.

[4] A heterogeneous air quality sensor network in Augsburg, Germany. `https://smartaq.net`, [Bu17]

3 Overall Kernel Structure

Gaussian Process Regression uses a covariance function (kernel) to map covariances between input points, which are then used to generate predictions [RW05]. Our approach is based on the idea that we can split the total kernel into a product of a spatial distance kernel, a wind strength kernel and a wind angle kernel whose covariances are independent of each other.

$$k_{total}(v_1, v_2, w_1, w_2) = k_{spatial}(v_1, v_2) \cdot k_{strength}(w_1, w_2) \cdot k_{angle}(v_1, v_2, w_1, w_2) \quad (1)$$

The spatial distance kernel depends only on the spatial coordinates of the two input points (v_1, v_2). The wind strength kernel, equivalently, depends only on the wind vector coordinates (w_1, w_2) of the two input points. The wind angle kernel depends only on the relative angles between the position vectors and the wind vectors. For very short distances, concentrations are likely to be correlated regardless of wind direction and strength. To account for this case, we add a complementary kernel of the wind direction and wind strength part with a very short length scale hyperparameter in the spatial kernel. We denote the spatial kernel to this complementary kernel as $k_\perp$ (k orthogonal) and the original spatial distance kernel as $k_{||}$ (k parallel). This implies $\ell_{||} >> \ell_\perp$ for the length scale hyperparameters of the respective kernels. The structure of the full kernel is thus

$$\begin{aligned} k_{total}(v_1, v_2, w_1, w_2) = {} & k_{||}(v_1, v_2) \cdot k_{strength}(w_1, w_2) \cdot k_{angle}(v_1, v_2, w_1, w_2) \\ & + k_\perp(v_1, v_2) \cdot \left[1 - k_{strength}(w_1, w_2) \cdot k_{angle}(v_1, v_2, w_1, w_2)\right] \end{aligned} \quad (2)$$

In the limit of identical spatial kernel length scales ($\ell_\perp = \ell_{||}$), the full kernel reduces to a purely spatial distance kernel.

We chose a kernel with a Gaussian profile (Radial Basis Function, RBF) for both the spatial distance kernels and the wind strength kernel. These are of the form

$$k_{spatial}(v_1, v_2) = \exp\left[-\frac{(v_1 - v_2)^2}{2\ell_{xy}^2}\right] \qquad k_{strength}(w_1, w_2) = \exp\left[-\frac{(w_1 - w_2)^2}{2\ell_{uv}^2}\right] \quad (3)$$

For the spatial distance kernel, $v_1 \equiv (x_1, y_1)$ and $v_2 \equiv (x_2, y_2)$ are two component spatial position vectors, while for the wind strength kernel w_1 and w_2 are two component wind field vectors.

4 Wind Angle Kernel

Only the relative angles of the input vectors $v := v_1 - v_2$, w_1 and w_2 are relevant for the construction of the angle kernel. Therefore, all vectors in this subsection are considered normalised vectors. We refrain from adding another symbol to indicate this for readability reasons.

4.1 Defining Support Points

Covariance functions are constrained by

1. a scalar output $k(\star_1, \star_2) \in \mathbb{R}^1$ (the covariance)
2. symmetry under exchange of the two input points $k(\star_1, \star_2) = k(\star_2, \star_1)$

Constraint 1) led us to consider only scalar products of the normalised input vectors v, w_1 and w_2. The scalar products in question are then $w_1 w_2$, $v w_1$ and $v w_2$. We can discard one of these angles as a degree of freedom by eliminating the rotational symmetry of the system and remain with $\theta_1 = \arg(v w_1)$ and $\theta_2 = \arg(v w_2)$. In this way we can visualise the angle kernel function in the $\theta_1 - \theta_2$ plane, which makes it much easier to construct.

We want to use the critical points {-1,0,1} of the scalar products as support points for the covariance function we seek. To this end, we look at angular values of $\theta_i \in \{0, \frac{1}{2}\pi, \pi, \frac{3}{2}\pi\}$. For two angles θ_1, θ_2, there are $4^2 = 16$ configurations which we show in figure 1, as well as the desired covariance we have assigned to each configuration. The assignment of these desired covariances is subjective in that the configuration of position and wind vectors must be interpreted as a physical situation.

