Bayesian Source Identification with Dual Hierarchical Neural Networks for Urban Air Pollution

Elissar Al Aawar, Sofien Resifi, Hatem Jebari, and Ibrahim Hoteit

Healthy People and Healthy Planet: AI for Decarbonized, Healthy, Inspiring, and Energy Positive Cities

July 25th, 2024

Introduction

- Air pollution is a growing concern with many negative impact on public health and the environment.
- Identifying the sources of pollution can help address air quality issues and inform policies and regulations for reducing harmful emissions.
- We propose a novel hierarchical framework for urban air pollution source identification, leveraging deep learning (DL) within an efficient Bayesian inference framework.

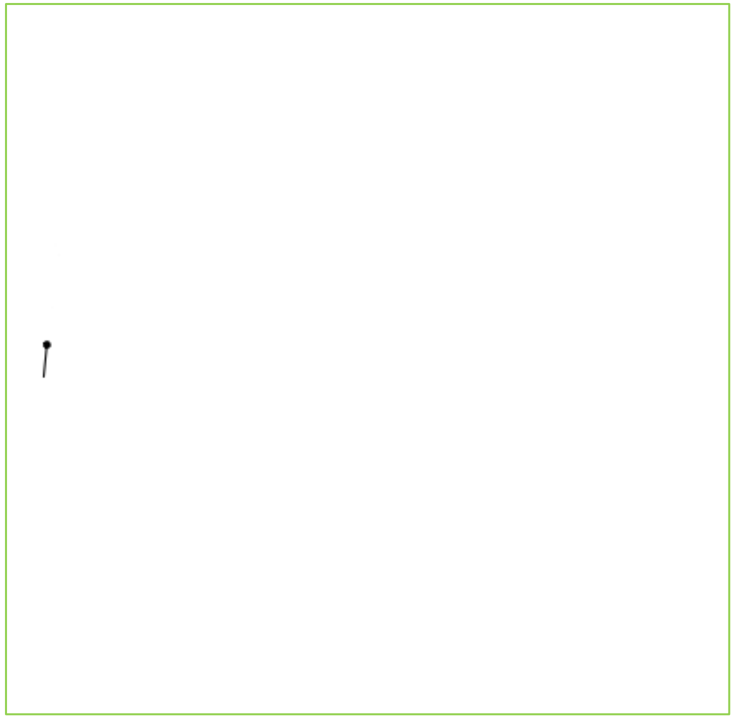
Bayesian Inference

- Baye's rule

$$\pi(M|D) \propto \pi(M)\pi(D|M)$$

Posterior probability ∝ prior probability × likelihood probability

- Advantages:
    - Incorporating prior knowledge
    - Adapting to new information
    - Uncertainty quantification
- Requirement: stochastic sampling tools → Monte Carlo Markov Chain (MCMC) family



Metropolis Hastings-MCMC Algorithm

1. Sample a candidate point from the posterior distribution.
2. Run the forward model.
3. Compute the model-data discrepancy, and the likelihood.
4. Accept or reject to build the corresponding chain.

MH-MCMC in Air Pollution Source Identification

Inverting for 4 parameters (M)

- Longitude (x)
- Latitude (y)
- Emission Rate (q)
- Emission Duration (d)

Uniform prior distribution $\pi(M)$

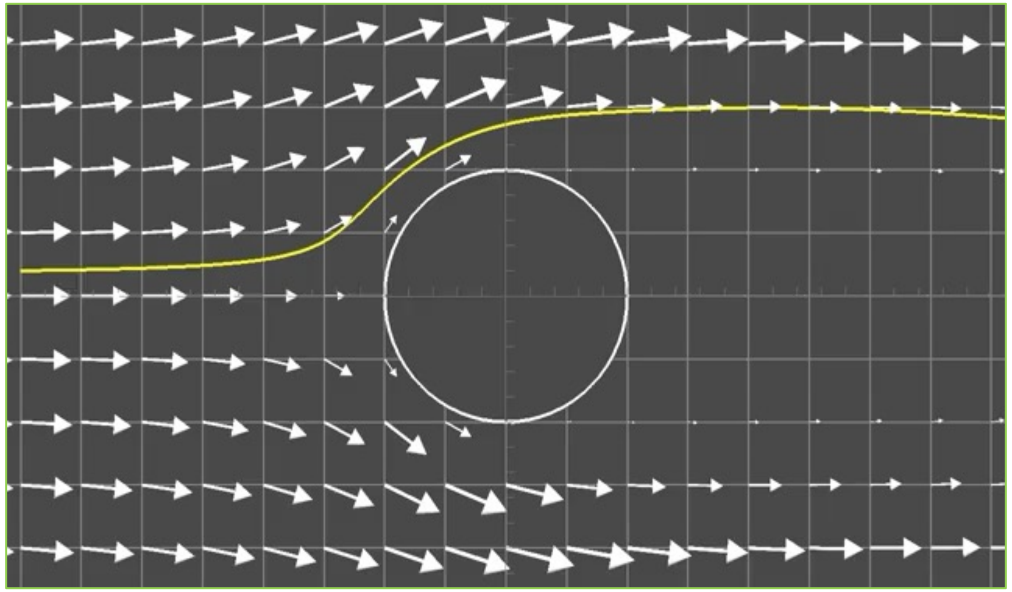
MH-MCMC in Air Pollution Source Identification

- Available environmental monitoring techniques:
  - Sensors
  - Satellite images
  - Remote sensing techniques
- Two forms of pollution observations:
  - Discrete point-wise concentration values
  - Concentration fields

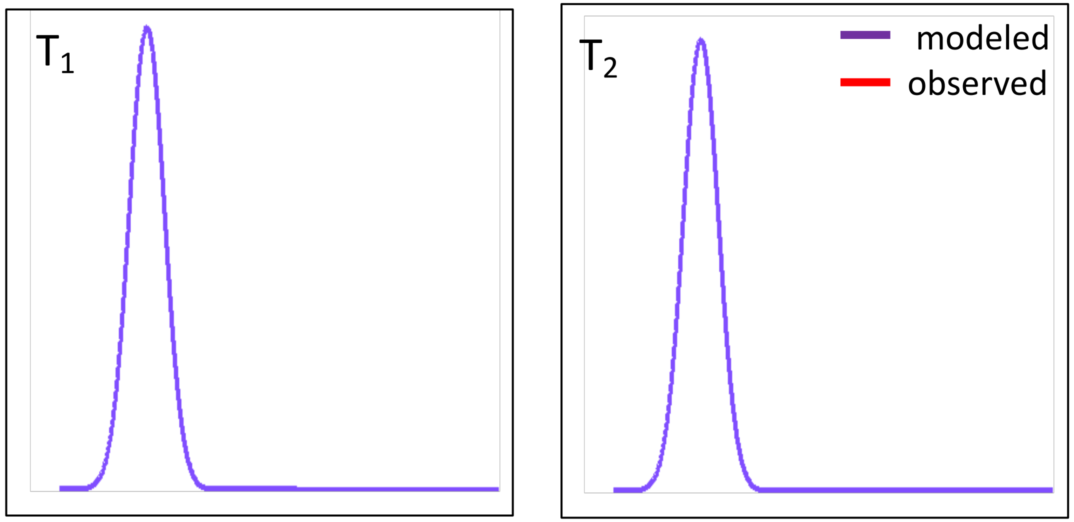
Al Aawar, et al, (2023). Bayesian source identification of urban–scale air pollution from point and field concentration measurements. Computational Geosciences. (in production).

