***Anton Sokolov*[(1)], *Hervé Delbarre*[(1)], *Daniil Boldyriev*[(2)], *Tetiana Bulana*[(3)], *Bohdan Molodets*[(2)], *and Dmytro Grabovets*[(4)]**

Industrial Atmospheric Pollution Estimation Using Gaussian Process Regression

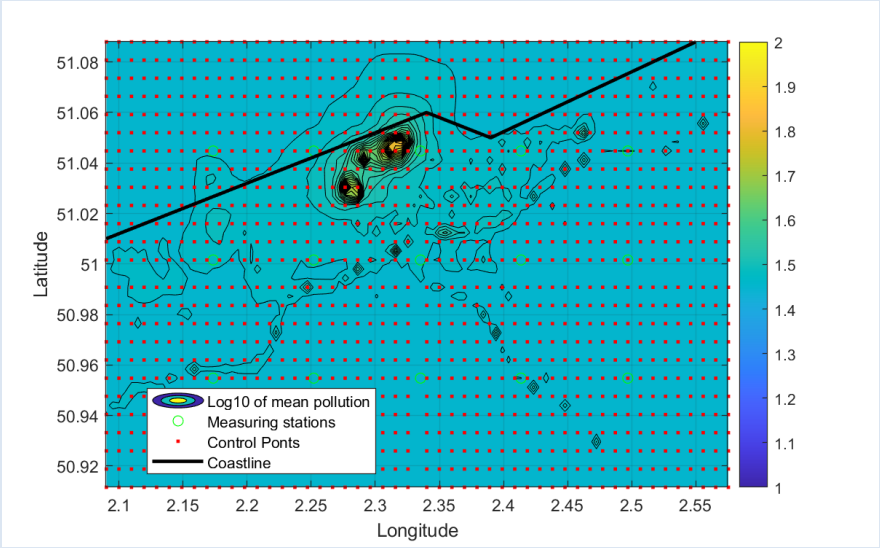
Session ESSI1.1 - Strategies and Applications of AI and ML in a Spatiotemporal Context
Oral presentation: **PICO spot 2** on Wednesday, 26 April 2023, 14:23
PICO **Screen: PICO2.10**, Wednesday, 26 April 2023, 14:33-15:45

(1) Laboratory of Physics And Chemistry of Atmosphere, University of Littoral Côte d'Opale & Univ Lille Nord de France, Dunkirk, France
(2) Department of Mathematical Support of Calculating Machines, Oles Honchar Dnipro National University, Dnipro, Ukraine
(3) Department of Information Technology and Computer Engineering, Dnipro University of Technology, Dnipro, Ukraine
(4) Alfred Nobel University, Noosphere Engineering School, Dnipro, Ukraine

Problem: Spatial Interpolation of Air Pollution

Case of dispersed pollution:

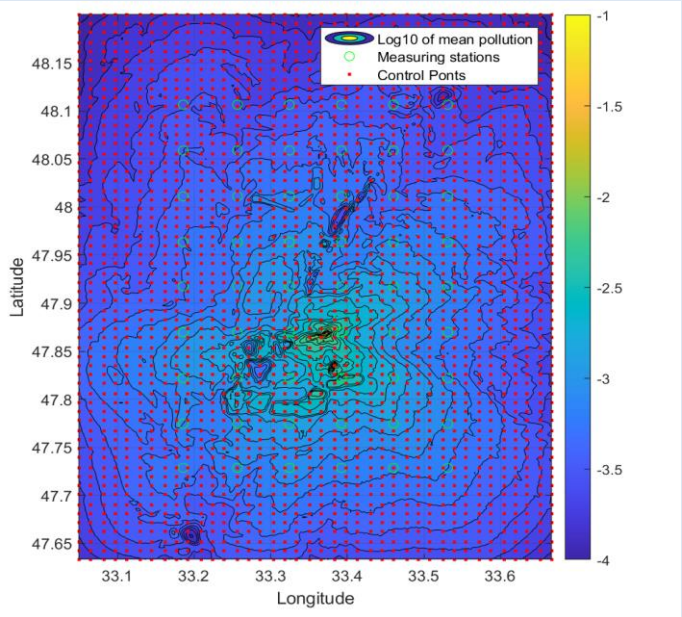
- **PM10** (µg/m^3) **simulated by (**coarse**) ADMS model**