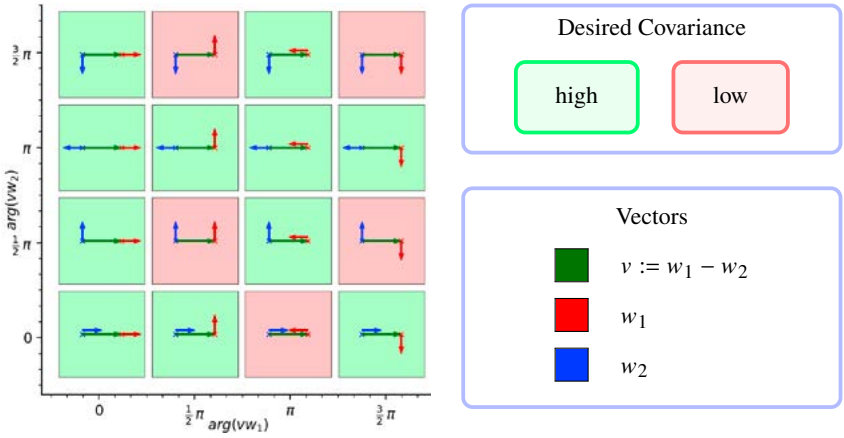

Fig. 1: All vector configurations in consideration as support points as well as a color coding of their desired covariance.

4.2 Constructing Building Blocks

Constraint 2) implies the symmetries shown in figure 2. These symmetries led us to combine the scalar products $w_1 w_2$, $v w_1$ and $v w_2$ into combinations that are invariant under the exchange of v_1 and v_2. These combinations serve as building blocks to construct

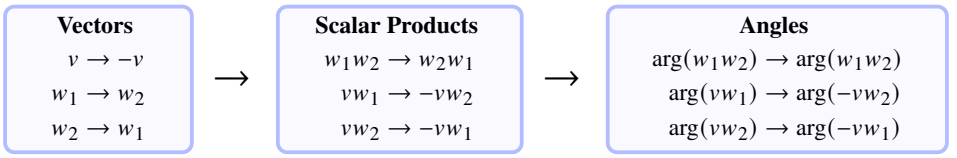

Fig. 2: Symmetry constraints of the covariance function.

covariance functions that automatically satisfy constraints 1) and 2). From the possible building blocks we chose

$$f_w \equiv f_w(w_1, w_2) := (w_1 w_2)^2 \tag{4}$$

$$f_+ \equiv f_+(v, w_1, w_2) := \left(\frac{vw_1 + vw_2}{2}\right)^2 \tag{5}$$

$$f_- \equiv f_-(v, w_1, w_2) := \frac{1}{4}\left(1 + \frac{vw_1 - vw_2}{2}\right)^2 \tag{6}$$

where the numerical factors are normalisation factors to force the outputs into the [0,1] interval. The squares provide smoother transitions between the parts of the function and ensure that f_+ is invariant. Combining these functions, we construct the angle kernel in the following way

$$\begin{aligned} k_{angle}(v, w_1, w_2) &= (1 - f_w) + f_w\left[f_+ + (1 - f_+)f_-\right] \\ &= 1 - f_w(1 - f_+)(1 - f_-) \end{aligned} \tag{7}$$

The functional behavior of the full angle kernel is shown in figure 6 in the $\theta_1\theta_2$-plane. The desired high/low covariances of the support points show up as green/red circles. We can see, that the function succeeds quite well in modelling the desired behavior.

