ECML PKDD
Dublin, Ireland
10-14 SEPT 2018

Evaluation procedures for forecasting with spatio-temporal data

MARIANA OLIVEIRA[1,3], LUÍS TORGO[1,2,3] AND VÍTOR SANTOS COSTA[1,2]

14 SEPTEMBER 2018

1 U.PORTO FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO

2 INESCTEC TECHNOLOGY & SCIENCE | ASSOCIATE LABORATORY PORTUGAL

3 DALHOUSIE UNIVERSITY

Motivation

Motivation

Fig 1: Remote monitoring equipment
(Source: NDSU)

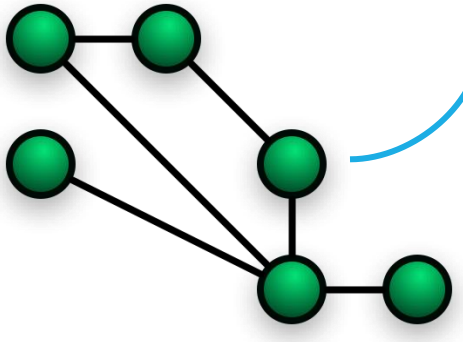


Fig 2: Sensor network
(Source: Wikimedia Commons)

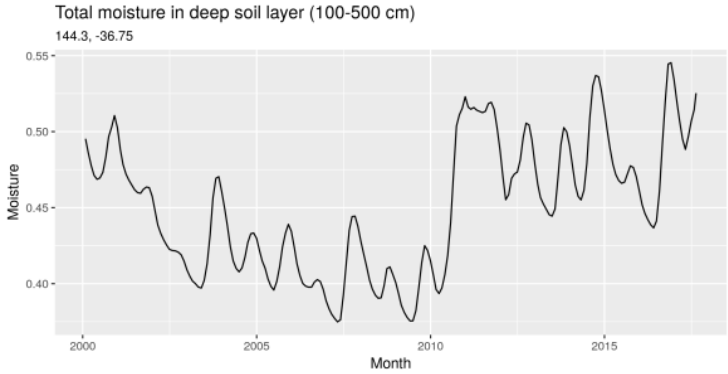
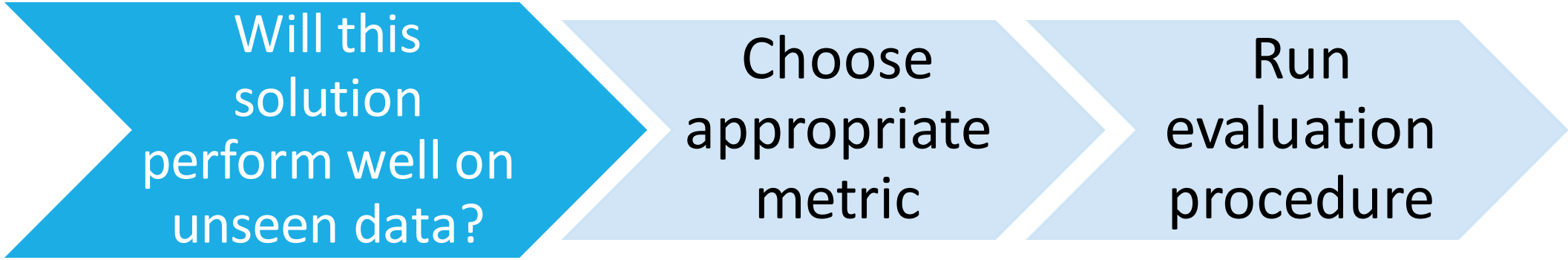
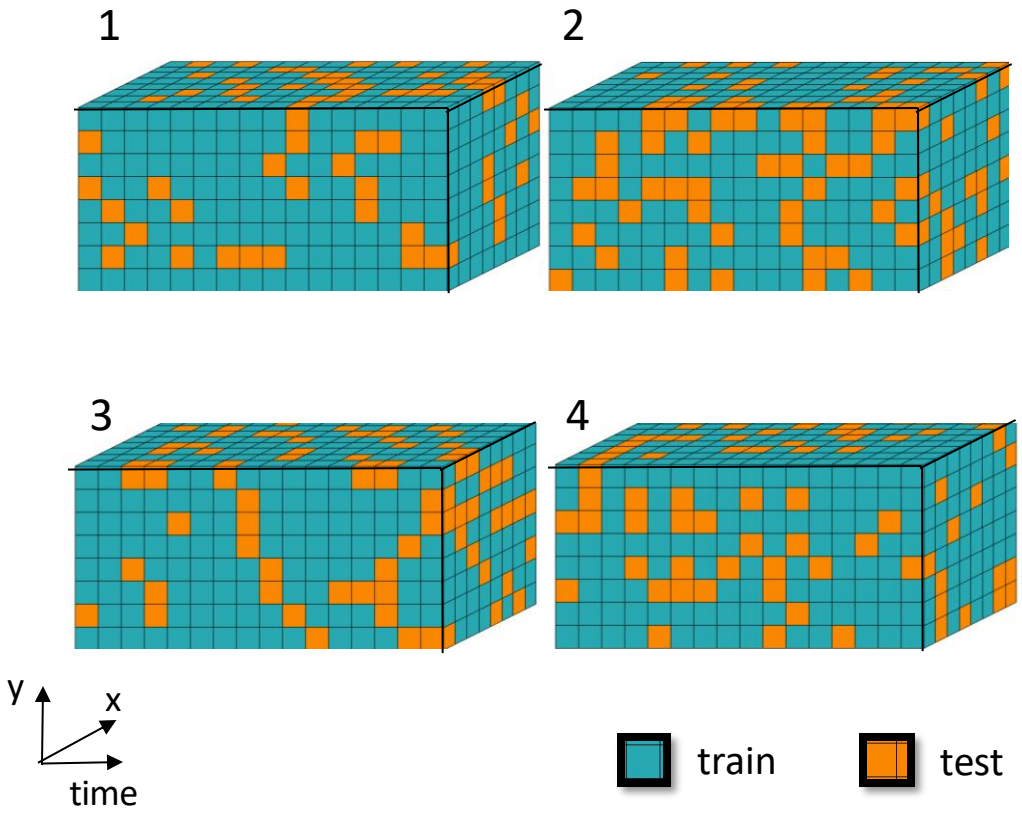


Fig 3: Soil moisture time series
(Source: The Lucid Manager)

Motivation


Will this solution perform well on unseen data?
Choose appropriate metric
Run evaluation procedure

