Air Quality Forecasting in South African Urban Areas Using LSTM Models and SACAQM Data to Capture Temporal Dynamics and Patterns

L. Moloi, S. Gumede, M. Kumar, R. McKenzie, B. Mellado, D. Ngobeni, E. Nkadimeng, L. Phakathi, T. Pule

School of Physics, University of the Witwatersrand

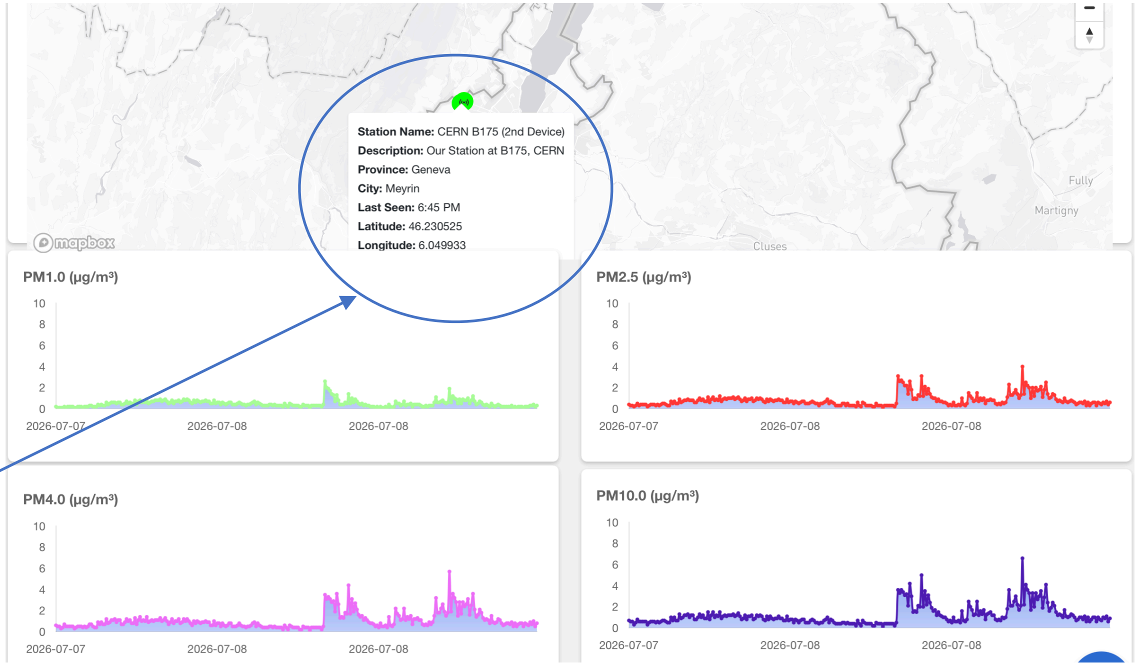
70th SAIP Annual Conference, 2026

Contents

Part 1 — Some Background on air quality monitoring

Why air quality monitoring matters

- 7 million people die from air pollution annually worldwide.
- This includes more than 1.7 million children deaths a year worldwide.
- In 2019, 99% of the world's population was found to breath air that exceeds the **World Health Organisation** limits.
- Moreover, air quality monitoring (temperature, humidity, radon, dust) matters in detector caverns for electronics reliability in research infrastructure such as the ATLAS at CERN.



Identification of air pollutants of concern

- According to the National Ambient Air Quality Standards, Particulate Matter (PM) 10, 2.5 and SO_2 are common proxy indicators for air pollution.

PM_{10}

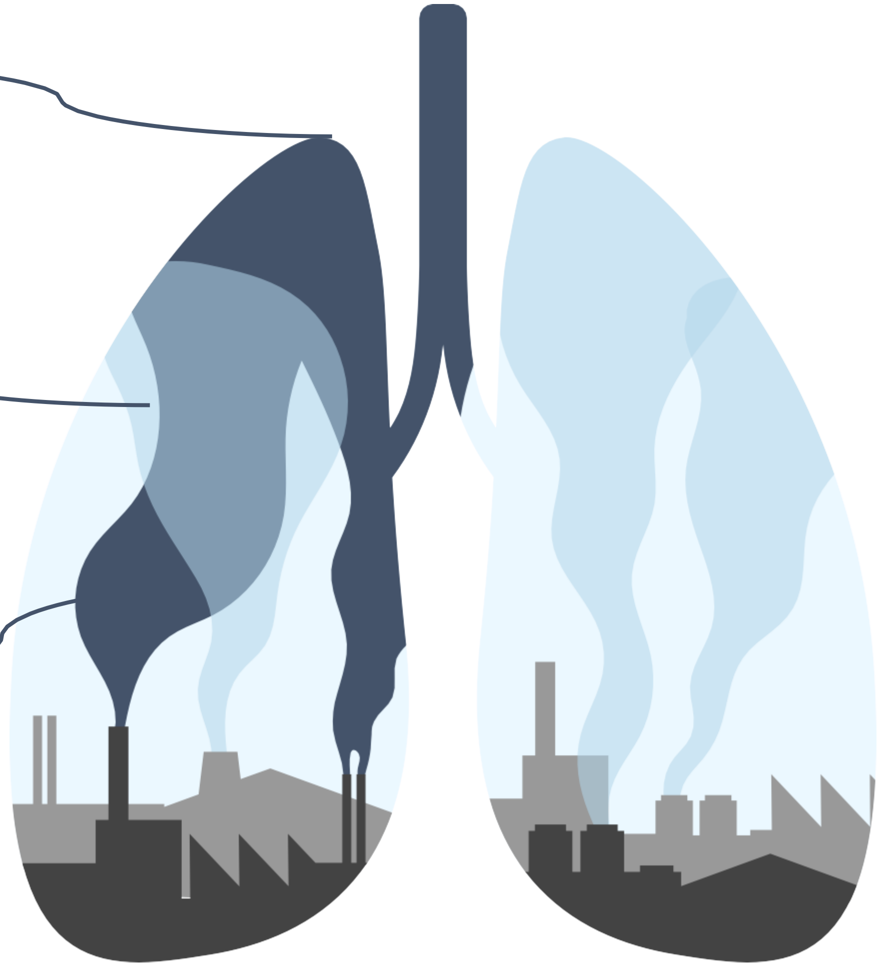
PM_{10} : Particles with a diameter of 10 microns or less, $\leq PM_{10}$ can penetrate and lodge deep inside the lungs

$PM_{2.5}$

$PM_{2.5}$: With a diameter of 2.5 microns or less, ($\leq PM_{2.5}$). $PM_{2.5}$ can penetrate the lung barrier and enter the blood system.