The Lagrangian Transport Model

- The Graz Lagrangian model **GRAL**:
  - Microscale wind field computations.
  - Lagrangian particle tracking.
- Navier-Stokes equation along with the k-ϵ turbulence closure model.
- Inputs: topography, land use, buildings, meteorological conditions and emission sources.

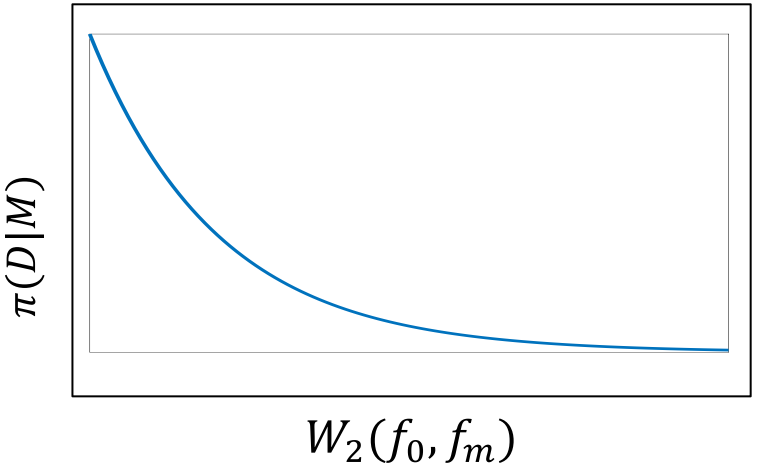


Modeling the Likelihood



- Each concentration field is considered as a distribution that could be displaced.
- The Wasserstein distance is the cost of displacing the predicted model output to the observation.
- This cost physically represents the required work as the product of the mass to be moved and the distance to be traveled.

