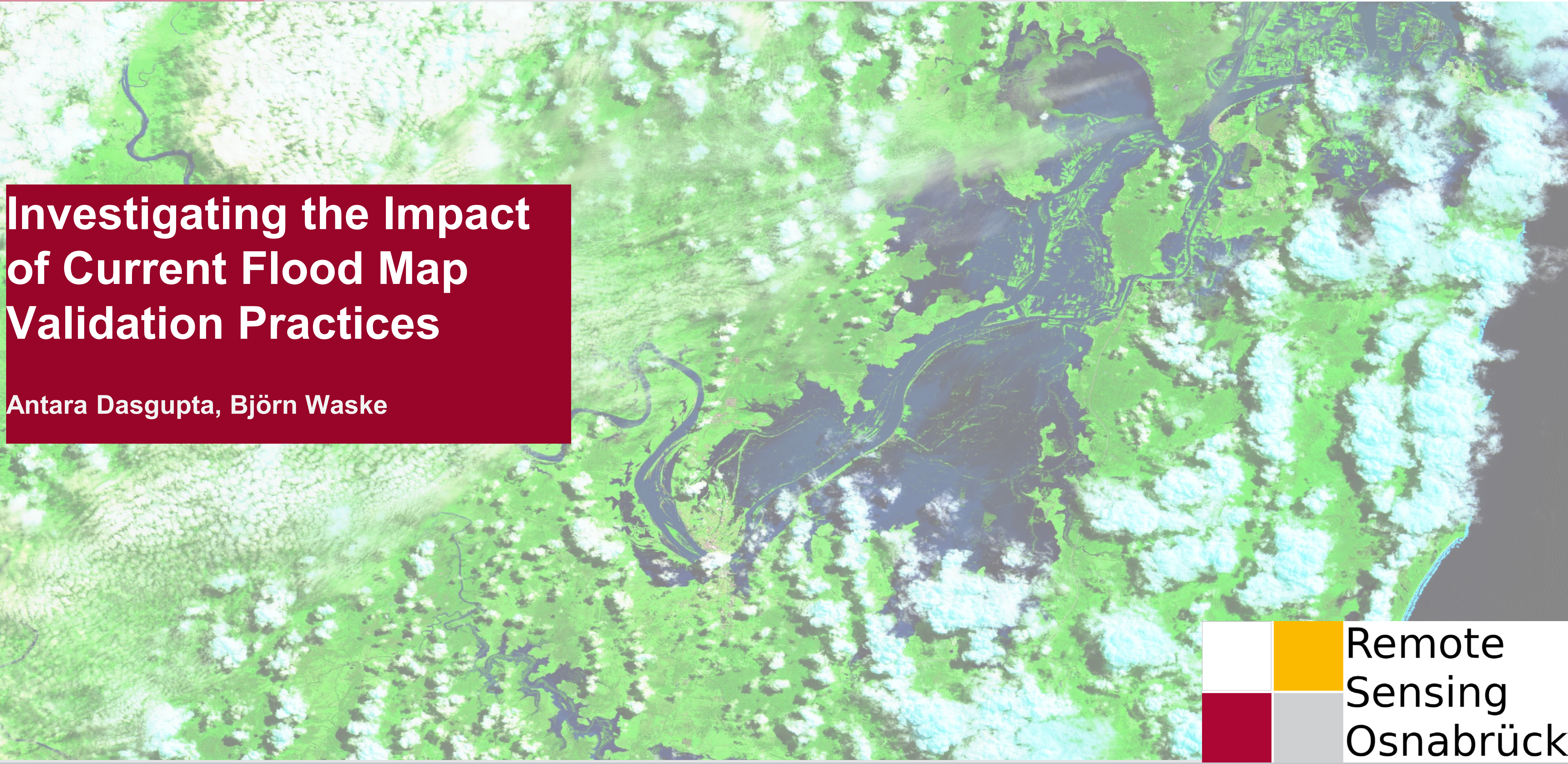
Investigating the Impact of Current Flood Map Validation Practices