- Industrialized Dunkerque region, North of France
- 15 measurement stations are simulated
- Interpolation based on Delaney triangulation
- **R2 = 0.85** ☺

Smoke-like narrow-directed industrial pollution:

- **SO2** (µg/m^3) **simulated by (**fine**) CALPUFF model**



- Industrialized Dnipropetrovsk Oblast, Ukraine
- 54 measurement stations are simulated
- Interpolation based on Delaney triangulation
- **R2 = 0.13** ☹

Screen: PICO2.10

PICO
PRESENTING INTERACTIVE CONTENT®

- Interpolation with a **GPR based on temporal covariance**
- **R2 = 0.65** ☺

Plan of the talk

- Regions to be explored and simulated pollution data
- Pollution modelling: a few snapshots
- Simple Interpolation techniques in 2D
- Simple interpolation: results
- Gaussian Process Regression Interpolation
- GPR with explicitly estimated kernel: results
- GPR implementation: discussion
- Conclusions and perspectives

Regions to be explored and simulated pollution data

Dunkirk region, North of France

Sources of PM10 pollution:

- Local industry
- Sea port and English channel boat circulation
- Local transport
- Far away sources

Data on pollution: air quality agency ATMO-Nord (https://www.atmo-hdf.fr/)

Dunkirk with superposed population density

Pollution dispersion model **ADMS** (40 x 20 km) was used to simulate PM10 pollution
One year of modeling, pollution at surface level, resolution ~one hour, ~500 m

Dnipropetrovsk Oblast, Ukraine

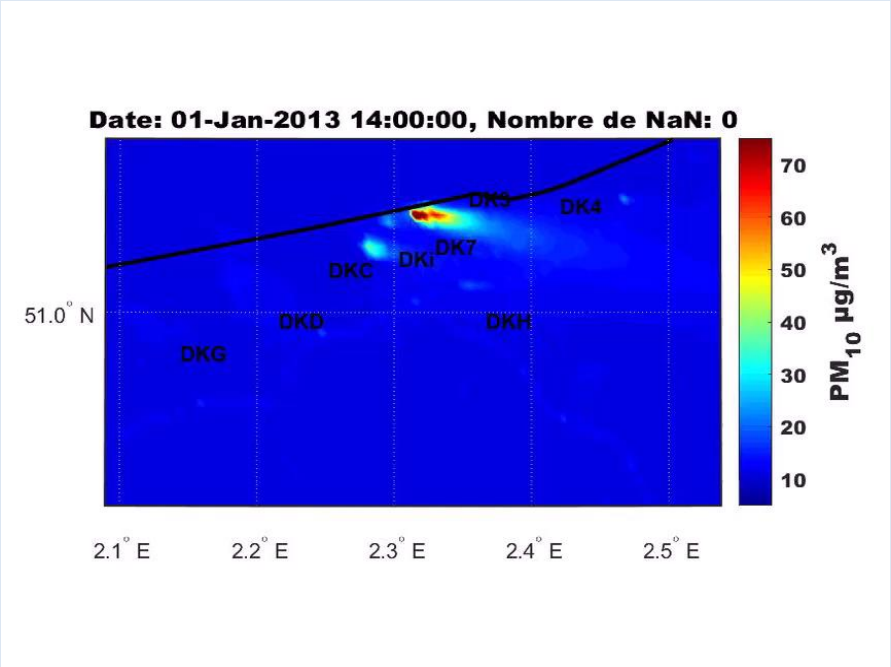
Sources of SO2 pollution:

- Local industry is a principal source of pollution
- Data on pollution: (https://partner.youairtest.com/)
- Transport and dispersion model **CALPUFF** was used to simulate **SO2** pollution,
- 40 x 60 km
- One year of modeling (2019), pollution at surface level, resolution ~one hour, ~100 m

Pollution modelling: a few snapshots

Dispersed pollution:

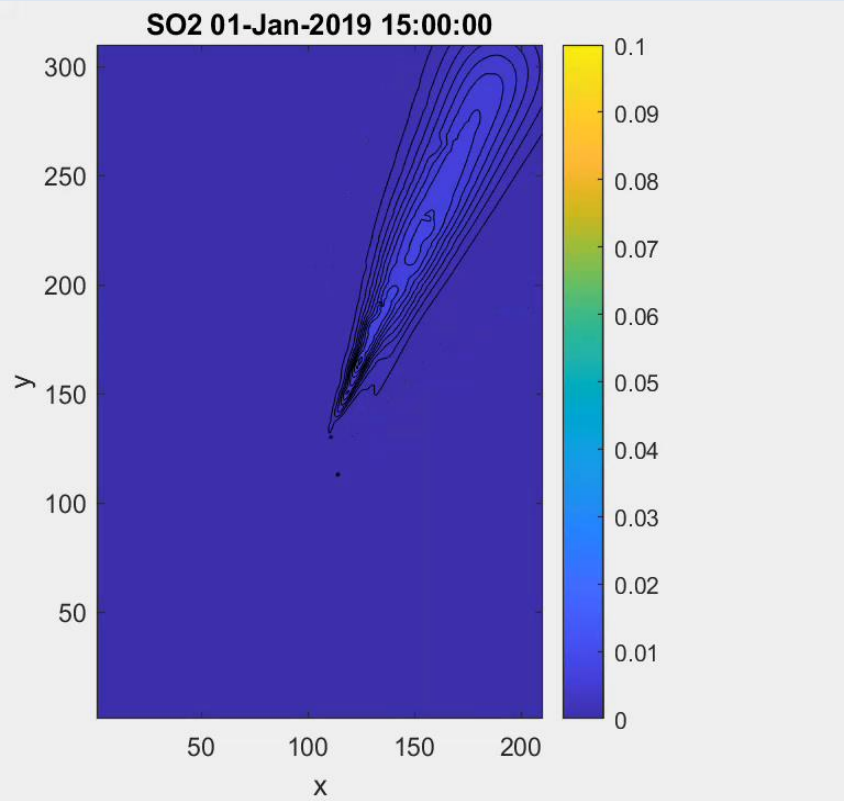
- **PM10** (µg/m^3) **simulated by ADMS model** (*)
- Industrialized Dunkerque region, North of France



(*) Sokolov et al. Optimization of environmental sensors placement in geophysical research. Proceedings of the 7th International Conference on Sensors Engineering and Electronics Instrumentation Advances (SEIA' 2021)

Smoke-like narrow-directed industrial pollution:

- **SO2** (µg/m^3) **simulated by CALPUFF model**
- Industrialized Dnipropetrovsk Oblast, Ukraine



Simple Interpolation techniques in 2D

- ***These techniques do not take into account a time correlation between values of pollution at different points***

Inverse Distance Weighting

$$y(x) = \frac{\sum_{i=0}^{N} w_i(x) y_i}{\sum_{i=0}^{N} w_i(x)} \qquad w_i(x) = \frac{1}{\mathrm{d}(x, x_i)^p}$$

