

Spatial Cross-Validation for Machine Learning Systems

Towards comprehensive geospatial dataset evaluation

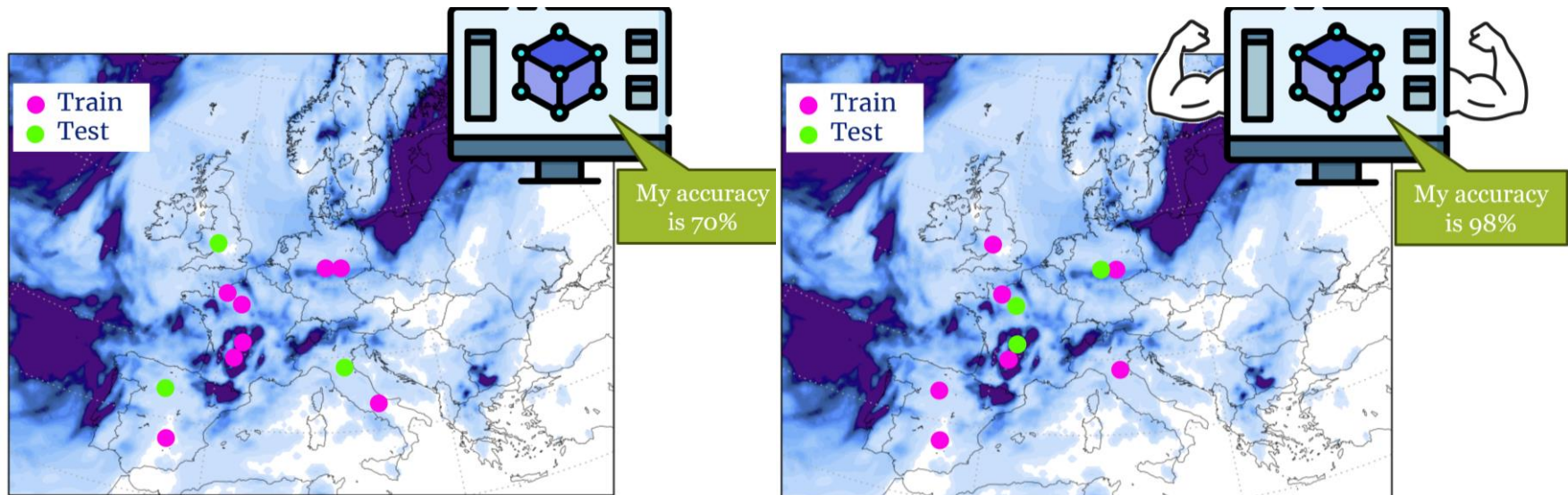
Abygail Stegenga, Julia Wąsala, Jan van Rijn, Mitra Baratchi