Motivation



Implicit data dependence

+

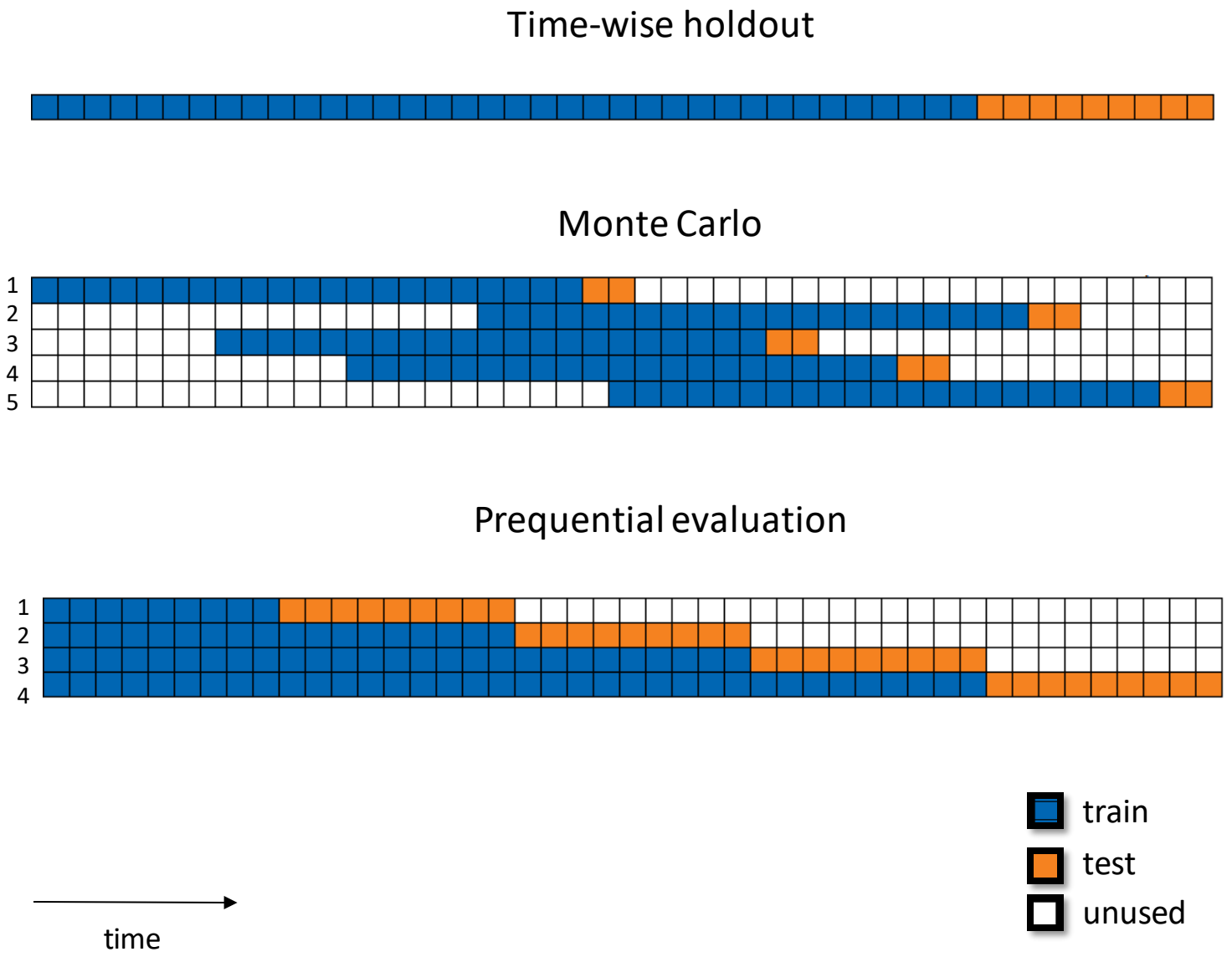
Assumption that data is i.i.d.

→

Evaluation breakdown

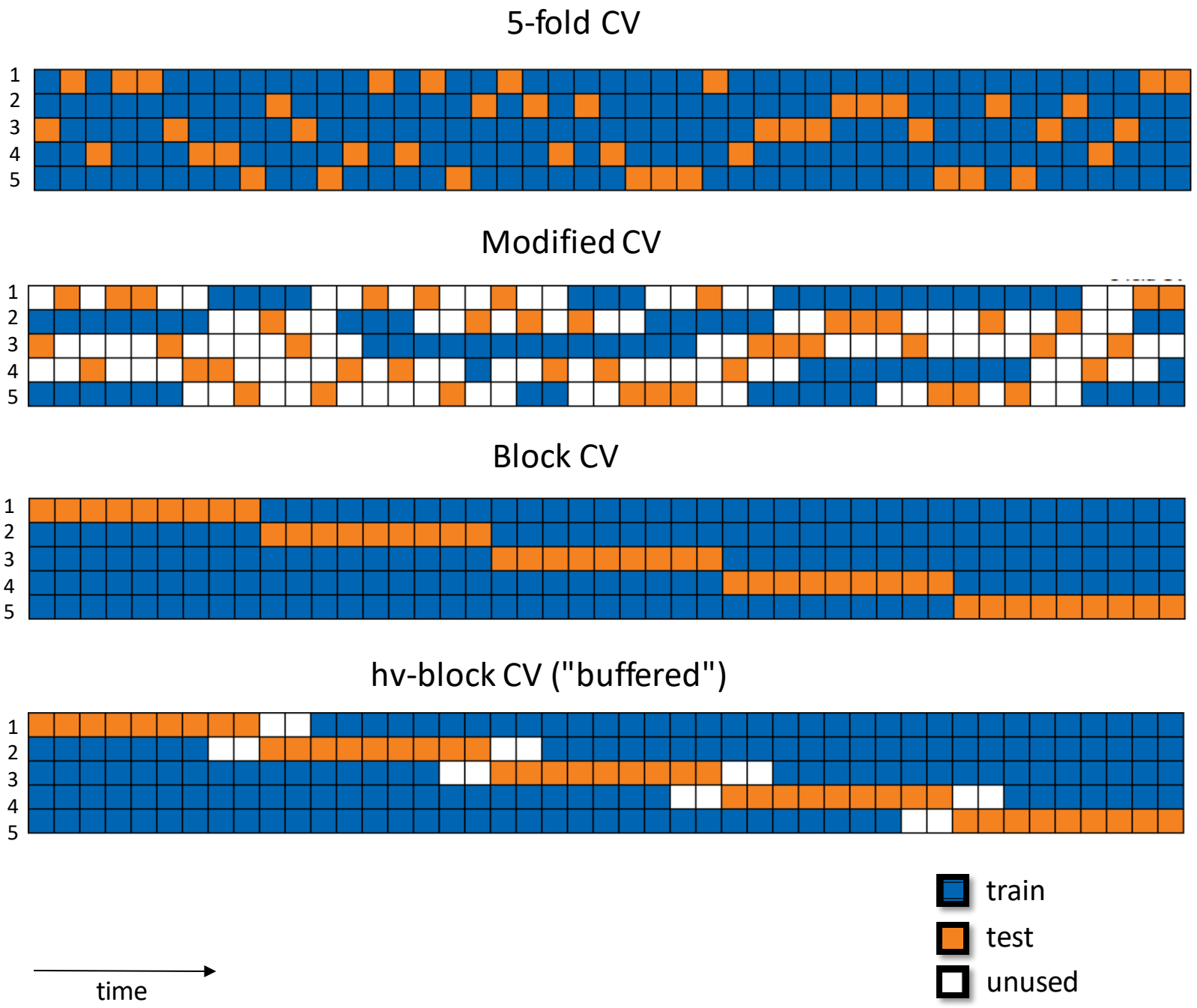
Background

OOS procedures for time series



CV procedures for time series

- Block CV works better than CV, especially for small stationary series (Bergmeir et al., 2012)
- In real-world scenarios, OOS methods perform better than CV (Cerqueira et al., 2017)



Bergmeir, Christoph, and José M. Benítez. "On the use of cross-validation for time series predictor evaluation." *Inf Sci* 191 (2012): 192-213.