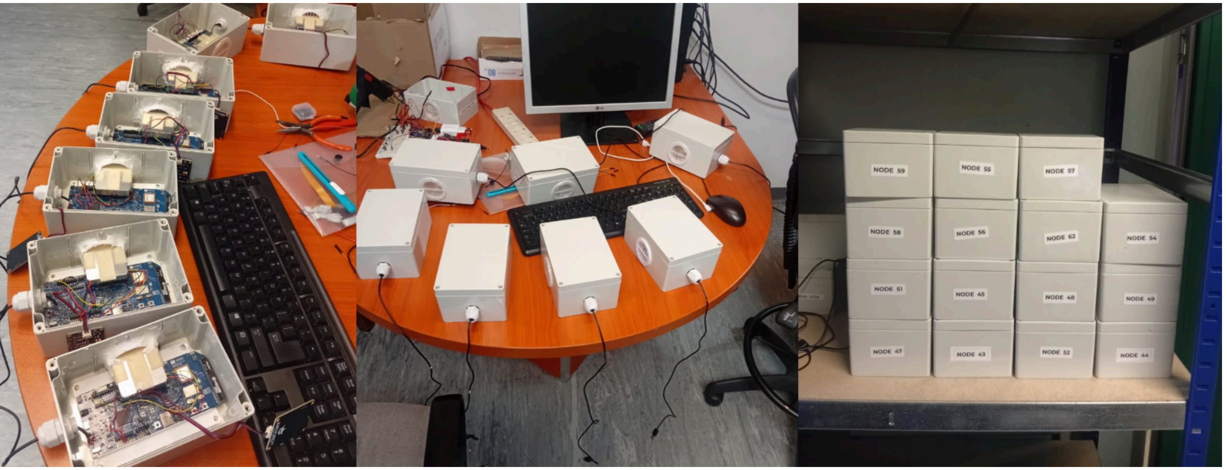
SO_2

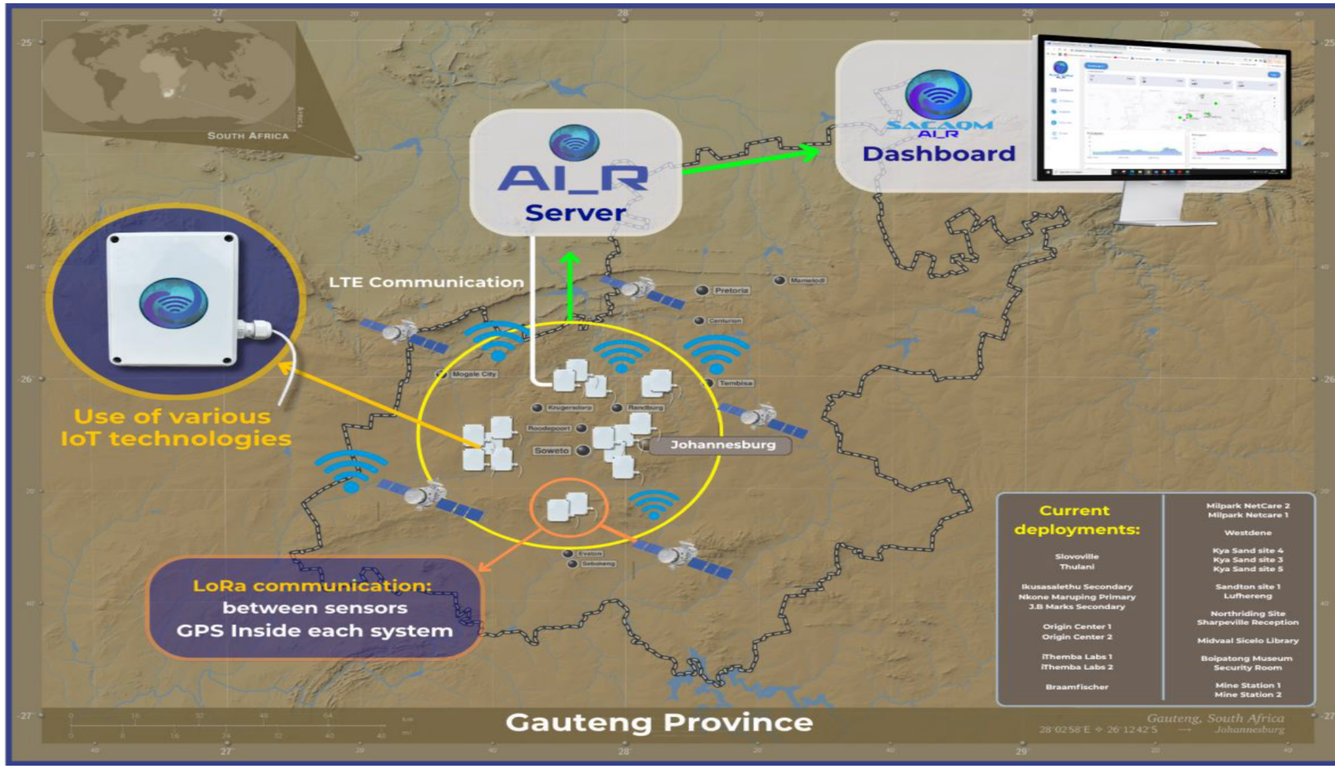
SO_2 : Colourless gas, sharp odour. Produced from the burning of fossil fuels and smelting of mineral ores that contain sulphur. Affect the respiratory system and the functions of the lungs and causes irritation of the eyes.



Artificial Intelligence (AI_r) and the Internet of things (IoT) workflow

- The **South African Consortium of air quality monitoring** is envisioned to address shortcomings in currently available commercial systems.



