Geographic Imbalance and Spatial Generalisation in Machine Learning for Global Water Quality Prediction

Sushma Kumari[1] Krista Alikas[2] Soumyabrata Dev[1,3]

[1]School of Computer Science, University College Dublin, Dublin, Ireland [2]Tartu Observatory, University of Tartu, Toravere, Estonia [3]School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland

1 · Motivation

Global lake water quality monitoring via satellite remote sensing is critical for tracking eutrophication, harmful algal blooms, and drinking water safety. ML models trained on the **GLORIA** dataset (n = 6,675 in-situ samples) promise global applicability but are they truly generalisable?

Two compounding failure modes:

1. **Spatial leakage** - nearby samples leak into test sets, inflating R^2 artificially.
2. **Optical mismatch** - models trained on clear blue water fail on CDOM-rich brown water.

2 · GLORIA Sampling Bias

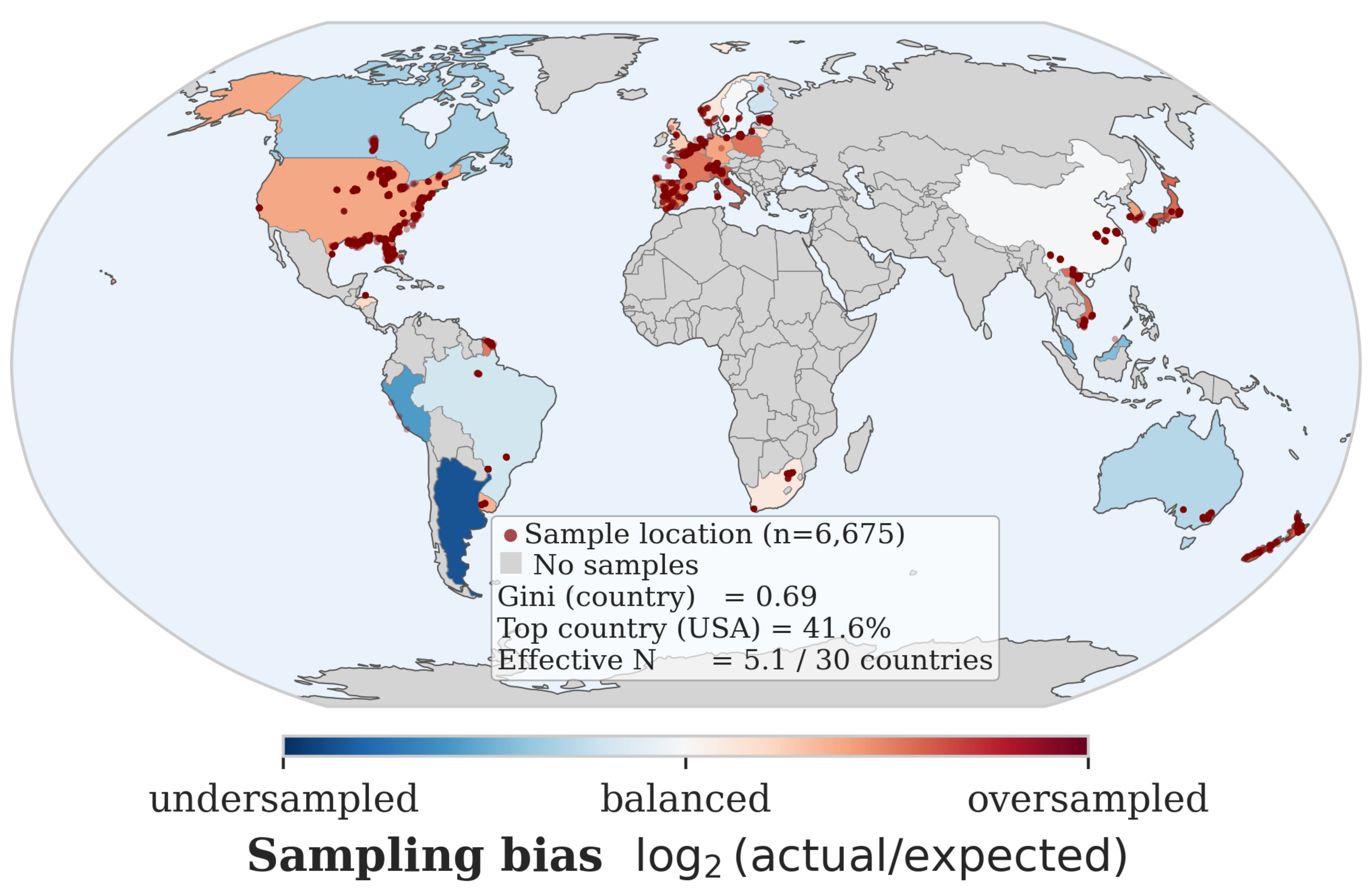

Fig. 1. Sampling bias ($\log_2$ actual/expected). USA (41.6%) and Europe dominate; Africa and Asia are severely underrepresented (Gini = 0.69, effective $N = 5.1/30$).