Cerqueira, Vitor, et al. "A comparative study of performance estimation methods for time series forecasting." *DSAA, 2017.*

Cross-validation for spatial data

- Block CV and buffered LOO CV are recommended by Roberts et al. (2017)

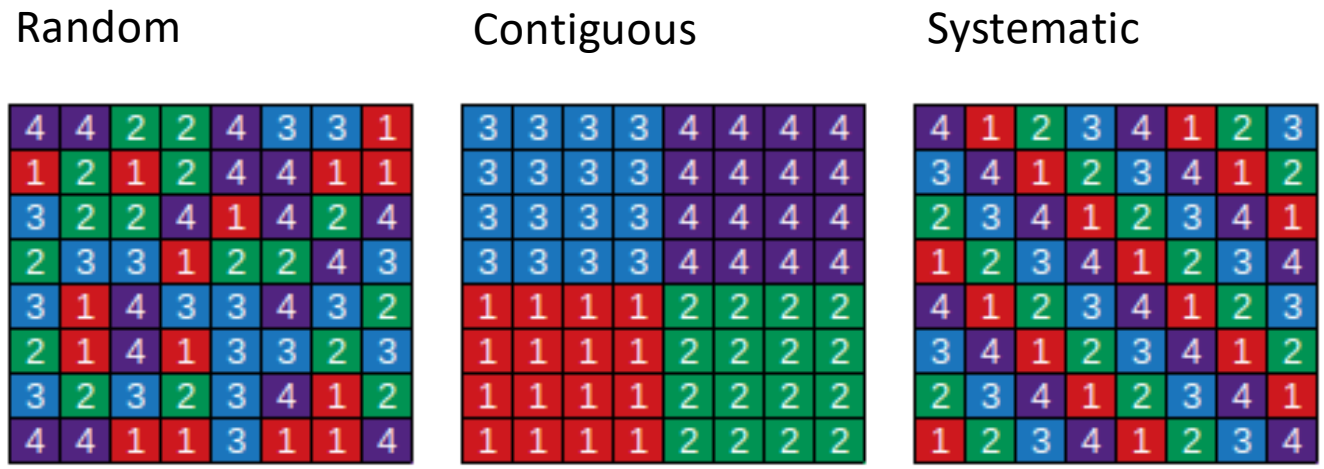


Fig: Spatial cross-validation.
(Adapted from Roberts, David R., et al. "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure." *Ecography* 40.8 (2017): 913-929.)

What about geo-referenced time series?

Most treat data as if It was spatial-only or temporal-only

Recent work by Meyer et al. (2018) shows how results using "target-oriented" CV (LLO, LTO and LLTO) differ from conventional CV in spatio-temporal interpolation problems

Our focus is on **forecasting** – making predictions about the **future** –, not interpolation.

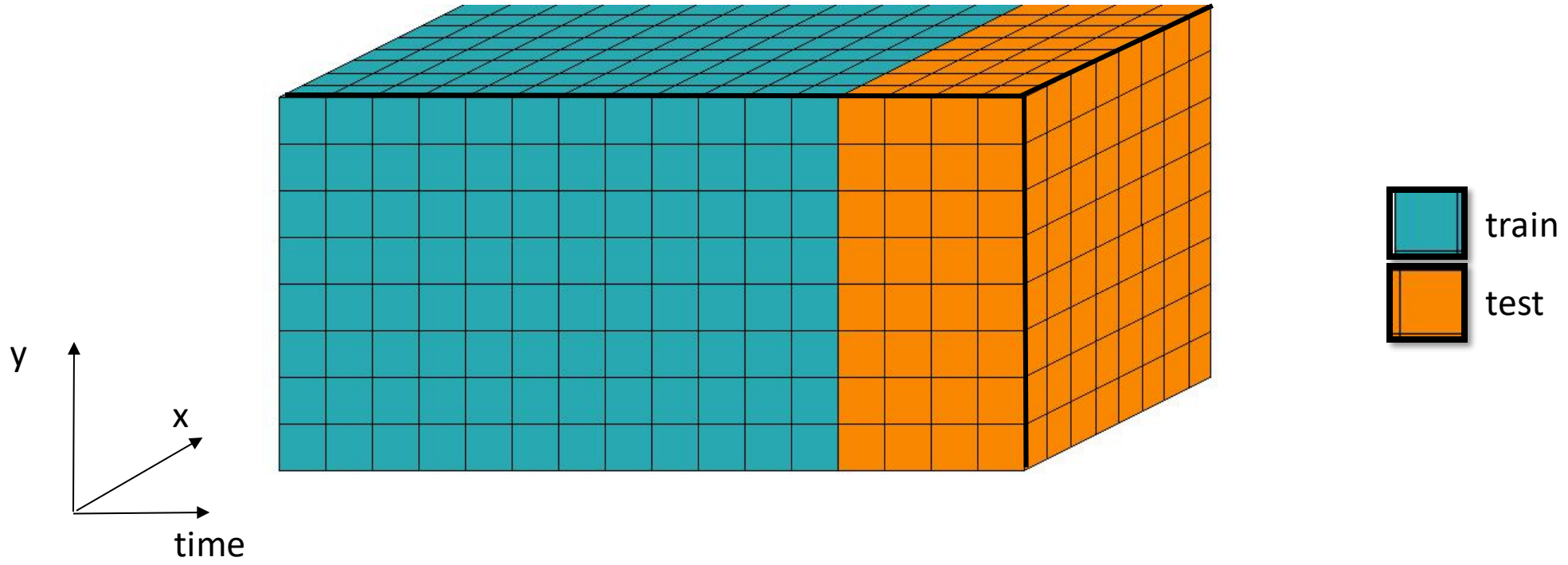
- The best evaluation procedure to make predictions about unseen locations might not be the same as when the aim is to predict in known sites!

Meyer, Hanna, et al. "Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation." *Environmental Modelling & Software*101 (2018): 1-9.

