Reduced-cost EnKF for parameter estimation of microscale atmospheric pollutant dispersion models

Eliott Lumet[1], Mélanie Rochoux[1], Thomas Jaravel[1] and Simon Lacroix[2]

[1]CECI, CNRS, CERFACS, Toulouse, France
[2]LAAS, CNRS, Toulouse, France

18/06/2024

Context

Pollutant dispersion is involved in many applications related to safety/public health (industrial accident, traffic air pollution, wildfire smoke)

Lubrizol factory fire in Rouen in September 2019
LP/JEAN PIERRE MAUGER

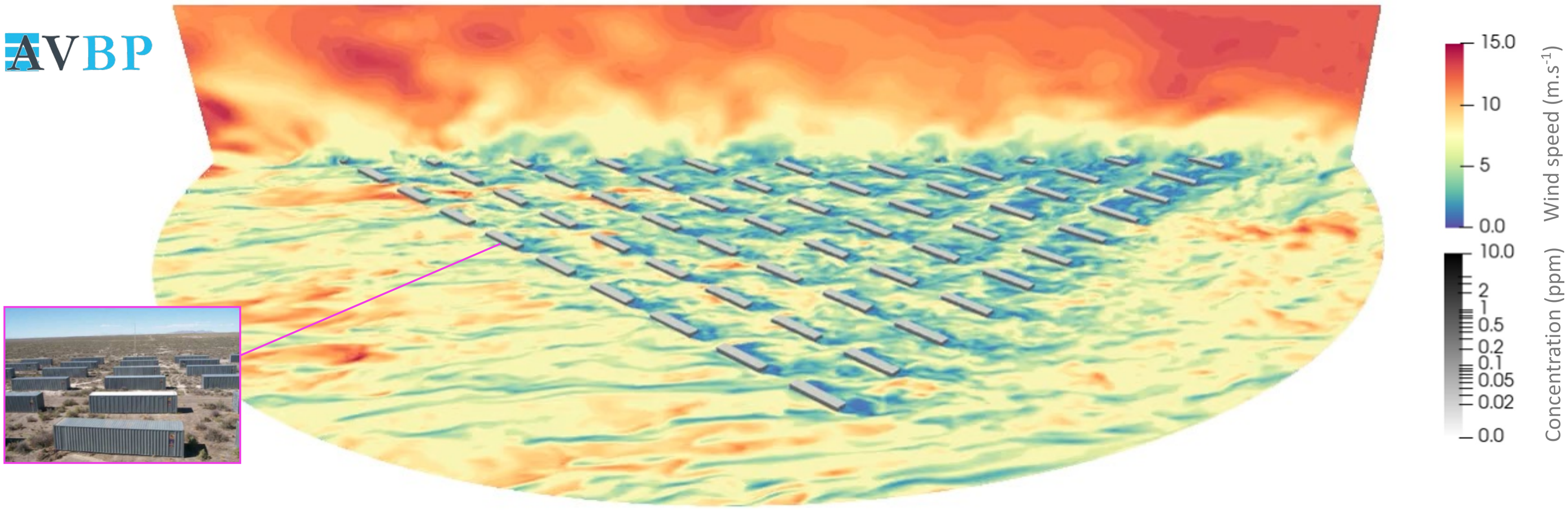
- **Focus on microscale dispersion in urban environments**
- **Large-eddy simulation (LES) modeling**
  - ⊕ Resolves the largest turbulence scales
  - ⊕ Explicitly accounts for the effect of urban buildings on the atmospheric flow and dispersion
  - ⊕ Provides a spatio-temporal description of the phenomenon
  - ⊖ LES still has **large uncertainty** despite its **substantial computational cost** (Dauxois et al. 2021)

General objective

Implement a reduced-cost DA system to reduce uncertainty in LES modeling of urban pollutant dispersion

Context

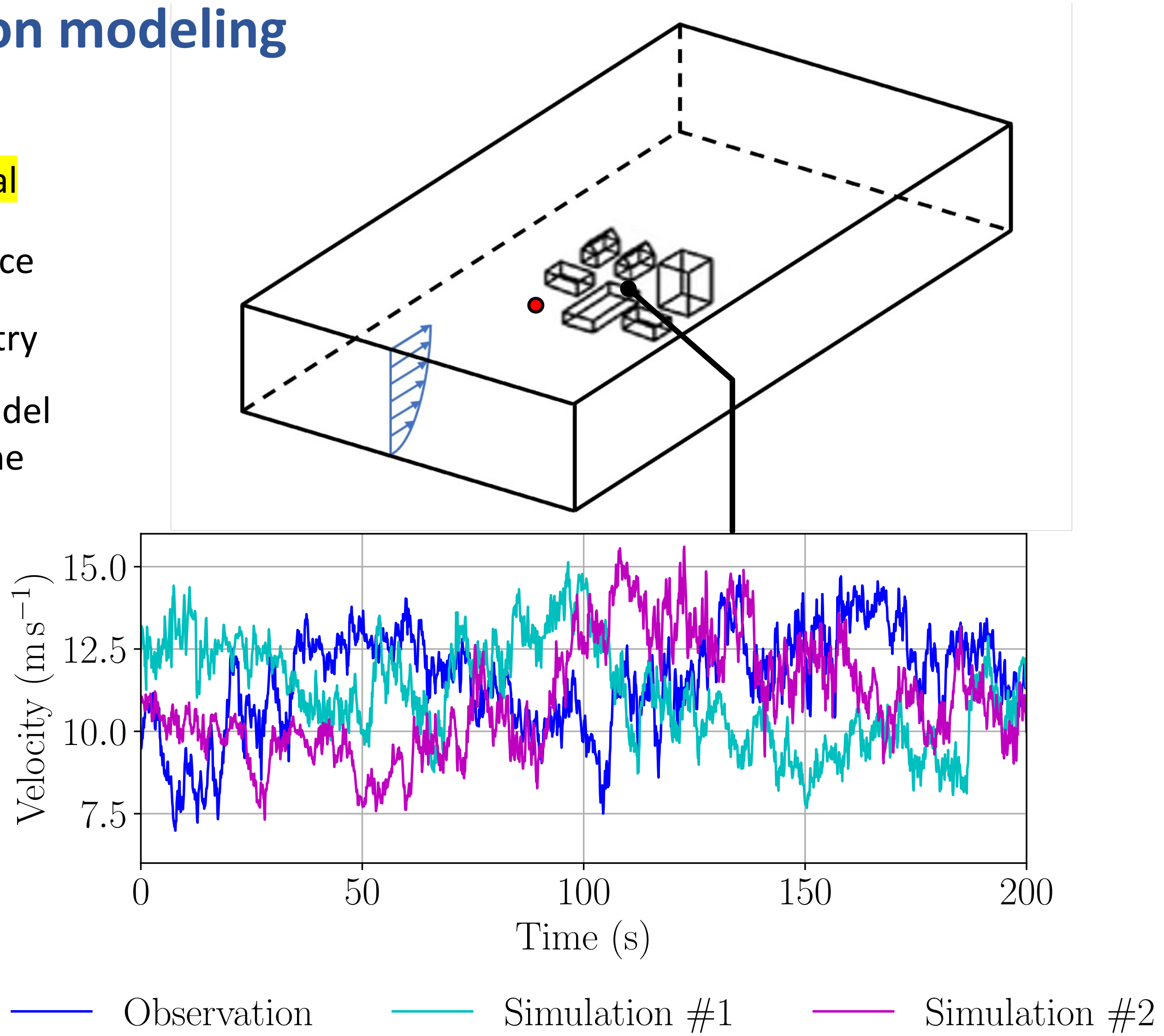
❖ **Case study:** LES of the MUST field campaign trial #2681829 (Yee and Biltoft. 2004)



Context

❖ Uncertainty in LES microscale dispersion modeling

- **Epistemic uncertainty**
  - Boundary conditions
    - Meteorological
    - Pollutant source
    - Urban geometry
  - Structural
    - Subgrid-scale model
    - Numerical scheme
- **Aleatory uncertainty**
  - Microscale internal variability of the ABL



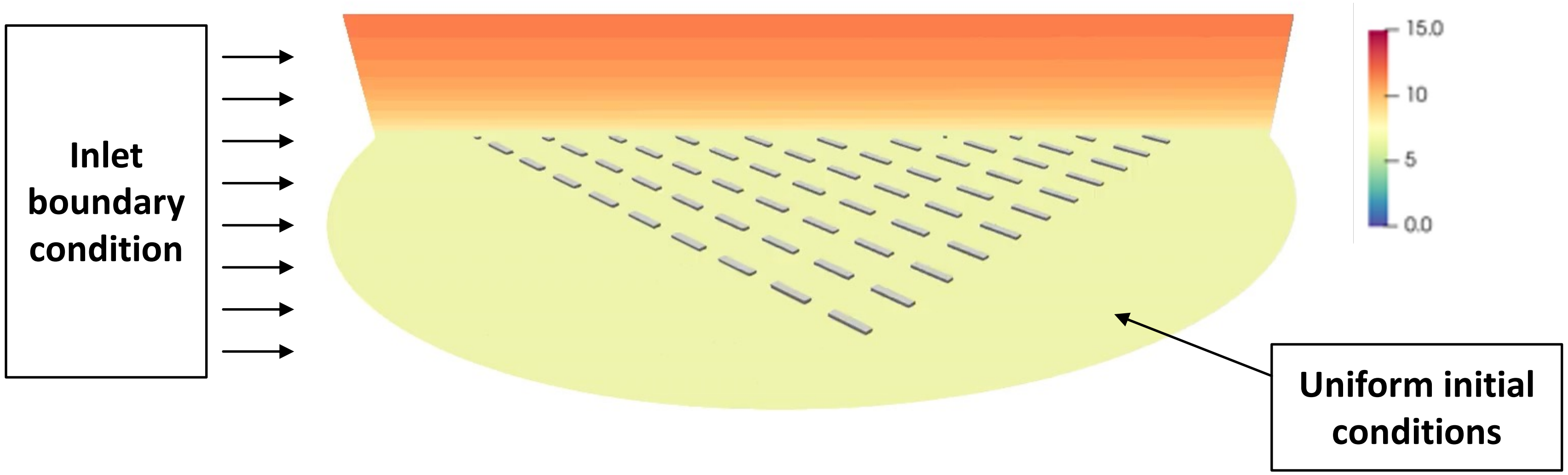
Focus on uncertainties related to atmospheric conditions

Approach

❖ State estimation versus parameter estimation

At the microscale, initial conditions have low persistence

⇒ More relevant to estimate boundary condition parameters*



Mons et al. 2017, Sousa et al. 2018, Sousa et Gorlé. 2019, Defforge et al. 2019, Defforge et al. 2021.