Data mismatch: GLORIA over-represents oligotrophic (clear, blue) lakes concentrated in USA & Europe. CDOM-rich and hypereutrophic lakes are undersampled, causing models to fail on out-of-distribution water types.

3 · Global Trophic States

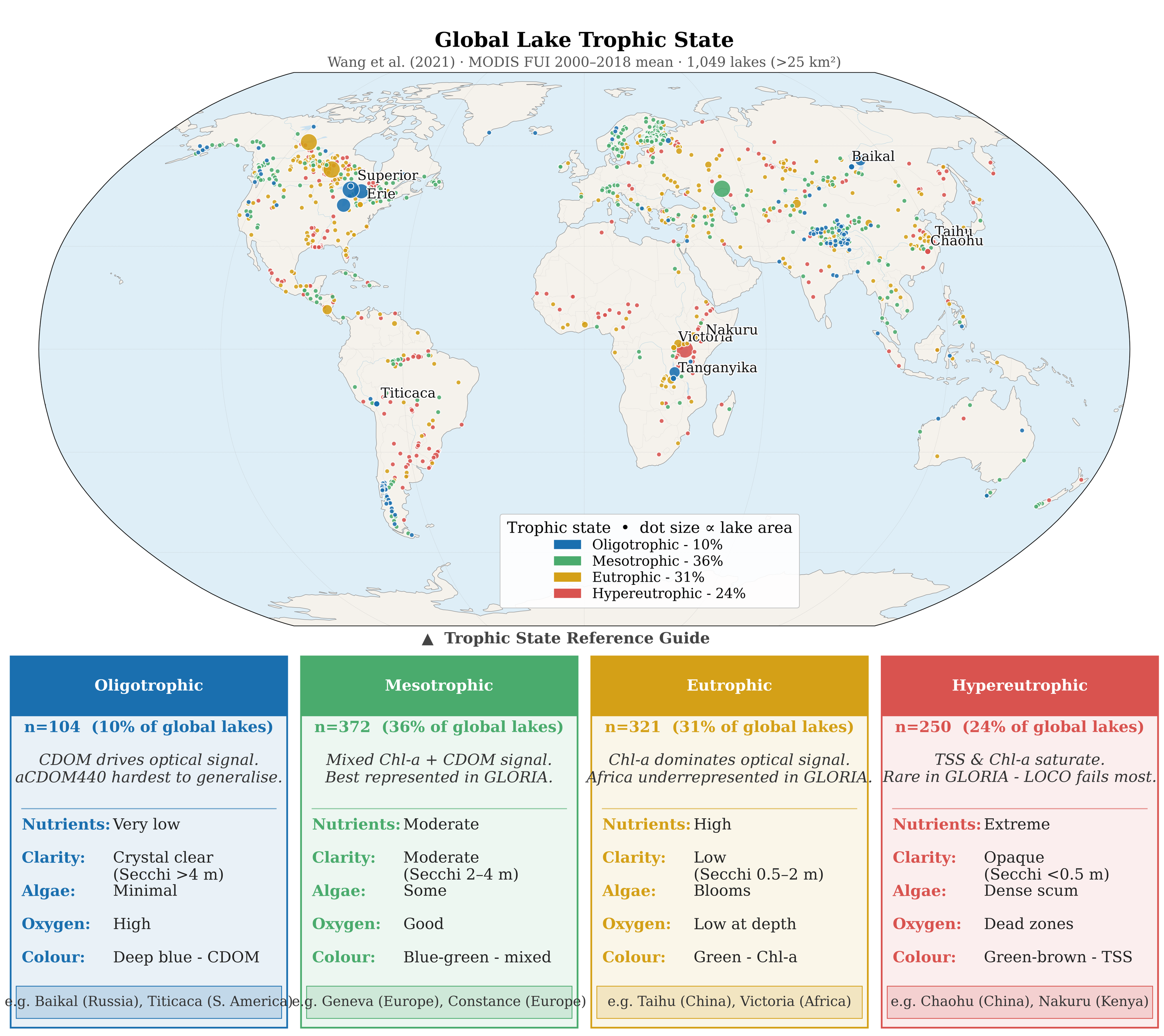

Fig. 2. Trophic state distribution across 1,049 lakes (Wang et al. 2021, MODIS FUI 2000 - 2018). Hypereutrophic lakes (24%) and CDOM-rich oligotrophic lakes are rare in GLORIA, creating optical gaps that spatial CV reveals.