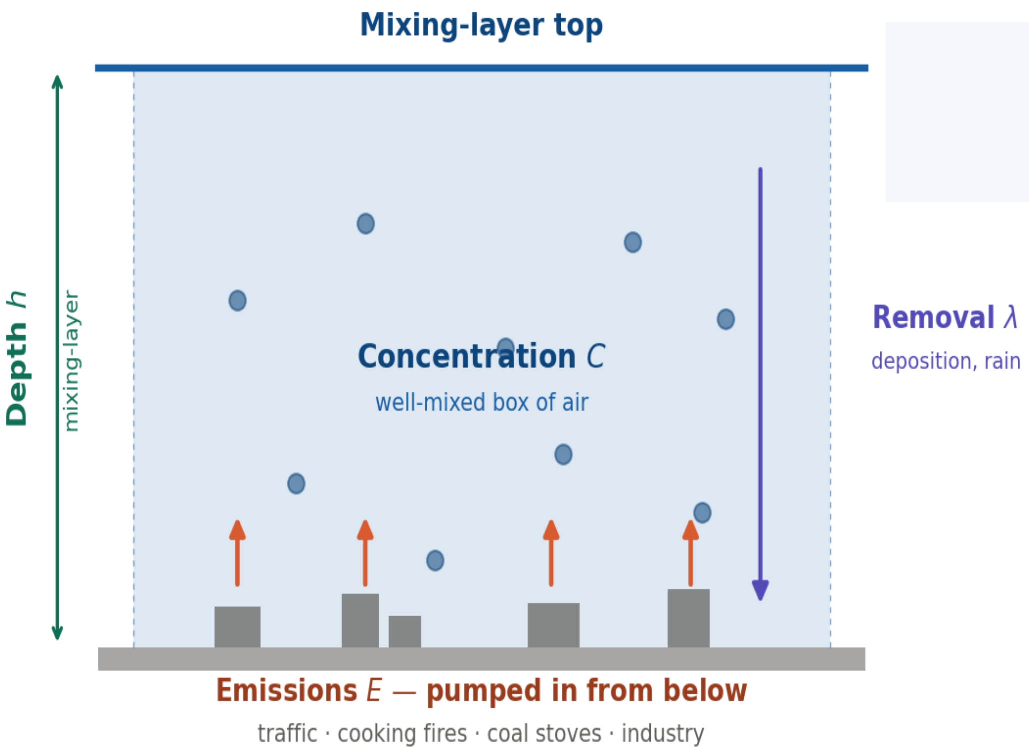


Deployment → Sensor → Dashboard → Predictive Modelling

Part 2 — The physics, before the predictive model(LSTM)

A boundary-layer budget, and what it dictates

The governing equation: the boundary-layer box model



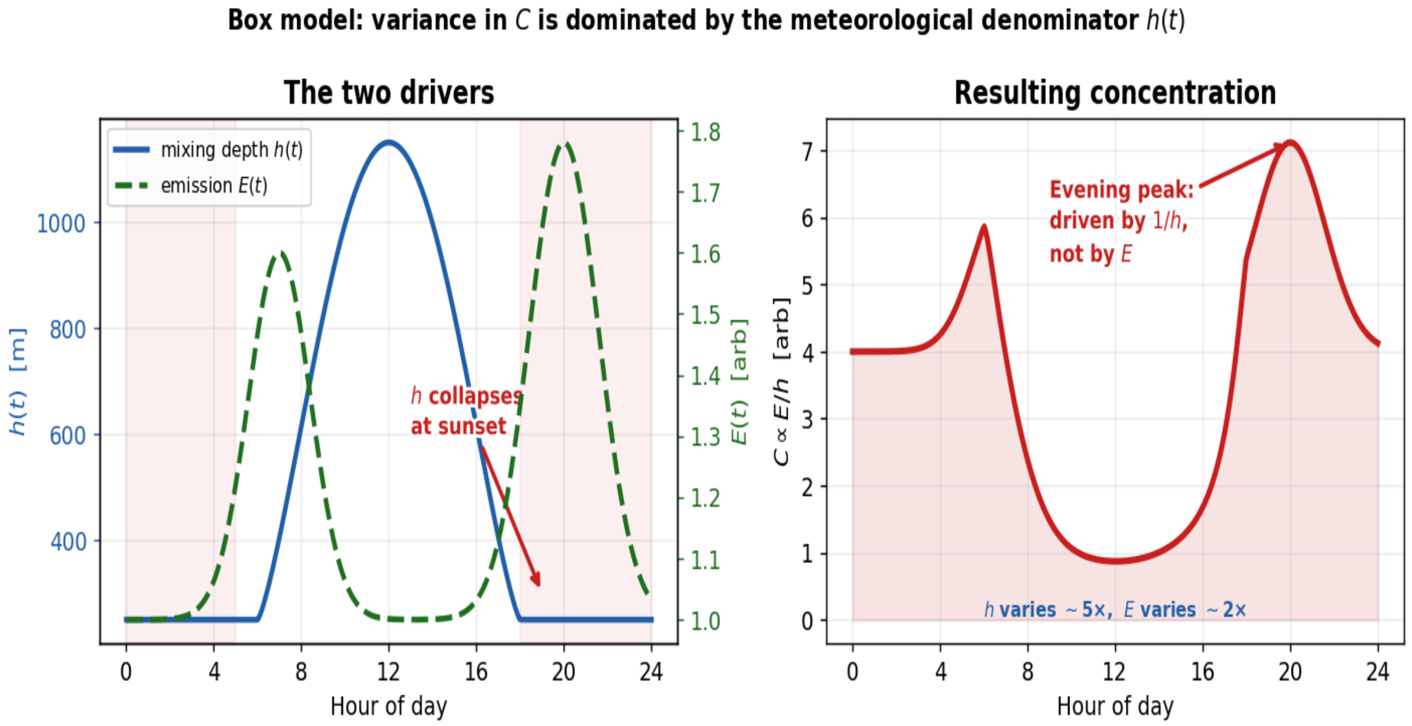
For emission flux E mixed into a layer of depth h, with loss rate λ and background C_0:

$$\frac{dC}{dt} = \frac{E}{h} - \left(\lambda + \frac{1}{h}\left(\frac{dh}{dt}\right)_+\right)(C - C_0)$$

The key term is $\dfrac{E}{h}$:

- concentration is inversely proportional to mixing depth
- collapse h $\Rightarrow$ same mass in a smaller volume $\Rightarrow$ C rises

What really drives the concentration? (Not emissions)



Over a day, h swings ~5× while emissions vary only ~2–3×.