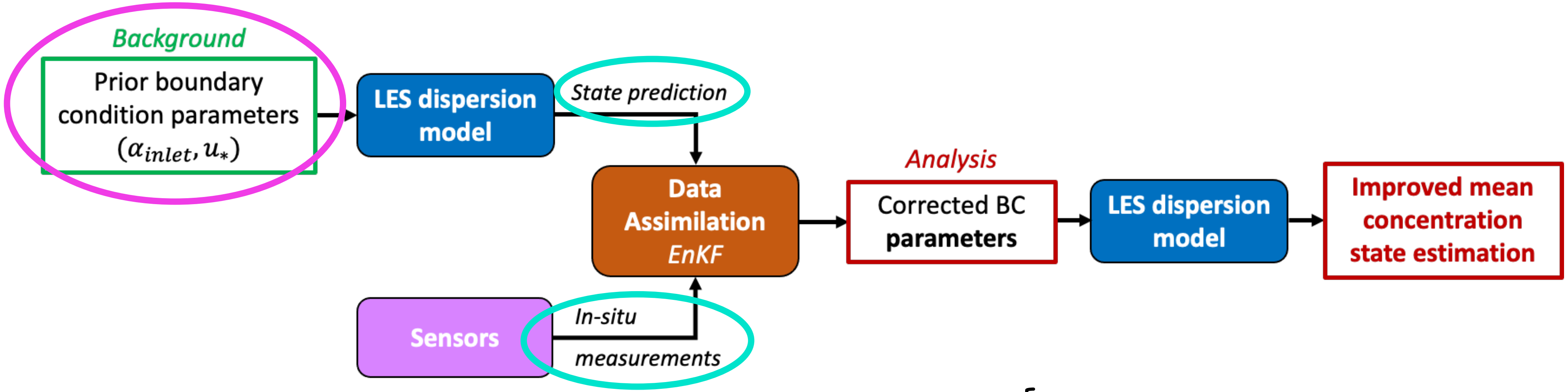
❖ Control vector definition:

$$\boxed{\boldsymbol{\theta} = (\alpha_{inlet}, u_*)}$$

Inlet wind direction *Inlet wind friction velocity*

Objectives

❖ Data assimilation system design



- **Epistemic** uncertainty on boundary condition parameters: ✓ { *Mons et al. 2017, Sousa and Gorlé. 2019, Defforge et al. 2021*
- **Aleatory** uncertainty associated with the internal variability: ✗

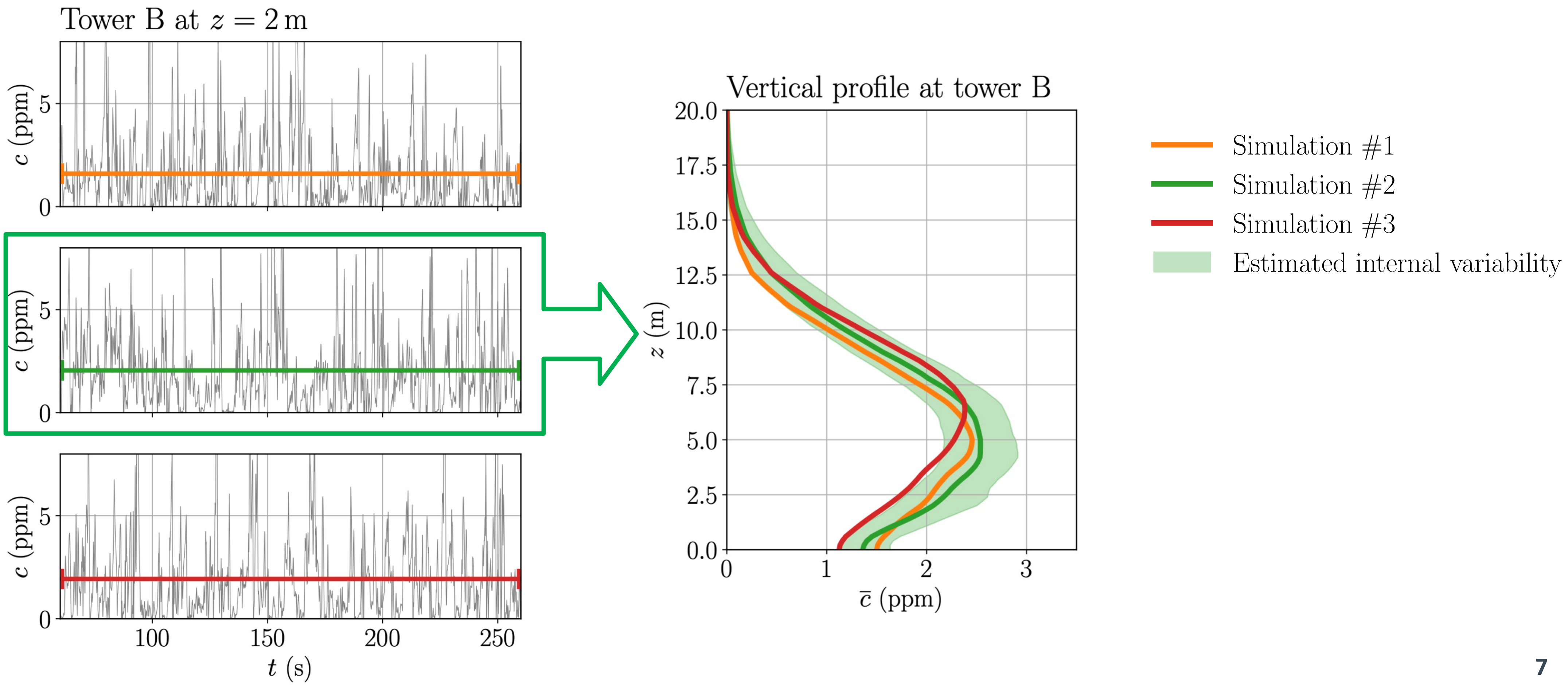
Objective #1

Take account of aleatory uncertainty linked to internal variability in the assimilation system

Objectives

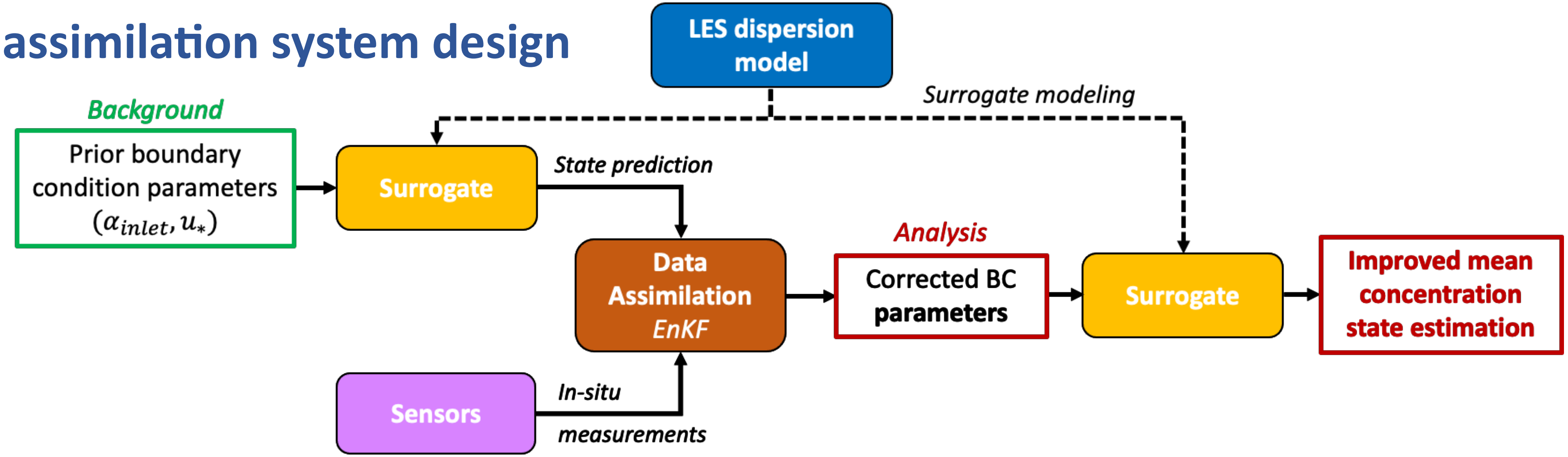
❖ **Effect of the internal variability of the ABL on the mean concentration predictions**

➢ Estimation using stationary bootstrap of LES sub-averages (Lumet et al. 2024)



Objectives

❖ Data assimilation system design



Issue: LES computational cost
1 prediction ≈ 20 000 h_{CPU}

Objective #2

Build a surrogate model to reduce the cost of the DA system without compromising the accuracy of its estimates

Contents

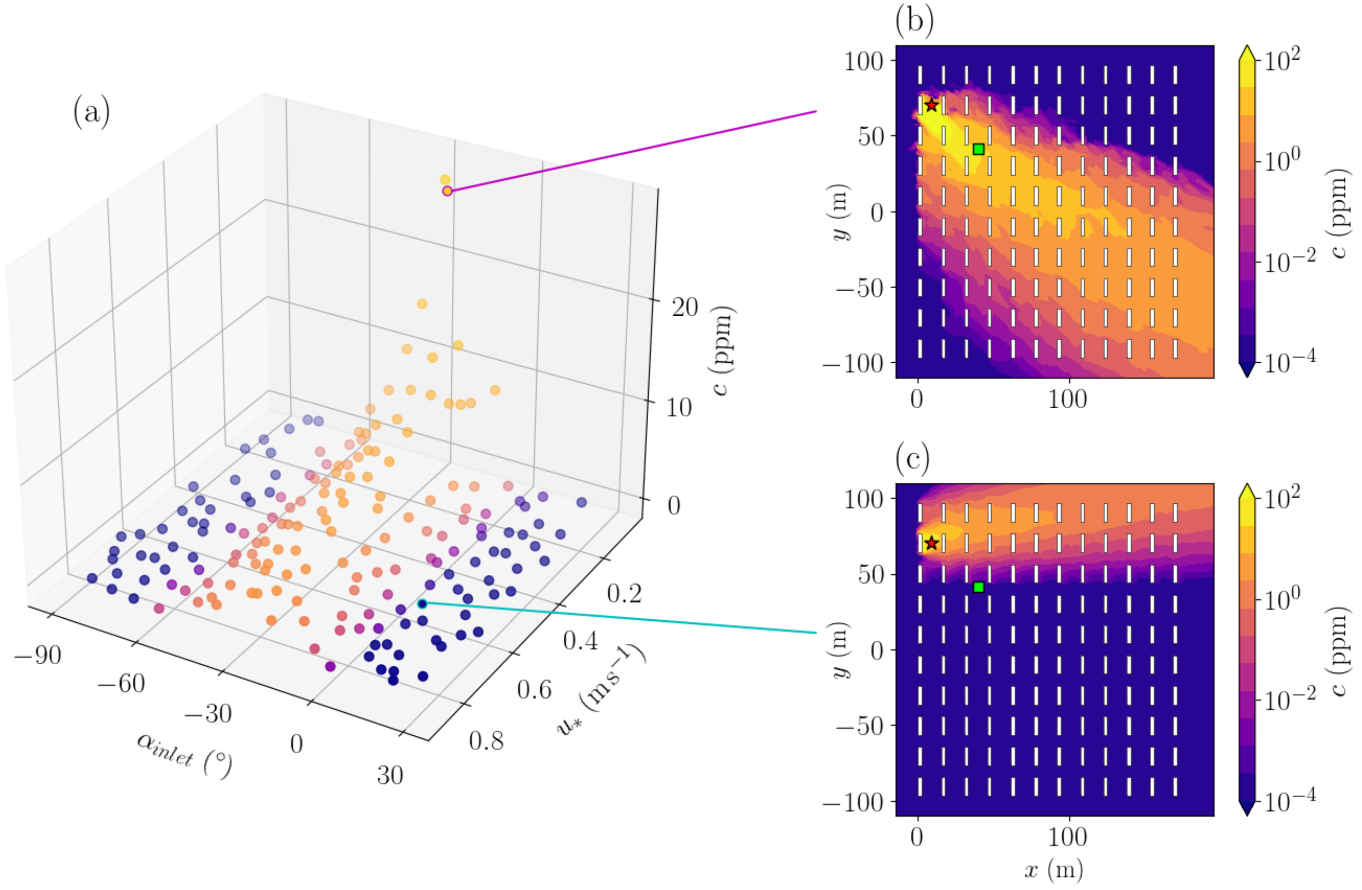
I. Reduced-cost DA system

II. Assimilation of the real field measurements

I – Reduced-cost DA system

❖ Construction of the surrogate model

1) Computation of a dataset of 200 LES using Halton's sequence to sample the control vector space



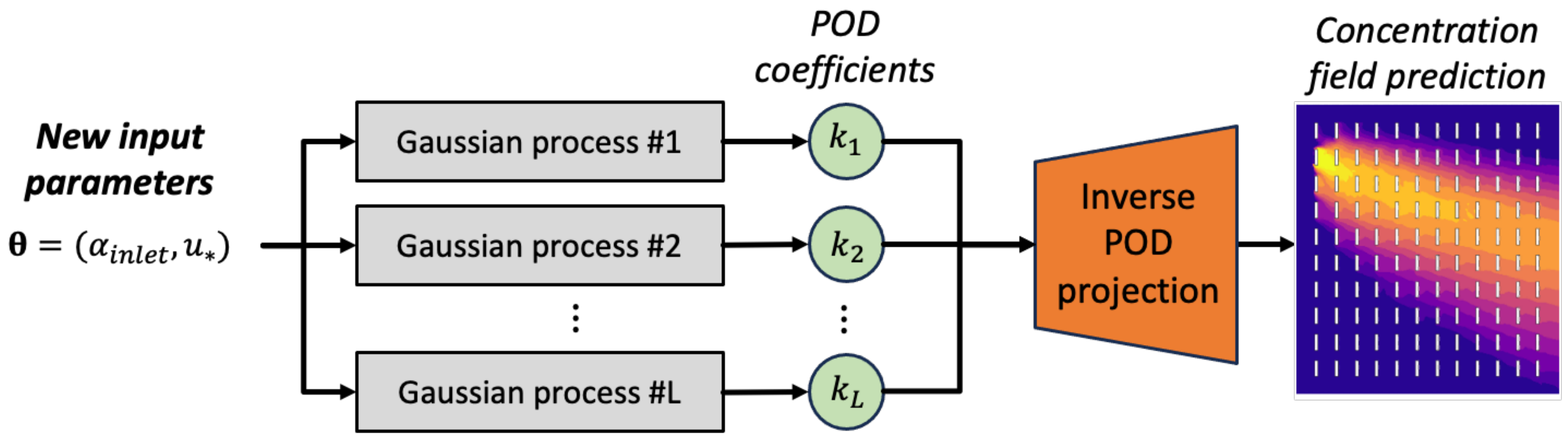
The dataset is being put online in open access on Zenodo