y : estimated PM_{10} pollution value at x position,

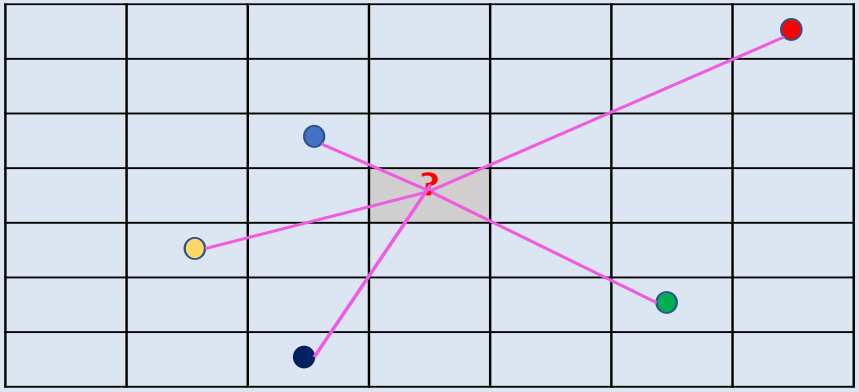
y_i: known values of the pollution

x_i: position of sensors,

$\mathrm{d}(x, x_i)$: the distance between x and x_i,

n : number of measurements,

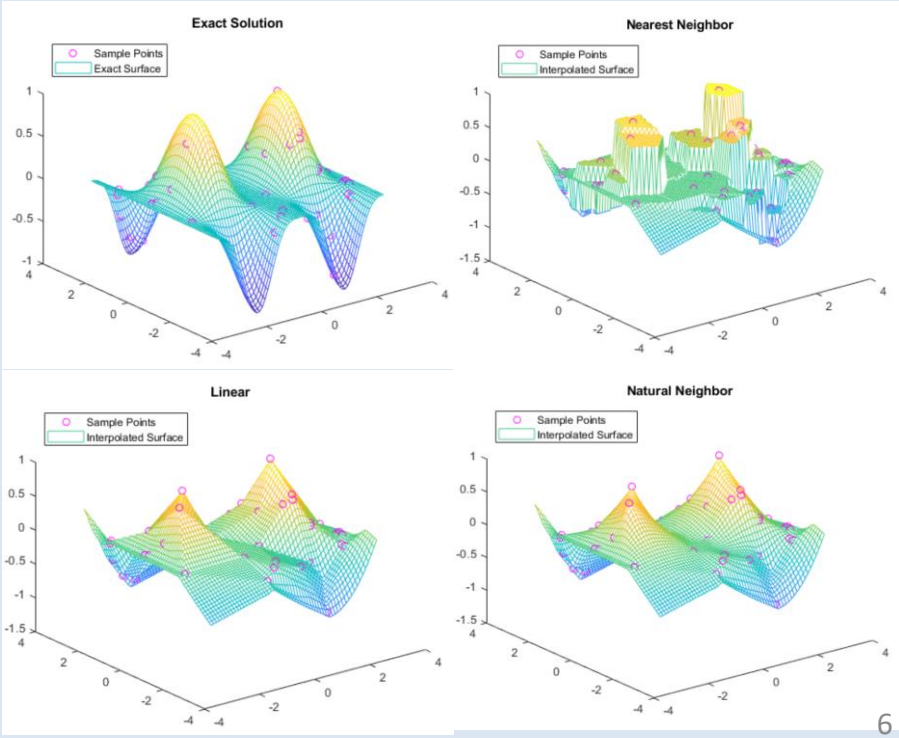
p : some hyperparameter



Delaunay Triangulation-based methods:

- Nearest neighbor (NN)
- Linier interpolation*
- Natural neighbor *

* NN method was used for extrapolation

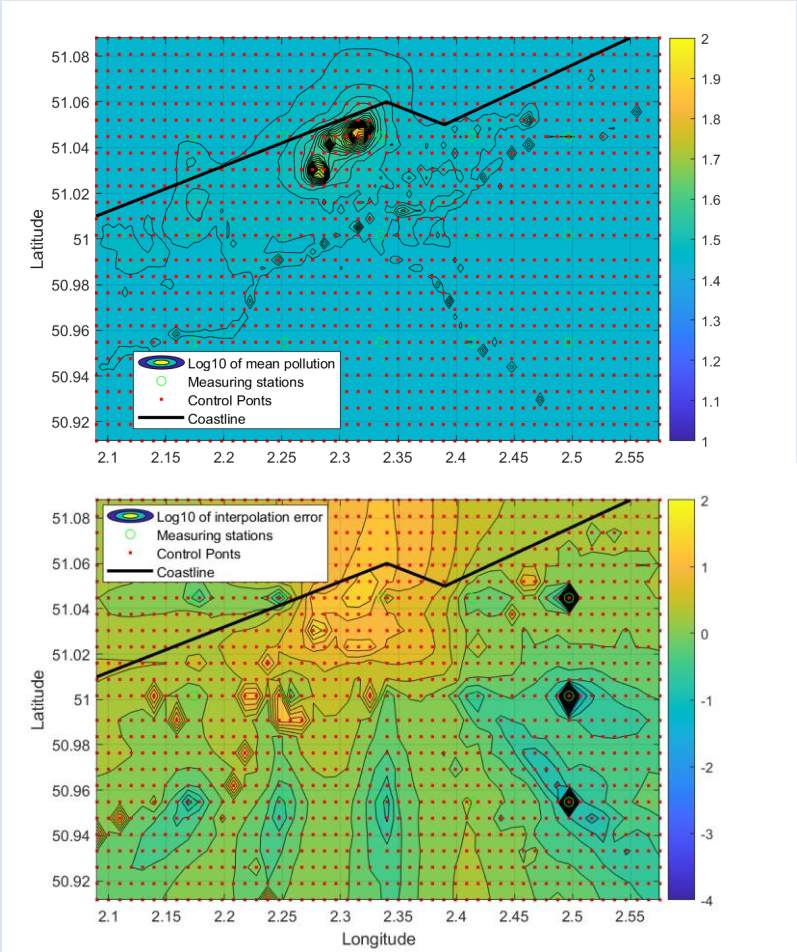


Examples are taken from Matlab software help for function $v(x,y)=\sin^4(x)\cdot\cos(y)$