Universiteit
Leiden
The Netherlands

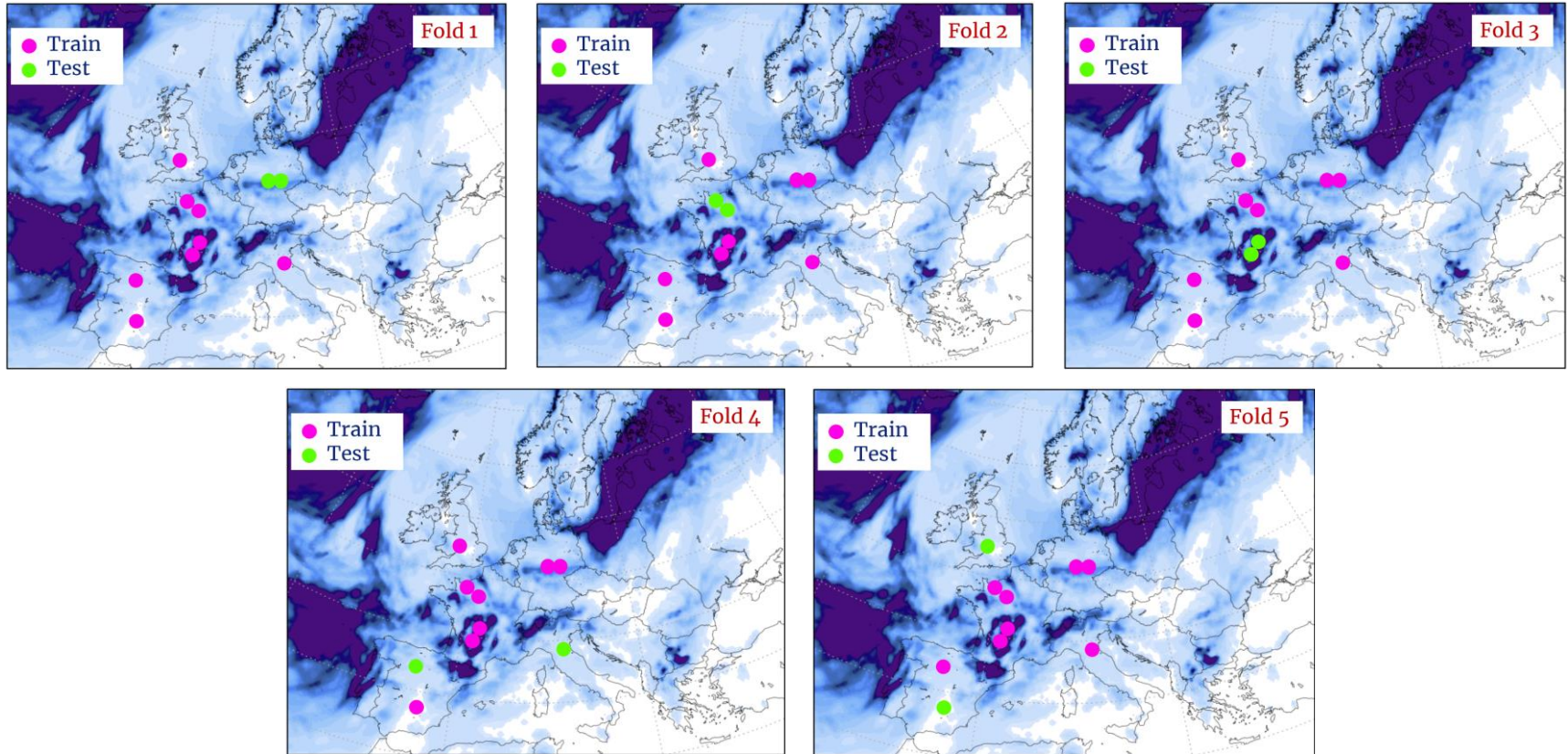
Discover the world at Leiden University

Spatial autocorrelation leads to performance inflation



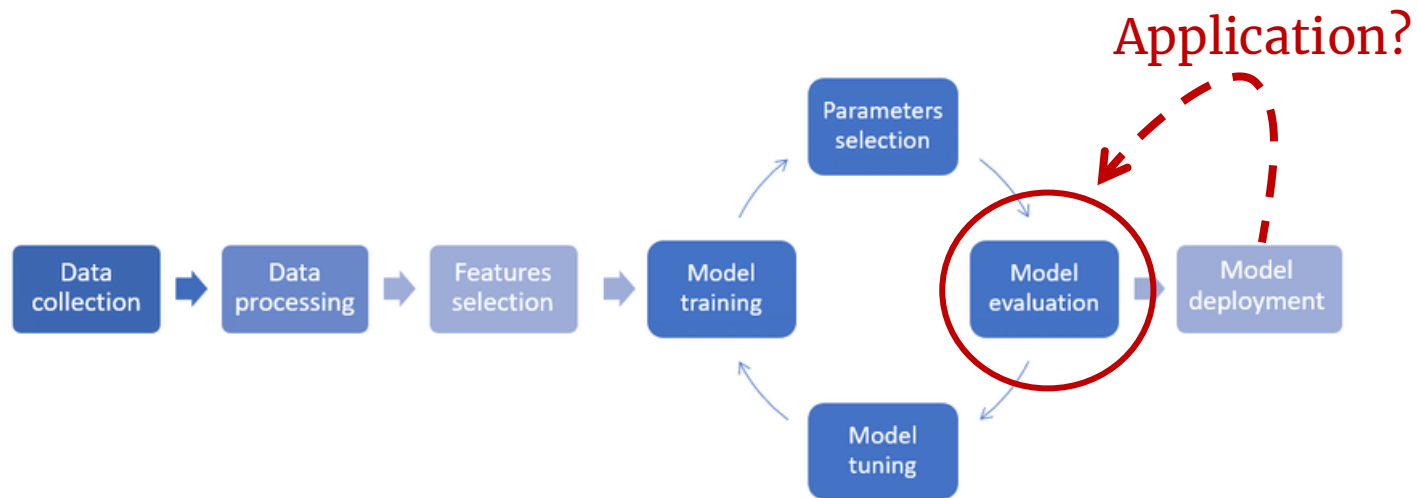
Up to ~~40~~% overoptimistic