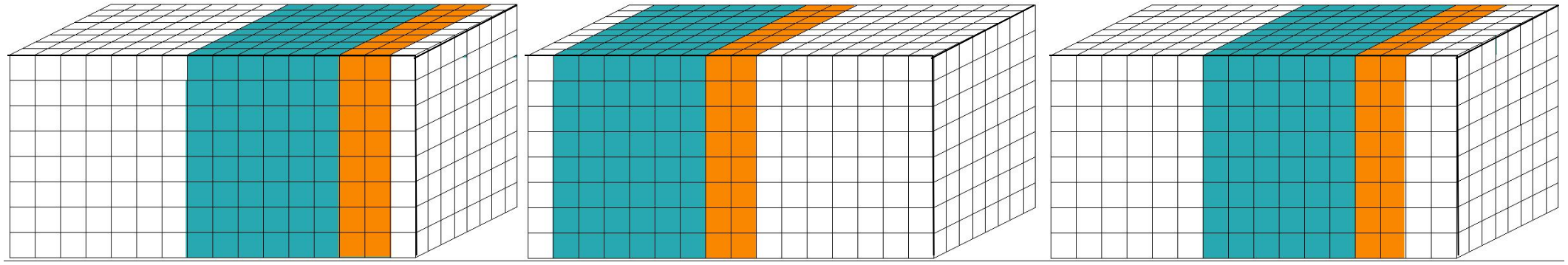
Estimation procedures

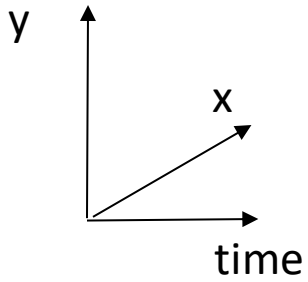
OUT-OF-SAMPLE

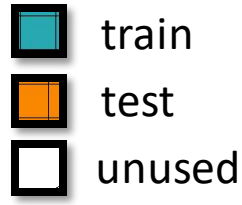
Holdout (H)



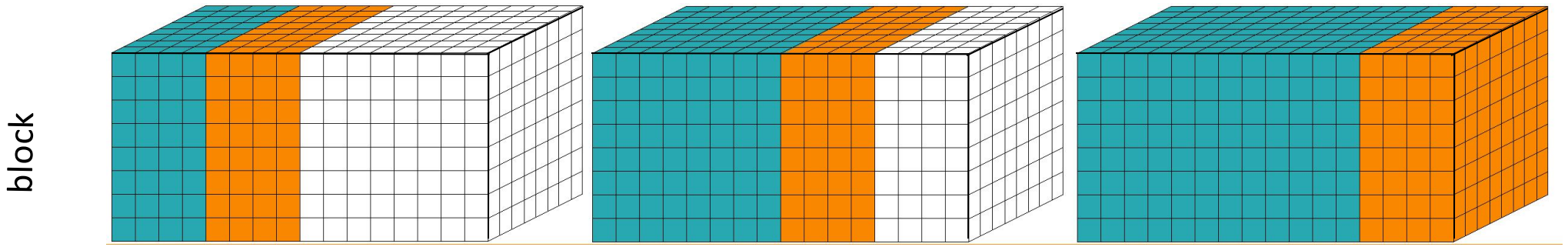
Monte Carlo (MC)



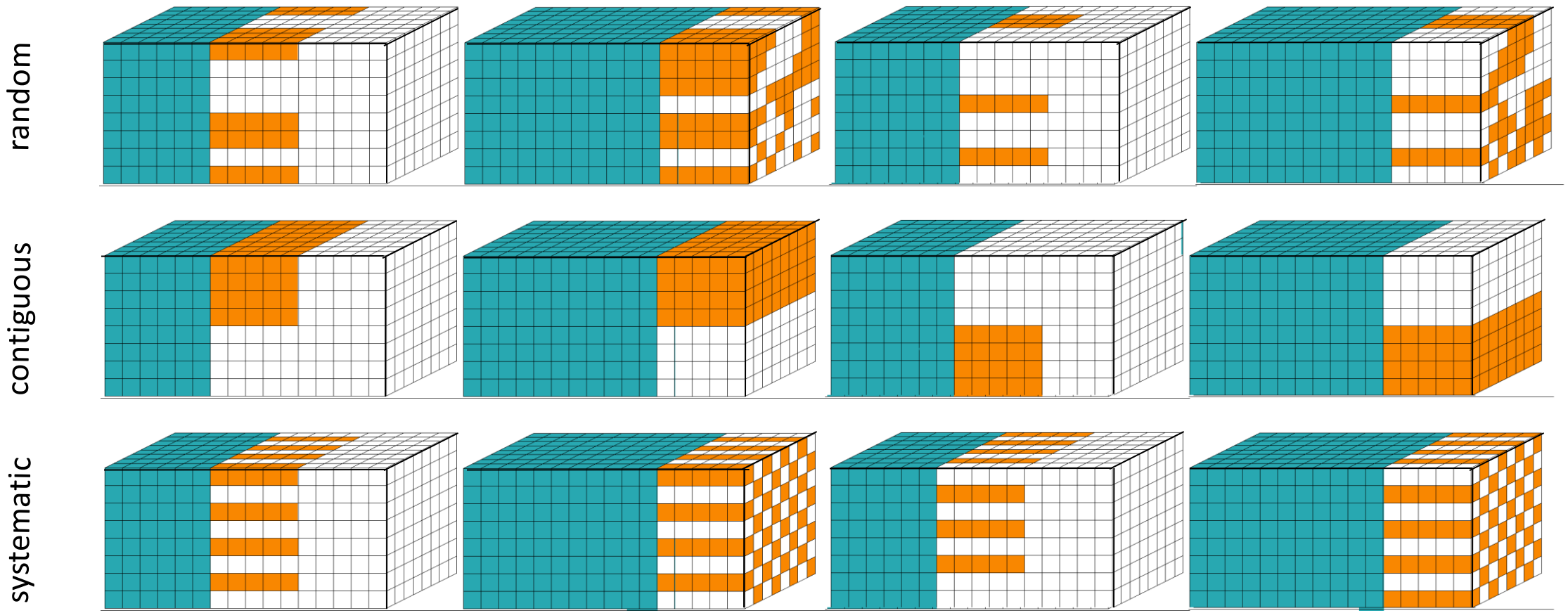




Prequential evaluation (P)



Spatio-temporal prequential

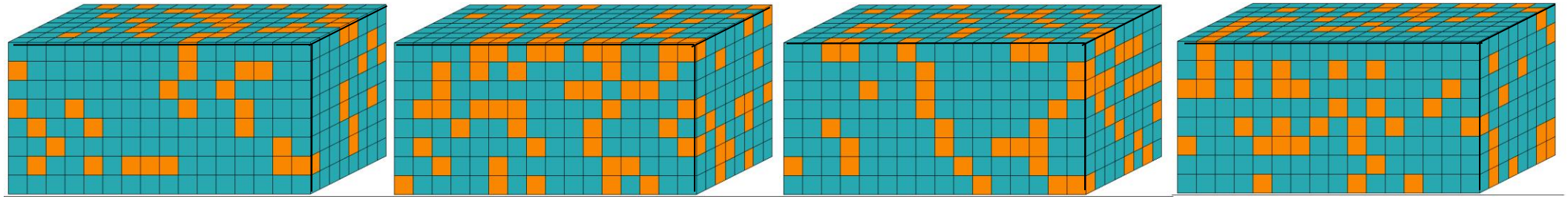


Estimation procedures

CROSS-VALIDATION

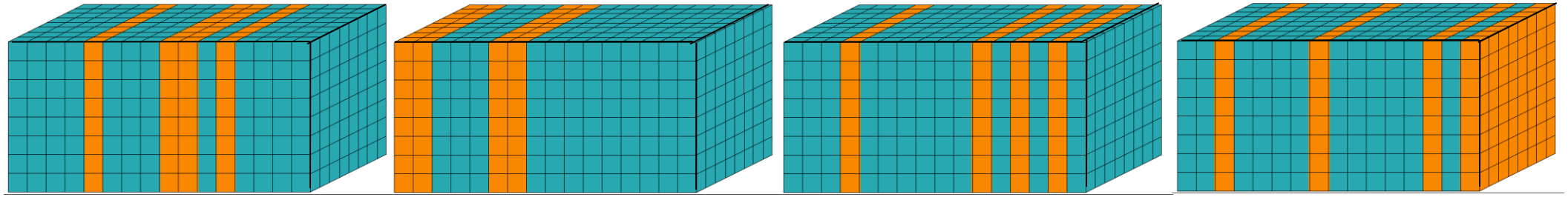
Cross-validation (CV)