Contact

Sushma Kumari

sushma.kumari@ucdconnect.ie

School of Computer Science, University College Dublin, Dublin, Ireland

GitHub Code

Linkedin Profile

Poster presented at ICML 2026

4 · When Does Random CV (Cross-Validation) Fail?

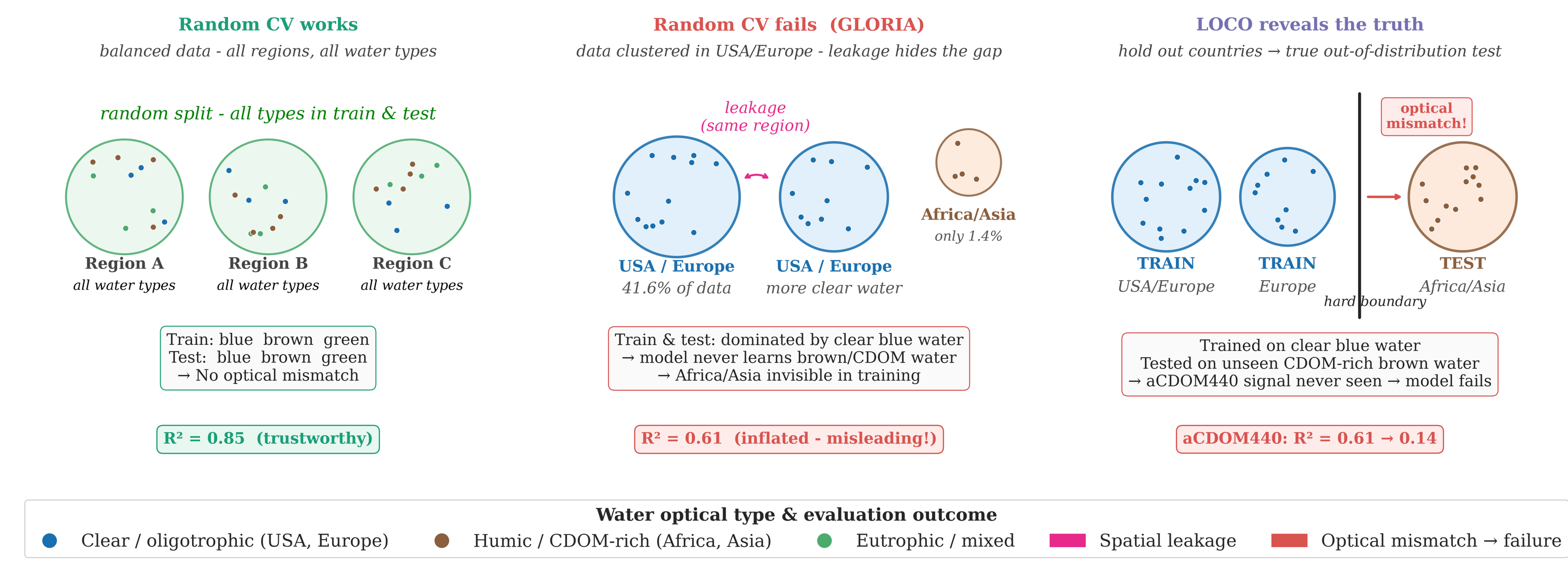

Fig. 3. **Left:** Random CV trustworthy ($R^2 = 0.85$). **Centre:** Geographic bias inflates $R^2 = 0.61$. **Right:** LOCO reveals aCDOM440 $R^2 = 0.14$.

Headline result - aCDOM440: Random CV reports $R^2 = 0.61$. LOCO reveals $R^2 = 0.13$ - 0.14. $\approx$**4**$\times$ **overestimation of true model skill.**

5 · R^2 Results Across All Splits

We compare four cross-validation splits across two models and three water quality targets (Chla, TSS, aCDOM440):

Split	Description
Random	Standard 80/20 (baseline)
LOCZO	Leave-One-Climate-Zone-Out
LORO	Leave-One-Region-Out
LOCO	**Leave-One-Country-Out (strictest)**