Mitigating information leakage through spatial cross-validation

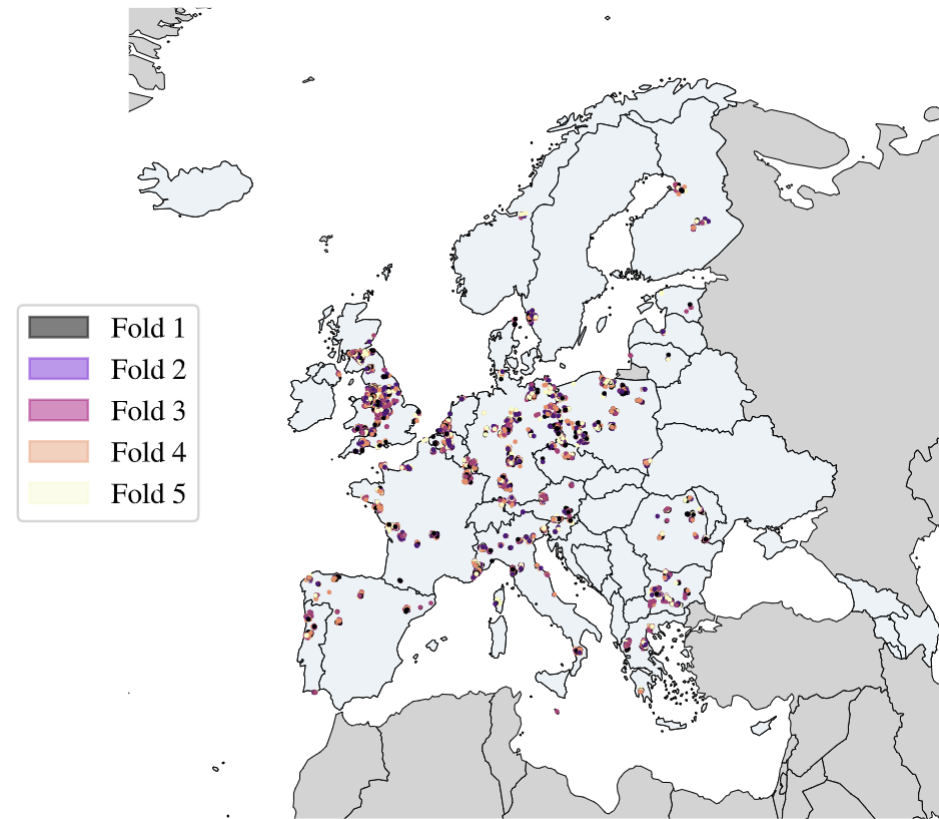
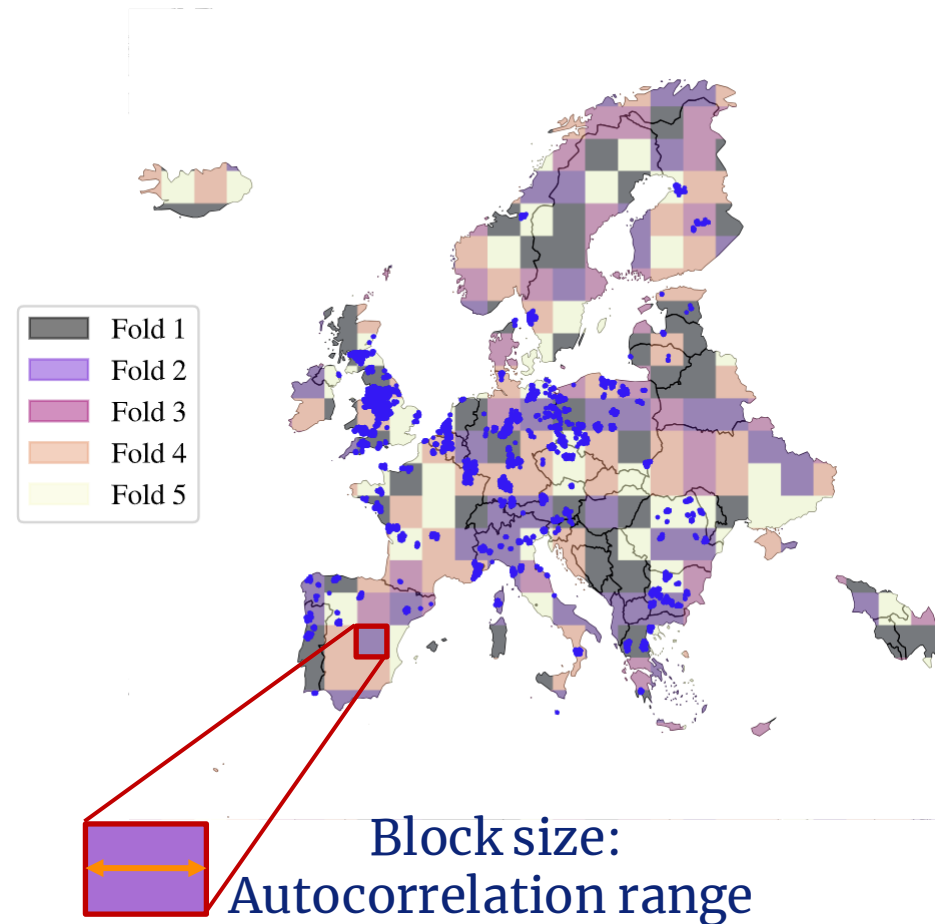


Spatial cross-validation not in general ML pipeline

- No generalised way of integrating SAC mitigation into the ML pipeline
- Unknown effects of information leakage on remote sensing image classification benchmark datasets
- ML pipeline currently assumes interpolation, but extrapolation is often taking place