random

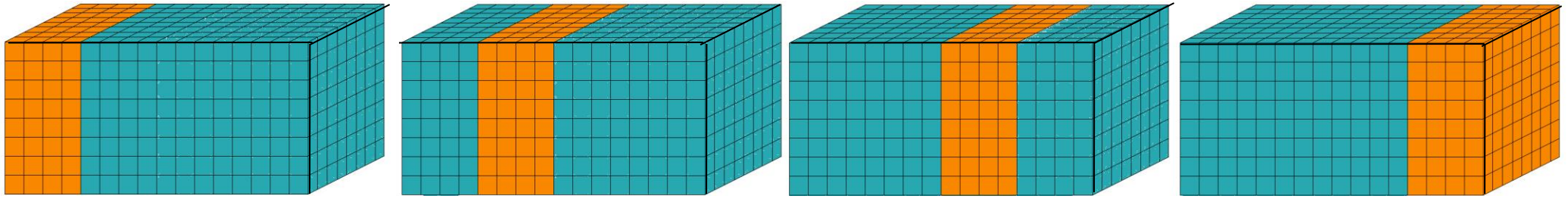


Temporal CV

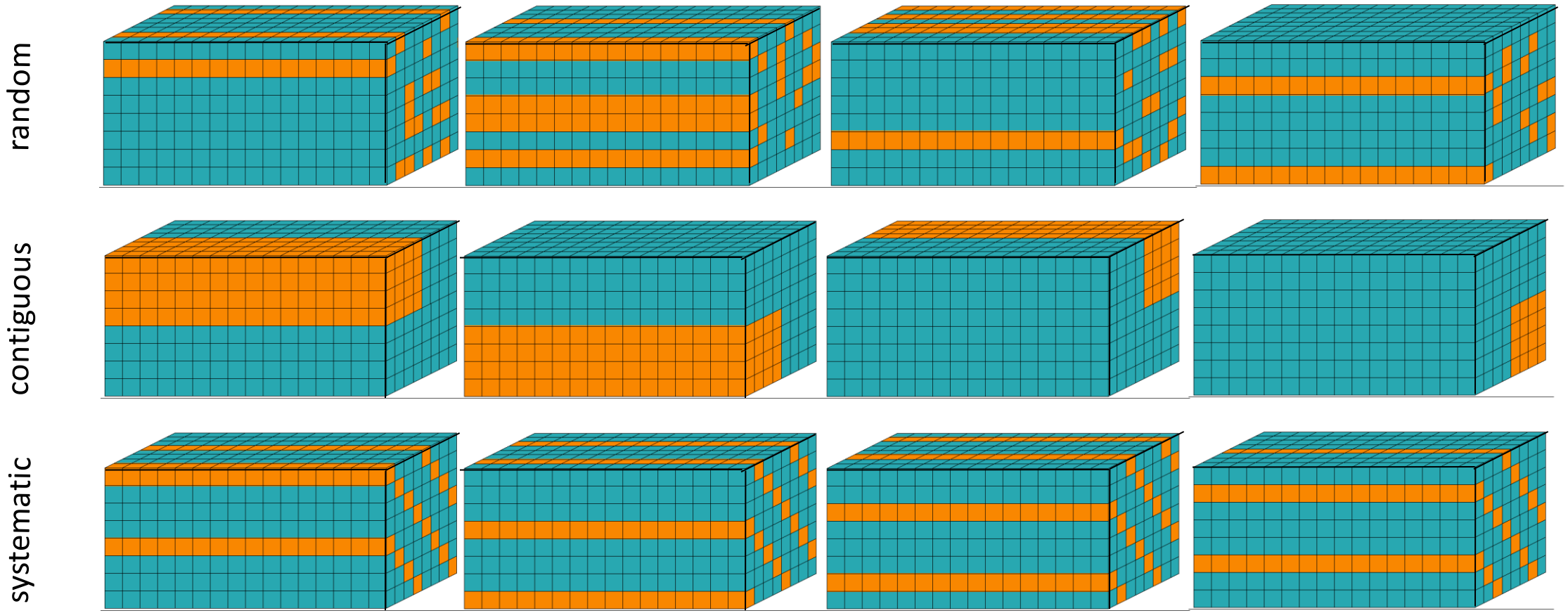
sliced



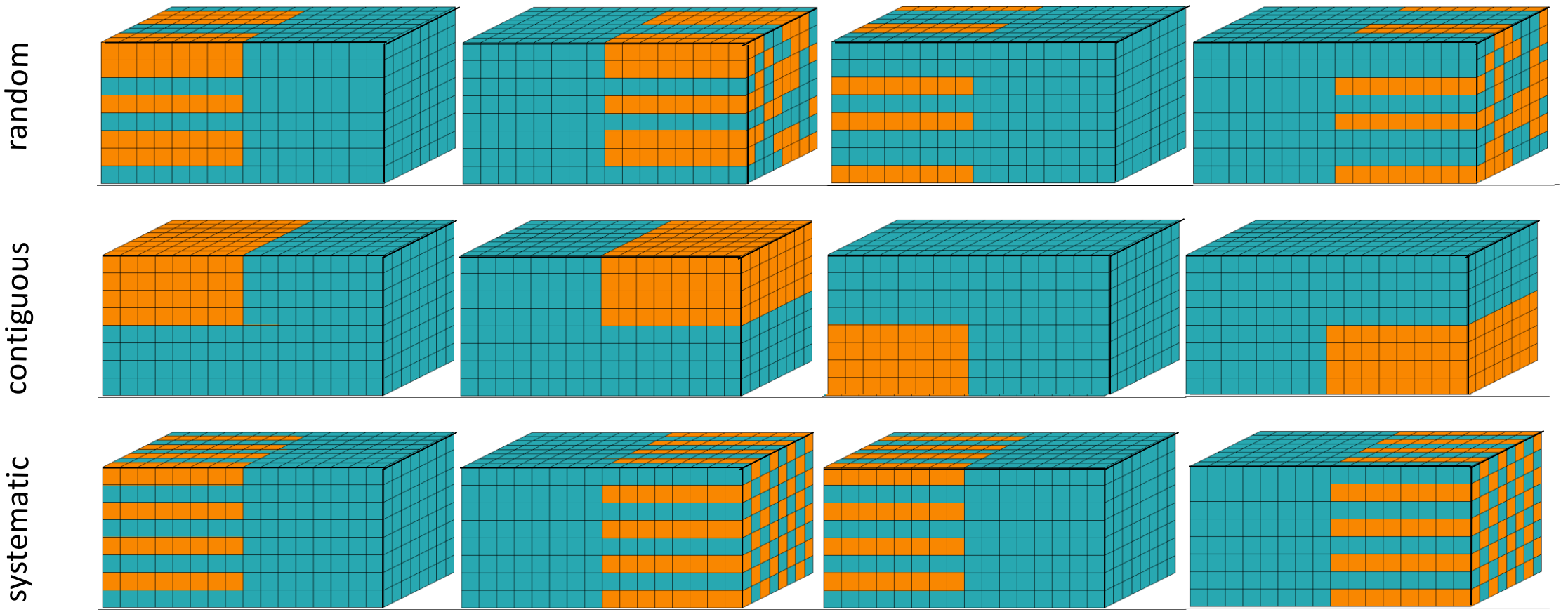
block