Simple interpolation: Results

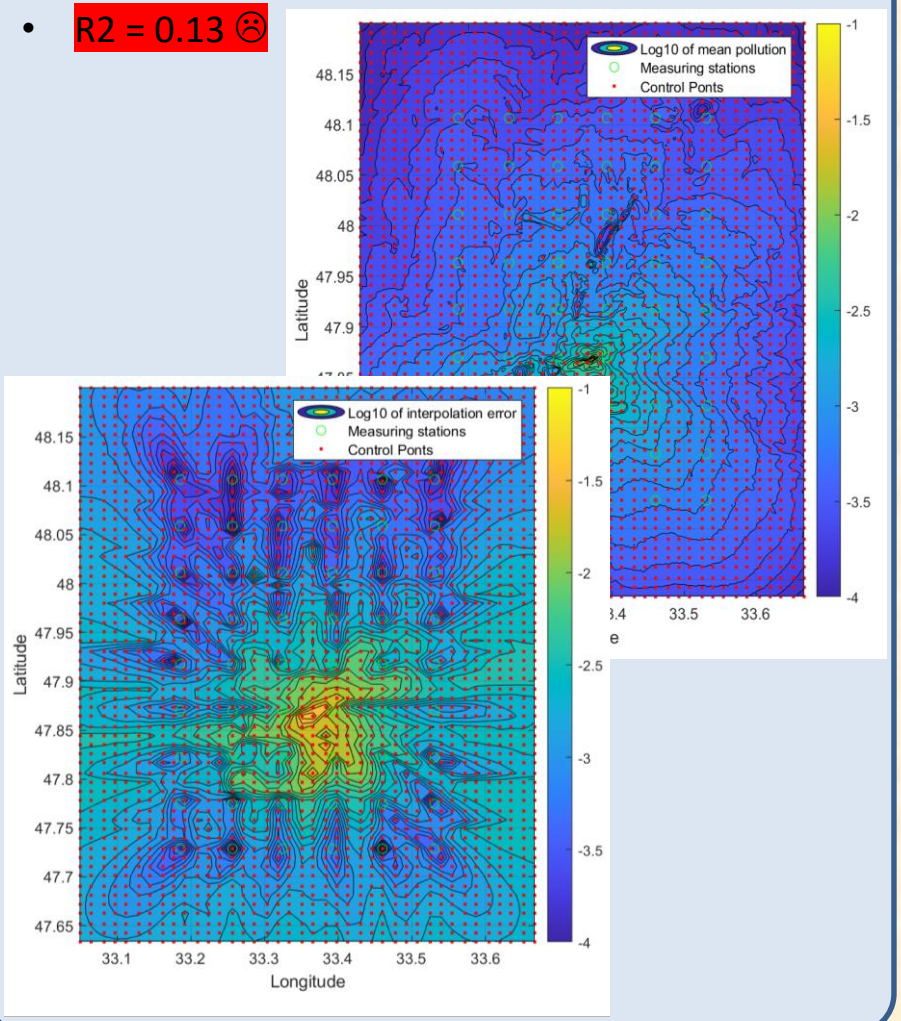
Case of dispersed pollution:

- **PM10** (μg/m^3) **simulated by ADMS model**
- 15 measurement stations, 1250 control points
- R2 = 0.85 ☺



Smoke-like narrow-directed industrial pollution:

- **SO2** (μg/m^3) **simulated by CALPUFF model**
- 54 measurement stations, 2400 control points
- R2 = 0.13 ☹



Gaussian Process Regression Interpolation

Gaussian process (GP) is a collection of random variables, any finite number of which follows a multivariate normal distribution (MVN):

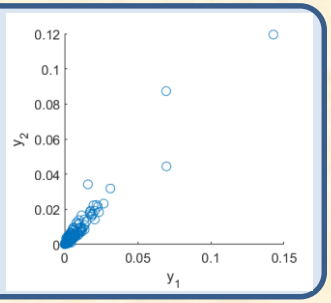
$$y = f(\boldsymbol{x}) \sim \mathcal{GP}\big(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}')\big),$$