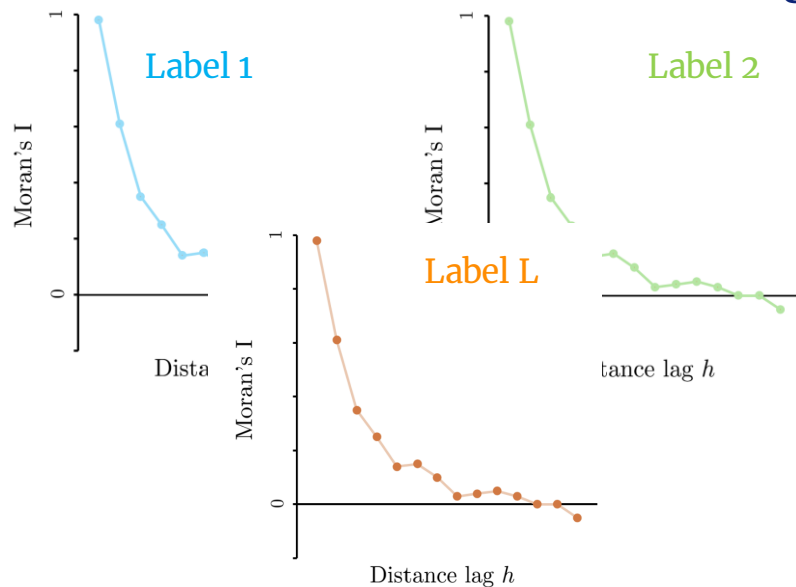


Blocked Spatial Cross-Validation (BSCV) VS. Random Cross-Validation (RCV)