Spatial CV



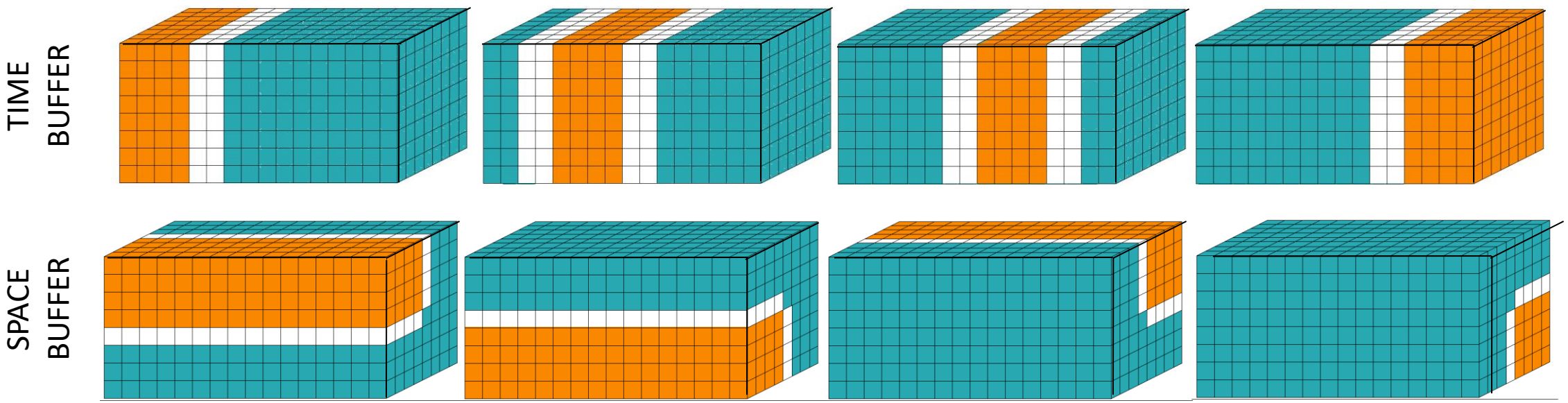
Spatio-temporal CV



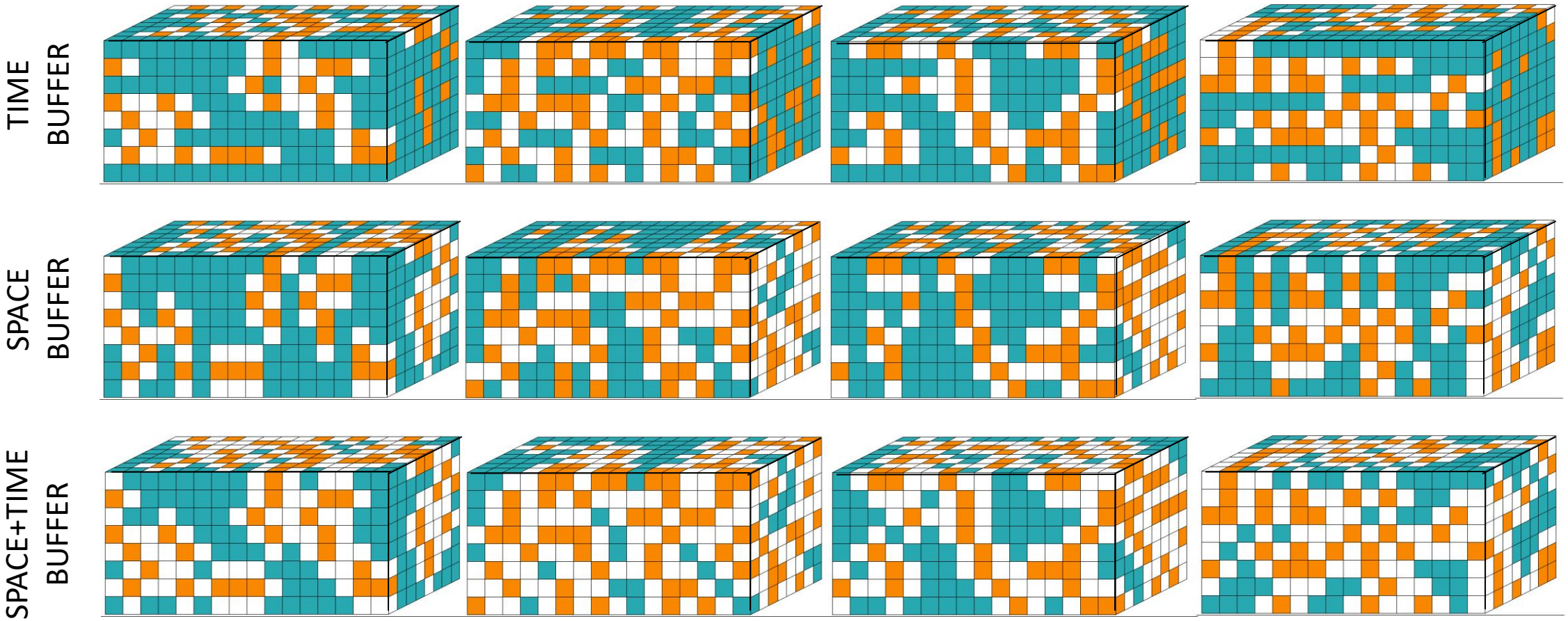
Estimation procedures

ADDING BUFFERS...