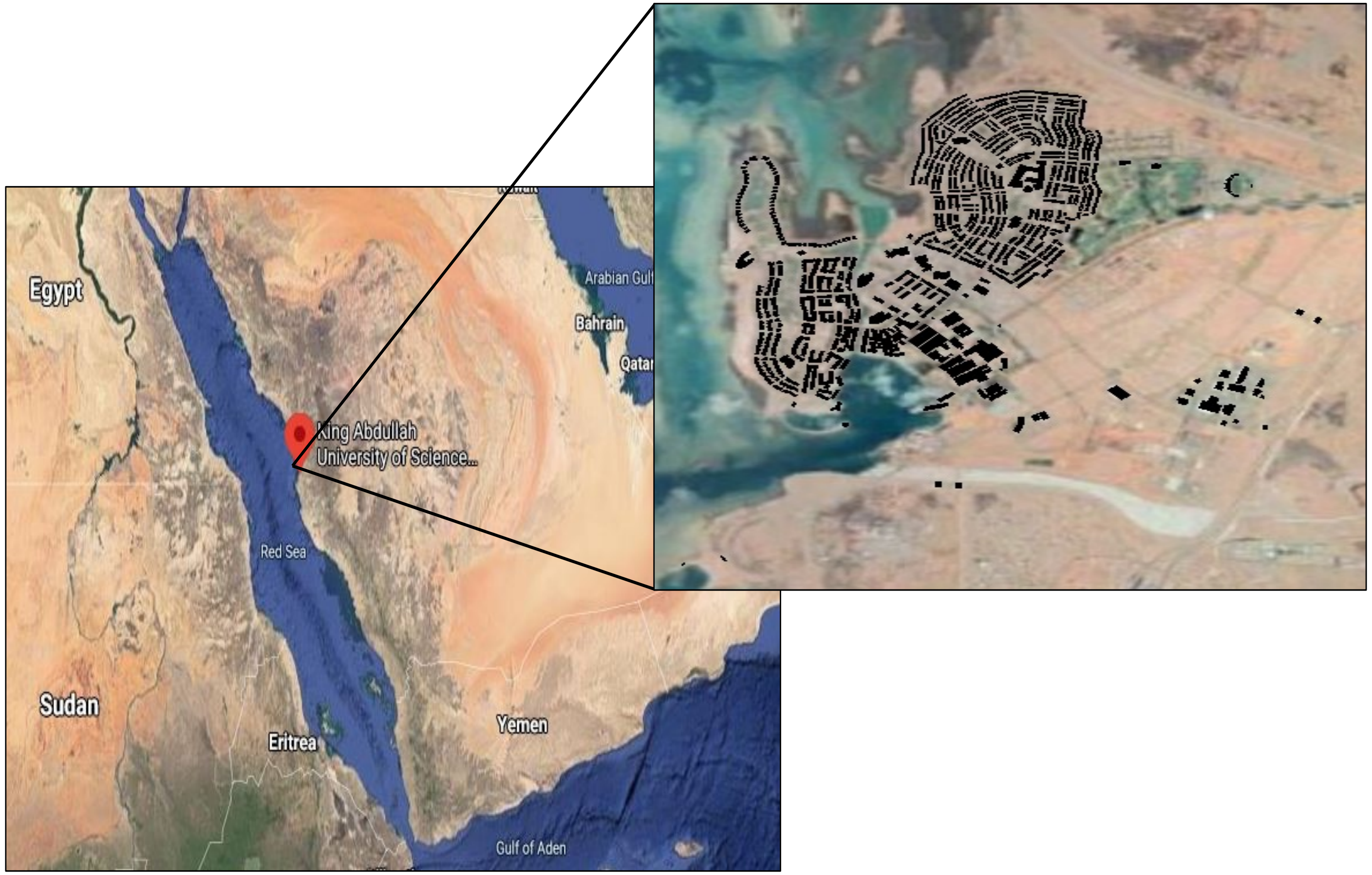
Exponential likelihood

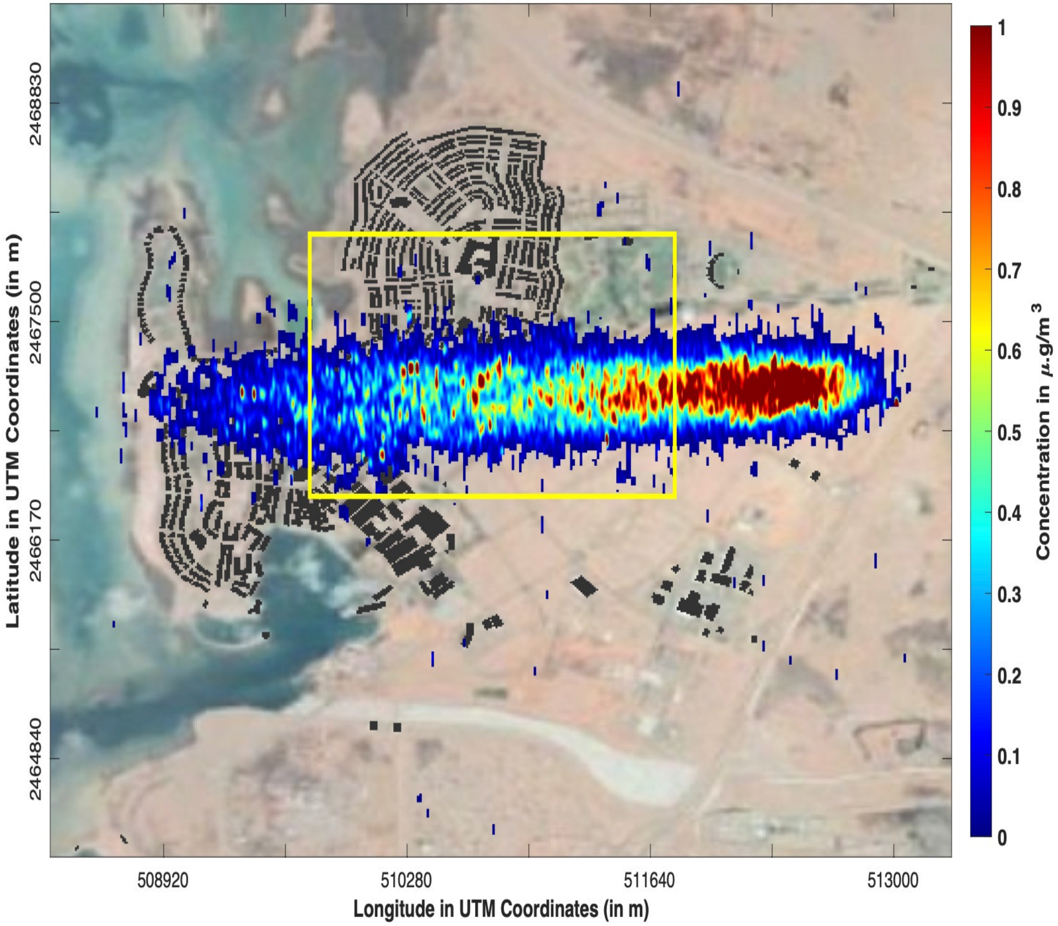


$$W_2(f_0, f_m) = \min_{T^* \in \mathcal{T}(f_0, f_1)} \mathcal{L}(u, v) f_0(u)$$

- f_0 and f_m: observed and modeled concentration fields, respectively.
- $\mathcal{T}$: set of regular bijections mapping f_0 to f_m.
- $\mathcal{L}(u, v)$: Euclidean distance from point u to point $v = T(u)$.

Case Study: KAUST Synthetic Scenario





Al Aawar, E., El Mohtar, S., Lakkis, I., Alduwais, A. K., & Hoteit, I. (2023). Bayesian source identification of urban-scale air pollution from point and field concentration measurements. Computational Geosciences, 27(4), 605-626.

Results and Challenges

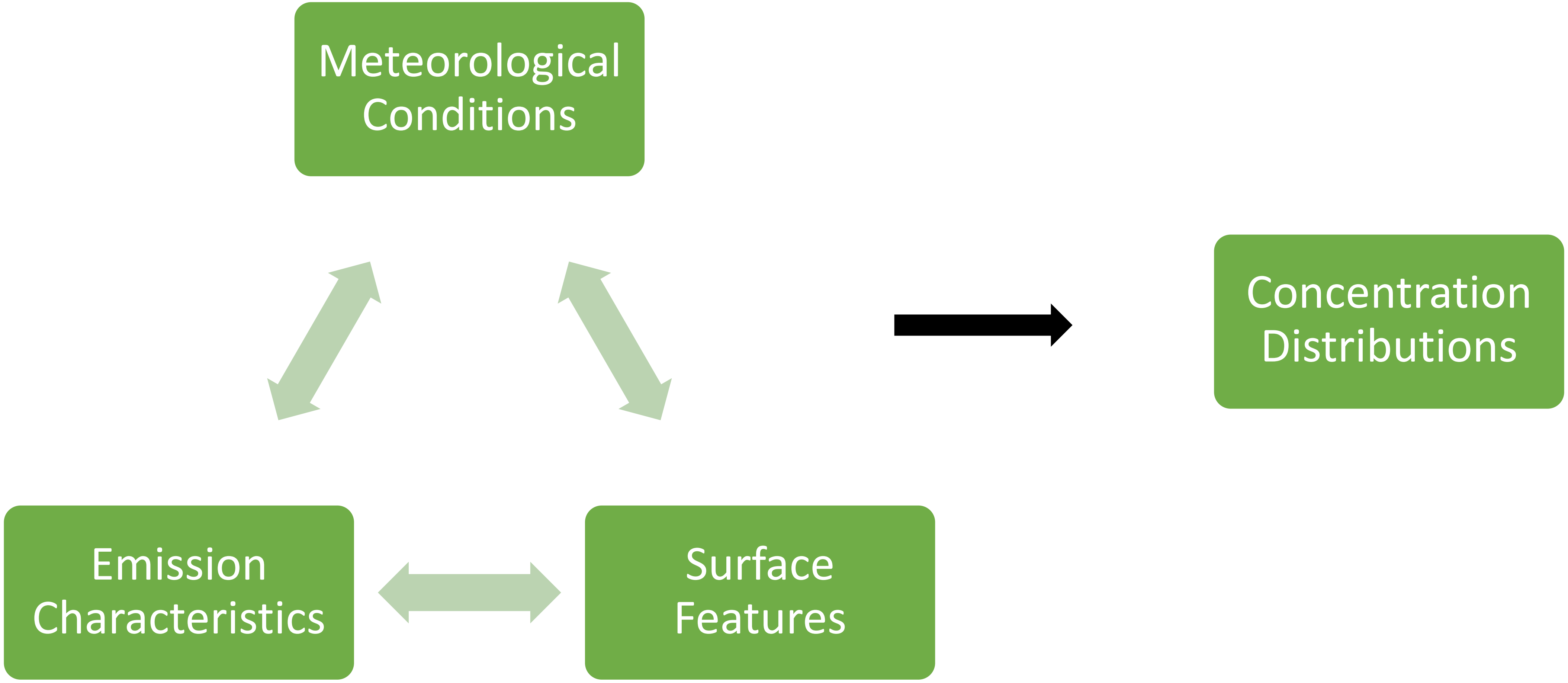
- Excellent solutions are obtained with the global W_2 dissimilarity metric.