$$m(\boldsymbol{x}) = E[f(\boldsymbol{x})],$$

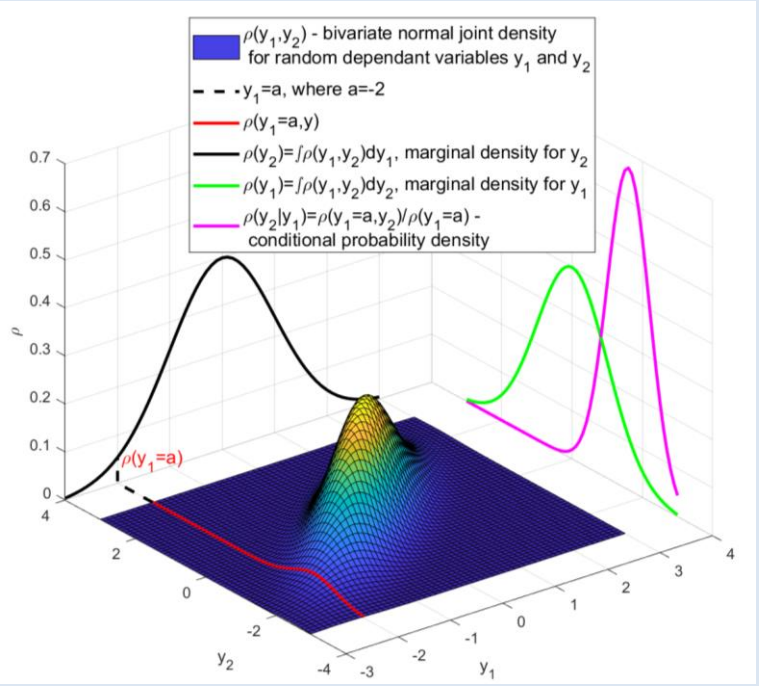
$$k(\boldsymbol{x}, \boldsymbol{x}') = E\big[\big(f(\boldsymbol{x}) - m(\boldsymbol{x})\big)\big(f(\boldsymbol{x}') - m(\boldsymbol{x}')\big)\big]$$

- $\boldsymbol{x}$ corresponds to position (latitude, longitude),
- $y = f(\boldsymbol{x})$ corresponds to a pollution value at $\boldsymbol{x}$
- mean $m(\boldsymbol{x})$ and kernel $k(\boldsymbol{x}, \boldsymbol{x}')$ should be somehow estimated from the available data
  - often a family of kernels $k(\boldsymbol{x}, \boldsymbol{x}'|\theta)$ depending on an optimized hyperparameter θ are used
  - here, the data of the temporal dimension allows estimating mean and kernel explicitly

- Suppose that we have a measurement of pollution at some point y_1
- The pollution value at a neighboring point y_2 will be highly correlated with y_1



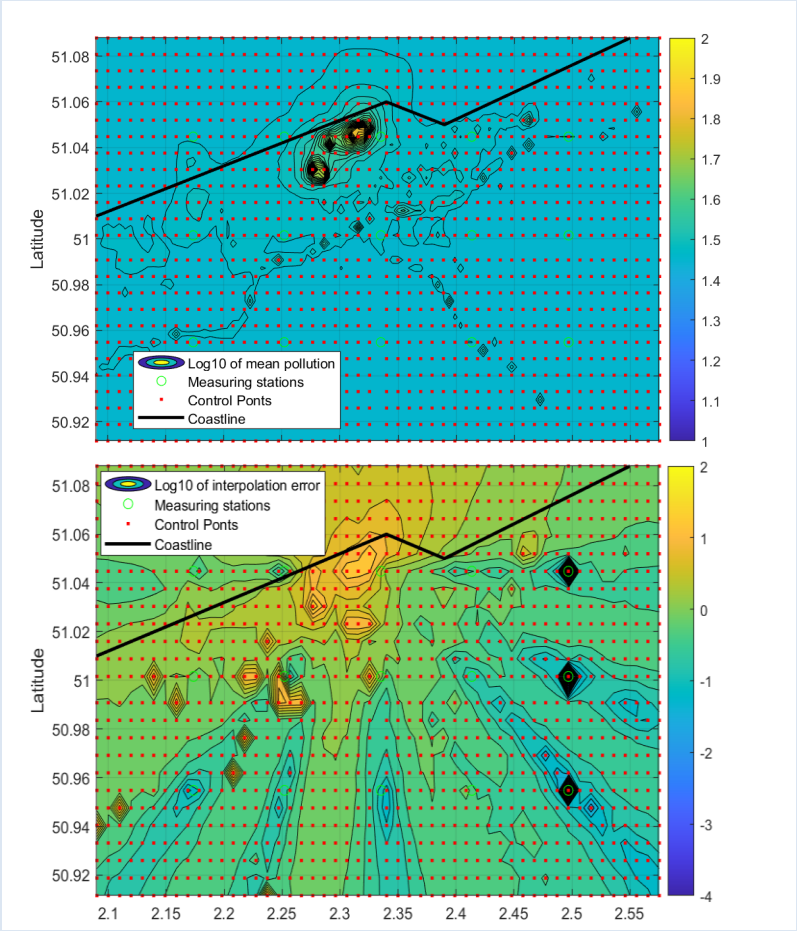
- If the joint probability density function (PDF) $\rho(y_1, y_2)$ is known, we can determine the posterior probability for y_2
- In particular, If y_1 and y_2 follows MVN, a posterior (conditional) $y_2|y_1$ will follow a normal low with known parameters
- Note below, that magenta curve (posterior) is narrow compared to green (marginal)
- Thus, the information on y_1 allows to specify y_2