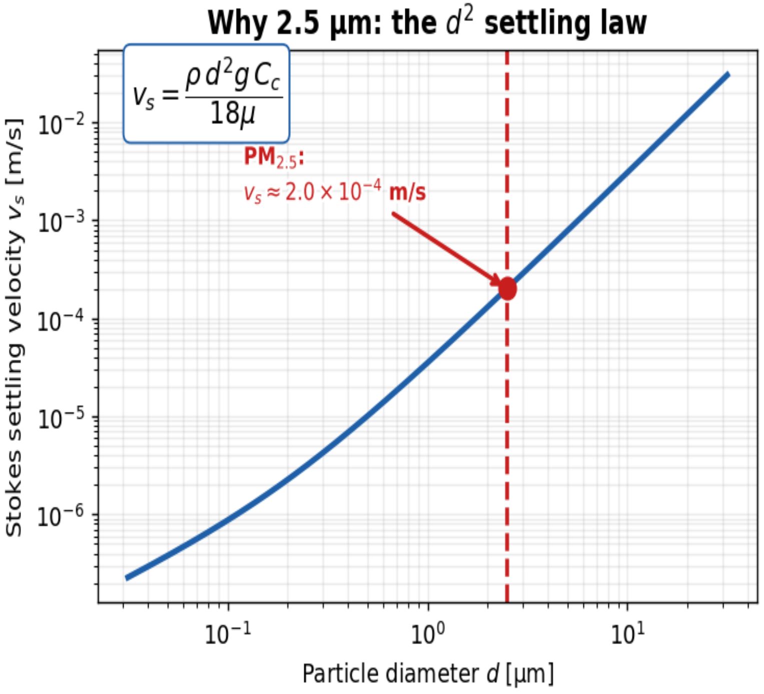
Over a single day the mixing depth varies by a **factor of five**, driven by:

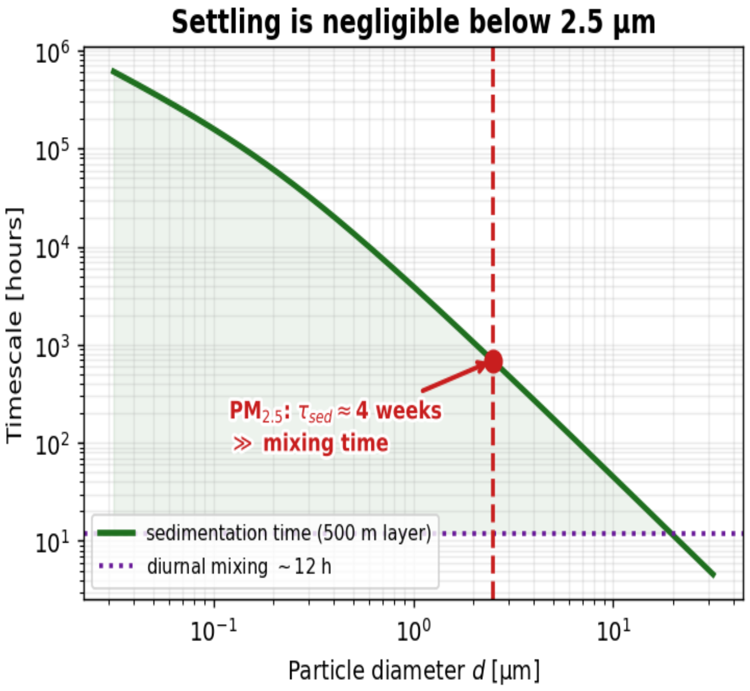
- daytime thermal convection (deep mixing)
- night-time decay of turbulent mixing

[INTERPRETATION] Variance in C is dominated by the meteorological **denominator** — not an emission at all.

A model trained only on past $PM_{2.5}$ sees the symptom, not the driver — the mixing-layer height h.

Why “2.5” is interesting physics, and why we can ignore the loss term





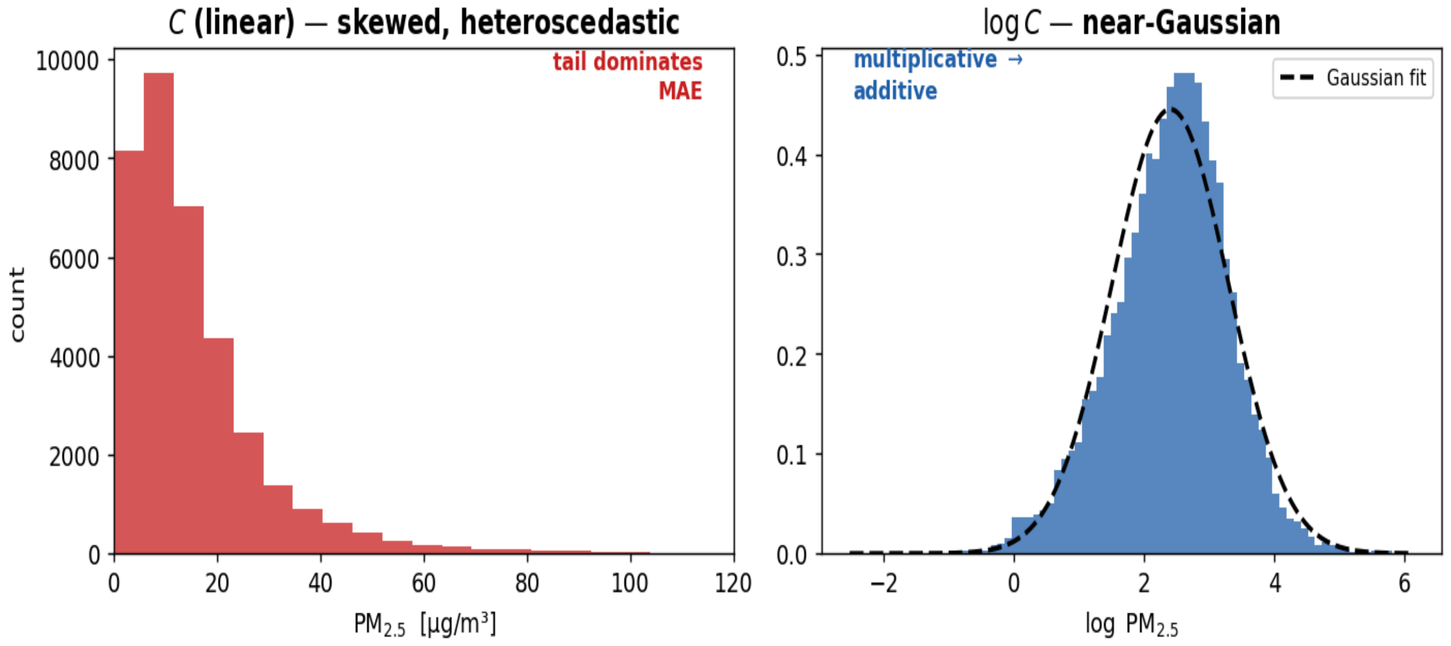
A small sphere falling through air reaches a terminal speed (Stokes’ law):

$$v_s = \frac{\rho d^2 g}{18\mu}$$

- at d = 2.5 µm: $v_s \approx 2\times10^{-4}$ m/s
- settling time ≈ 4 weeks (500 m layer)
- mixing time ≈ 12 h — ~600× faster

Dropping λ is justified. Below 2.5 µm the particle is a quasi-passive tracer of the air itself.