I – Reduced-cost DA system

❖ Construction of the surrogate model

2) Train a POD—GPR surrogate model (Marrel et al. 2015) over 160 LES samples

- Dimension reduction step using Proper Orthogonal Decomposition (**POD**, a.k.a PCA)
- Learn the dependency of POD coefficients on the control vector using Gaussian Process Regression (**GPR**)



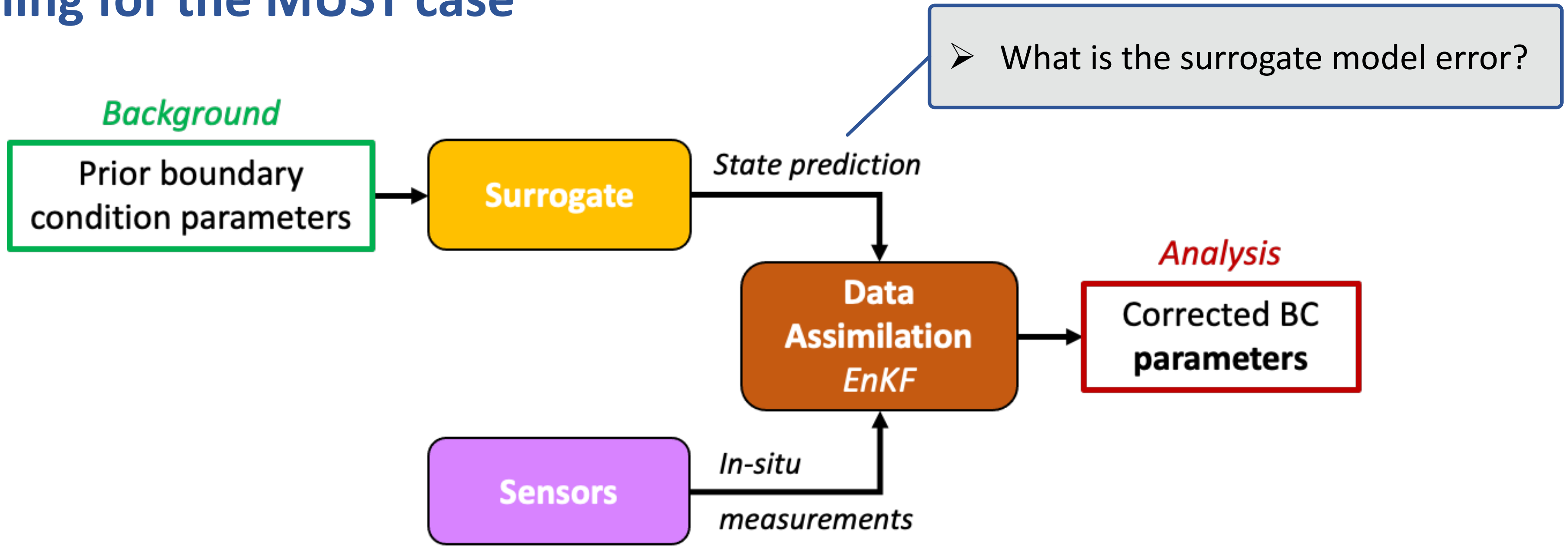
3) Validation against 40 independent LES test samples

- Surrogate error close to the minimal level of error reachable given the LES internal variability

Article presenting the validation of the surrogate model currently being submitted to Building and Environment

I – Reduced-cost DA system

❖ Errors modeling for the MUST case



I – Reduced-cost DA system

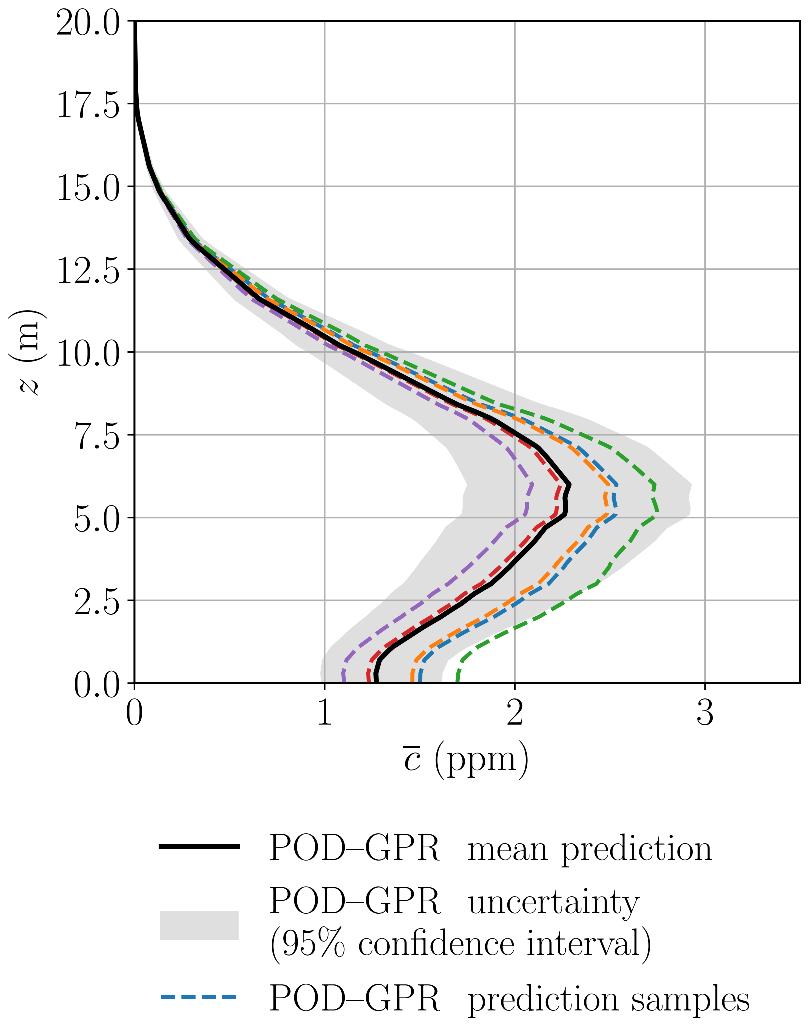
❖ Taking into account model error

- **During the EnKF prediction step:**

$$\boxed{\mathbf{x}^{f}_{(i)} = \mathcal{M}_{(i)}\big(\boldsymbol{\theta}^{b}_{(i)}\big), \qquad 1 \leq i \leq N_e}$$

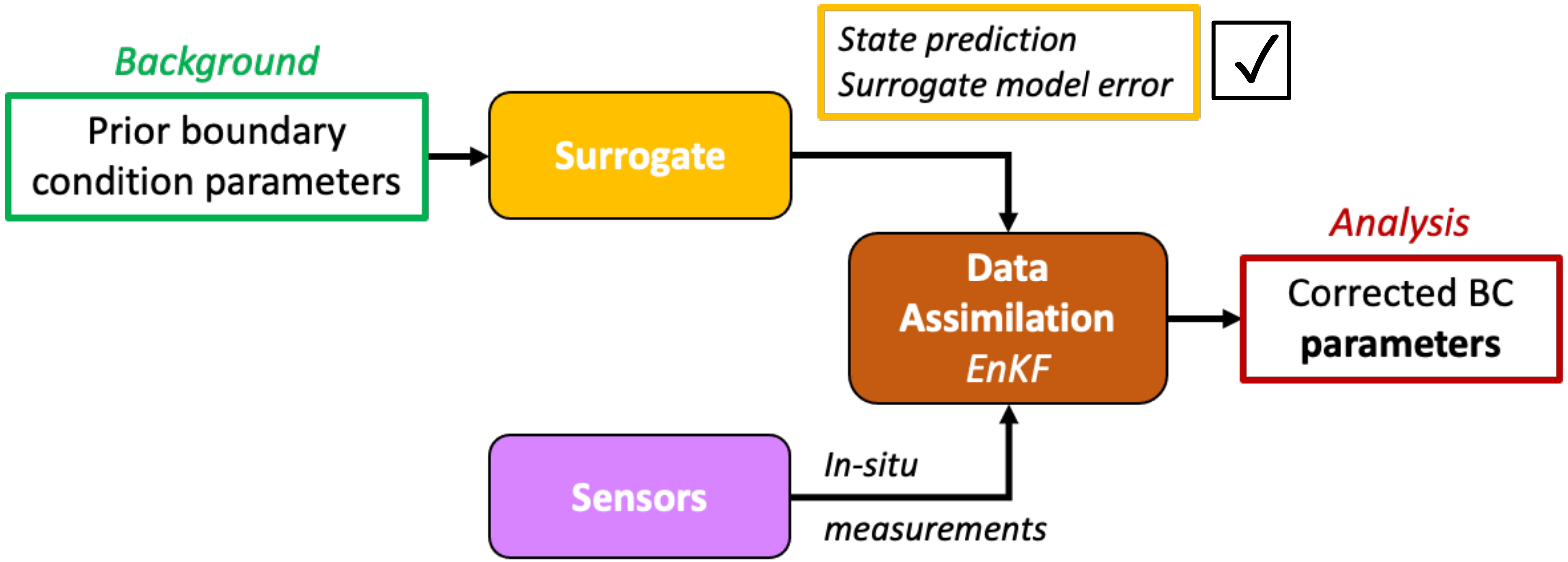
With $\mathcal{M}_{(\mathrm{i})}$ a random sample from the POD—GPR distribution

- We show that the POD—GPR variance covers
  - The regression error of each GP
  - The aleatory uncertainty associated with internal variability
- This approach integrates the spatial correlations of the errors



I – Reduced-cost DA system

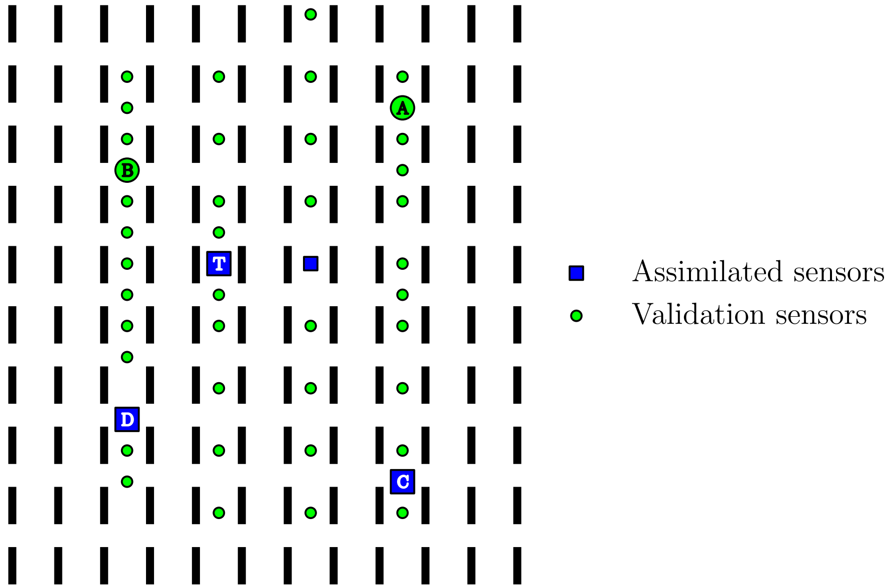
❖ Errors modeling for the MUST case

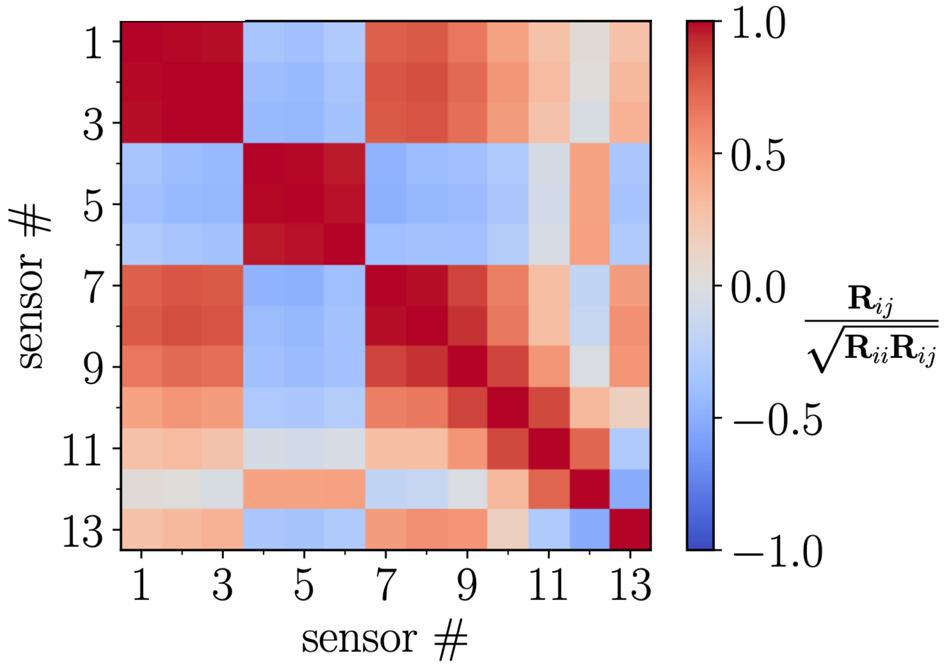


- ➢ Which measurements do we assimilate?
- ➢ What is their uncertainty $\mathbf{R}$?