Antara Dasgupta, Björn Waske

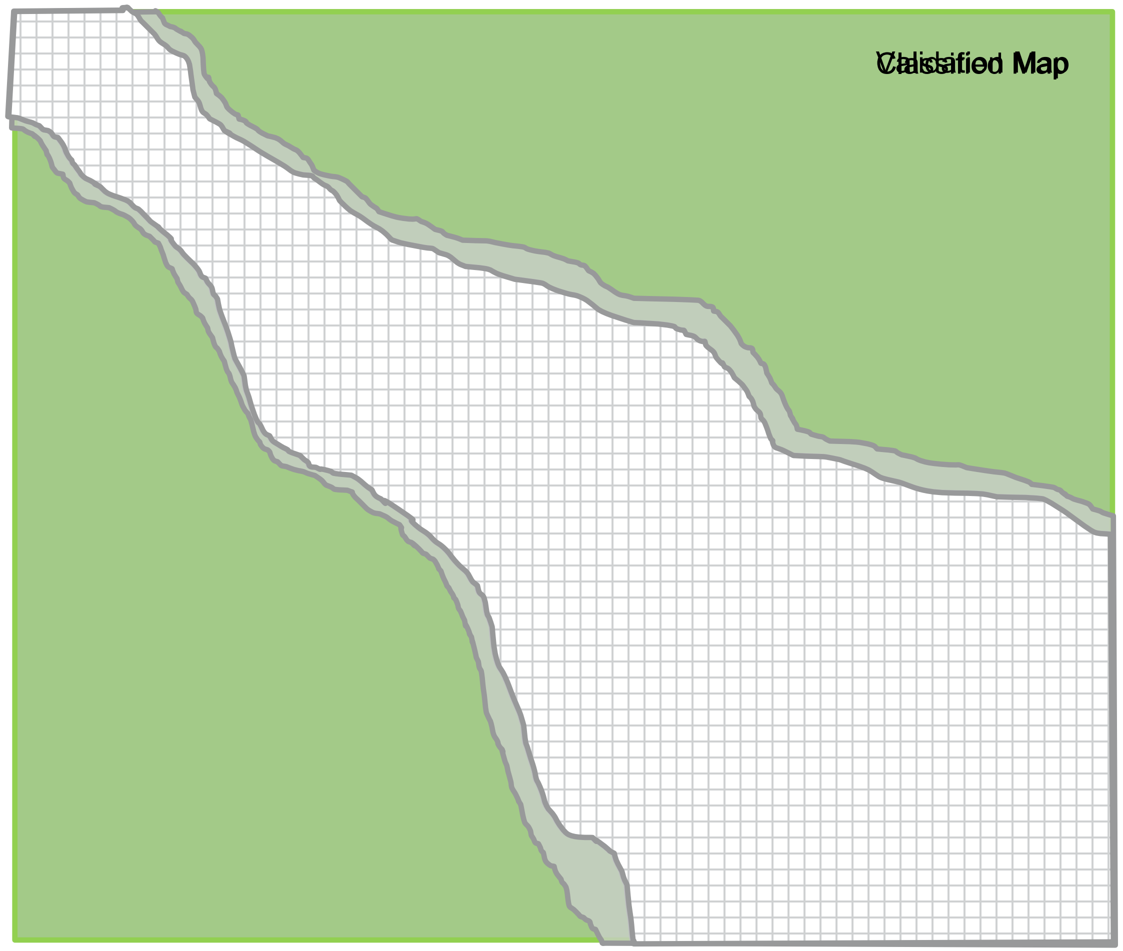


Remote Sensing Osnabrück

Metric sensitivity, sampling strategies, class imbalance

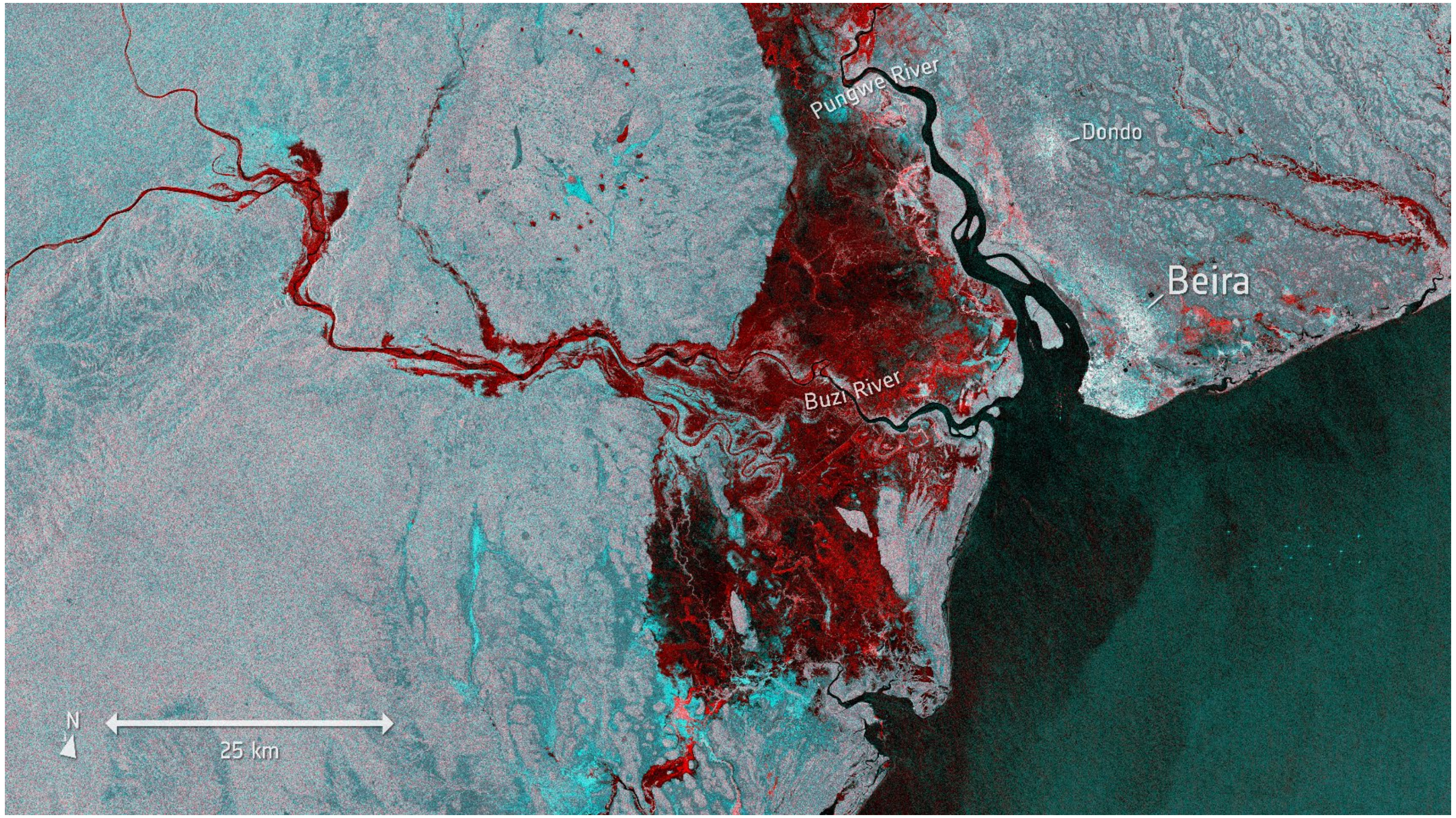
PROBLEMS WITH SATELLITE-BASED FLOOD MAP VALIDATION