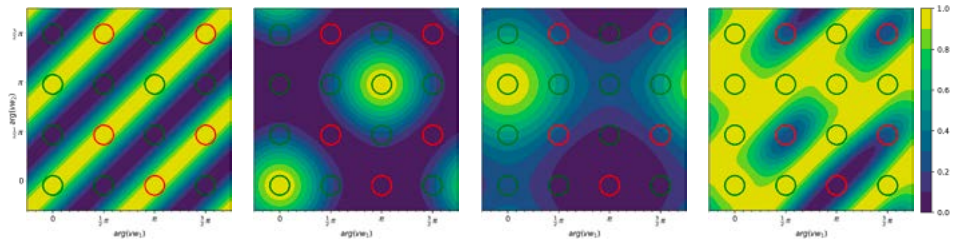

Fig. 3: f_w Fig. 4: f_+ Fig. 5: f_- Fig. 6: k_{angle}

Fig. 3-6: The angle kernel k_{angle} and its constituent building blocks f_w, f_+, f_- in the $\theta_1\theta_2$-plane. Regions with high (low) covariances are colored yellow (purple). Green (red) circles indicate a high (low) desired covariance.

Note that if $v_1 = v_2 \leftrightarrow v = 0 \leftrightarrow w_1 = w_2$. Hence $f_w = 1$, $f_+ = 0$, $f_- = 1/4$ and therefore $k_{angle}(0, w_1, w_1) = 1/4$. Thus the diagonal of the angle part of the covariance matrix is $1/4$, which has to be set to 1 manually. This is a consequence of $\arg(vw_1)$ and $\arg(vw_2)$ becoming undefined in this case.

5 Evaluation

We used a GRAMM/GRAL[5] dispersion simulation as train/test data for our model since this should sufficiently capture the physical processes that govern pollutant dispersion. The dataset used in our study encompasses an area of an Austrian city of 656×458 meters with grid cells of 2×2 meters, leading to 75,670 datapoints per run. Each data point provides information about u and v coordinates of the wind field, as well as the simulated concentration c at 2 m height.

Spatial RBF Kernel Model (RBFxy): Since the spatial distance kernel of our model is an RBF kernel, comparing the full model with a standard RBF kernel model as baseline is the obvious choice. We call this model the RBFxy baseline model.

$$k_{RBFxy}(v_1, v_2) = \exp\left[-\frac{(v_1 - v_2)^2}{2\ell_{xy}^2}\right] \tag{8}$$

Wind-Spatial RBF Kernel (RBFxyuv): Another useful baseline model consists of an RBF kernel in spatial coordinates multiplied by an RBF kernel in wind vector coordinates. This construction mirrors the spatial distance kernel and the wind strength kernel of our full model, but without the angle kernel and the split into a parallel and an orthogonal part. We call this model the RBFxyuv baseline model.

$$k_{RBFxyuv}(v_1, v_2, w_1, w_2) = \exp\left[-\frac{(v_1 - v_2)^2}{2\ell_{xy}^2}\right] \cdot \exp\left[-\frac{(w_1 - w_2)^2}{2\ell_{uv}^2}\right] \tag{9}$$

5.1 Evaluation Methodology

We performed a grid search to determine the hyperparameters of the kernels. We used 20 randomly chosen points with concentration $c > 2$ as training data (sensors). The reason for this constraint was twofold: 1) it prevents from picking points inside of buildings and 2) it prevents from picking from the outer regions of the domain, where the simulation does not account for sources outside of the domain. Other than that, $c > 2$ is chosen to act as a conservative lower bound for realistic sensor placements. Subsequenly, we used all datapoints within $150 <= x <= 550$ and $50 <= y <= 400$ (to avoid errors of unaccounted for sources from outside of the domain) with a concentration > 0 (to exclude unphysical values and buildings) as test data.