I – Reduced-cost DA system

- **Observation vector $\mathbf{y^o}$** $=$ {13 mean concentration measurements from 4 masts as in Defforge et al. 2021}
- **Observation error $\mathbf{e}^o$** $=$ ~~*measurmeent error*~~ $+$ *internal variability error*
- **Observation error covariance matrix $\mathbf{R}$** $=$ estimated using stationary bootstrap (Lumet et al. 2024)

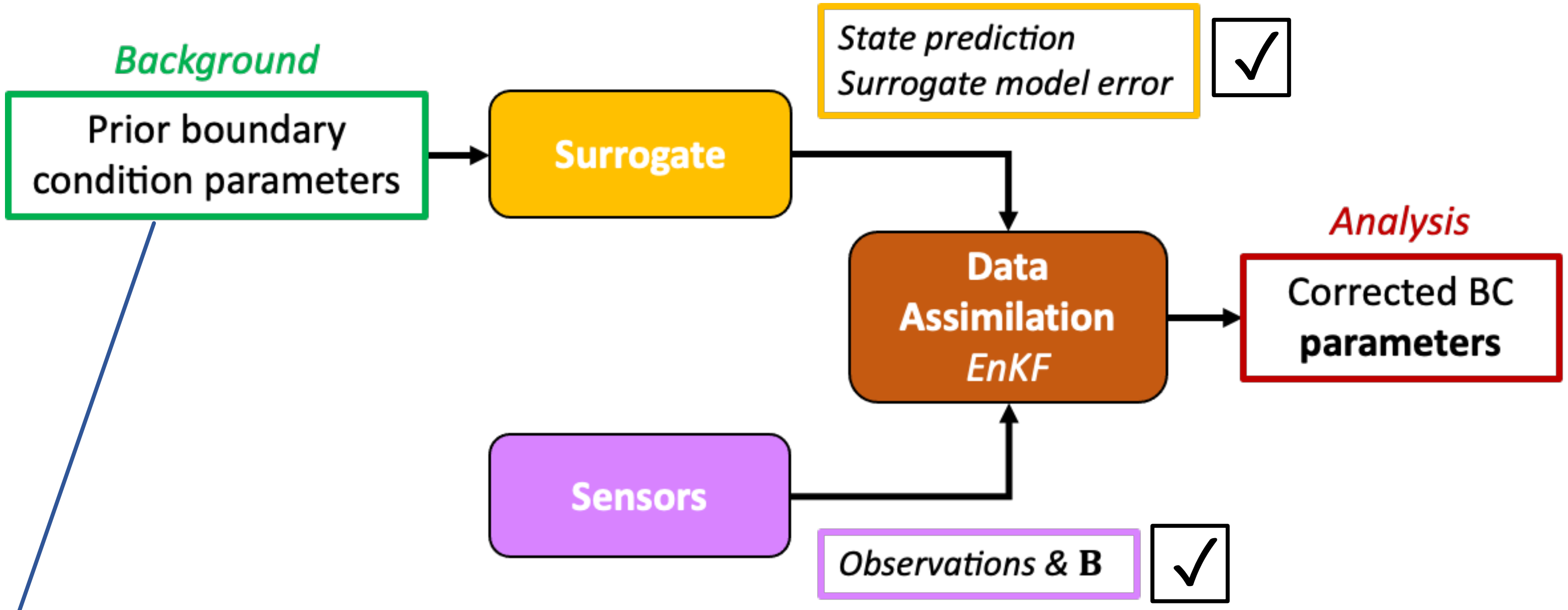




- We can also take into account for spatial correlations of observation errors

I – Reduced-cost DA system

❖ Errors modeling for the MUST case



- **Background error covariance matrix $\mathbf{B} = \mathbf{Cov}(e^b(\alpha_{inlet}), e^b(u_*))$** estimated by a microclimatology
- Log anamorphosis for the friction velocity: $\widetilde{u_*} = \ln(u_* + u_t)$
- ⚠ The hypothesis of a normally distributed error for the wind direction is rejected ($\alpha_{inlet} \in]-\pi, \pi]$)

I – Reduced-cost DA system

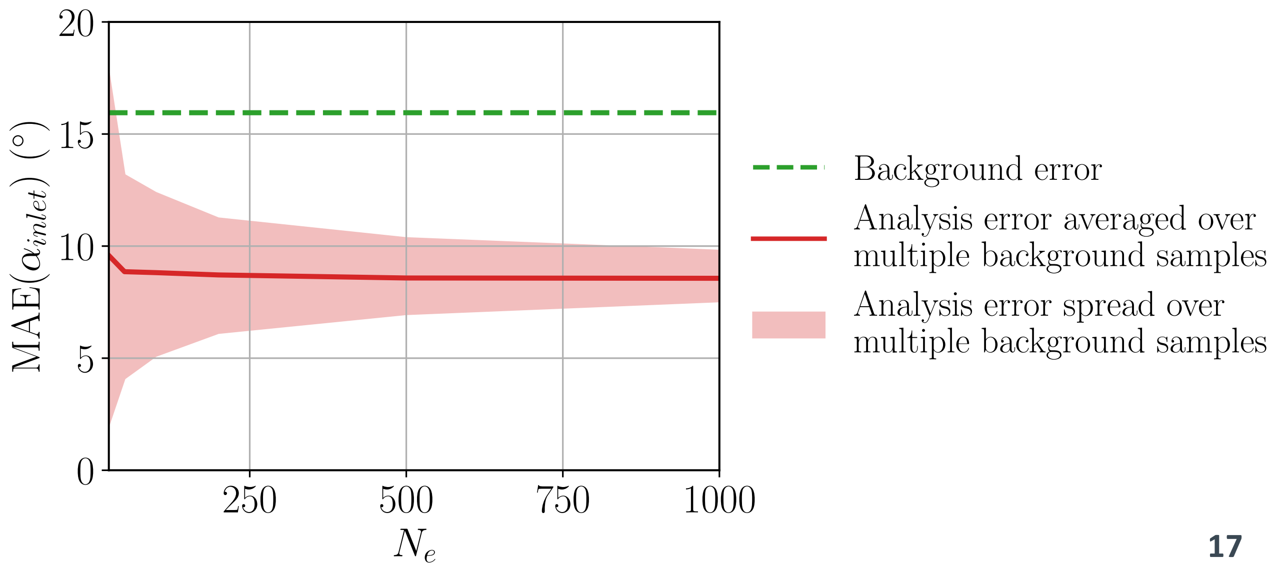
❖ Ensemble size selection using OSSEs

- Compromise between accuracy and computational cost ⇒ $\boxed{N_e = 500}$

1 cycle with 500 surrogate members ≈ **50 s (1 CPU)**

1 cycle with 10 LES members ≈ **200 000 CPU hours**

- Using a surrogate model and large ensemble also allows us to:
  - Reduce sensitivity to background sampling
  - Perform a large number of tests to optimize the DA system and investigate its sensitivities



Contents

I. Reduced-cost DA system

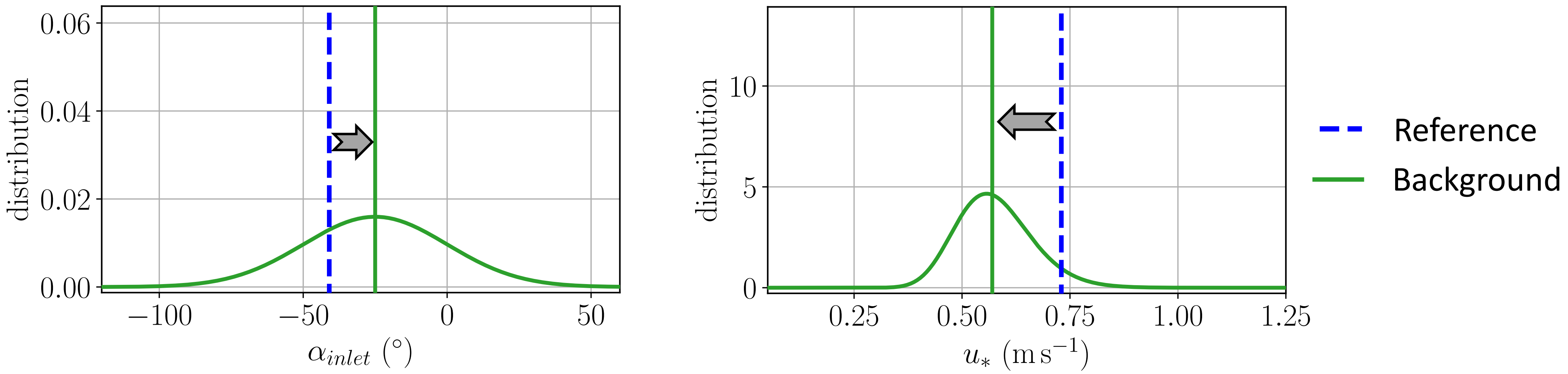
II. Assimilation of the real field measurements