Problem 1: Metric Sensitivity

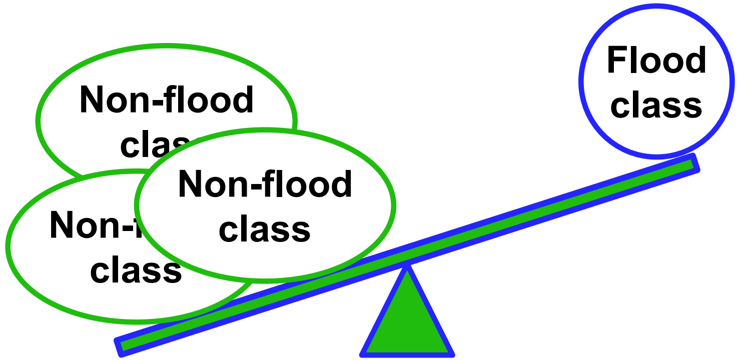


Common area > Map differences (mostly!)

Problem 2: Class Imbalance



Copernicus Sentinel-1 Image showing the flooding from Cyclone Idai in red, around the port town of Beira in Mozambique on 19 March 2019, provided by the Copernicus Emergency Mapping Service (CEMS).

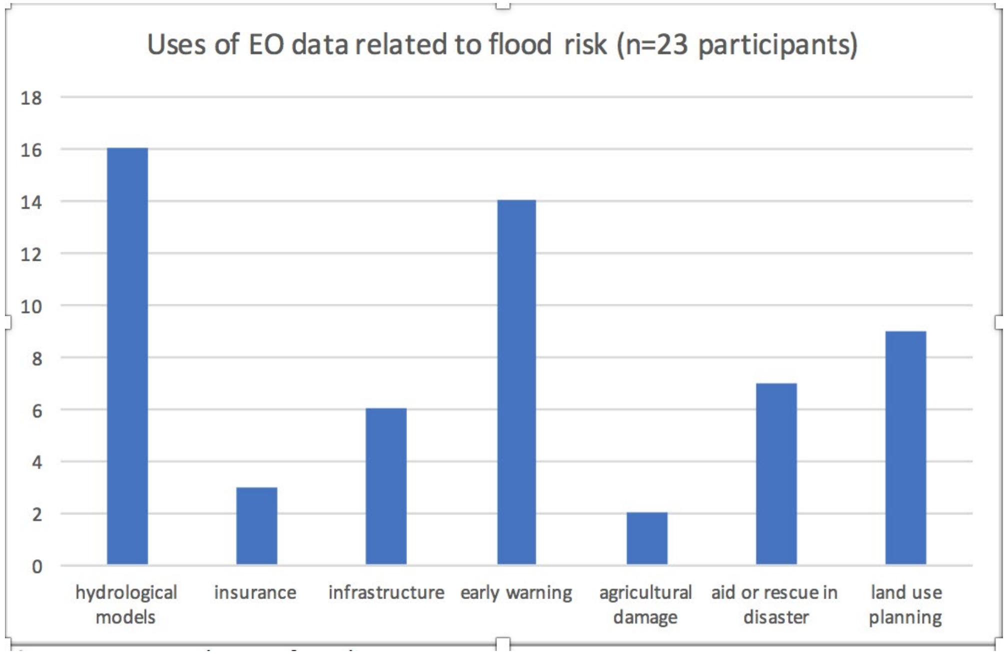


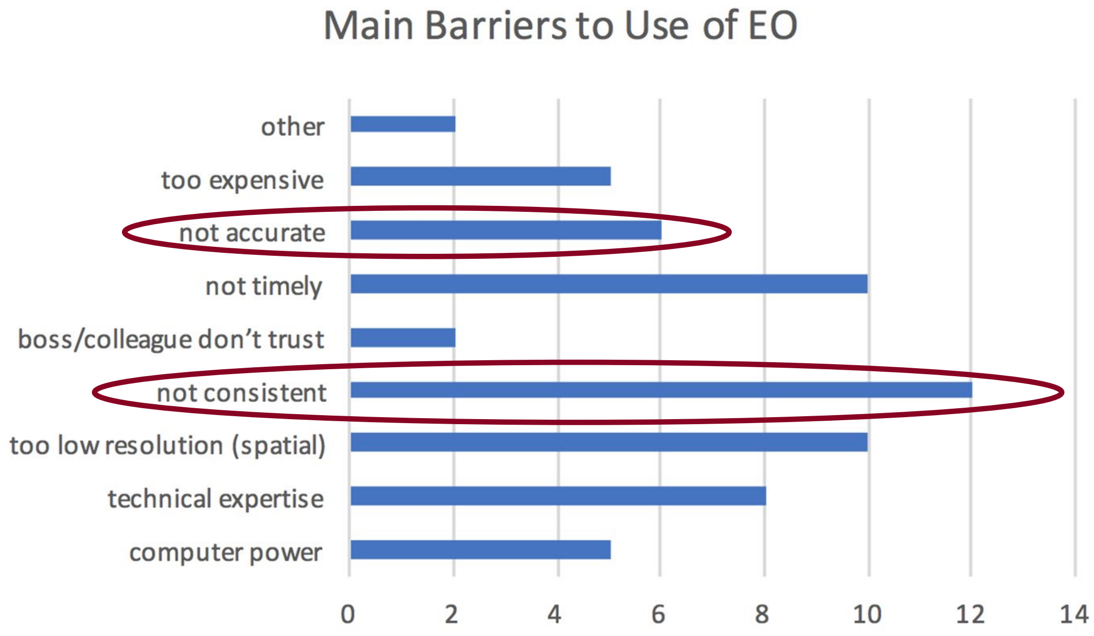
Common metrics designed for LULC assessment