C is not the natural variable — its logarithm is

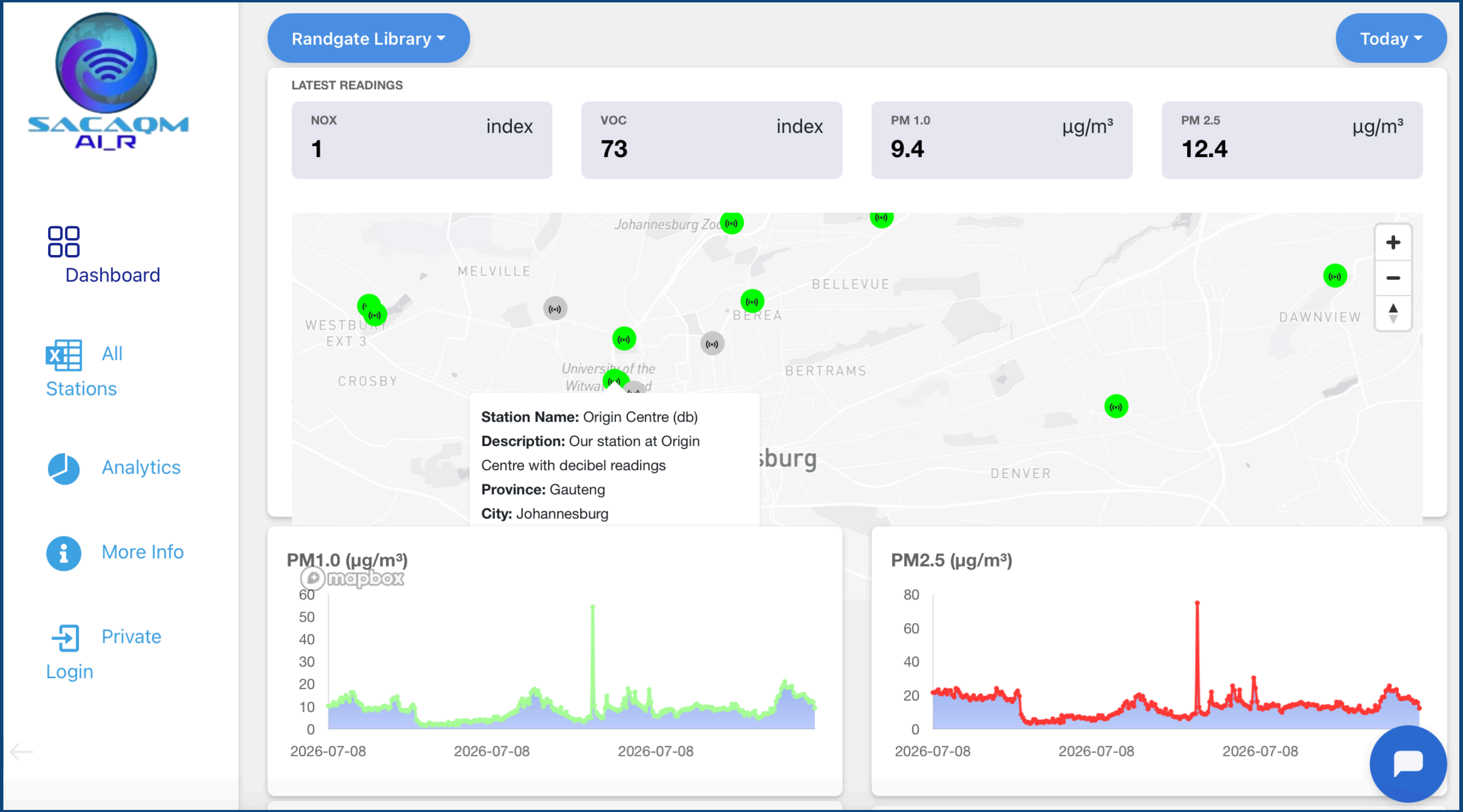


$C \sim \dfrac{E}{h}$ combines by **multiplication, not addition**: halve h ⇒ double C.

Many independent factors multiplying together give a log-normal distribution (central limit theorem in log-space).