We found the following minima for the hyperparameters:

Wind Kernel Model	$\ell_\perp = 22$, $\ell_{\|\|} = 100$, $\ell_{uv} = 1.5$
RBFxy baseline model	$\ell_{xy} = 50$
RBFxyuv baseline model	$\ell_{xy} = 134$, $\ell_{uv} = 1$

[5] https://gral.tugraz.at/

Using these hyperparameters, we computed 1000 runs, each time randomly choosing a new set of 20 training points. We report the MSE statistics when compared against the ground truth in table 1.

c>0	mean ± std	(min, max)	c>2	mean ± std	(min, max)
Wind Model	14.01 ± 3.16	(8.53, 38.02)	Wind Model	13.47 ± 2.71	(8.68, 29.44)
RBFxy	11.90 ± 0.84	(10.01, 15.39)	RBFxy	19.15 ± 3.25	(11.48, 34.11)
RBFxyuv	14.08 ± 2.62	(8.97, 26.06)	RBFxyuv	13.82 ± 2.23	(8.72, 25.06)

Tab. 1: MSE Comparison of the baseline models (RBFxy, RBFxyuv) and our wind model. 1000 runs, each with 20 randomized sensor positions at concentration values c>2, evaluating the central area of the domain (see section 5). c>0 and c>2 refer to leaving out concentrations equal to zero (buildings) and smaller than 2 μg/m^3 (to match the training points).

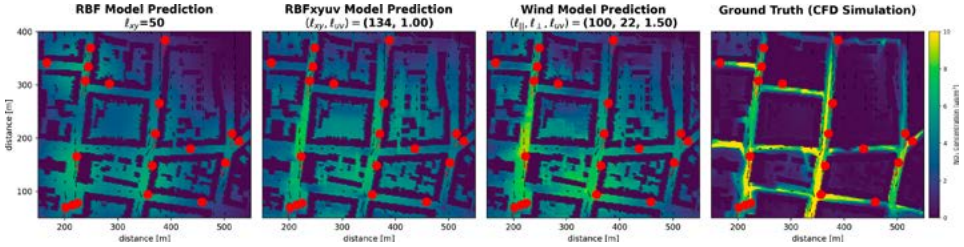

Fig. 7: Comparison of a random set of 20 training points (red dots) between baseline (RBF, RBFxyuv), our wind model and and GRAMM/GRAL simulated ground truth. In all cases, the wind field stems from the GRAMM/GRAL simulation.

5.2 Preliminary Evaluation Results

Although the predictions of our model look very different compared to the baseline kernels (see Fig. 7), standard metrics such as MSE or MAE show no significant improvement. Table 1 shows that the wind model and the RBFxyuv model have a much higher variance than the purely spatial RBFxy model. Although their mean MSE is generally higher, indicating a poorer fit, their minimum values are lower, indicating that they are more sensitive to sensor placement and have the potential for better predictions.

Indeed, Fig. 7 shows the potential of including wind information in the street level interpolation, as both the RBFxyuv model and our wind model are able to model pollutant concentrations along street canyons to some extent. All of the models shown suffer from overestimating concentrations in small streets and backyards, which would require model extensions to address. One possibility would be to include land use information in the model or in a wind interpolation to suppress wind-protected areas.

We did not perform any evaluation against other state of the art prediction models since our model to date does not outperform these simple baselines.

6 Challenges and Future Work

Because our model uses covariance functions, it cannot represent aligned wind vectors separated by an obstacle. Our model also only works in stationary (steady-state) situations. When the wind field becomes dynamic in time, information about the individual transport trajectories becomes much more important, which significantly complicates matters. Eg. if a training point picks up a wind direction caused by a small eddy to point in a drastically different direction than the larger wind flow of the area, the prediction will be significantly distorted.

6.1 Challenges

Wind Strength Kernel: We chose an RBF kernel for the wind speed, a standard choice when detailed information about the system is lacking. A kernel with a more nuanced functional behaviour to capture the covariances between wind strengths as well as a non-stationary kernel component to suppress low wind values could prove beneficial.

More nuanced Angle Kernel: We chose to go with a rather simple version of the angle kernel function (see 4). This could be changed: 1. Different, additional and/or more nuanced support points. 2. Change the functional behavior that interpolates between them 3. Introduce hyperparameter to the angle kernel. A more comprehensive systematization of the construction process would be a prerequisite.

Factorization of Kernel Parts: The assumption that the distance , wind strength and angle kernel factorize might not hold well enough. In reality, the different kernel parts are likely interwoven. A numerical analysis (see paragraph below) further insight.