Unable to deal with large class imbalance in binary classifications

 https://www.esa.int/ESA_Multimedia/Images/2019/03/Floods_imaged_by_Copernicus_Sentinel-1#.X5w0pnb7xYc.link

Problem 3: Over-confidence in maps confuses users!





Tellman, B. (2019). What flood event map accuracy is required to enable governments, aid agencies, and insurance companies to protect vulnerable lives and livelihoods? Global Flood Partnership 2019 Conference – 11 - 13 June 2019, Guangzhou (China).

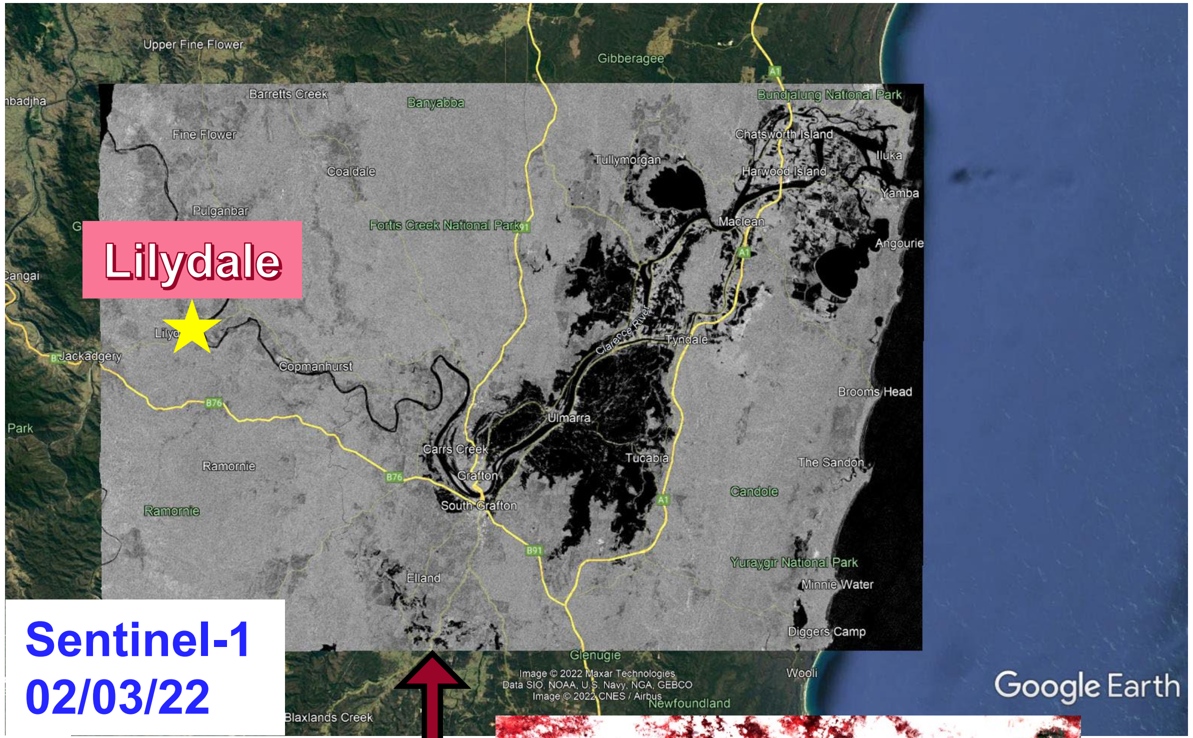
Kettner, A. J.,Schumann, G. J.-P., and Tellman, B. (2019), The push toward local flood risk assessment at a global scale, *Eos, 100*, https://doi.org/10.1029/2019EO113857. Published on 14 January 2019.