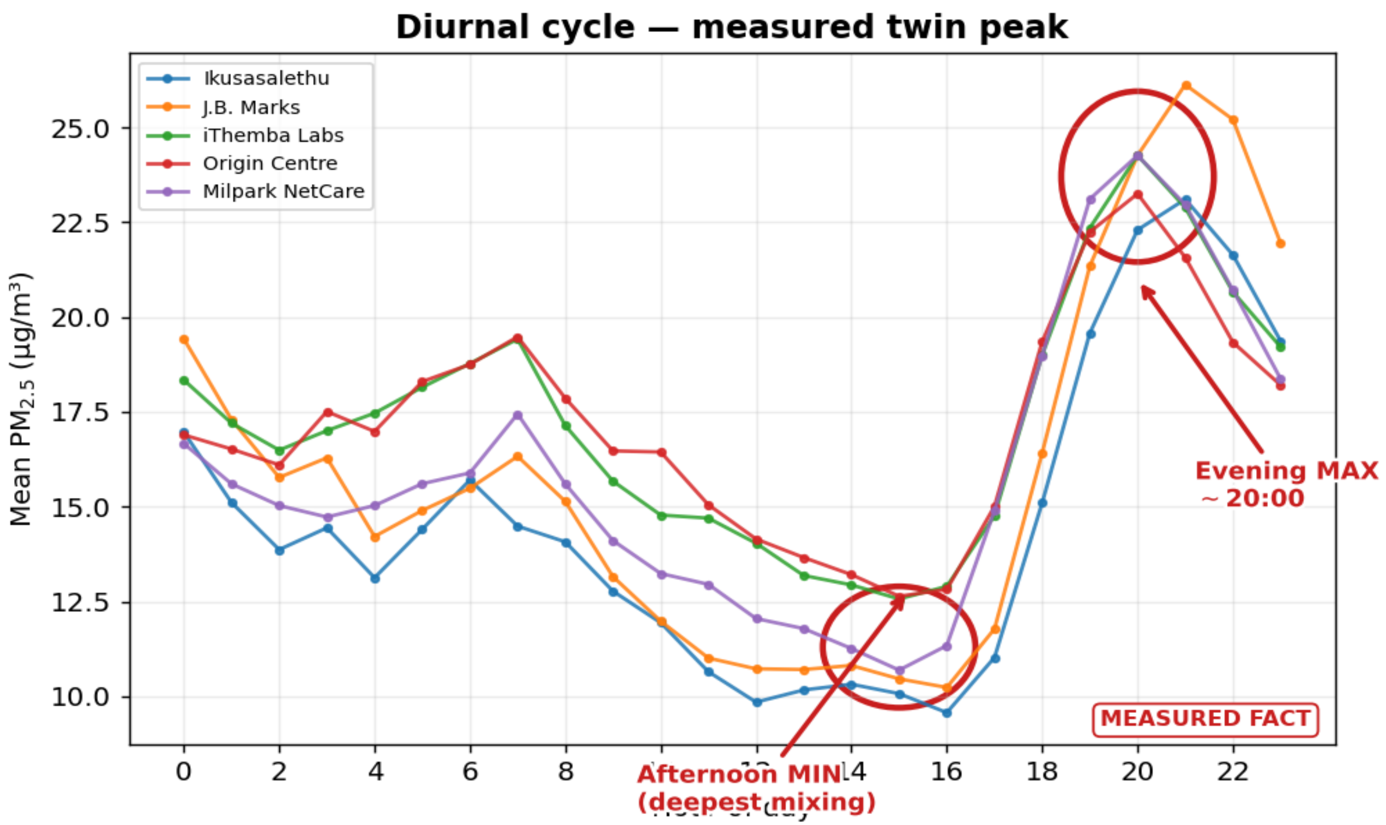
Predict the logarithm of $PM_{2.5}$, not $PM_{2.5}$ itself.

Part 3 — The data as evidence (Five Gauteng Sensors)

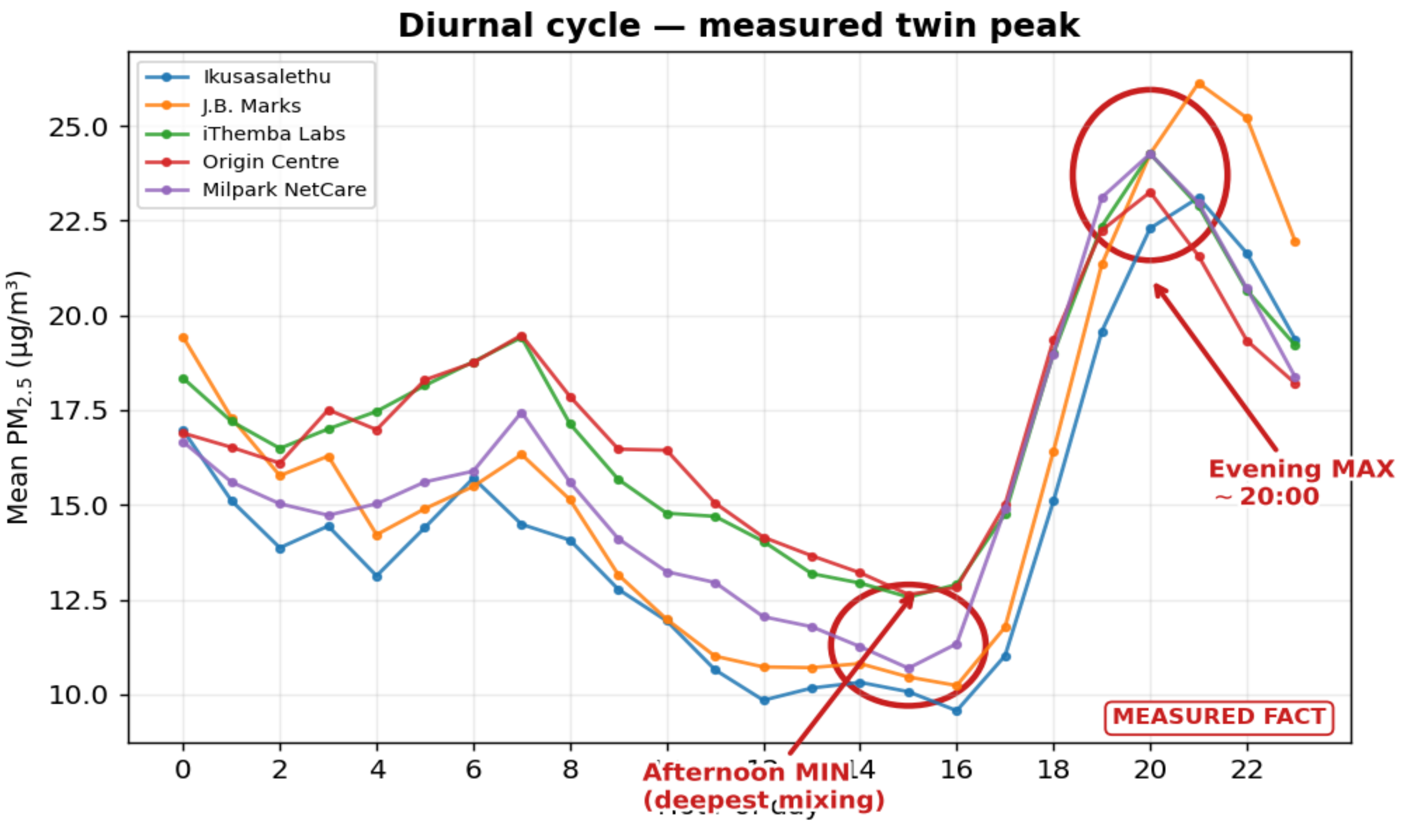


The daily cycle — raw measurement

- Ikusasalethu (Residential)
- J.B Marks (Residential)
- IThemba labs (Research Facility)
- Origin Centre (University)
- Milpark NetCare(Hospital)