… to block CV

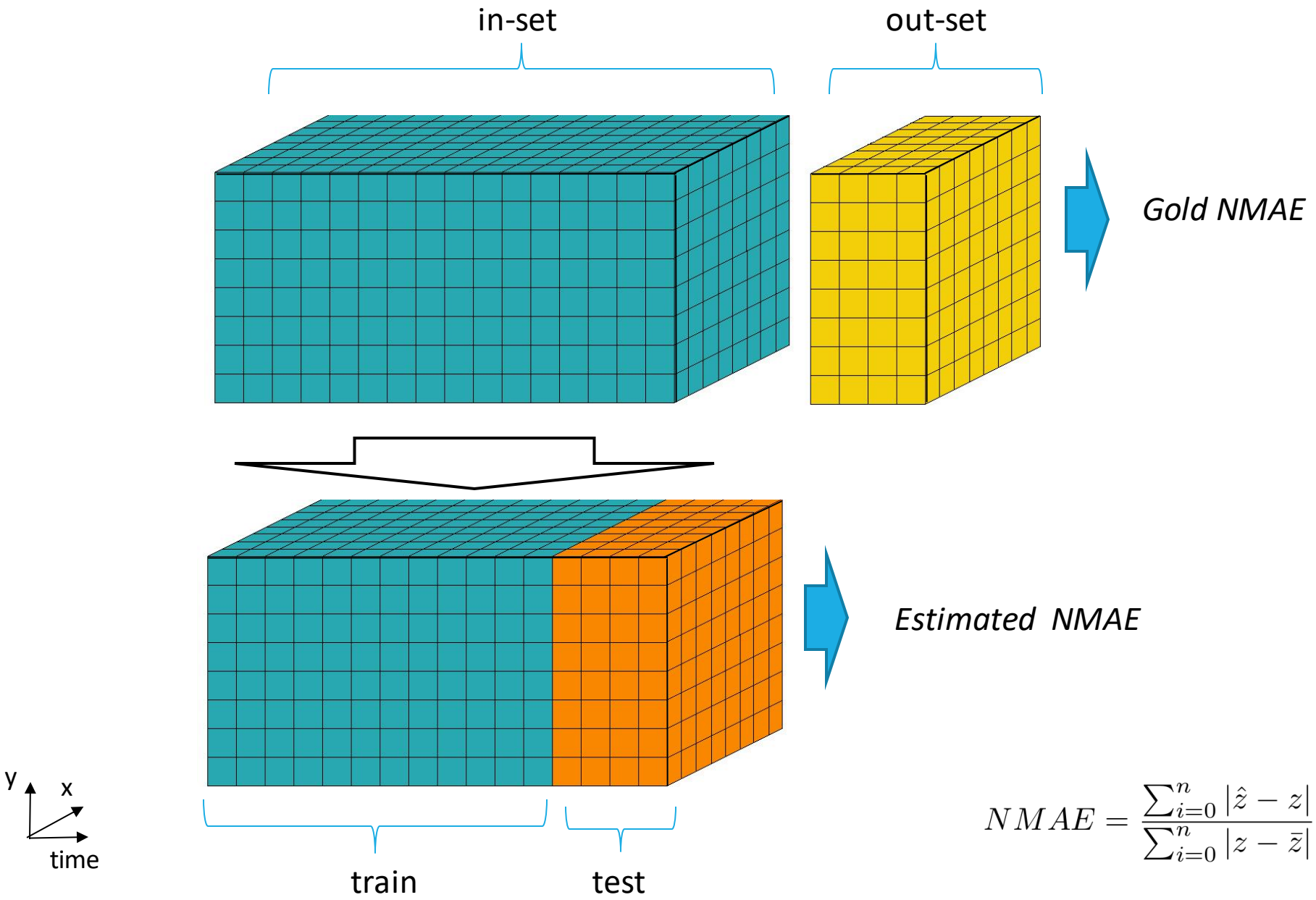


… to random CV



Experimental design

Experimental design



$$NMAE = \frac{\sum_{i=0}^{n} |\hat{z} - z|}{\sum_{i=0}^{n} |z - \bar{z}|}$$

Datasets

ARTIFICIAL

STARMA, STMA, STAR and NL-STAR

Orders 2(10), 2(01) and 2(11)

8x8 and 20x20 regular grids

150 and 300 time points

---

96 datasets with embed 3(110)

REAL-WORLD

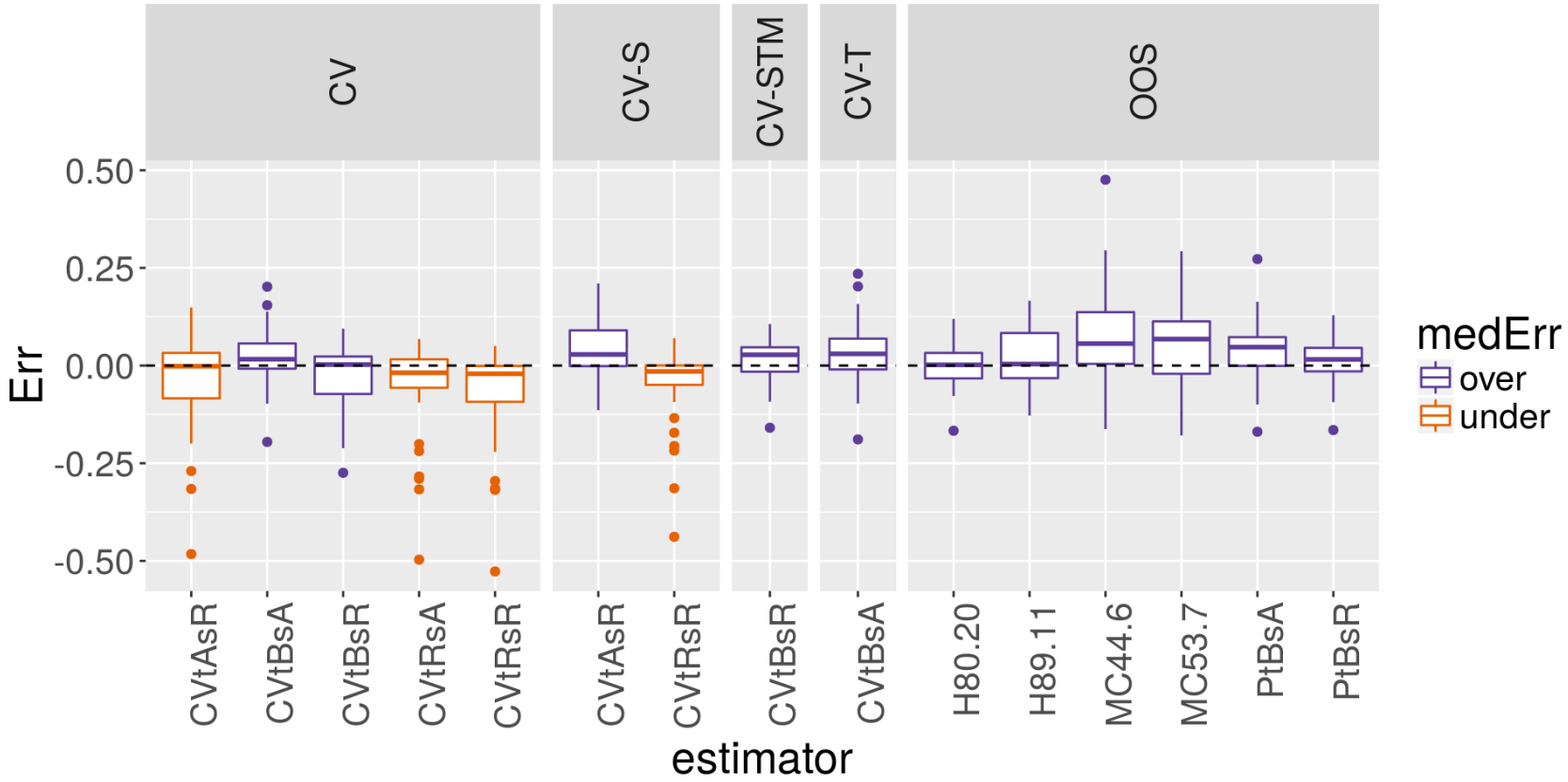
Data	#	# timeIDs	# locIDs	% avail.
MESA	1	280	20	100
NCDC	2	105	72	100
TCE	3	330	26	100
COOK	3	729	42	~73
SAC	1	144	900	100
Rural	1	4382	70	~49
Beijing air	6	11357	36	~40