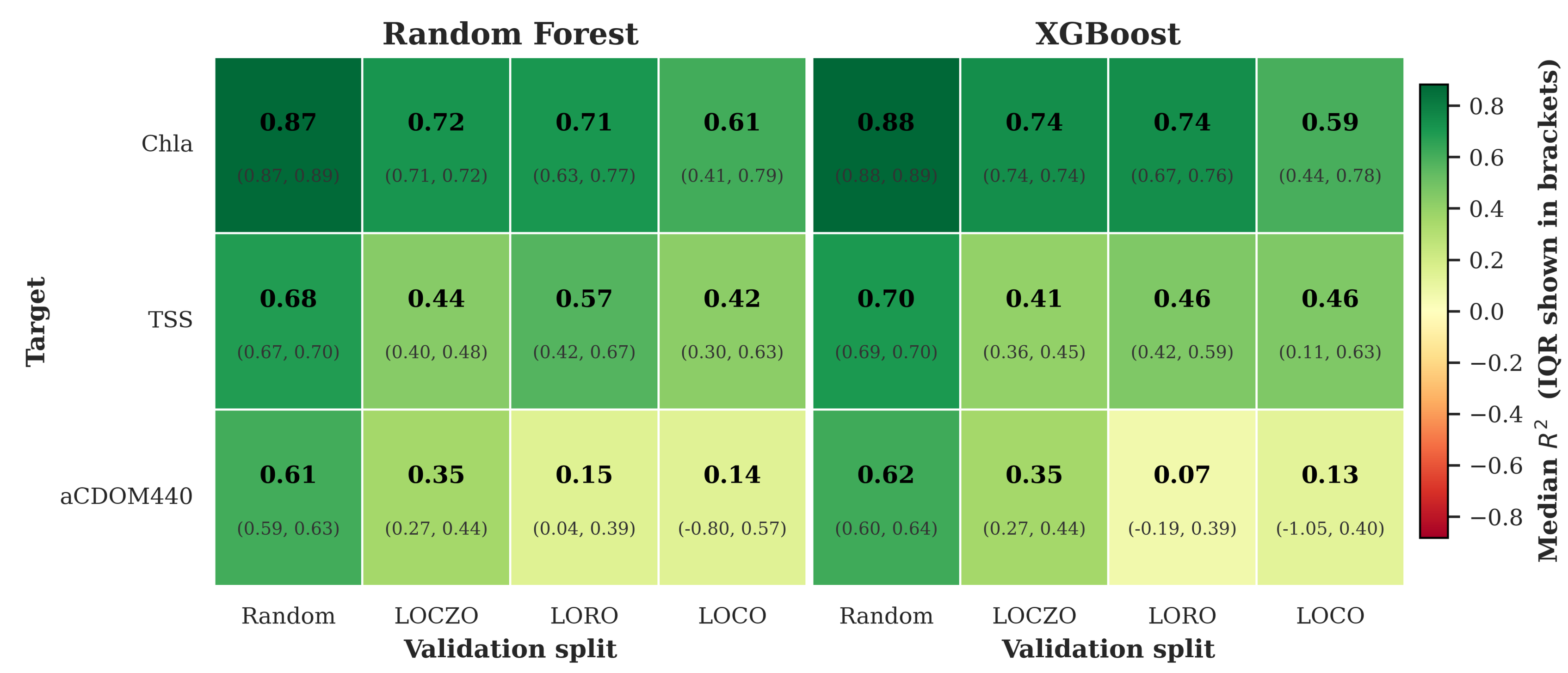

Fig. 4. Median R^2 (IQR in brackets) for Random Forest and XGBoost. Green = high skill; red = low. Spatial splits consistently reveal lower but more *honest* performance than random CV.

6 · Key Findings by Target

Chla - Most Robust

Best R^2 across all splits (0.59 - 0.88). LOCO: RF = 0.61, XGB = 0.59. Best represented target in GLORIA. Strong Chla signal survives spatial holdout.

TSS - Moderate Degradation

Random CV: $R^2 \approx 0.68$ - 0.70. LOCO drops to 0.42 - 0.46. Turbidity signal degrades in eutrophic, Africa-type lakes.

aCDOM440 - Fails OOD

Random CV: $R^2 = 0.61$ (deceptively high). LOCO: 0.13 - 0.14 (near chance). IQR spans negative values: (-1.05, 0.40). CDOM signal absent from most training lakes.

Both models behave almost identically under all splits ($\Delta R^2 < 0.02$ in most cells). XGBoost marginally outperforms RF on Chla. **Neither model overcomes the data bias problem the bottleneck is data, not model architecture.**

7 · Conclusions

1. **Spatial leakage is pervasive:** Random CV inflates R^2 by up to 4$\times$ vs. LOCO.
2. **Optical mismatch is the root cause:** Models trained on clear water cannot retrieve aCDOM440 in CDOM-rich lakes.
3. **LOCO is essential:** Country-level holdout is the minimum honest benchmark for global generalisation.
4. **Data beats model:** Both RF and XGBoost fail equally - diverse data is the real fix.

Take-home message

A model reporting $R^2 = 0.61$ on random CV may achieve only $R^2 = 0.14$ on unseen countries. **Always validate spatially.**

References

1. Lehmann et al. (2023). *Earth Syst. Sci. Data*
2. Wang et al. (2021). *Remote Sens. Environ.*
3. Roberts et al. (2017). *Ecography*
4. Breiman, L. (2001). Random Forests. *Mach. Learn.* 45(1):5–32.
5. Chen, T. & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *KDD.*