- Challenge:
  - Each chain generation with **10,000** samples requires **143** hours.
  - Numerous runs of the expensive physical dispersion model.
  - High Cost of the W_2 Distance: Half of the computational time is spent to calculate the W_2 distance for each sample.

Solution: Use DL

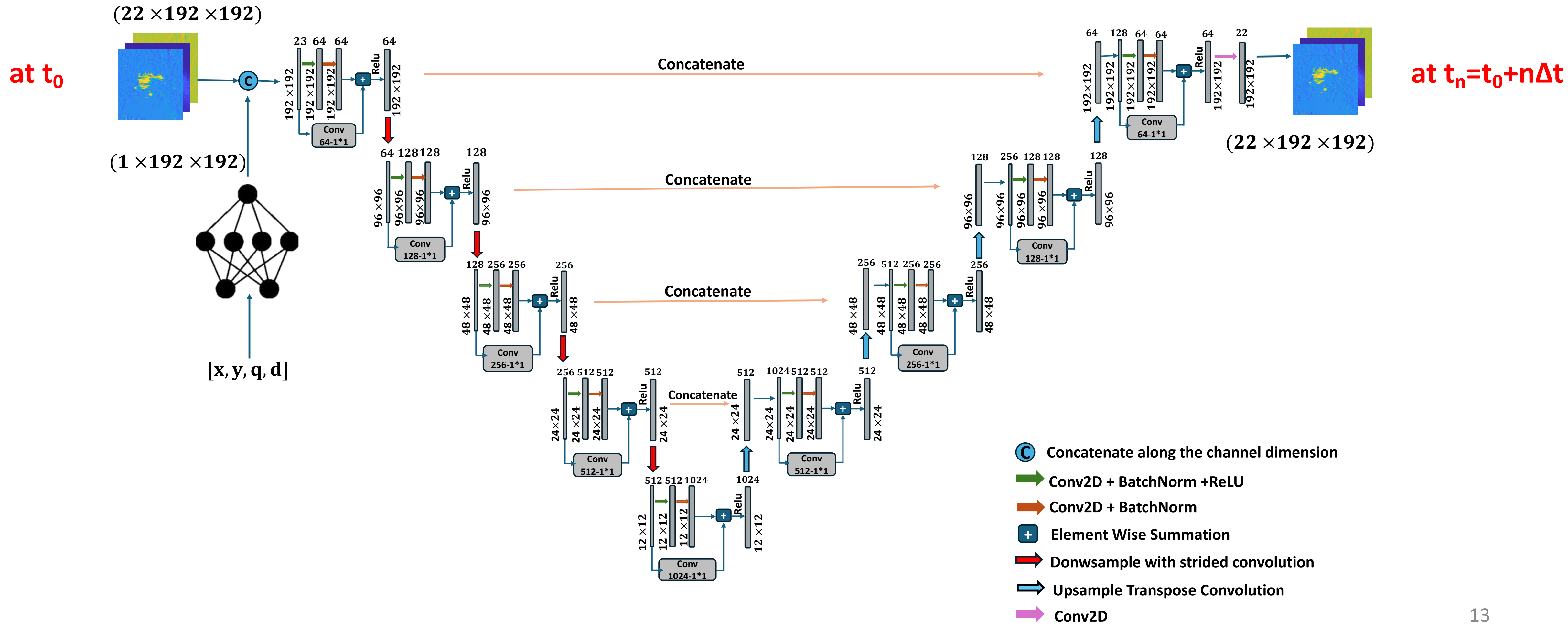
- The acceleration of tradition Bayesian inference by training a NN to predict air pollutant concentrations based on given flow conditions and emission characteristics.
- Coupling this emulator with a NN approximation of the likelihood distribution to synergistically accelerate computations.
- Full operation on GPUs, leveraging parallel computing architectures to expedite computational costs.

The Lagrangian Transport Surrogate Model: Learning Task



Al Aawar, E., Resifi S., Jebari H. & Hoteit I. (2024). Bayesian Source Identification with Dual Hierarchical Neural Networks for Urban Air Pollution. (Under Revision)

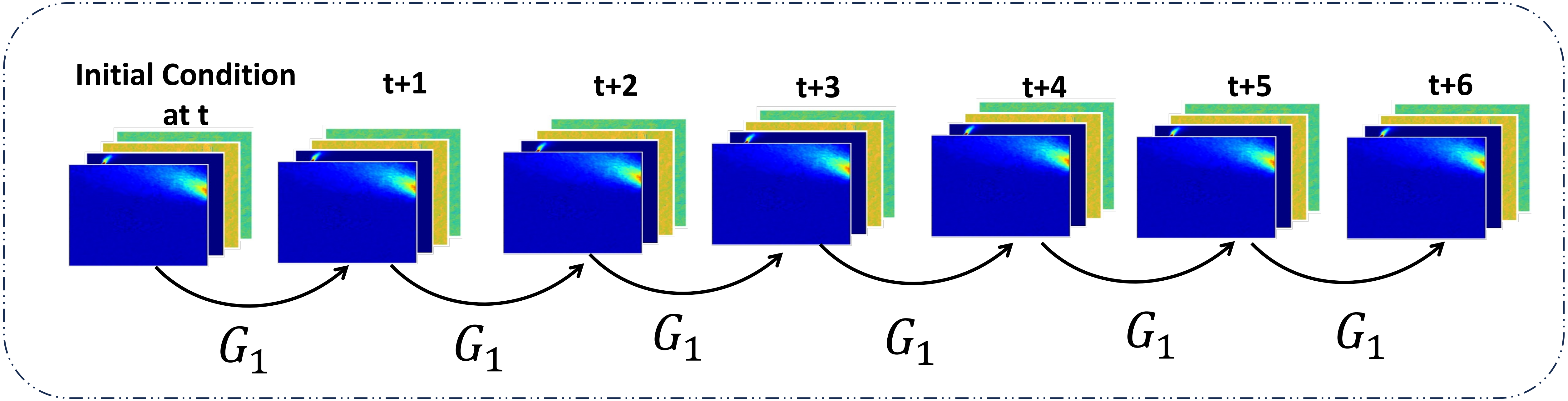
The Lagrangian Transport Surrogate Model: Architecture



Al Aawar, E., Resifi S., Jebari H. & Hoteit I. (2024). Bayesian Source Identification with Dual Hierarchical Neural Networks for Urban Air Pollution. (Under Revision)