The twin-peak daily cycle



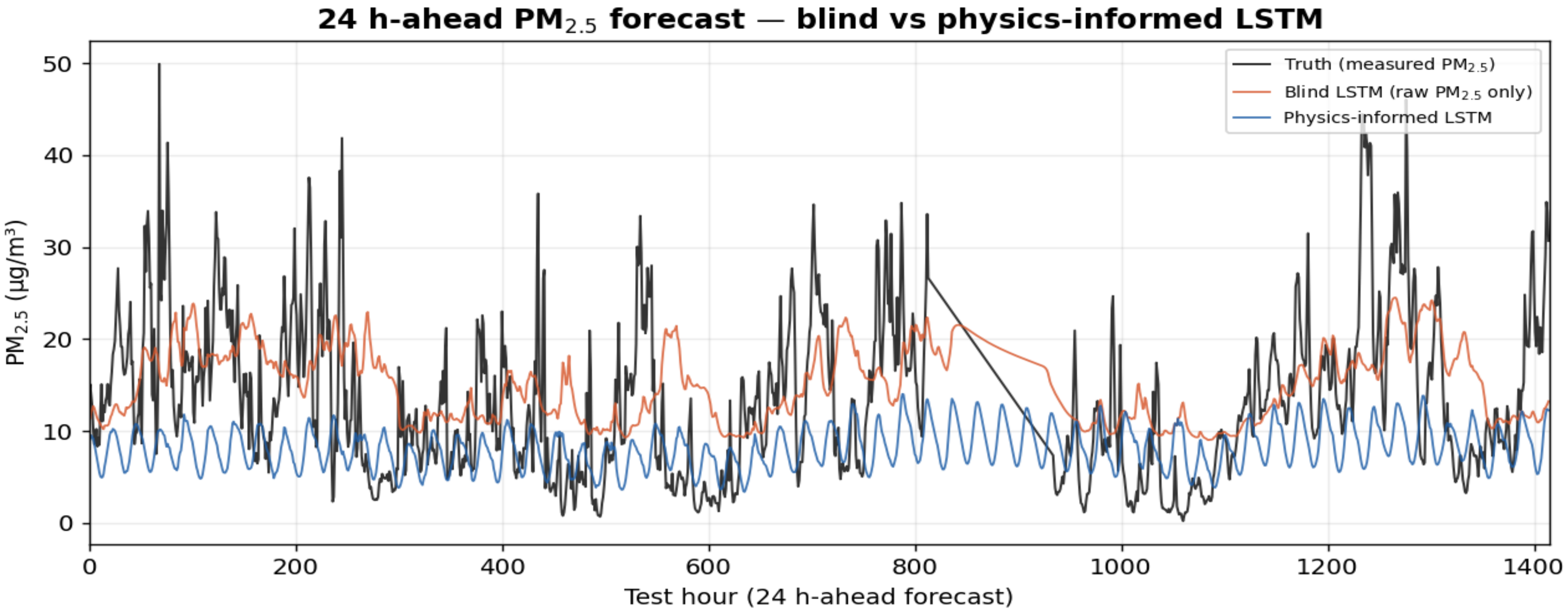
Afternoon minimum (~15:00, deep mixing) and evening maximum (~20:00), at all five stations.

The concentration peak ***lags*** the emission activity — the tell-tale fingerprint that mixing depth, not emission timing, is the driver.

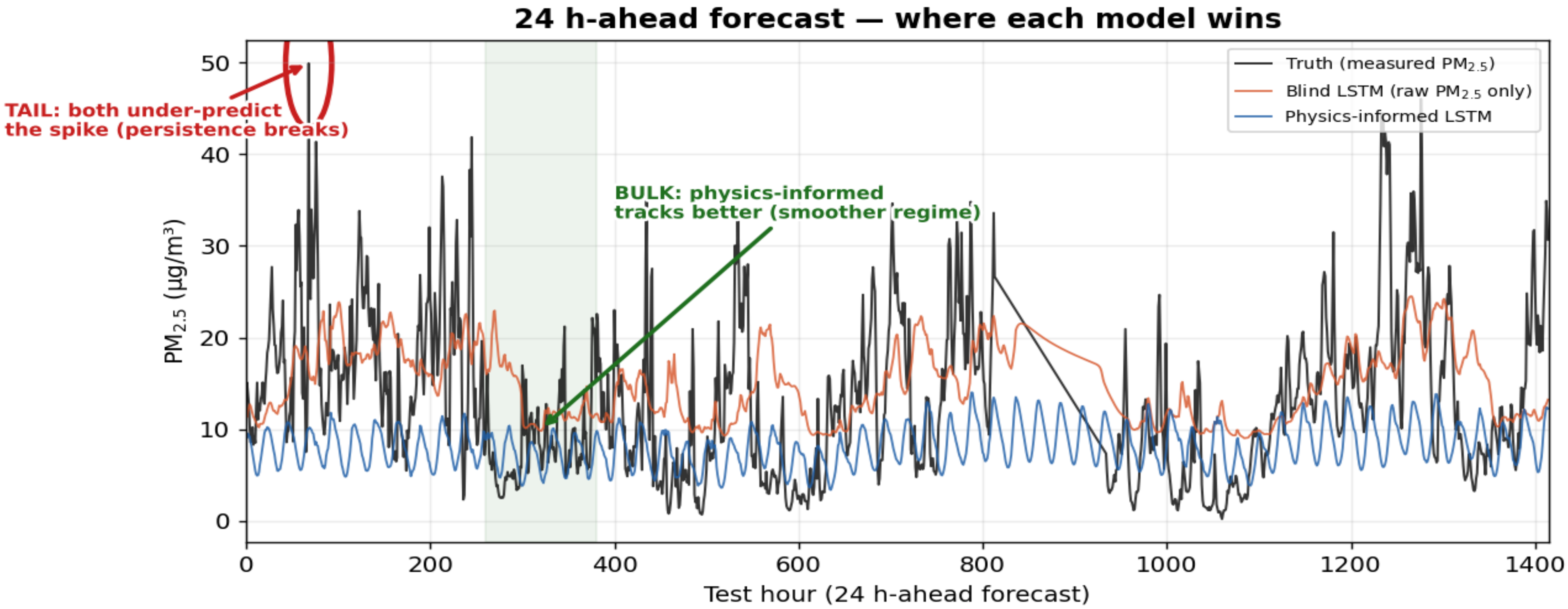
Part 4 — The test (Results)