II – Assimilation of the real field measurements

❖ Prior parameters

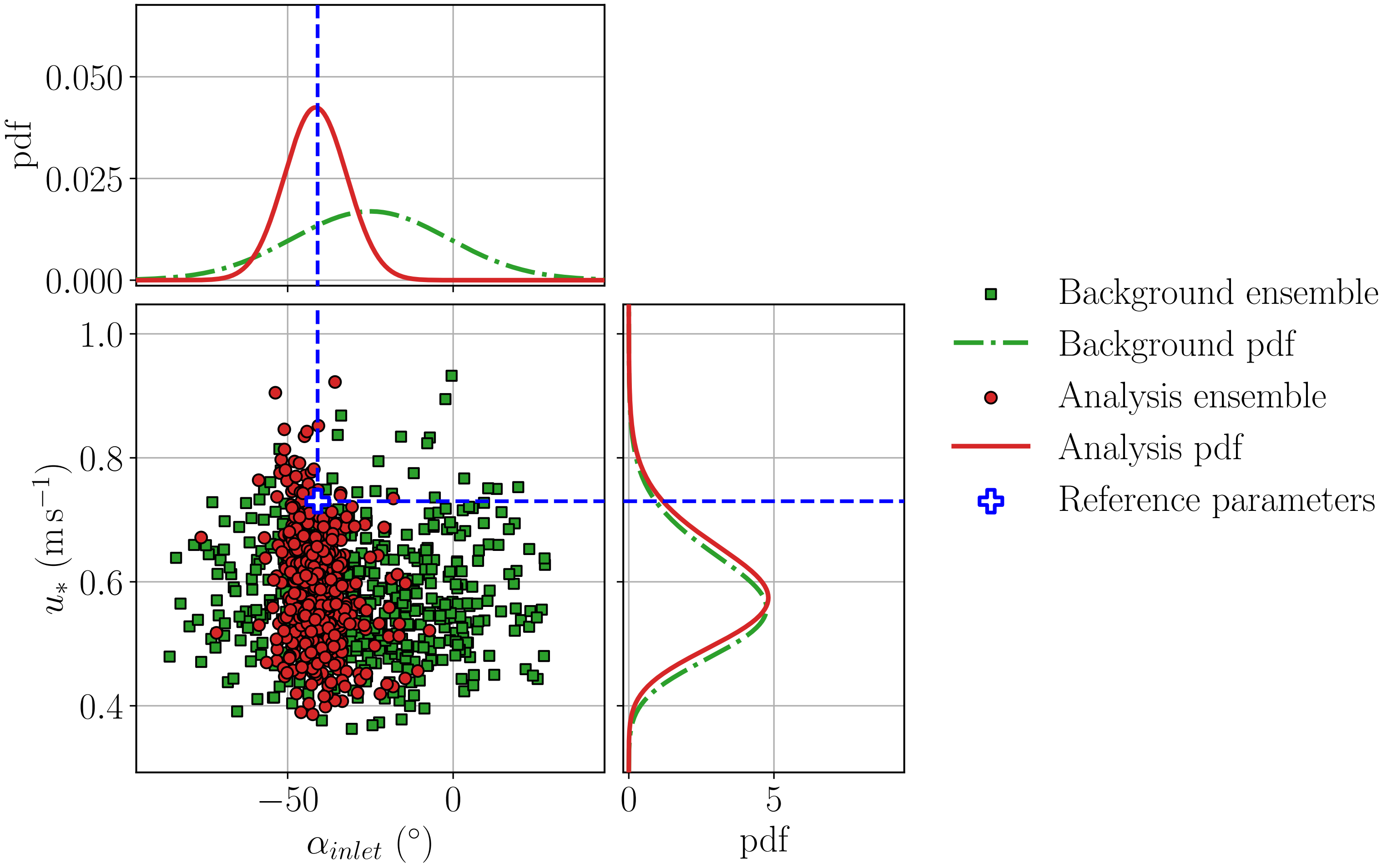
➢ $\boldsymbol{\theta}^b$ obtained by biasing the reference measurements $\boldsymbol{\theta}^{ref}$ defined using the nearest meteorological measurement masts (non-assimilated)

$$\begin{cases} \alpha_{inlet}^{ref} = -41° \\ u_*^{ref} = 0.73 \text{ m.s}^{-1} \end{cases} \Rightarrow \begin{cases} \alpha_{inlet}^{b} = -25°. \\ u_*^{b} = 0.57 \text{ m.s}^{-1} \end{cases}$$



II – Assimilation of the real field measurements

❖ Control vector estimation

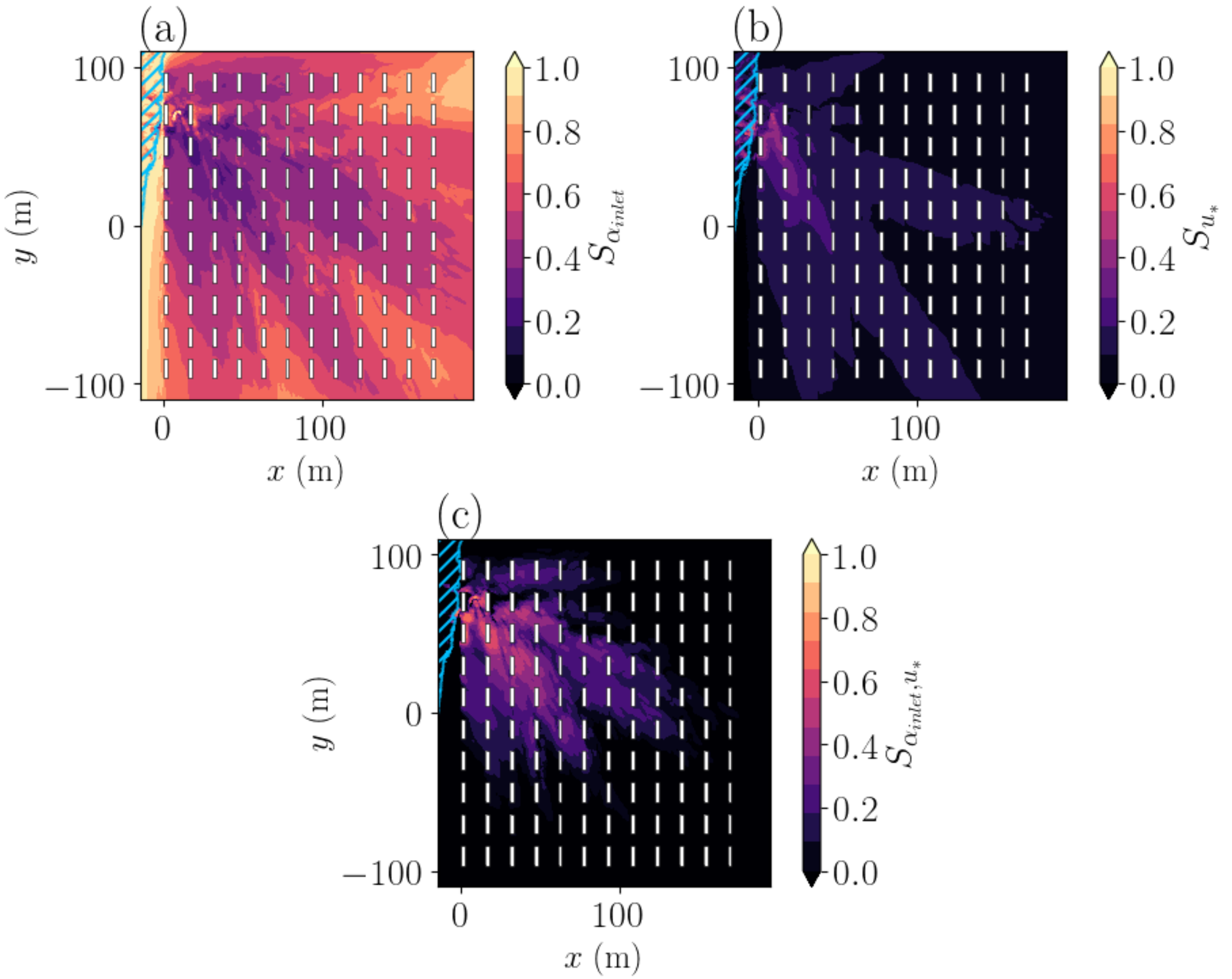


- ➢ The EnKF estimates very well α_{inlet}
- ➢ The analysis does not improve u_* estimation
- ➢ The analysis error covariance is coherent:
  - ○ Uncertainty on α_{inlet} is reduced
  - ○ Uncertainty on u_* is unchanged

II – Assimilation of the real field measurements

❖ Sensitivity of the mean concentration to the control vector

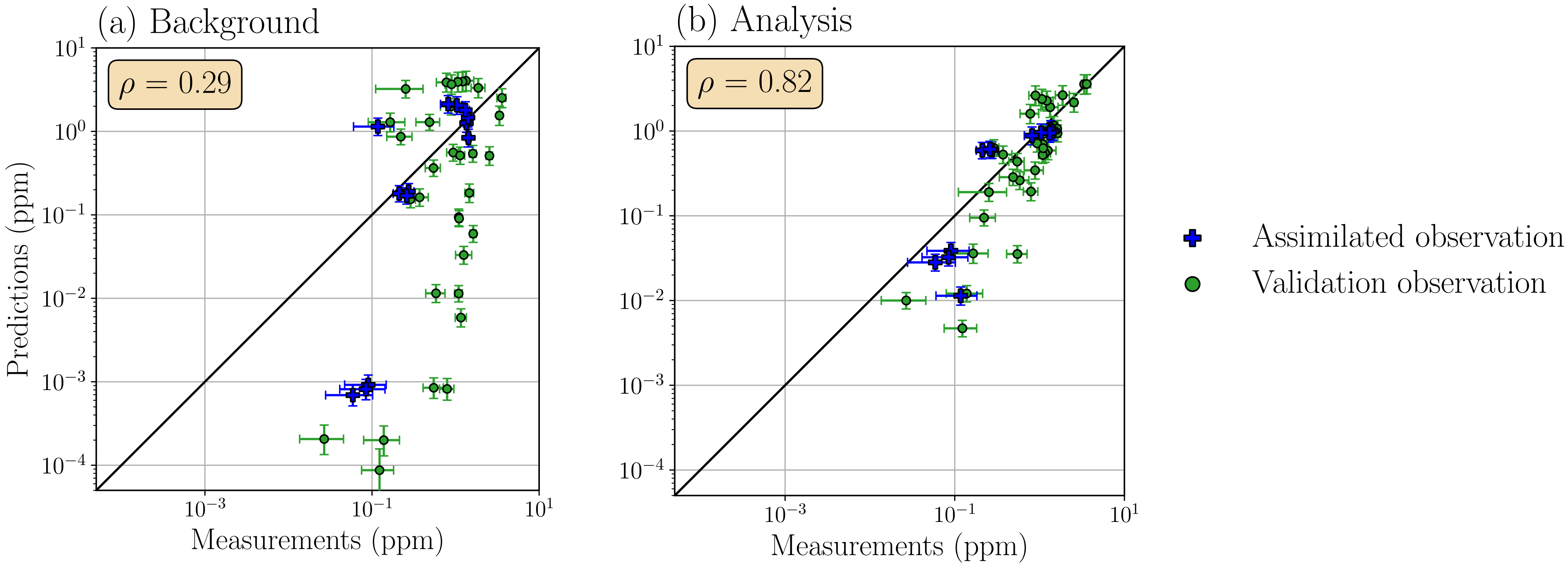
➢ Sensitivity analysis using Sobol' indices



- ➢ The EnKF fails to estimate u_* because the observation space is less sensitive to this parameter
- ➢ The dependence on u_* is mostly conditioned by α_{inlet}
- ➢ **Perspective:** using a an iterative estimation procedure?

II – Assimilation of the real field measurements

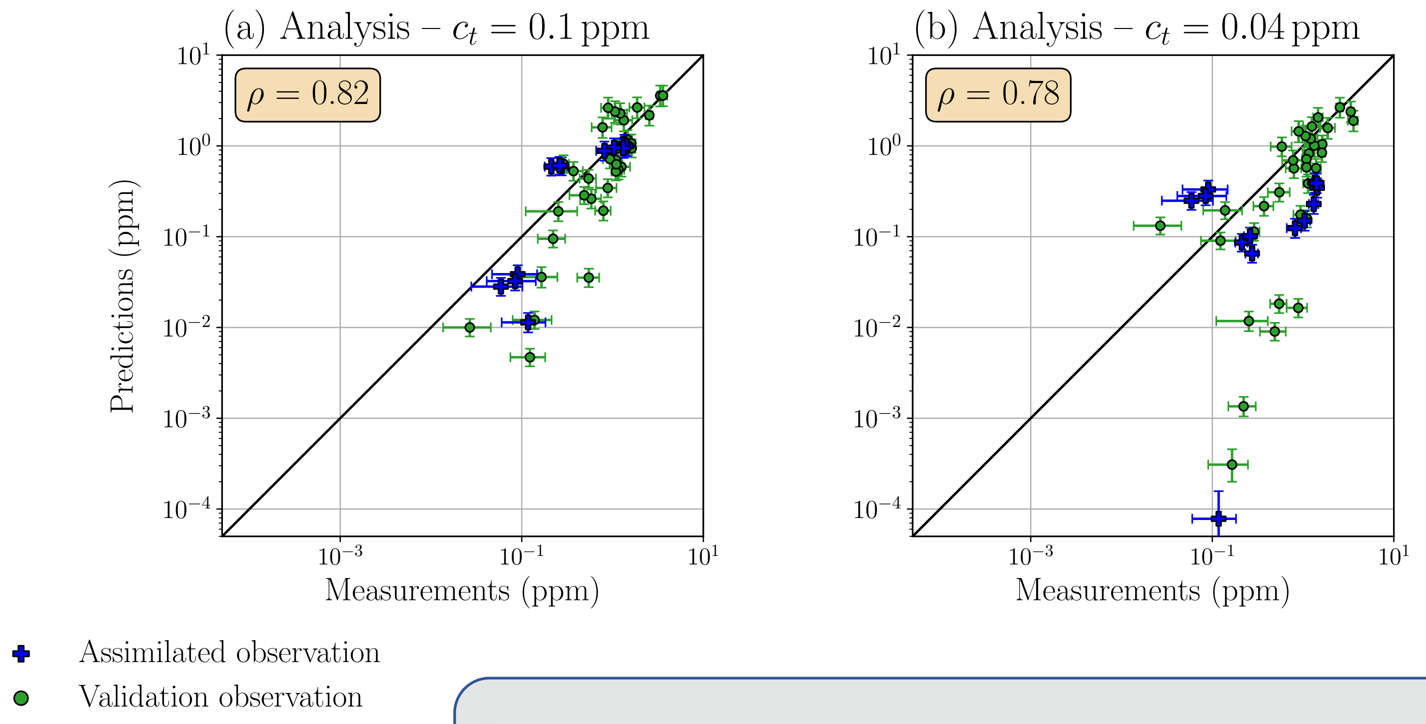
❖ **Propagation to state estimation** – Validation against measurements