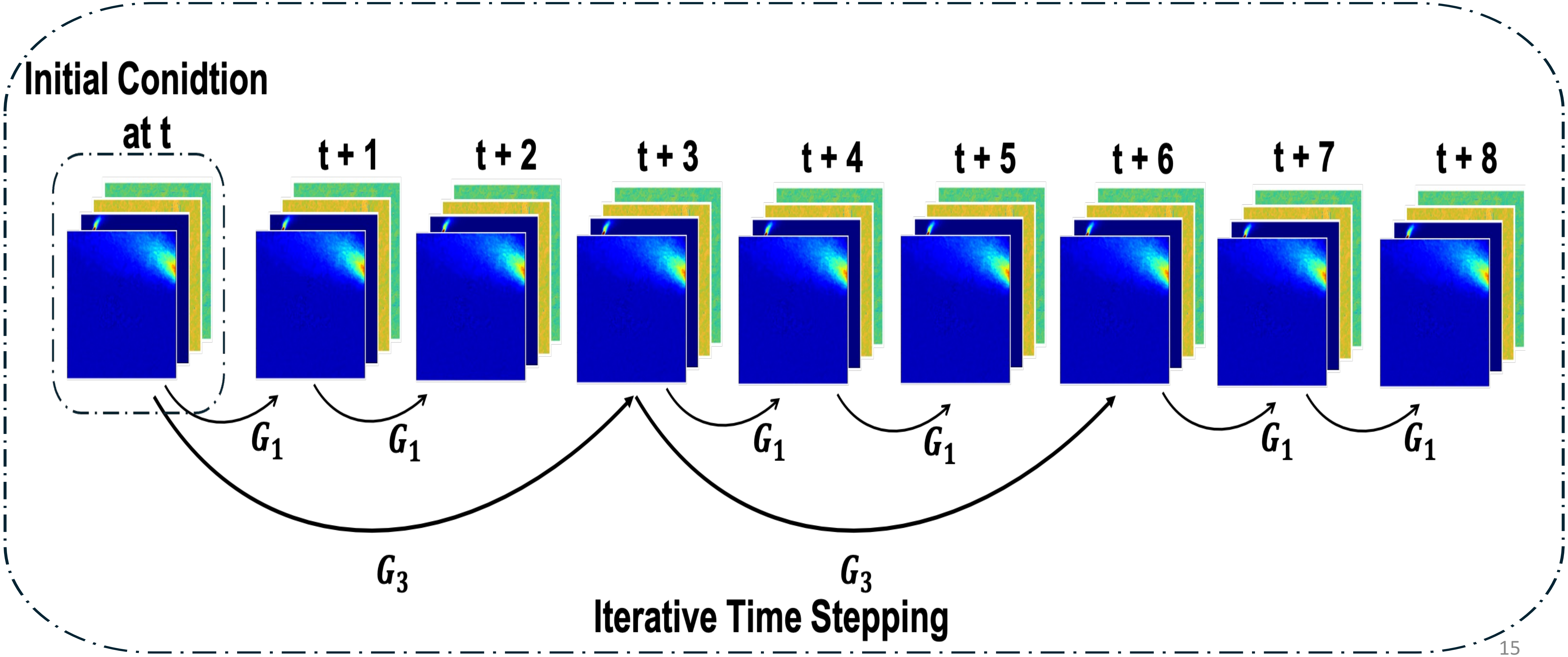
The Lagrangian Transport Surrogate Model: Training

- Prediction in time at each time step.
- Cumulative error growth for each time step.



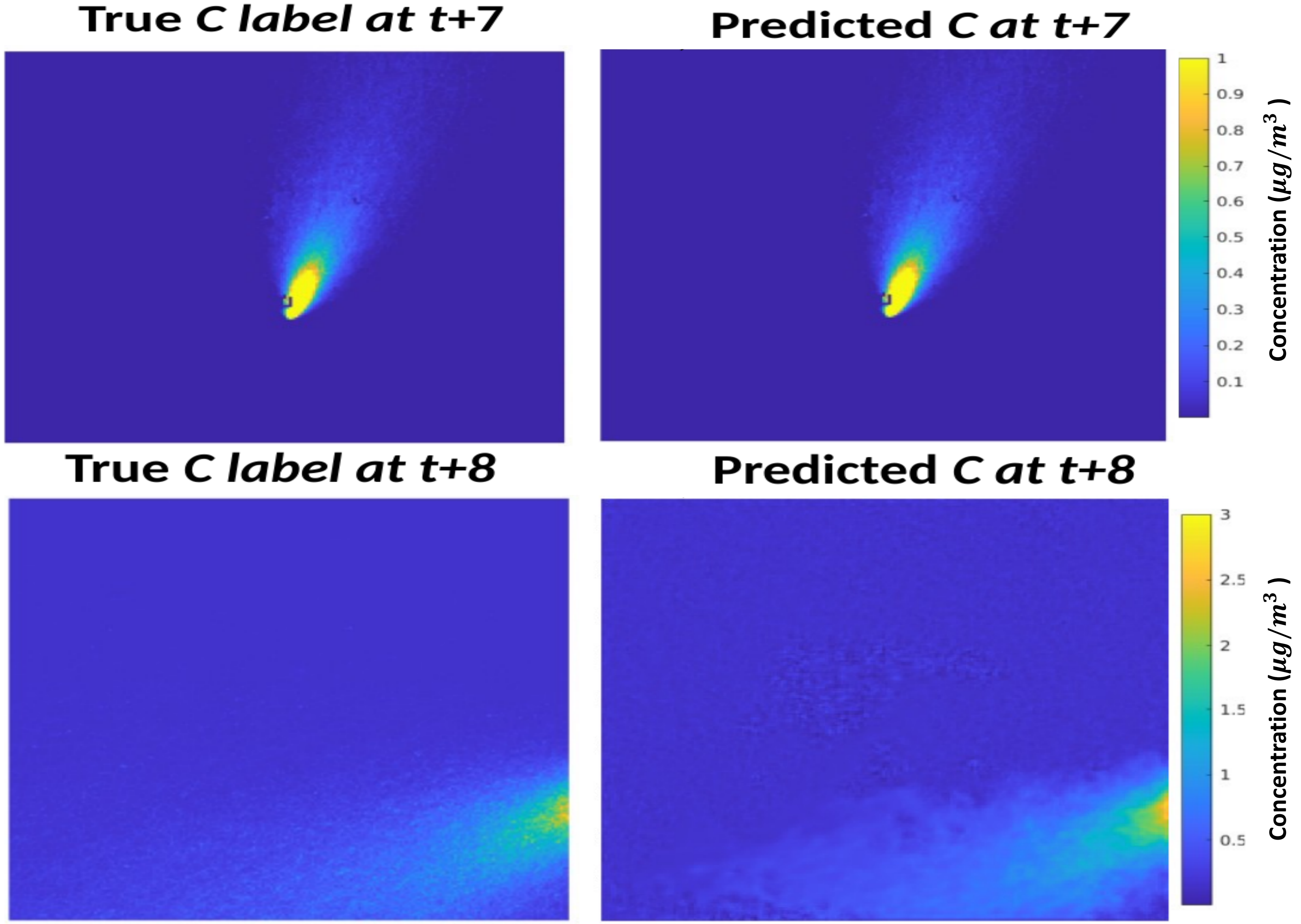
Al Aawar, E., Resifi S., Jebari H. & Hoteit I. (2024). Bayesian Source Identification with Dual Hierarchical Neural Networks for Urban Air Pollution. (Under Revision)