Numerical Kernel: A covariance analysis of an existing dataset could be used as basis for a numerical kernel function. Since such a numerical approach would represent the true covariance within the dataset, it could be used to validate an analytical approach such as ours. Comparing the numerical kernel from different datasets would also shed light whether or not the underlying physical processes govern the covariances at all.

Evaluation Metric: The value of the model might not lie in spatial coverage (as described by the MSE), but, e.g., in directly predicting individual exposure in certain scenarios. This would require defining meaningful metrics for the specific use case.

Sharpening the use case: Although an evaluation of an entire domain at high resolution does not show a significant improvement in prediction, there may be specific use cases where our model works generally better. For example, forecasting away from sources, areas with more homogeneous wind fields or different pollutants.

Wind information from sensors: For our evaluation we used the wind field from the CFD simulation. For an application, we would take the wind field from an interpolation of sensor

data. This should work as model only draws information from the general flow and not from finely resolved structures. In fact, we suspect that eddies have a negative impact on the predictions.

Improve Model Training: Currently, optimisers struggle to find minima of the loss function, leaving us with brute force grid search. In addition, GPR is known to have problems with hyperparameter optimisation when the covariance matrix becomes indefinite. Both problems probably require the kernel to be analysed for critical points and made numerically more stable.

Dataset: Due to the reactivity of NO_2, which is taken into account by GRAMM/GRAL, a small-scale prediction of PM may be more feasible. In addition, the small domain and high spatial resolution of the CFD simulation poses a problem for the interpolation method. We suspect that a gradual coarsening of the resolution could improve the prediction. Other studies focusing on the integration of wind information into pollutant interpolation consider much larger and much coarser urban domains, see for e.g. [Ar07] [LGZ14].

6.2 Future work

We see two promising paths forward for our model: 1. Stay on street level scales to push the boundaries of air quality interpolation methods for predicting personal exposure. 2. Take the model to coarser scales and compare our model with state of the art interpolation methods which use wind fields in their predictions. The latter likely requires changing the angle kernel function as larger area correlations might have different patterns than street level correlations. In both cases, the steps we plan to take are as follows:

- Specify a use case and determine a meaningful evaluation metric for that use case
- Perform a correlation analysis of a dataset to get information on how to improve the kernel and verify the results on a second dataset.
- Evaluate the improved kernel for the use case and interrelate with other wind incorporating methods such as LUR [Ar07], IDW [LGZ14] or ML models [Zh21b].
- In case of success, evaluate our model with interpolated wind data from sensors.

7 Conclusion

We presented a Gaussian Process Regression model capable of processing wind information to make predictions about pollutant concentrations. We explained the construction process of the kernel and its inconclusive evaluation so far. We then discussed challenges, possible improvements and our plans going forward. Despite the inconclusive evaluation, we stress that models constructed from physical considerations have value because of their straightforward interpretability. Gaussian Process Regression (GPR) offers a framework,

which employs the advantages of machine learning while at the same time offering physically interpretable hyperparameters. On top of that, GPR predictions natively come in form of gaussian distributions, offering a measure of error to the prediction. A Gaussian Process Regression based interpolation model, which is able to reliably predict pollutants at street level in real time with a native error measure would be a valuable tool to estimate personal exposure.

Acknowledgements

We would like to thank the Helmholtz European Partnership for Technological Advancement (HEPTA) for supporting this study. We also would like to thank Markus Kuntner (Amt der Tiroler Landesregierung) for helping with GRAMM/GRAL and providing validation data.

References

[Ar07] Arain, M. et al.: The use of wind fields in a land use regression model to predict air pollution concentrations for health exposure studies. Atmospheric Environment 41/16, pp. 3453–3464, 2007.

[Bu17] Budde, M. et al.: SmartAQnet: remote and in-situ sensing of urban air quality. In: Remote Sensing of Clouds and the Atmosphere XXII. Vol. 10424, International Society for Optics and Photonics, p. 104240C, 2017.

[CF16] Contreras, L.; Ferri, C.: Wind-sensitive Interpolation of Urban Air Pollution Forecasts. Procedia Computer Science 80/, International Conference on Computational Science 2016, ICCS 2016, 6-8 June 2016, San Diego, California, USA, pp. 313–323, 2016.