GPR with explicitly estimated kernel: Results

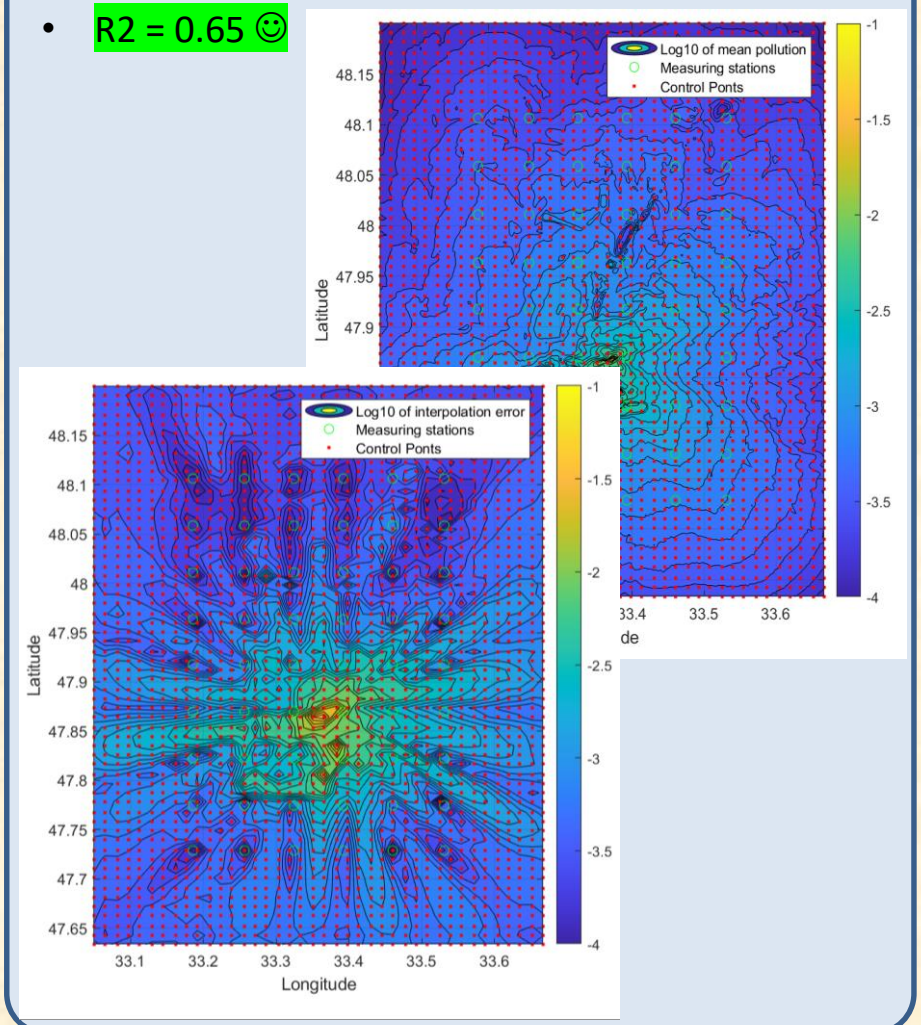
Case of dispersed pollution:

- **PM10** ($\mu g/m^3$) **simulated by ADMS model**
- 15 measurement stations, 1250 control points
- R2 = 0.96 ☺