The Lagrangian Transport Surrogate Model: Training



Al Aawar, E., Resifi S., Jebari H. & Hoteit I. (2024). Bayesian Source Identification with Dual Hierarchical Neural Networks for Urban Air Pollution. (Under Revision)

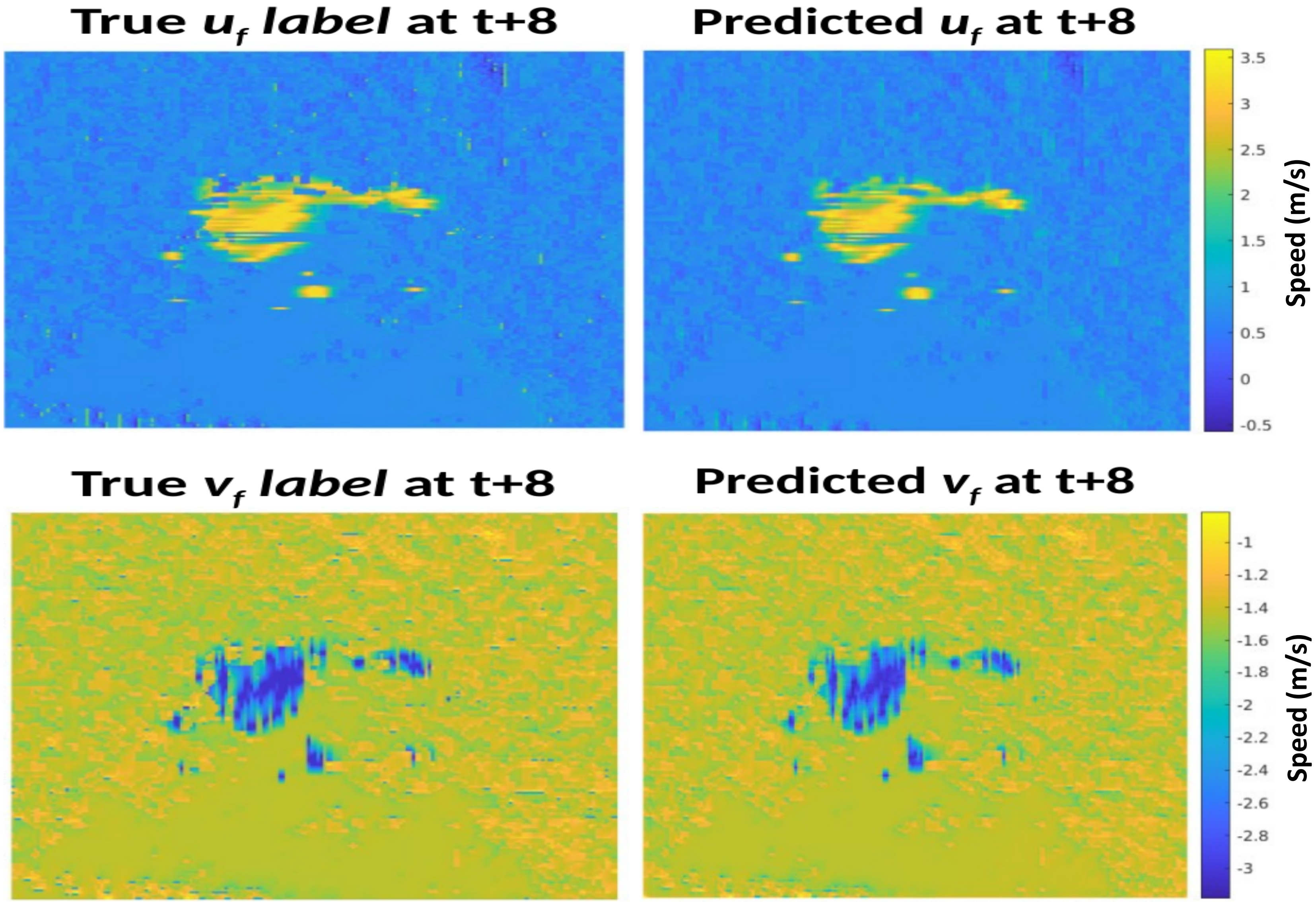
Examples of Predictions



Evaluation Results:

- RRMSE = 3.6%
- MBE = 0.015
- IOA = 98%

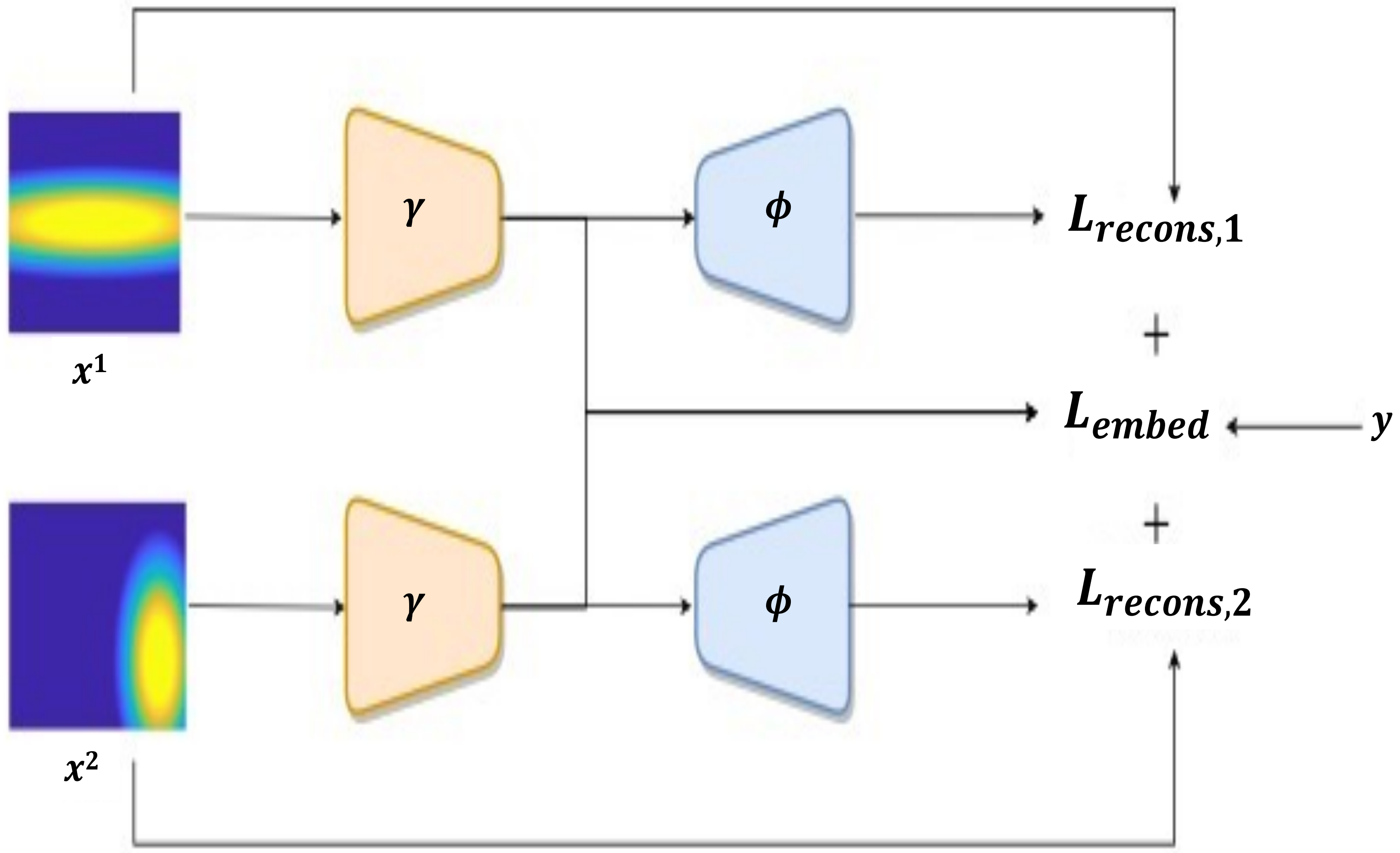
Examples of Predictions



Evaluation Results:

- RRMSE = 3.6%
- MBE = 0.015
- IOA = 98%

Siamese Network for W_2 Approximation

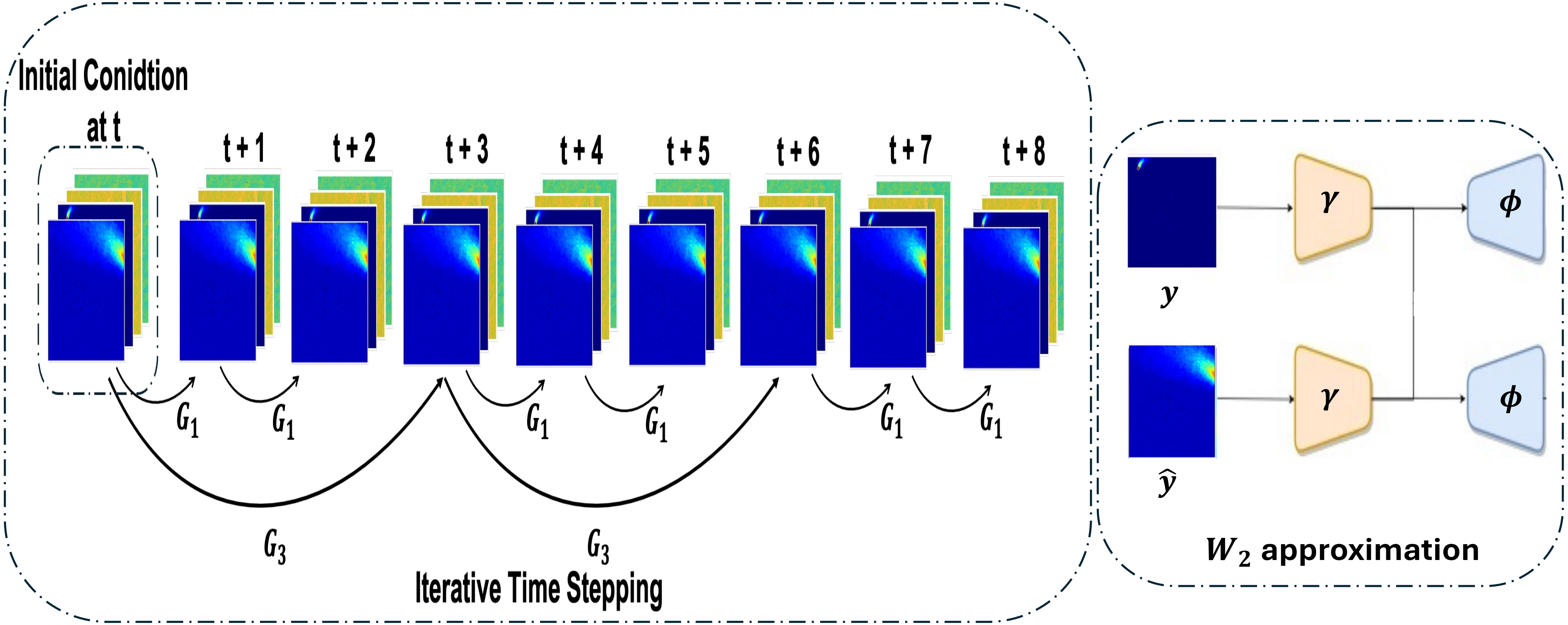


- x^1, x^2: two input fields.
- γ: encoder.
- ϕ: decoder.
- y: the corresponding W_2 distance.
- $L_{recons,1}$ and $L_{recons.2}$: reconstruction loss terms of each input as computed by the KL divergence.
- L_{embed}: L_2 norm of the difference between the Euclidean distance of the embedded features and y.

Normalized L_2 error = 4.35%

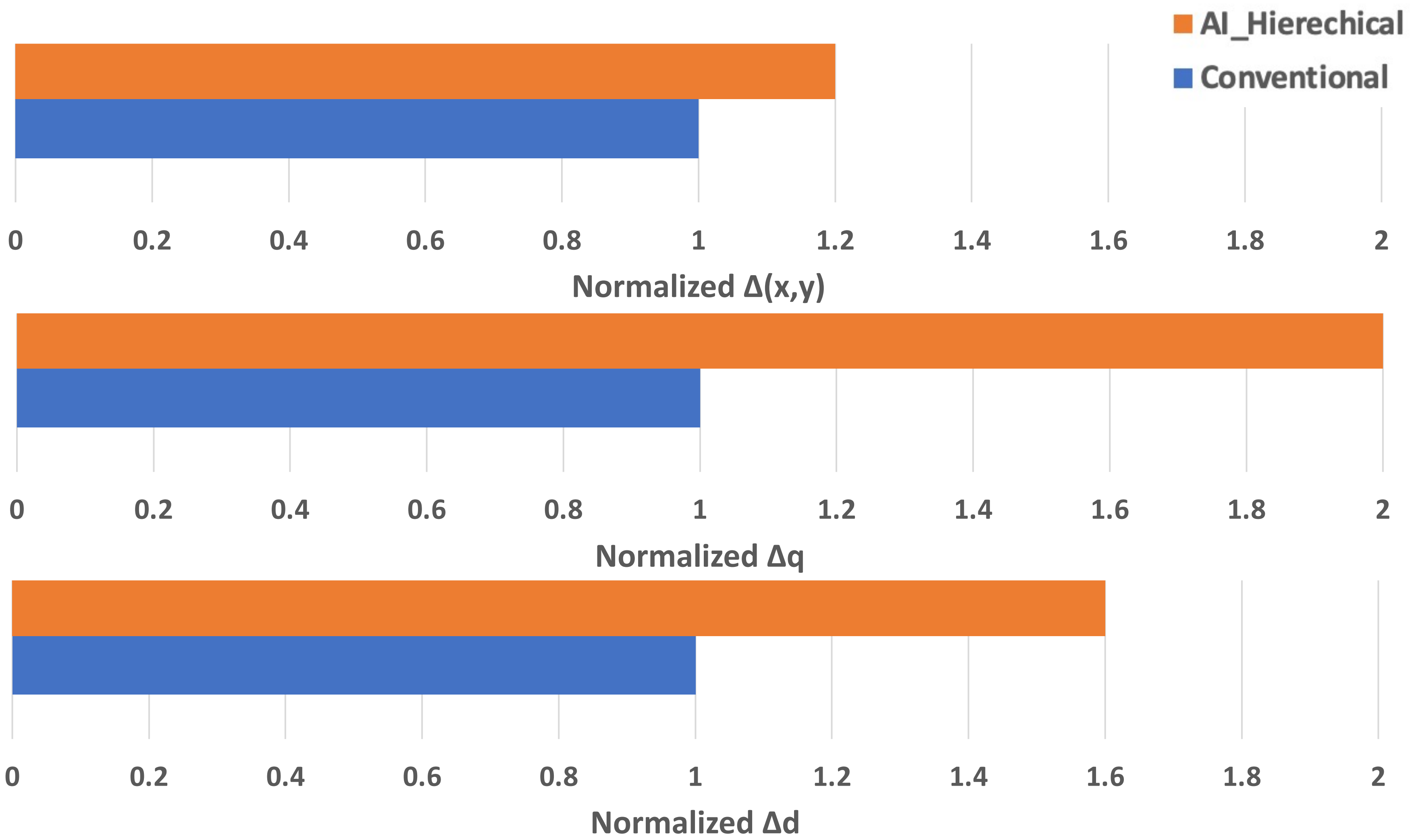
Al Aawar, E., Hammoud, M. A. E. R., & Hoteit, I. (2024). Two-step AI-aided Bayesian source identification of urban-scale pollution. Atmospheric Environment, 323, 120388.