OSNABRÜCK UNIVERSITY

Test Site, Data, and Workflow

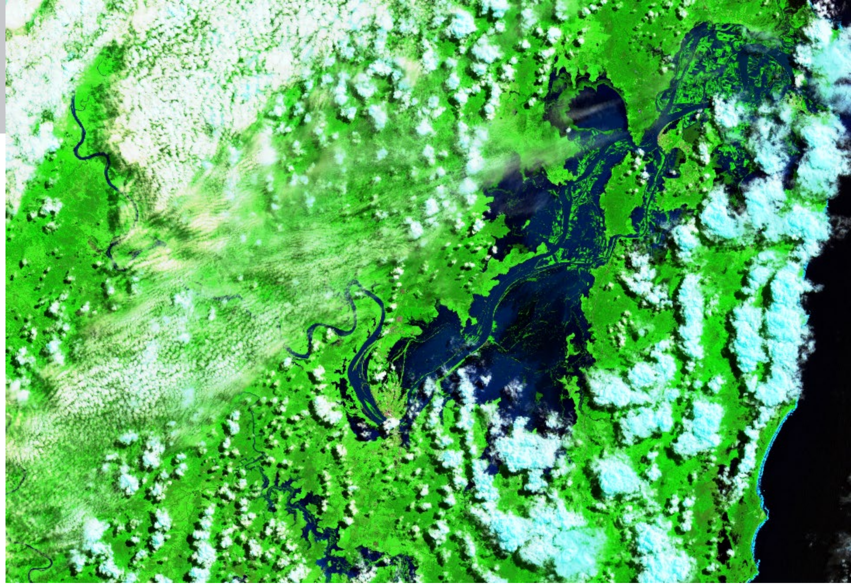
STUDY AREA

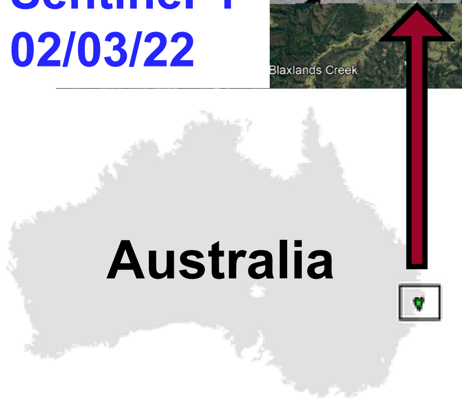
Test Case and Data: Clarence Valley Floods – 2022

