



ETIQUET – EXtrapolating Temperature and alr Quality Exposure in Space and Time: A Semi-Supervised Learning Framework for Environmental Intelligence in Data-Scarce Settings



Anotida Marrian Chapunza¹, Liberty Makacha¹, Yiqun Han³, Benjamin Barratt³, Peter von Dadelszen⁴, Marie Laure Volvert⁴, Matthew F Chersich⁵, Prestige Tatenda Makanga^{1,2}, Stanley Luchters², The PRECISE Network, The HE²AT Centre

Affiliations: ¹Midlands State University, Gweru, Zimbabwe, ²CeSHHAR, Harare, Zimbabwe, ³Imperial College London, United Kingdom, ⁴King's College London, United Kingdom, ⁵Wits Health Consortium, South Africa

Background

Ambient air pollution (PM_{2.5}) and temperature extremes are major maternal and child health risks in sub-Saharan Africa

Sparse monitoring and label-limited LMIC datasets lead to exposure misclassification and constrain traditional modelling

These data gaps hinder accurate large-scale exposure estimation, hence the need for robust modelling frameworks

Objectives

To develop and evaluate a machine learning pipeline to extrapolate PM_{2.5} and temperature exposure from a small cohort to a larger population across space and time

To assess the explanatory value of the environment and geography in exposure prediction and identify key predictors