Blind vs “physics-informed” LSTM

The forecast

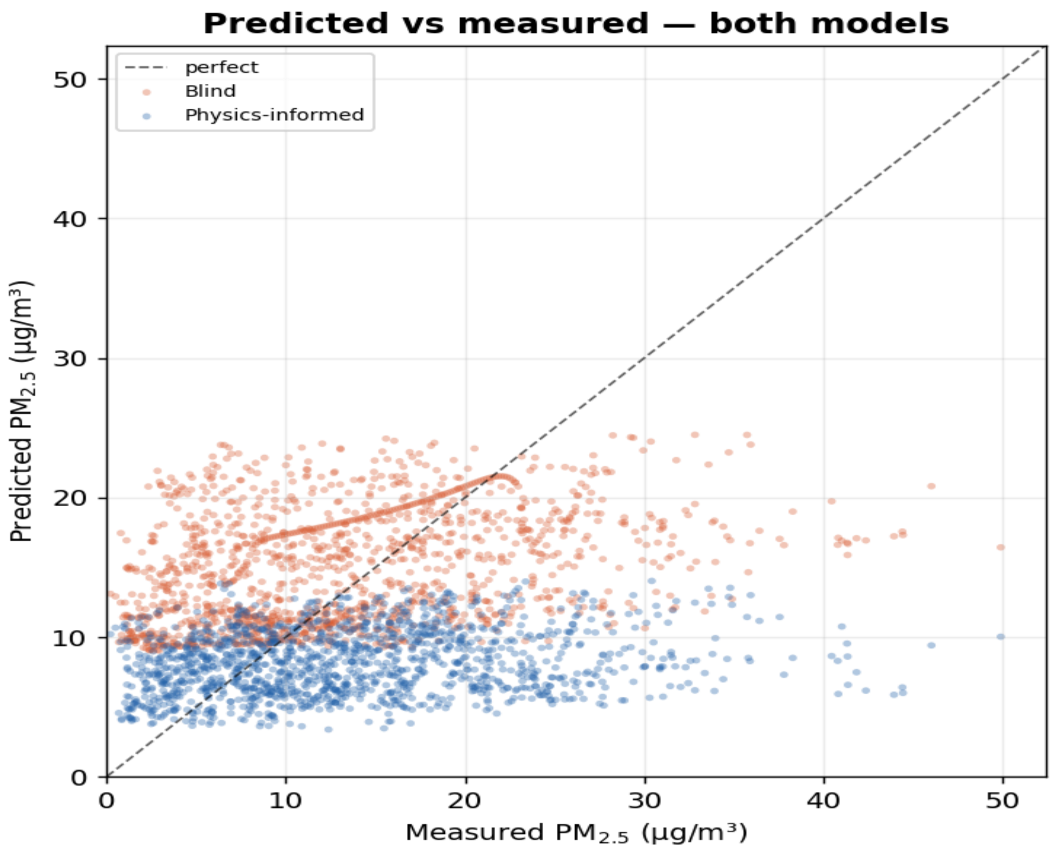


The forecast — where each model wins

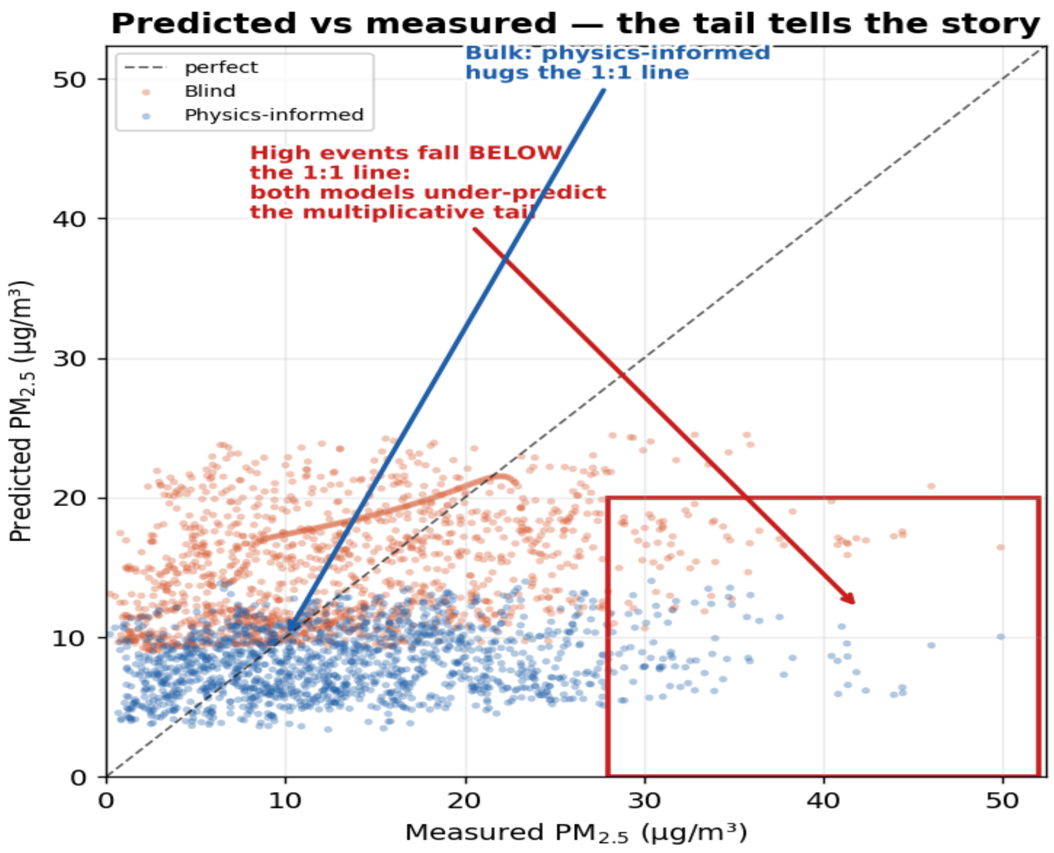


In the **bulk**, the physics-informed model tracks better. On the **tail spike**, both under-predict — persistence alone cannot anticipate the event.

Predicted vs measured



Raw scatter.



Annotated.

Summary

- Where that absorption is possible (bulk, single site) ⇒ the two tie. **(When the data is calm and comes from one station both models do equally well)**
- Where it is not (the multiplicative tail) ⇒ both fail, and the log-target's behaviour is itself a physical prediction of why. **(When there are sudden spikes both models struggle)**
- The physics does not merely improve a model, it explains when a model can work at all, and why it breaks. For instance:
  (i) The process is *multiplicative* → the data should be log-normal
  (ii) It's driven by a daily mixing cycle → there should be 24-hour periodic memory

Future outlook

- Use Spatial Temporal Graphic Neural Networks as the predictive model

Thank you

Questions?