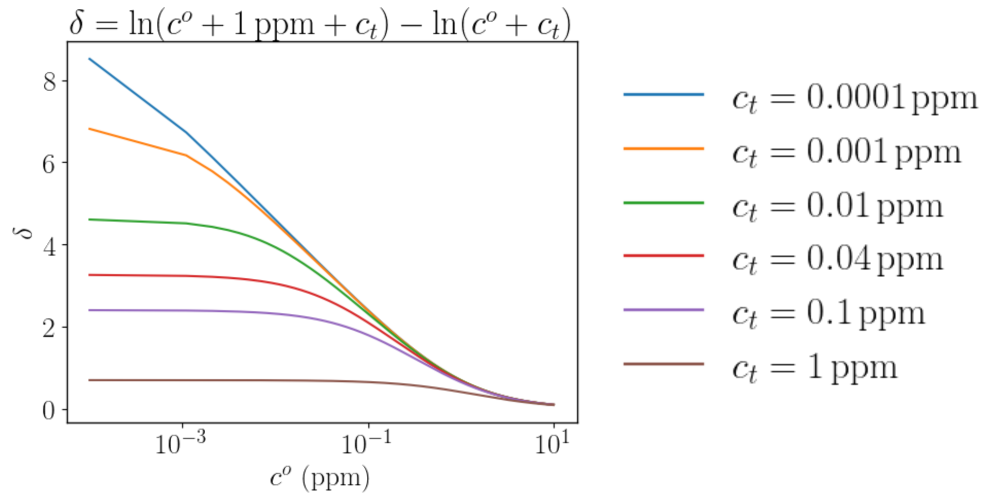


➢ Correction of boundary condition parameters significantly improves concentration estimation, even at unobserved locations

II – Assimilation of the real field measurements

Sensitivity to the concentration anamorphosis threshold: $\boxed{\mathbf{y}^o = ln(\mathbf{c} + c_t)}$

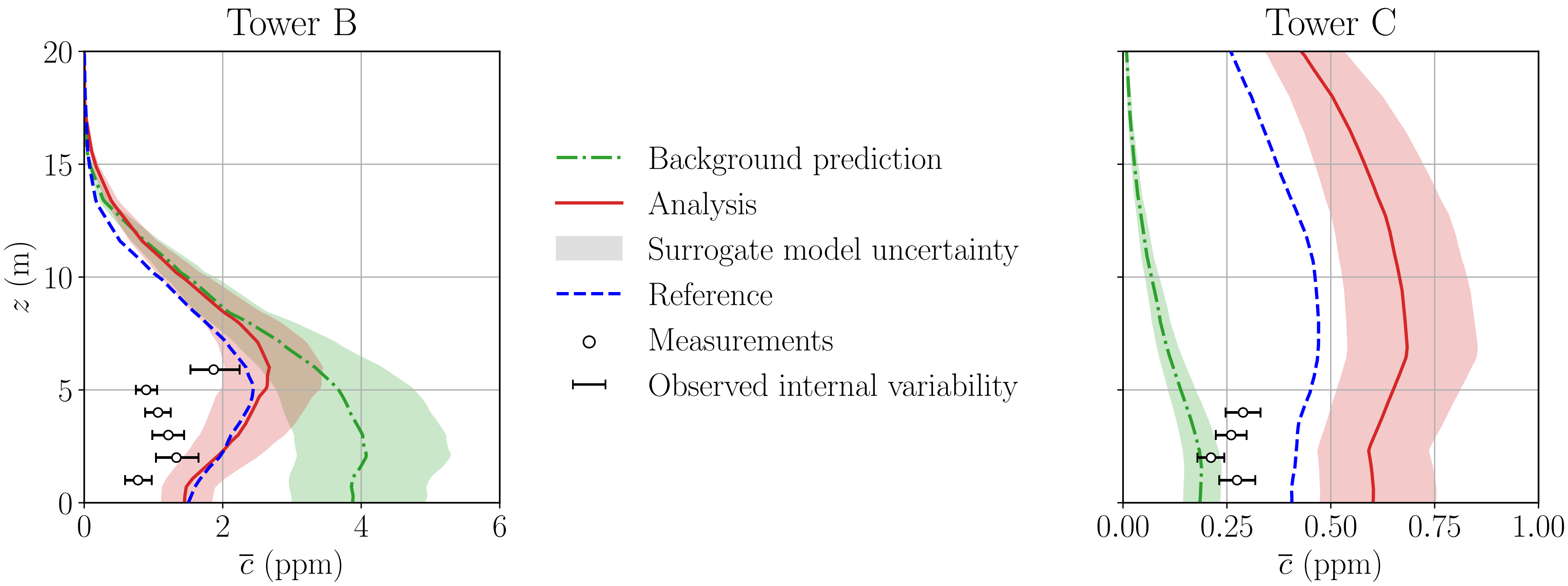




- The choice of c_t significantly affects the analysis: because the lower c_t, the greater the weight of low concentration deviations
- **Open question:** how to a priori select c_t?

II – Assimilation of the real field measurements

❖ **Propagation to state estimation** – Vertical profiles estimation



- ➢ The analysis may locally degrade concentration estimation
- ➢ Parameter estimation cannot compensate for internal LES model biases

Conclusion and perspectives

Summary

- Application of a reduced-cost EnKF for reducing uncertainty in microscale pollutant dispersion modeling
  - We provided realistic error models that account for internal variability
  - The use of a surrogate model allows us to provide large-ensemble analysis in a very short time (< 1min)
  - The DA system successfully corrects wind direction from real concentration measurements
  - It has difficulties for inferring friction velocity and cannot correct for internal LES model bias

Perspectives

I. Analyze the influence of using realistic error models

II. Move to joint state-parameter estimation to correct for internal LES model biases

III. Investigate the sensitivity to observation location to develop optimal sensor network design

References

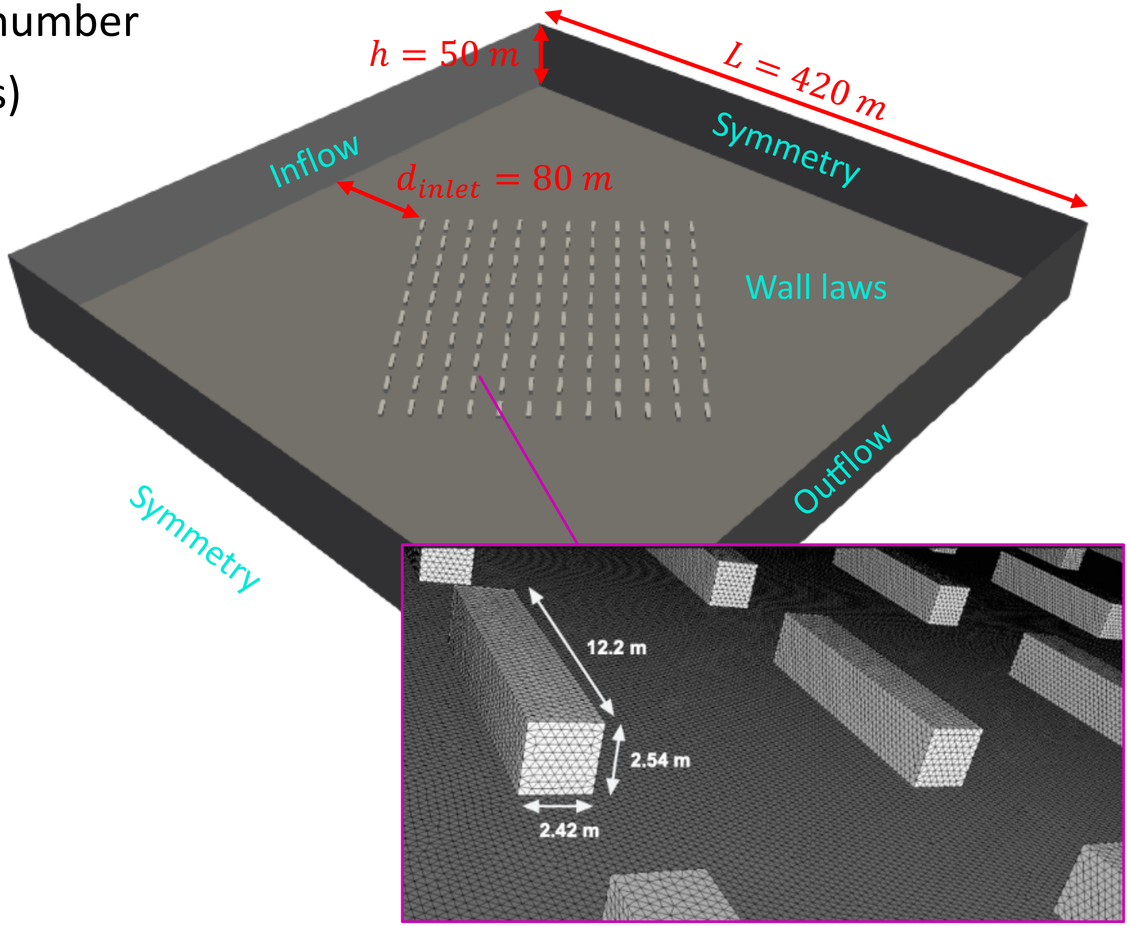
- Dauxois et al. (2021). Confronting Grand Challenges in Environmental Fluid Mechanics
- Lumet (2024). Assessing and reducing uncertainty in large-eddy simulation for microscale atmospheric dispersion
- Lumet et al. (2024). Assessing the Internal Variability of Large-Eddy Simulations for Microscale Pollutant Dispersion Prediction in an Idealized Urban Environment. Boundary-Layer Meteorology
- Marrel et al. (2015). Development of a surrogate model and sensitivity analysis for spatio-temporal numerical simulators
- Mons et al. (2017). Data assimilation-based reconstruction of urban pollutant release characteristics
- Schatzmann and Leitl. (2011). Issues with validation of urban flow and dispersion CFD models
- Sousa et al. (2018). Improving urban flow predictions through data assimilation
- Sousa et Gorlé. (2019). Computational urban flow predictions with Bayesian inference: Validation with field data
- Defforge et al. (2019). Improving CFD atmospheric simulations at local scale for wind resource assessment using the iterative ensemble Kalman smoother
- Defforge et al. (2021). Improving Numerical Dispersion Modelling in Built Environments with Data Assimilation Using the Iterative Ensemble Kalman Smoother
- Yee and Biltoft (2004). Concentration Fluctuation Measurements in a Plume Dispersing Through a Regular Array of Obstacles

Appendices

Appendices

❖ Model set-up

- **Solver:** AVBP - LES for compressible flows and low Mach number
- **Sub-grid Scale Model:** WALE (tailored for boundary-layers)
- **Numerical scheme:** Lax-Wendroff (2nd Order – FVM)
- **Pressure Gradient Scaling:** to reduce the CFL constraint
- **Turbulence injection:** Kraichnan-Celik method
- **Mesh:** unstructured, 90 million tetraedra
  - Refinement near the walls:
    - *30 cm ⇔ At least 8 cells by obstacle edge*
- **Computational cost:** {60s spin-up + 200s} ⇔ 20 000 hCPU