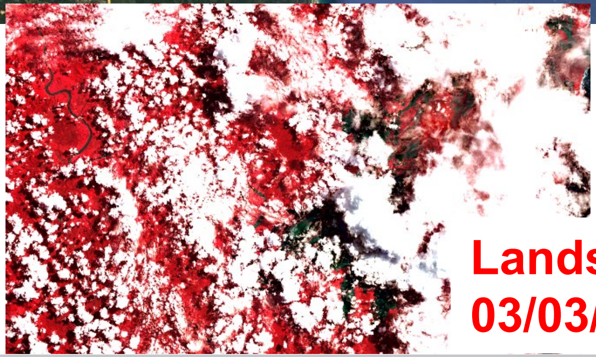


Sentinel-1
02/03/22

Sentinel-2
01/03/22

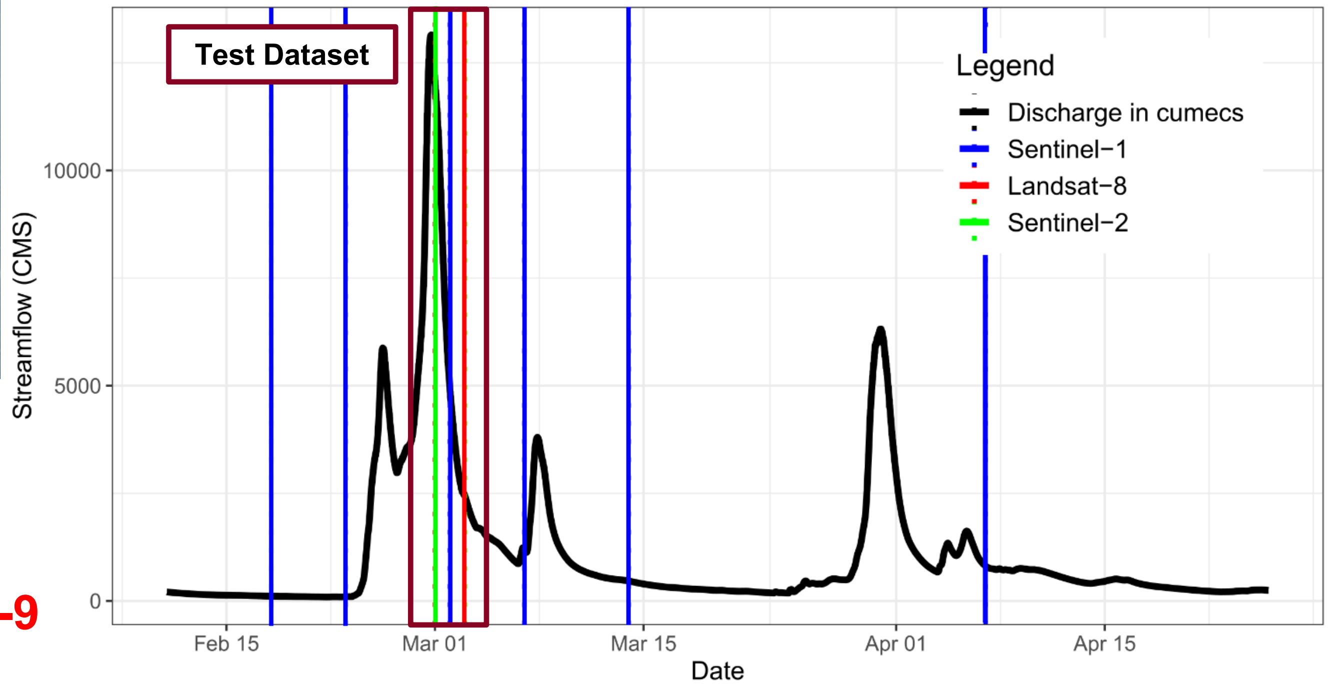


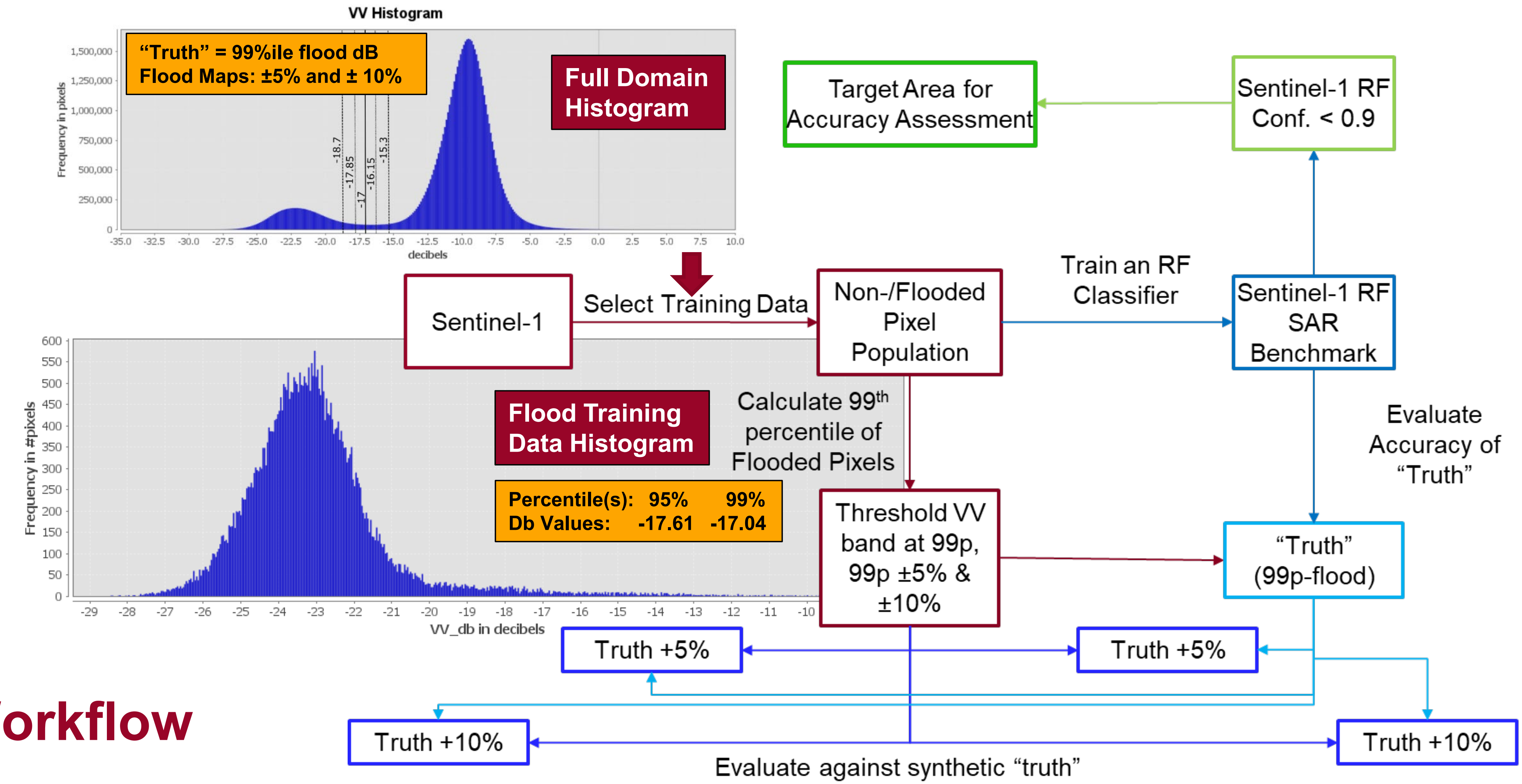




Landsat 8-9
03/03/22

Discharge Hydrograph at Lilydale (2022)




VV Histogram
"Truth" = 99%ile flood dB
Flood Maps: ±5% and ± 10%
Full Domain Histogram
Frequency in pixels
decibels
-18.7
-17.85
-17
-16.15
-15.3
Sentinel-1
Select Training Data
Non-/Flooded Pixel Population
Train an RF Classifier
Sentinel-1 RF SAR Benchmark
Target Area for Accuracy Assessment
Sentinel-1 RF Conf. < 0.9
Flood Training Data Histogram
Percentile(s): 95% 99%
Db Values: -17.61 -17.04
Frequency in #pixels
VV_db in decibels
Calculate 99th percentile of Flooded Pixels
Threshold VV band at 99p, 99p ±5% & ±10%
Evaluate Accuracy of "Truth"
"Truth" (99p-flood)
Truth +5%
Truth +5%
Truth +10%
Truth +10%
Evaluate against synthetic "truth"

Workflow

OSNABRÜCK UNIVERSITY

Validation Metrics Used

Confusion Matrix		Predicted condition	
	Total population = P + N	Positive (PP)	Negative (PN)
Actual condition	Positive (P)	True positive (TP), hit	False negative (FN), type II error, miss, underestimation
	Negative (N)	False positive (FP), type I error, false alarm, overestimation	True negative (TN), correct rejection