Methods

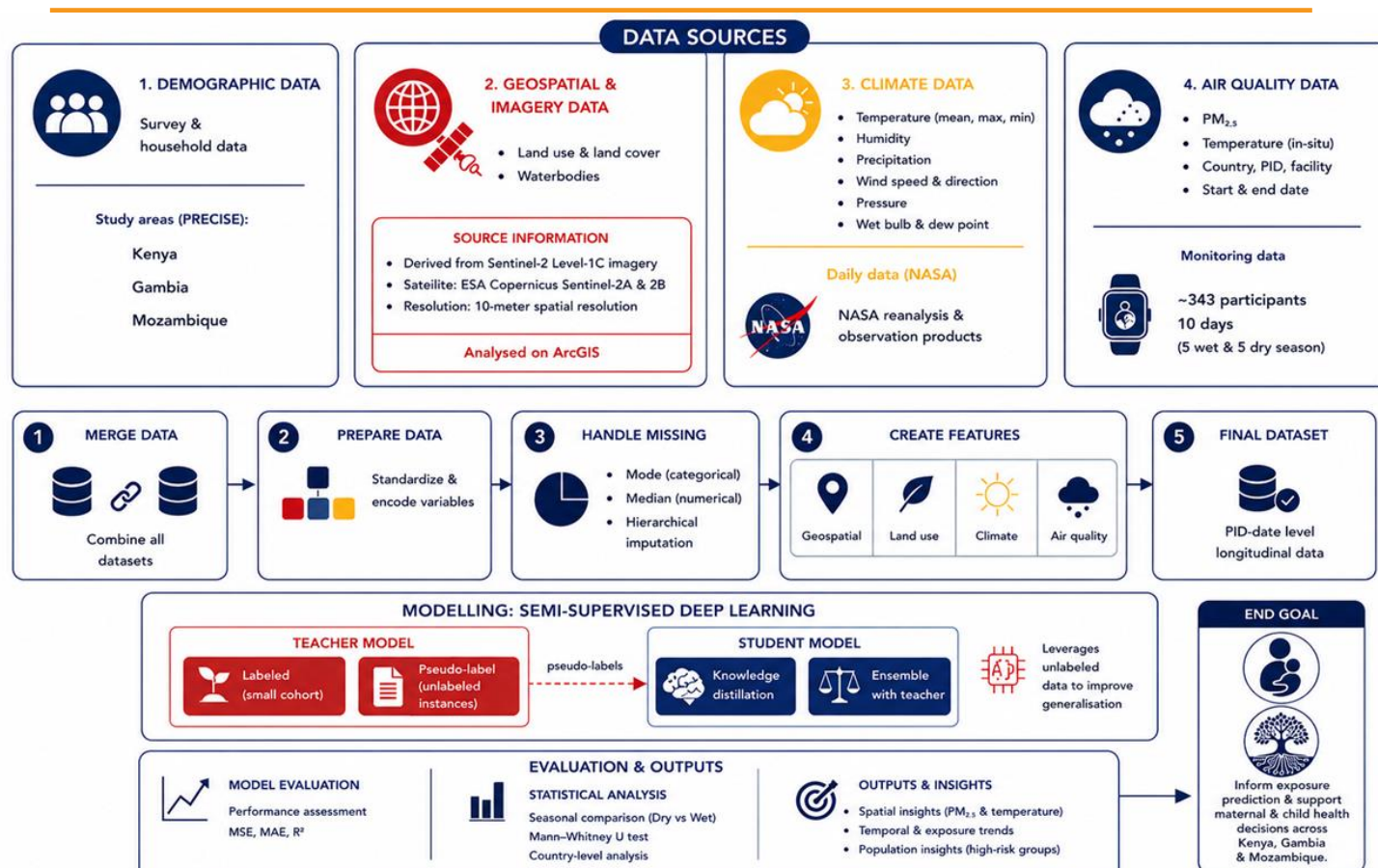


Fig 1: Project Methodology:

Results

Combining labelled and unlabelled data substantially improves cross-validation performance and stability, reducing error (RMSE) and increasing R² compared to labelled-only models.