Appendices

❖ Air quality metrics from Chang and Hanna. 2004

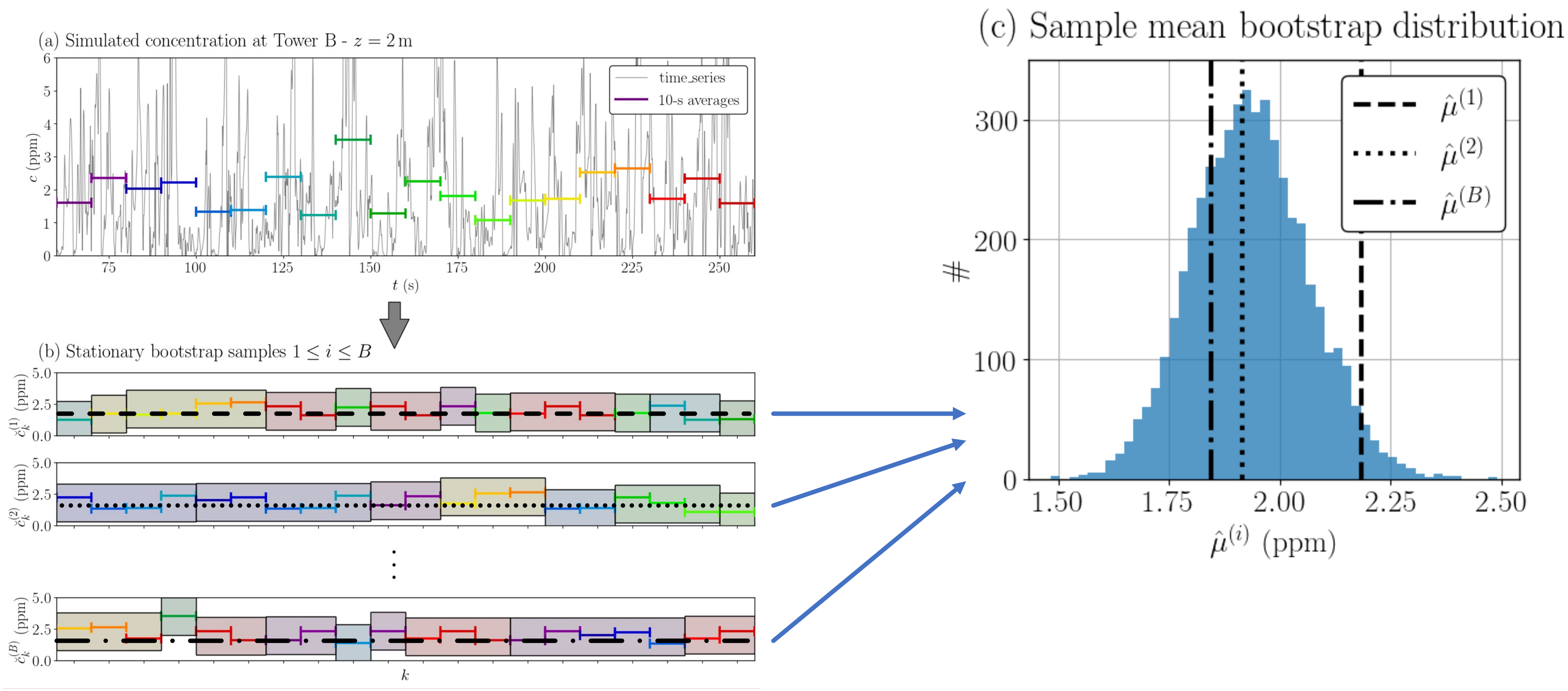
- **FAC2 :** Fraction of predictions that verify $0.5 \leq \frac{C_p}{C_o} \leq 2.0$
- **FB** *(Fractional Bias)* **:** $FB = \frac{(\overline{C_o} - \overline{C_p})}{0.5(\overline{C_o} + \overline{C_p})}$
- **MG** *(Geometric Mean Bias)* **:** $MG = \exp(\overline{\ln C_o} - \overline{\ln C_p})$
- **NMSE** *(Normalized Mean Square Error)* **:** $\mathrm{NM}SE = \frac{\overline{(C_o - C_p)^2}}{\overline{C_o}\,\overline{C_p}}$
- **VG** *(Geometric Variance)* **:** $\mathrm{VG} = \exp\overline{(\ln C_o - \ln C_p)^2}$

With :

- C_o: Measured concentrations
- C_p: Concentration predicted by the model at probes location
- $\bar{C}$: Averaged value over the dataset

Appendices

❖ **Stationary bootstrap method used to quantify internal variability**

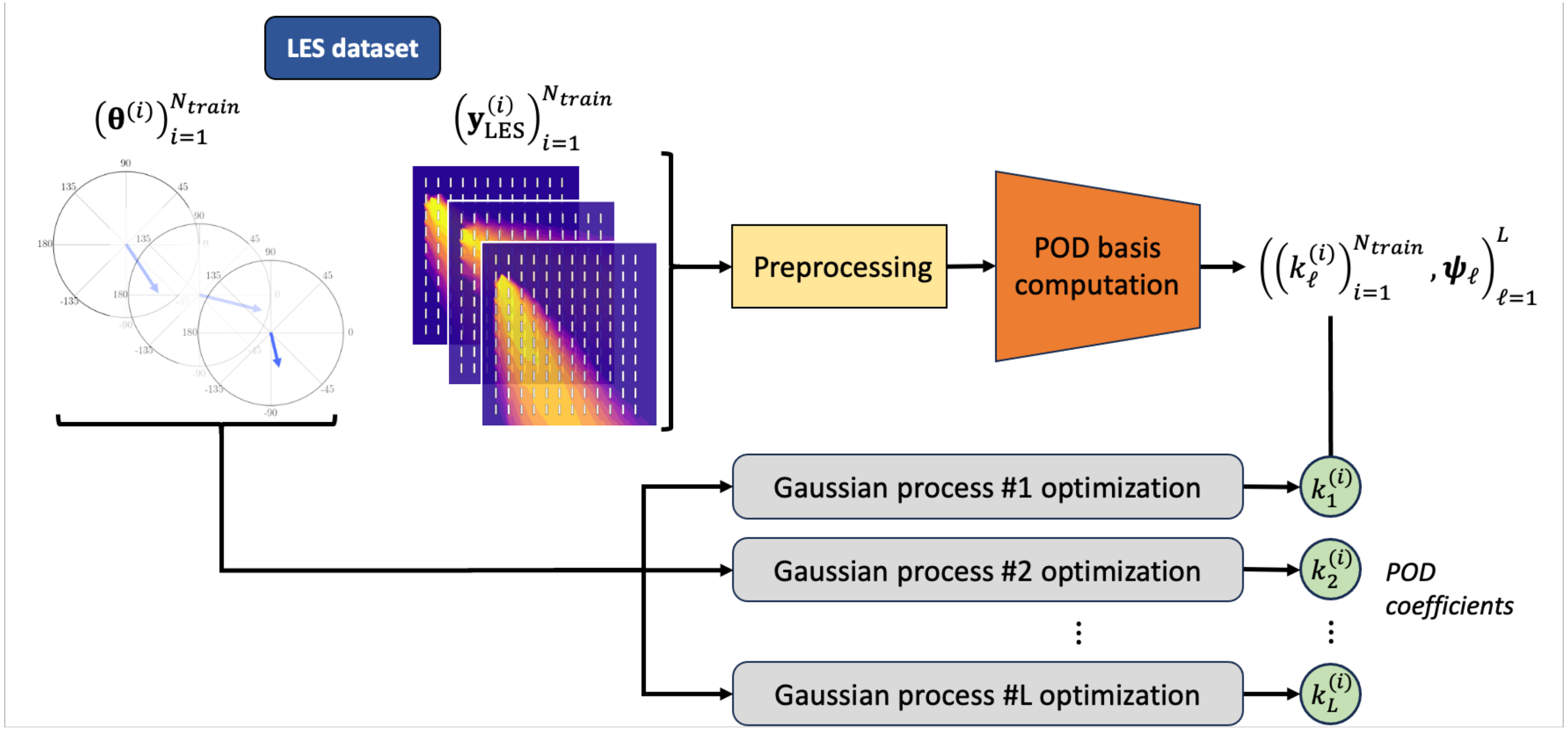


Appendices

❖ Construction of the surrogate model

2) Train a POD—GPR surrogate model (Marrel et al. 2015) over 160 LES samples

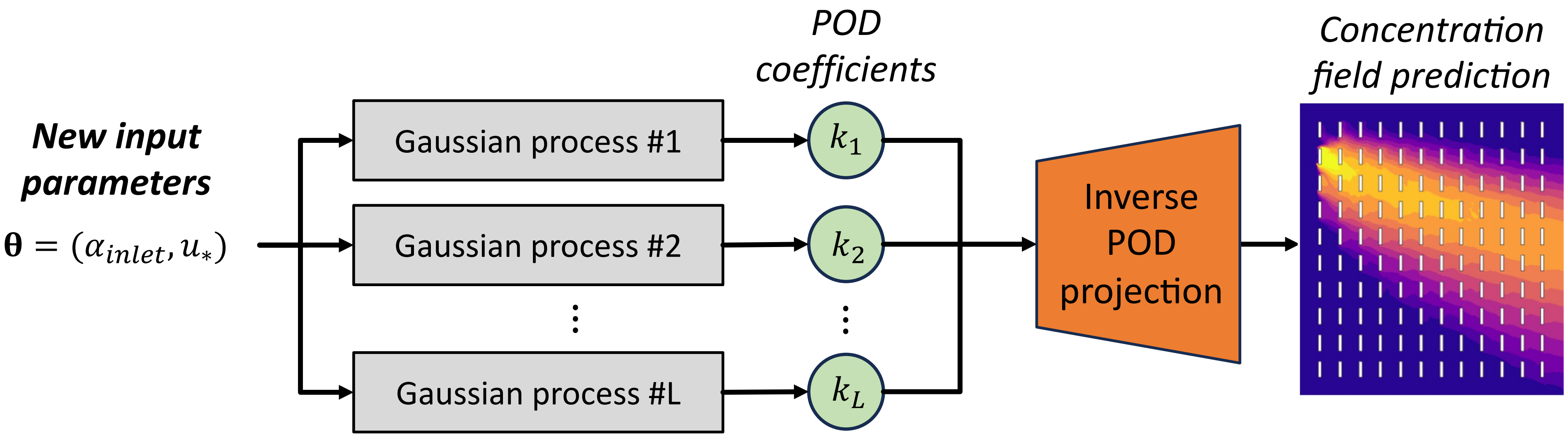
- Dimension reduction step using Proper Orthogonal Decomposition (**POD**, a.k.a PCA)
- Learn the dependency of POD coefficients on the control vector using Gaussian Process Regression (**GPR**)



Appendices

❖ Construction of the surrogate model

3) Compute new prediction with the POD—GPR surrogate model



4) Validation against 40 independent LES test samples

➢ Emulates well the LES response surface with an approximation error close to the minimal level of error reachable given the LES internal variability

Article presenting the validation of the surrogate model currently being submitted to Building and Environment

Appendices

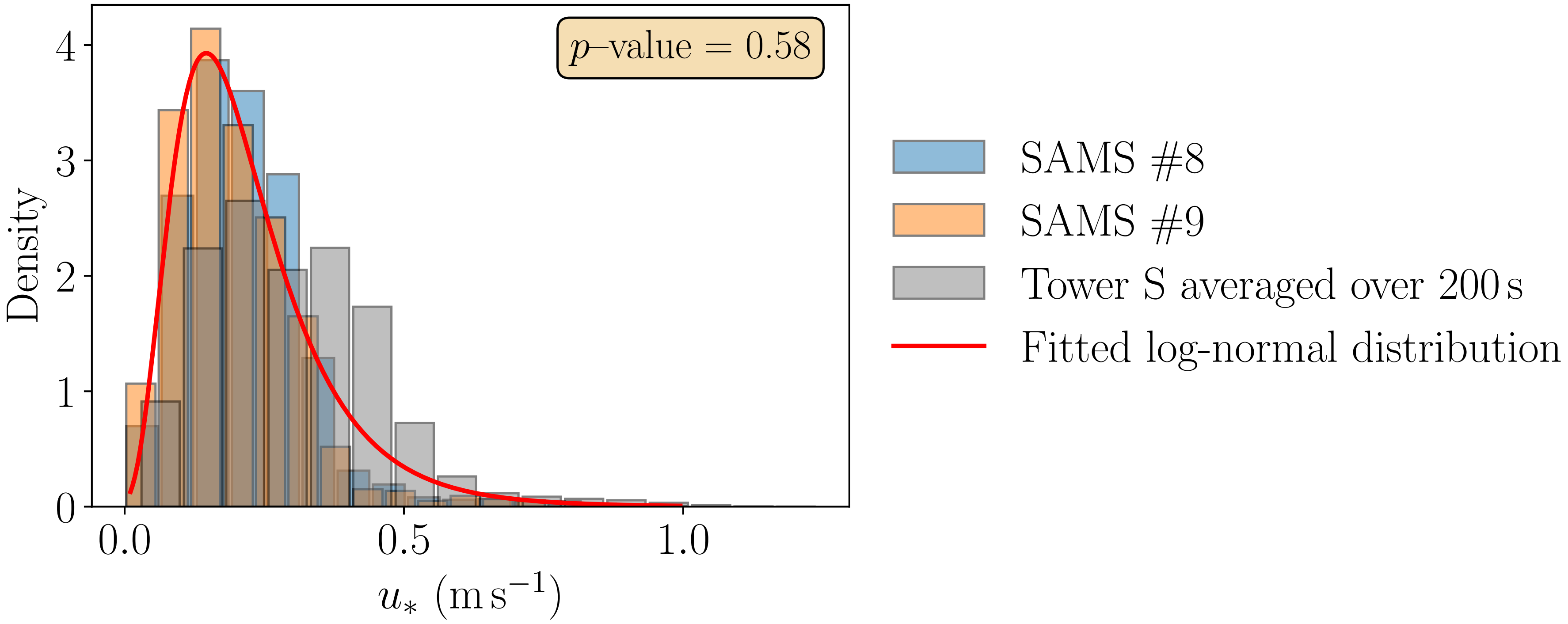
❖ Set-up used to assimilate the real field measurements