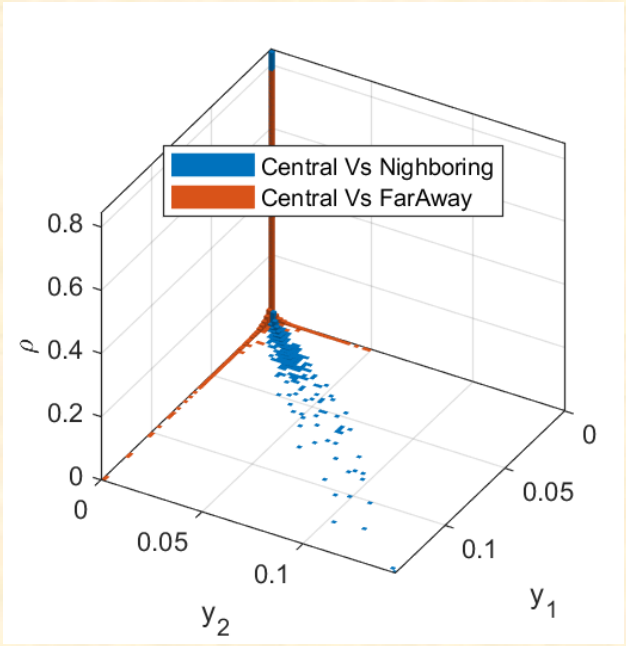
Smoke-like narrow-directed industrial pollution:

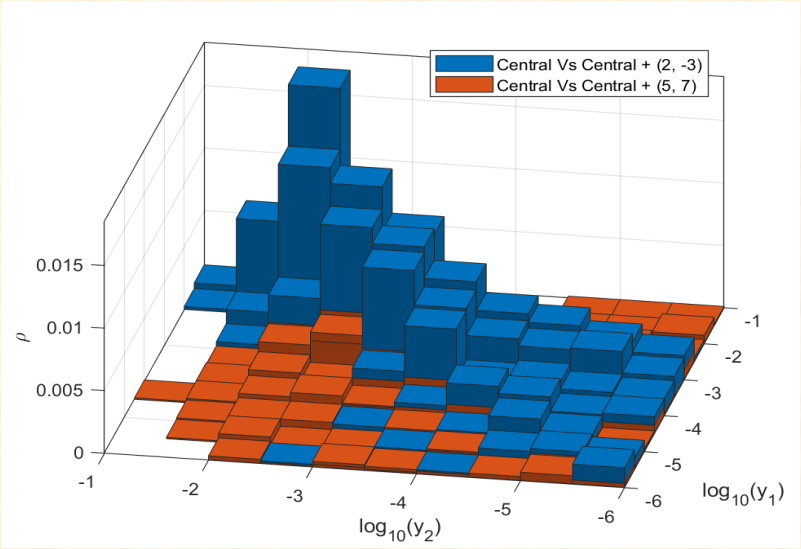
- **SO2** ($\mu g/m^3$) **simulated by CALPUFF model**
- 54 measurement stations, 2400 control points
- R2 = 0.65 ☺



GPR implementation : Discussion

- For kernel calculation pollution data has been smoothed and coarsened
- **Problem**: 5% of negative pollution values ☹
- **Problem**: in GPR variables should follow MVN distribution
- Is it the case?
  - The answer is no… ☹
- And if we use log transformation of the pollution?
  - What should we do with **zero values**?
  - Still not an MVN... ☹





- Two possible solutions could further be explored:

1) Find some other one to one transformation that allow getting an MVN distribution: Wrapping function. Unfortunately standard BoxCox transformation does not work…
2) Approximate the joint PDF by a (multivariate) gaussian mixture

Conclusions

- Simple 2D techniques are not suitable for efficient interpolation of smoke-like industrial pollution at scales of a few kilometers.
- This kind of problem can be tackled with GPR with the kernel explicitly estimated on the basis of time-depended pollution data.
- The proposed approaches was verified on two test-cases:
  - Dispersed PM10 pollution simulated by ADMS model for Dunkirk region
  - Smoke-like SO2 pollution simulated by CALPUFF model for Dnipropetrovsk Oblast
- A sufficient precision of interpolation was achieved for these test cases.

Perspectives

- Optimization of smoothing of the kernel function.
- Rigorous statistical validation
- Wrapping of pollution data to obtain MVN distribution in GPR
- Optimization of the observation network

Thank you for your attention…