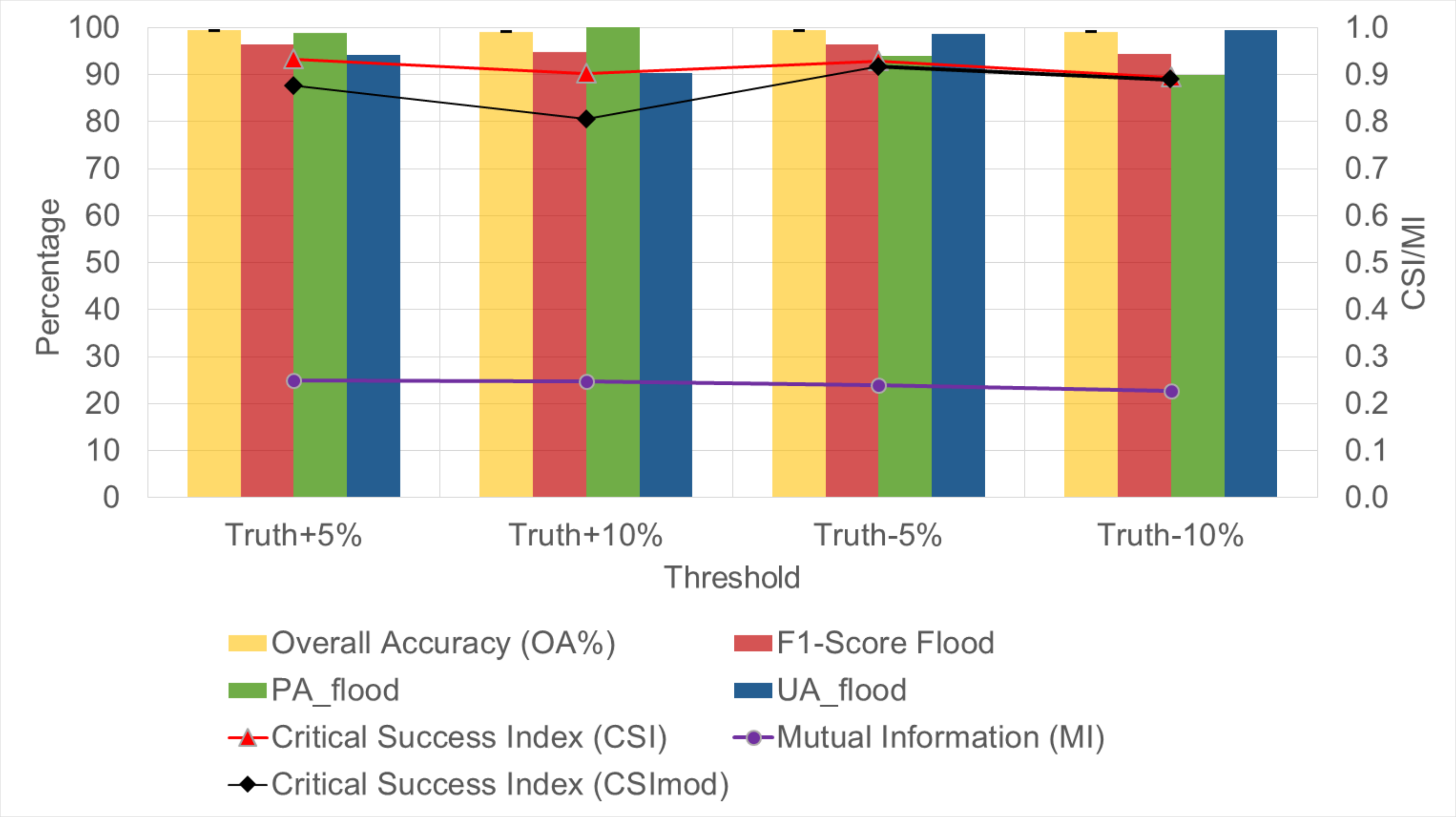
Metric Name	Formula
Prevalence	$\frac{P}{P+N}$
Overall Accuracy	$\frac{TP+TN}{P+N}$
User's Accuracy Flood/Precision	$\frac{TP}{TP+FP}$
Producer's Accuracy Flood/Recall/True Positive or Hit Rate	$\frac{TP}{TP+FN}$
F1-Score Flood	$\frac{2\times TP}{2\times TP+FP+FN}$ OR $\frac{2\times Precision\times Recall}{Precision+Recall}$
Critical Success Index/Intersection over Union	$\frac{TP}{TP+FP+FN}$
Critical Success Index modified	$\frac{TP-FP}{TP+FP+FN}$
Mutual Information	$I(y;x)=H(x)-H(x\|y)$ where $I(Y;X)=\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}p(x,y)\log(\frac{p(x,y)}{p(x)p(y)})$

*Rarely used metrics marked in red

Research Question 1: How do metric choice, sampling design and sample size, influence accuracy assessment?