The Dual Hierarchy



Al Aawar, E., Resifi S., Jebari H. & Hoteit I. (2024). Bayesian Source Identification with Dual Hierarchical Neural Networks for Urban Air Pollution. (Under Revision)

Bayesian Solution-10,000 samples

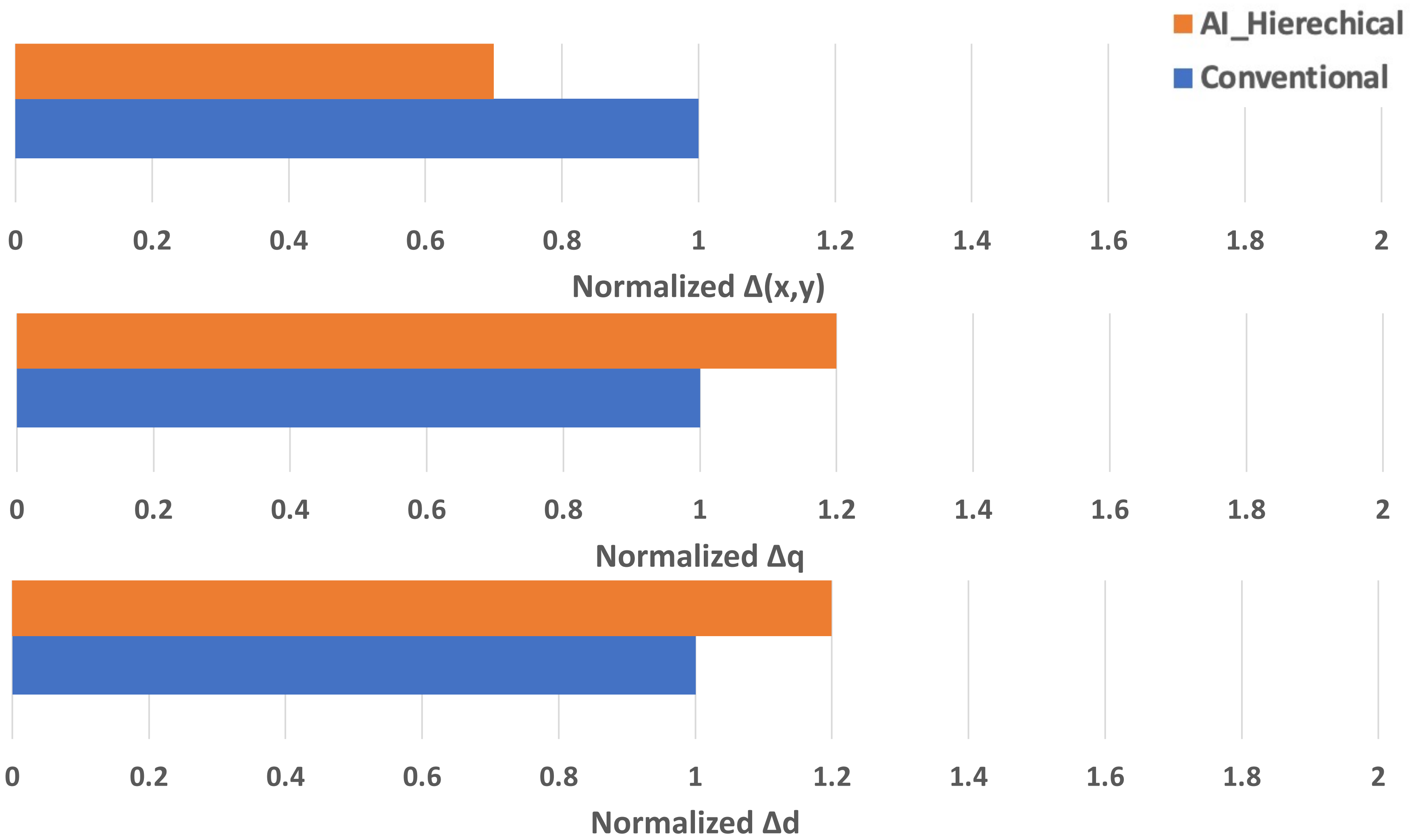


Computational Time

- Conventional: 6days!
- AI-Hierechical: 20minutes

More samples!

Bayesian Solution-100,000 samples



Computational Time

- Conventional: 60days!
- AI-Hierechical: 3.34hours

High Computational Savings
Better Convergence
Higher Confidence

Conclusion

- Successfully developed a NN surrogate model for Lagrangian dispersion, and a NN approximation for the likelihood estimation.
- Bayesian inference framework employs the dual NNs to infer the emission parameters in an urban environment.
- Suggested solutions results in appreciable reduction in computational requirements with minimal loss in performance.
- Our approach can accurately identify sources of air pollution, thus help in responding to harmful emissions and improve overall air quality.

Thank you for your time and attention!

Elissar Al Aawar
PhD Candidate, Earth Sciences and Engineering, KAUST
elissar.aawar@kaust.edu.sa