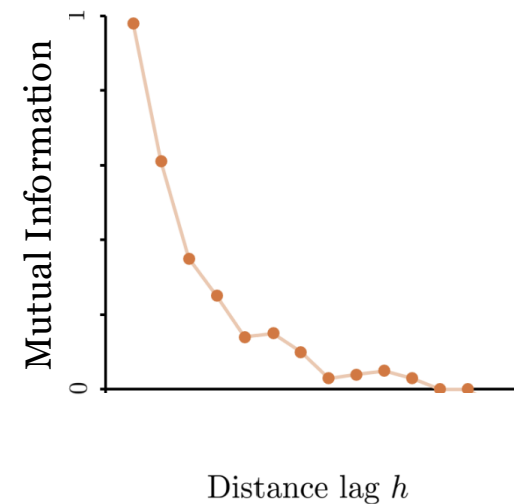


Generalised way of determining autocorrelation range

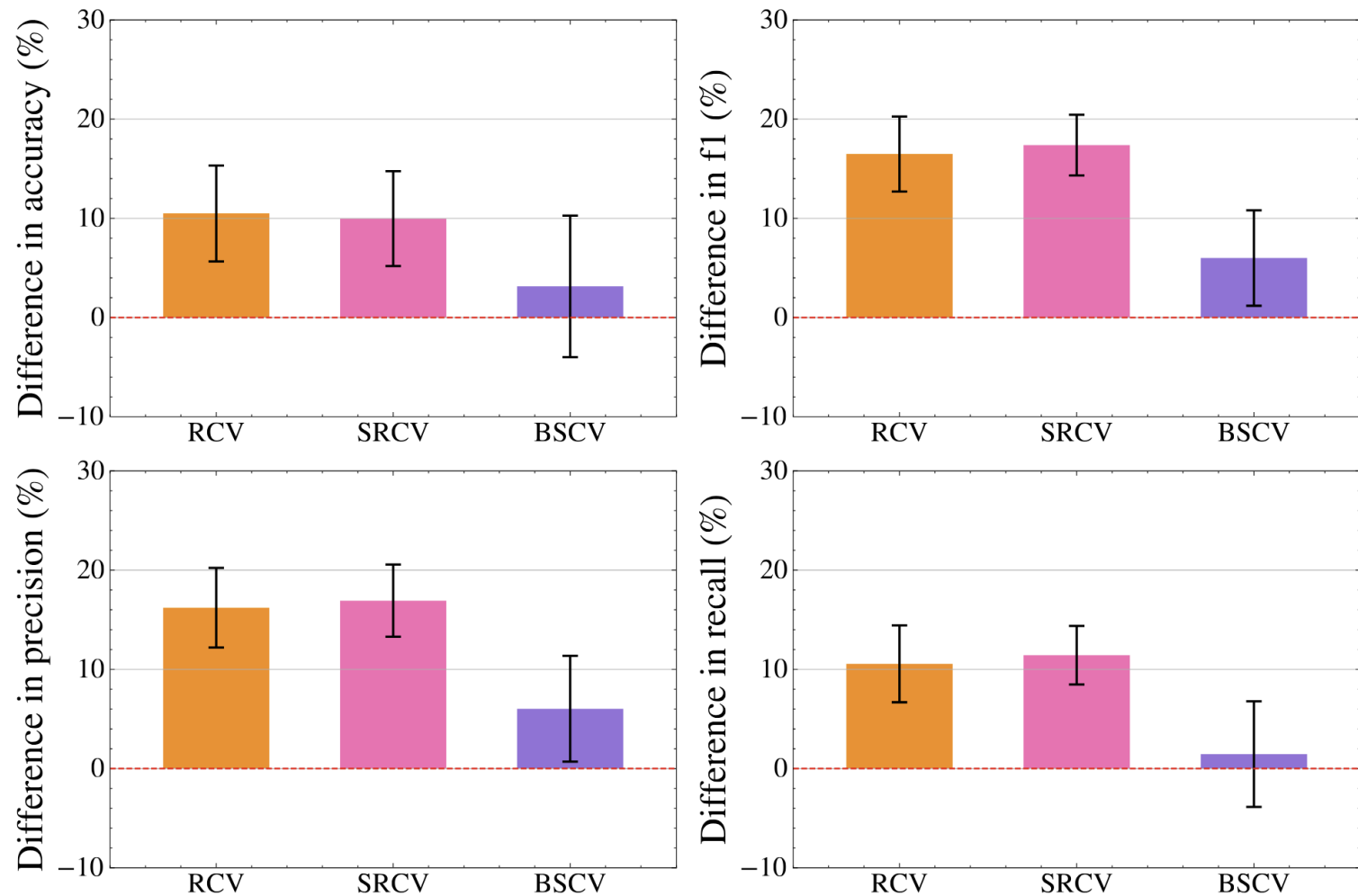
Global Moran's I indicator correlogram



Entropogram



BSCV accurate for interpolation, RCV overoptimistic



BrazilDAM (interpolation)



Closest to true generalisability R_{repr}

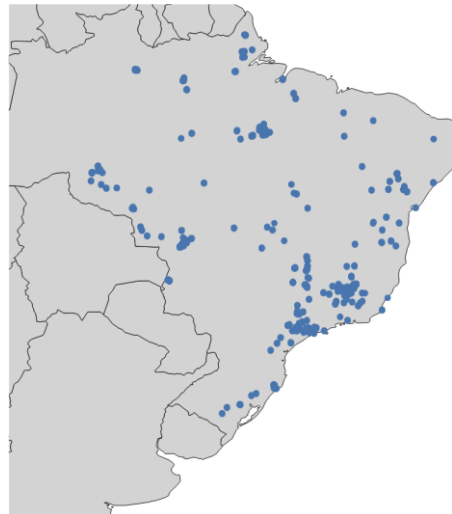


Suitable performance evaluation



Optimistic performance evaluation

Dataset	Method	Accuracy $\uparrow$	F1 $\uparrow$	Precision $\uparrow$	Recall $\uparrow$
BrazilDAM	R_{repr}	80.40 $\pm$ 9.98*	79.52 $\pm$ 11.04*	82.69 $\pm$ 7.40*	80.59 $\pm$ 8.83*
	RCV	81.48 $\pm$ 7.36*	80.41 $\pm$ 8.80*	82.95 $\pm$ 5.61*	81.02 $\pm$ 7.00*
	BSCV	80.60 $\pm$ 9.21*	79.80 $\pm$ 11.04*	81.70 $\pm$ 8.35*	80.27 $\pm$ 8.92*



So2SAT (extrapolation)