	Notation	Setup
Truth parameters	$\boldsymbol{\theta}^t = (\alpha^t_{inlet}, u^t_*)$	$(-41\,°, 0.73\,\mathrm{m\,s^{-1}})$
Background parameters	$\boldsymbol{\theta}^b = \left(\alpha^b_{inlet}, u^b_*\right)$	$(-25\,°, 0.57\,\mathrm{m\,s^{-1}})$
Background errors	$\mathbf{B} = \begin{pmatrix} \sigma^2_{\alpha_{inlet}} & 0 \\ 0 & \sigma^2_{u_*} \end{pmatrix}$	with $\sigma_{\alpha_{inlet}} = 25\,°$, $\sigma_{u_*} = 0.09\,\mathrm{m\,s^{-1}}$
Observation network	$\mathbf{y}$	13 observations of concentration at towers C (1, 2, 3 m), D (1, 2, 3 m) T (1, 2, 4, 6, 8, 10 m) and DPID #26
Observation error	$\mathbf{R}$	See Sect. V.3.2
EnKF ensemble size	N_e	500
Anamorphosis threshold	(y_t, u_t)	$(0.04\,\mathrm{ppm}, 0.04\,\mathrm{m\,s^{-1}})$

Appendices

❖ Anamorphosis for friction velocity

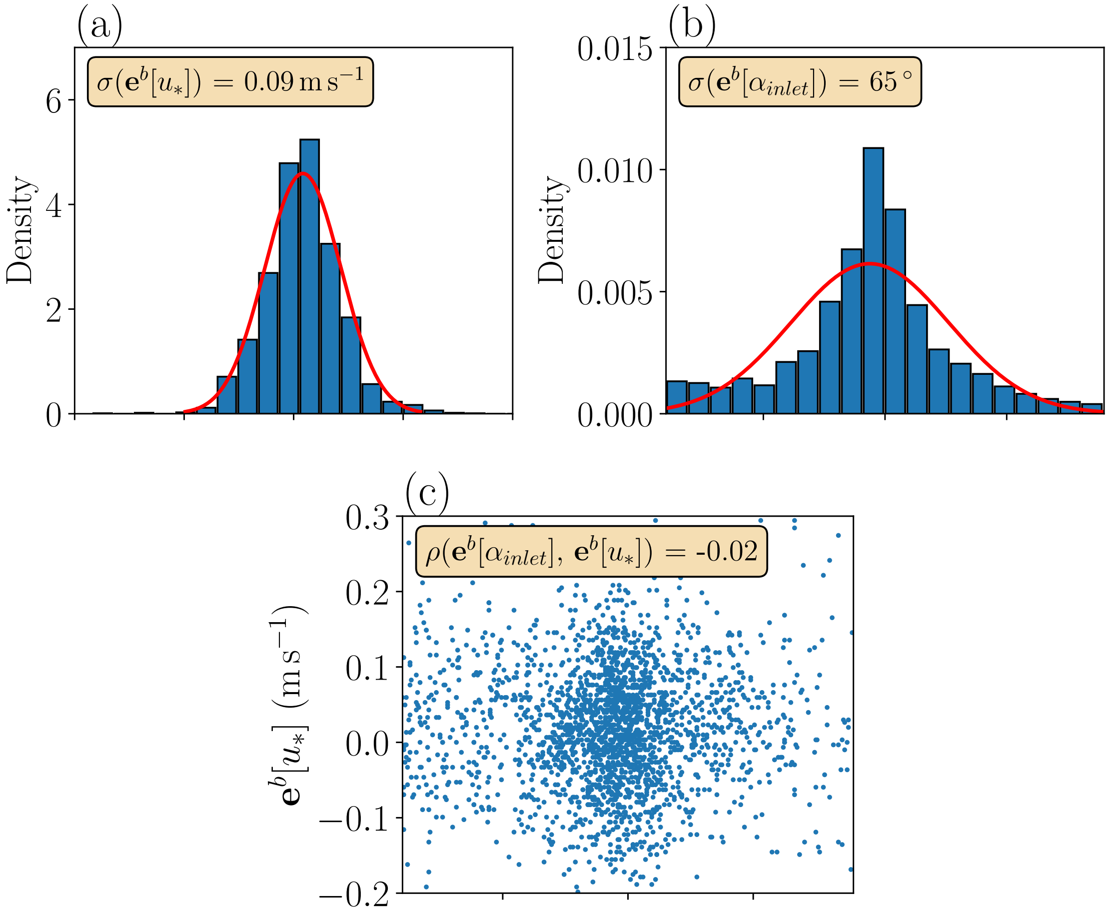
$$\widetilde{u_*} = ln(u_* + c_t)$$



Appendices

❖ Background error covariance matrix estimation

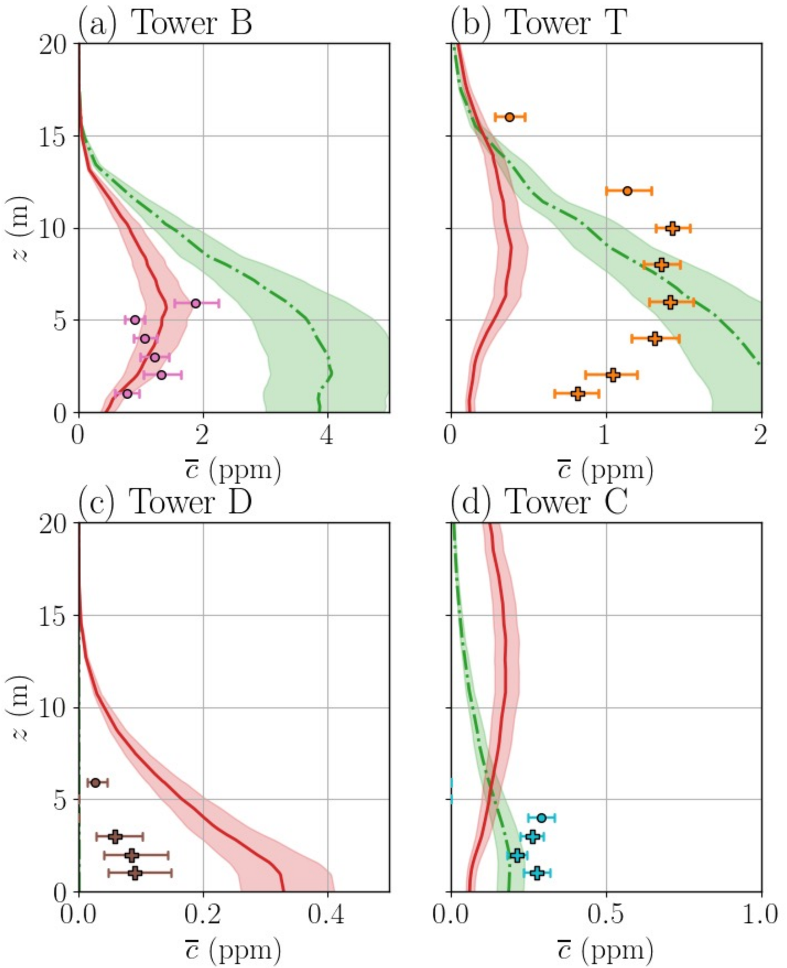
- Statistics based on 12 days of measurements of the difference between the two nearest masts



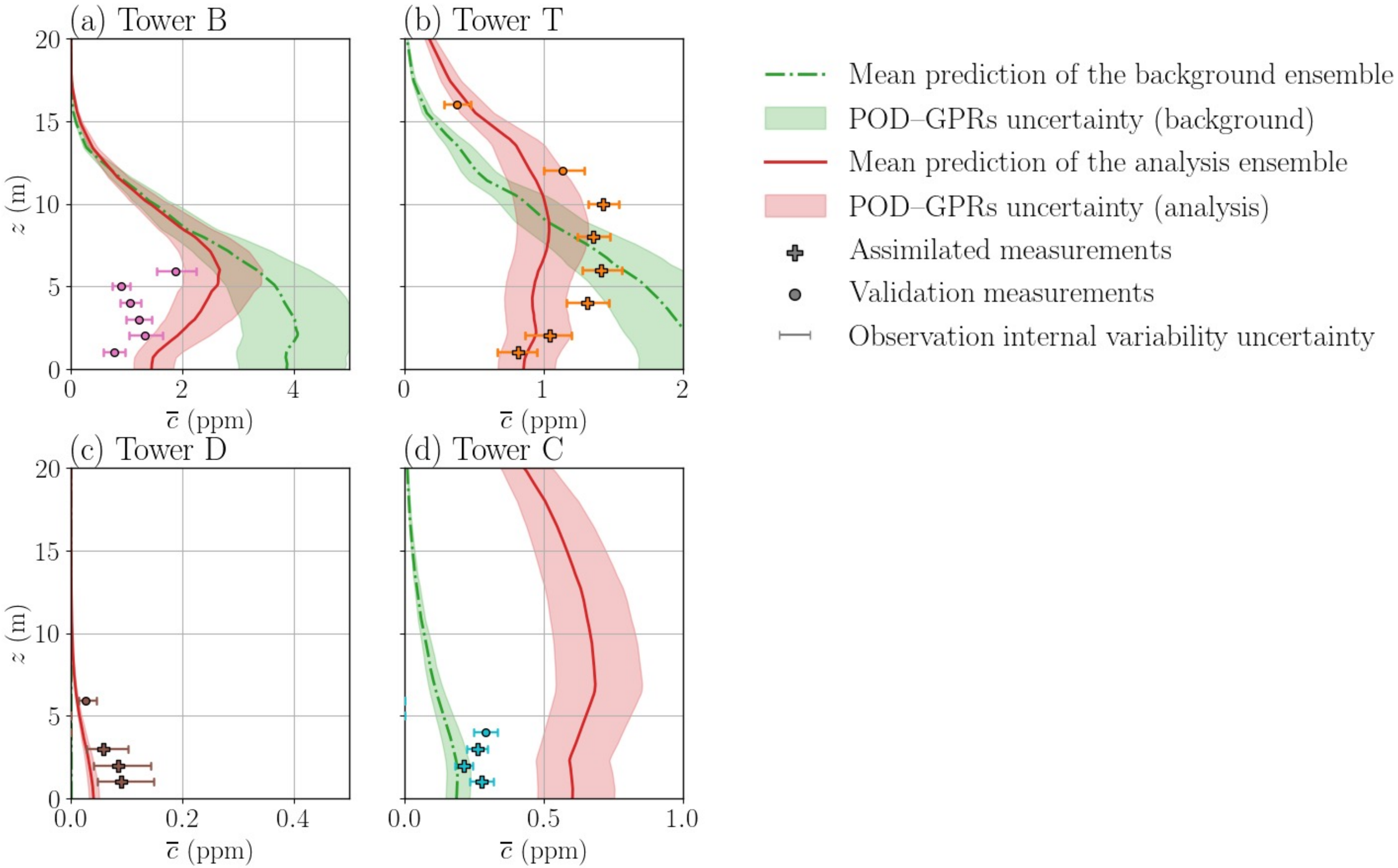
Appendices

❖ Effect of the concentration anamorphosis threshold

$c_t = 0.04$ ppm



$c_t = 0.1$ ppm



Appendices

❖ Effect of the concentration anamorphosis threshold

$c_t = 0.04$ ppm

$c_t = 0.1$ ppm

En concentration log