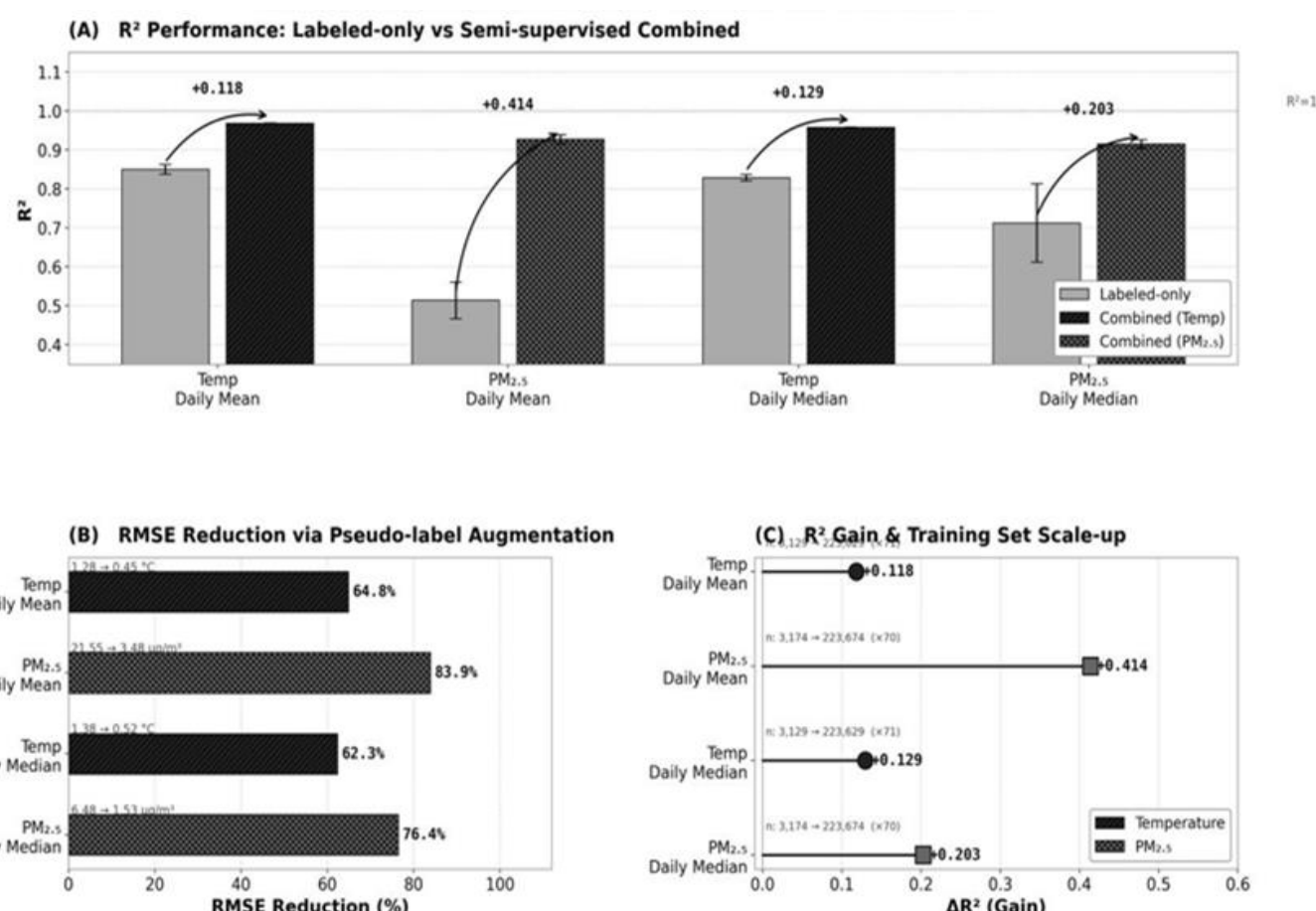


Fig 2 : 5-fold cross-validation on (1) labelled data only, and (2) labelled data + pseudo-labels (semi-supervised).

Results

Cross-country validation was performed training country specific models on one country and testing them on another.

Results reveals showed severe performance collapse between the two sets. PM_{2.5} models and temperature models.

	Gambia Dry	Gambia Wet	Kenya Dry	Kenya Wet	Mozambique Dry	Mozambique Wet
Temperature (Mean/Max)	2.50	2.47	0.12	0.59	0.10	1.18
Humidity / Precipitation	0.74	1.00	0.34	0.87	0.40	0.96
Wind (Speed/Direction)	0.82	0.57	0.07	0.76	1.96	0.52
Land Use / Environment	0.28	0.18	0.53	0.45	0.49	0.49
Socioeconomic / Demographic	0.61	0.36	0.29	0.36	0.34	0.34

Mean [SHAP] (Average Impact on PM_{2.5}) Low Impact High Impact

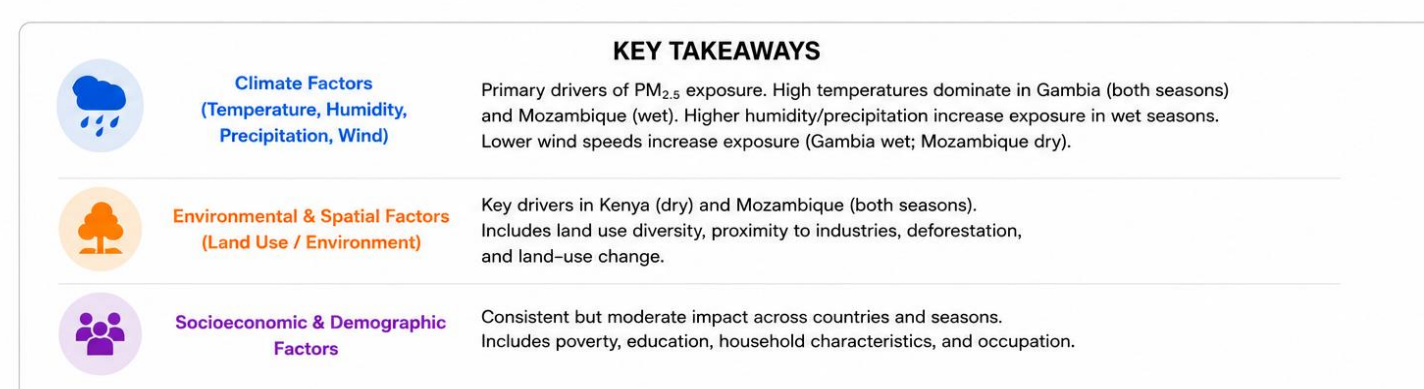


Fig 5: Seasonal PM_{2.5} Exposure by country

Climate and environmental factors dominate predictions across all countries and seasons.

Interpretation

Combining labelled and unlabelled data improves accuracy and stability over labelled-only approaches

Pronounced spatial heterogeneity reduces cross-country predictive performance and transferability across settings

PM_{2.5} prediction patterns are highly location-specific, limiting model generalization across countries.

Climate and environmental variables strongly influence exposure, with effects differing by country and season.

Conclusion

Semi-supervised frameworks bridge critical data gaps where traditional supervised approaches fail especially in low resource settings

Exposure patterns are geographically dependent, requiring place-specific interventions

Robust deployment requires local calibration rather than cross-country transfer

Key words:.

Semi-supervised learning, PM_{2.5} exposure modelling, Spatial heterogeneity, Temperature extremes, Data scarcity, LMIC.

Conflict of Interests: None

Corresponding Author: Email: chapunzaa@staff.msu.ac.zw