[DMŽ17] Dünnebeil, G.; Marjanović, M.; Žarko, I. P.: Approaches to Fuse Fixed and Mobile Air Quality Sensors. In (Hřebíček, J.; Denzer, R.; Schimak, G.; Pitner, T., eds.): Environmental Software Systems. Computer Science for Environmental Protection. Springer International Publishing, Cham, pp. 71–84, 2017.

[Jo22] Johansson, L. et al.: An operational urban air quality model ENFUSER, based on dispersion modelling and data assimilation. Environmental Modelling & Software 156/, p. 105460, 2022.

[Ka23] Kassandros, T. et al.: Machine learning-assisted dispersion modelling based on genetic algorithm-driven ensembles: An application for road dust in Helsinki. Atmospheric Environment 307/, p. 119818, 2023.

[Ki14] Kilibarda, M. et al.: Spatio-temporal interpolation of daily temperatures for global land areas at 1 km resolution. Journal of Geophysical Research: Atmospheres 119/5, pp. 2294–2313, 2014.

[LGZ14] Li, L.; Gong, J.; Zhou, J.: Spatial Interpolation of Fine Particulate Matter Concentrations Using the Shortest Wind-Field Path Distance. PLOS ONE 9/5, pp. 1–10, May 2014.

[No20] Nori-Sarma, A. et al.: Low-cost NO2 monitoring and predictions of urban exposure using universal kriging and land-use regression modelling in Mysore, India. Atmospheric Environment 226/, p. 117395, 2020.

[Or23] Organization, W. H.: WHO ambient air quality database, 2022 update: status report. World Health Organization, 2023.

[Oy22] Oyola, P.; Carbone, S.; Timonen, H.; Torkmahalleh, M.; Lindén, J.: Editorial: Rise of Low-Cost Sensors and Citizen Science in Air Quality Studies. Frontiers in Environmental Science 10/, 2022.

[Pa16] Park, N.-W.: Time-Series Mapping of PM 10 Concentration Using Multi-Gaussian Space-Time Kriging: A Case Study in the Seoul Metropolitan Area, Korea. Advances in Meteorology 2016/, pp. 1–10, Jan. 2016.

[RW05] Rasmussen, C. E.; Williams, C. K. I.: Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press, 2005.

[Sc17] Schneider, P. et al.: Mapping urban air quality in near real-time using observations from low-cost sensors and model information. Environment International 106/, pp. 234–247, 2017.

[So22] Sokhi, R. S. et al.: Advances in air quality research – current and emerging challenges. Atmospheric Chemistry and Physics 22/7, pp. 4615–4703, 2022.

[Su16] Sun, L. et al.: Impact of Land-Use and Land-Cover Change on urban air quality in representative cities of China. Journal of Atmospheric and Solar-Terrestrial Physics 142/, pp. 43–54, 2016.

[TRB21] Tremper, P.; Riedel, T.; Budde, M.: Spatial Interpolation of Air Quality Data with Multidimensional Gaussian Processes, INFORMATIK 2021, 2021.

[Yu16] Yu, R.; Yang, Y.; Yang, L.; Han, G.; Move, O. A.: RAQ–A random forest approach for predicting air quality in urban sensing systems. Sensors 16/1, p. 86, 2016.

[Zh18] Zheng, H.; Li, H.; Lu, X.; Ruan, T.: A Multiple Kernel Learning Approach for Air Quality Prediction. Advances in Meteorology 2018/, p. 3506394, 2018.

[Zh21a] Zhang, H. et al.: Using Kriging incorporated with wind direction to investigate ground-level PM2.5 concentration. Science of The Total Environment 751/, p. 141813, 2021.

[Zh21b] Zhou, H.; Zhang, F.; Du, Z.; Liu, R.: Forecasting PM2.5 using hybrid graph convolution-based model considering dynamic wind-field to offer the benefit of spatial interpretability. Environmental Pollution 273/, p. 116473, 2021.