Closest to true generalisability R_{repr}

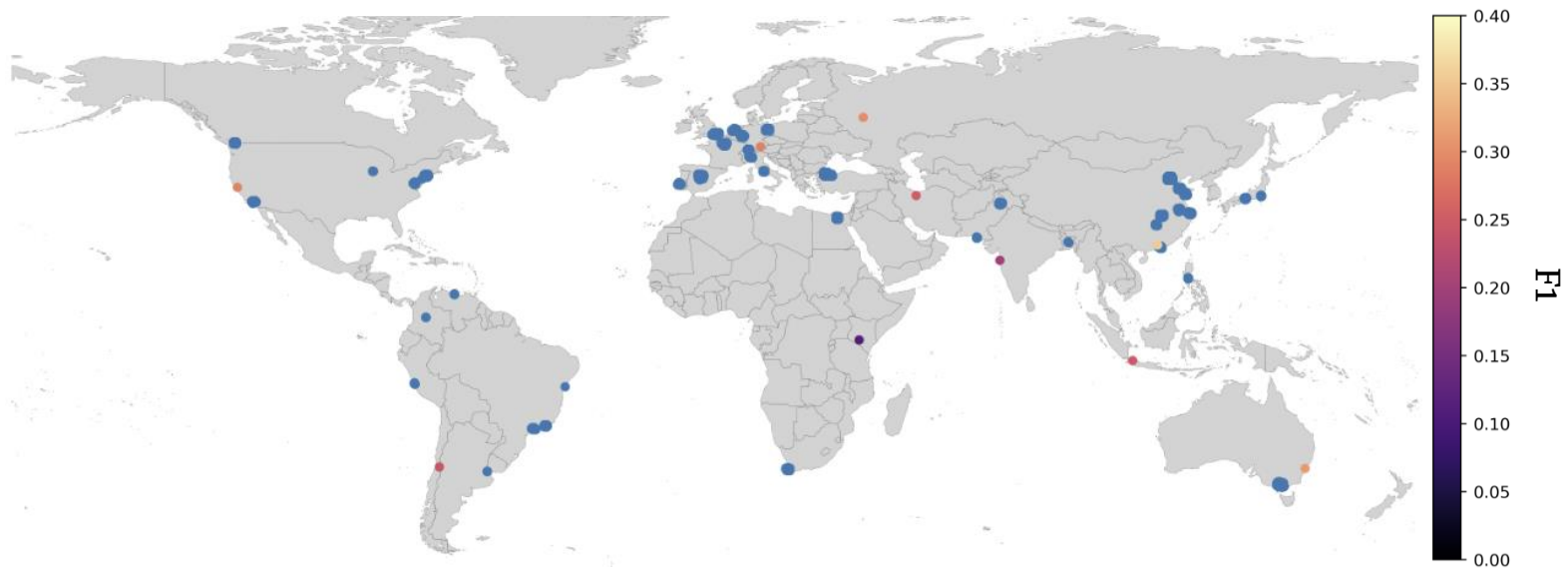


Suitable performance evaluation



Optimistic performance evaluation

Dataset	Method	Accuracy $\uparrow$	F1 $\uparrow$	Precision $\uparrow$	Recall $\uparrow$
So2SAT	R_{repr}	47.12 $\pm$ 0.48*	33.49 $\pm$ 0.10*	35.83 $\pm$ 0.12*	34.05 $\pm$ 0.14*
	RCV	78.52 $\pm$ 0.50*	71.81 $\pm$ 0.73*	73.42 $\pm$ 0.84*	70.79 $\pm$ 1.10*
	BSCV	78.28 $\pm$ 0.75*	70.32 $\pm$ 1.57*	71.60 $\pm$ 2.44*	70.20 $\pm$ 0.98*



BSCV is more suitable for computing performance estimates

- Blocked spatial cross-validation outperforms random cross-validation in both **interpolation** and **extrapolation scenarios**, but no strategy is able to represent extrapolation properly
- Blocked spatial cross-validation is **better** for **EuroSAT**, **BrazilDAM** and **So2SAT**

Future Directions

- Automated machine learning
- Foundation models
- Quantifying degree of extrapolation