---

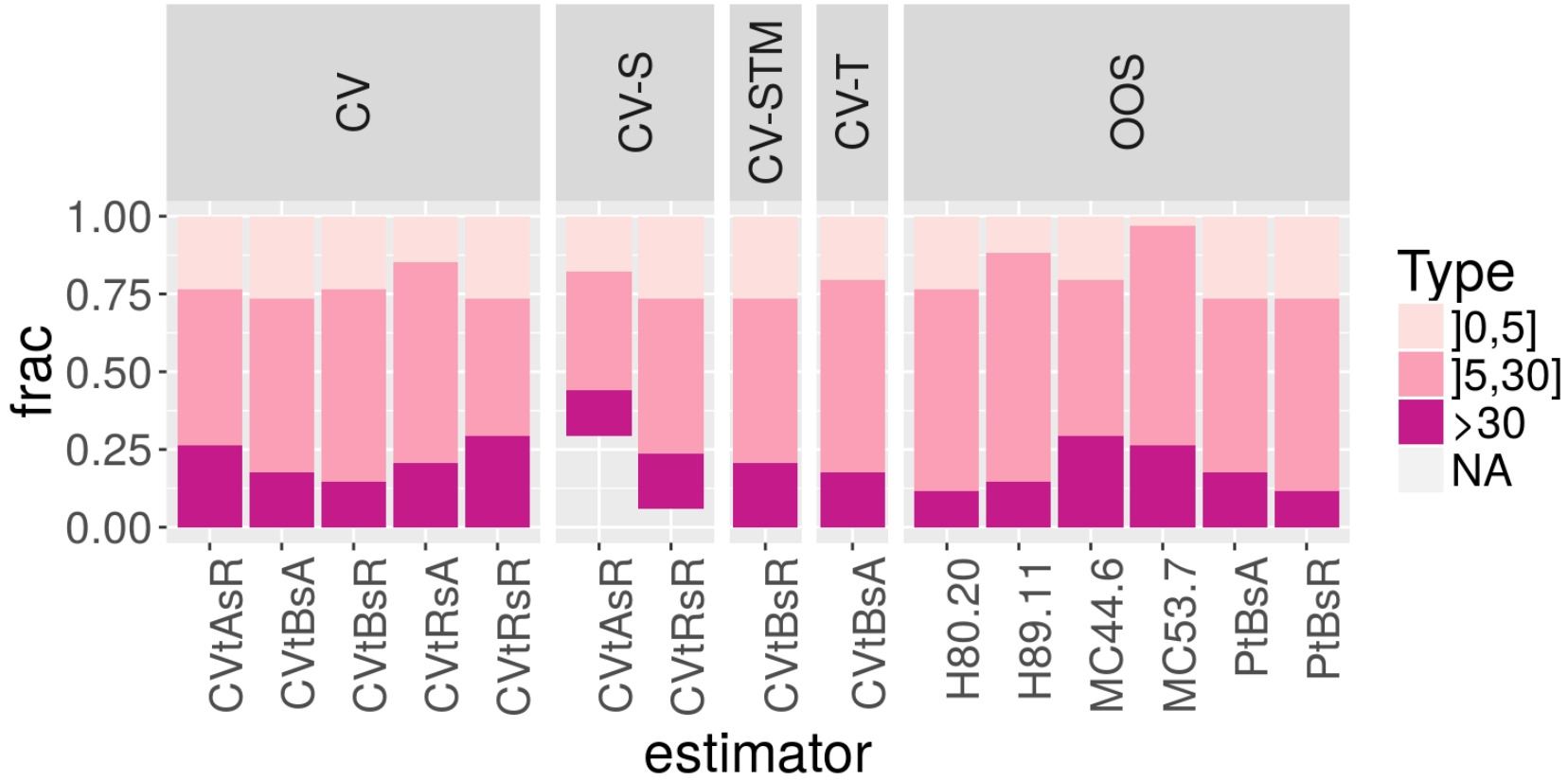
17 univariate datasets with spatio-temporal indicators

Real-world results

Median errors: *Est - Gold*

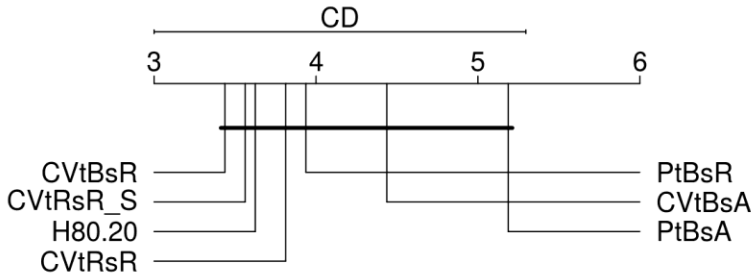


Relative errors: |*Est - Gold*|/ *Gold*

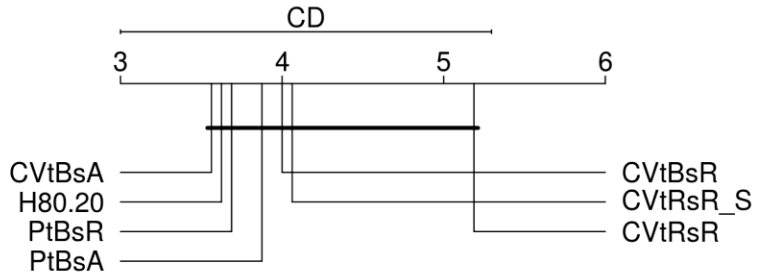


Absolute errors: |*Est-Gold*|

LINEAL MODEL



RANDOM FOREST



Conclusion

Conclusion

- Standard CV exhibits outliers of severe error underestimation;
- In real-world cases, spatio-temporal block and time block CV approximate the error better than other methods and avoid being overly optimistic;
- OOS procedures did not do as well, but they did avoid underestimation of the error in almost all real-world cases;
- Results seem to point to the temporal dimension being more important to respect during evaluation.

Thank you!

Code available at **http://bit.ly/STEvaluation**