DIFFERENT THRESHOLD FLOOD MAPS VS. SYNTHETIC "TRUTH"

Current Practice: Validation over the entire domain

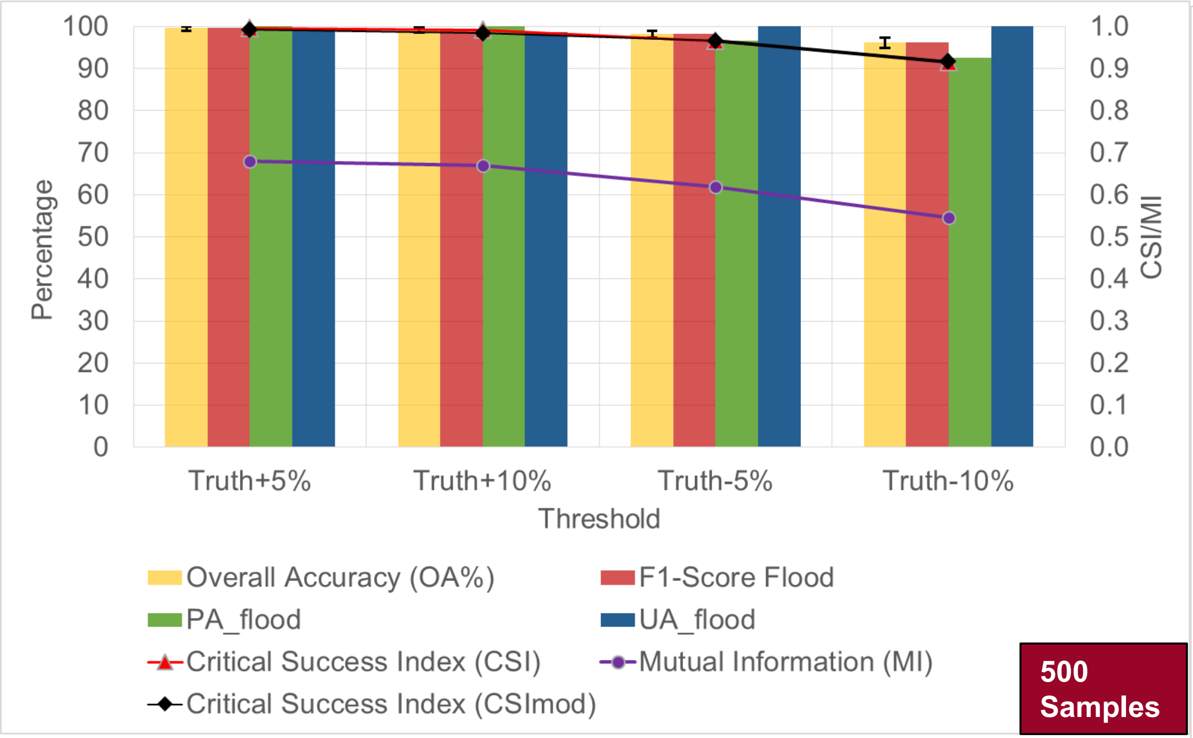


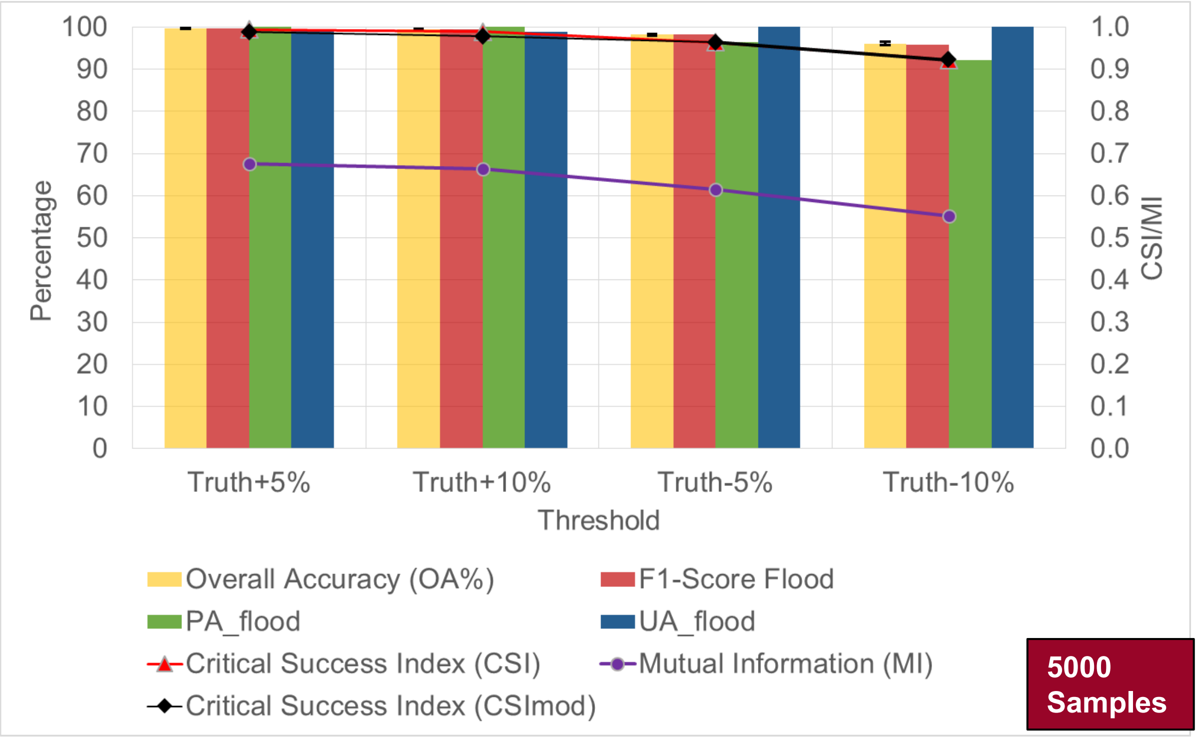
Low sensitivity of all metrics towards flooded area variations

All pixels in image used for the metric computation

Changes in maps hard to capture – overall accuracy remains almost the same (even for large drops in user's and producer's accuracy)

Stratified Random Sampling*



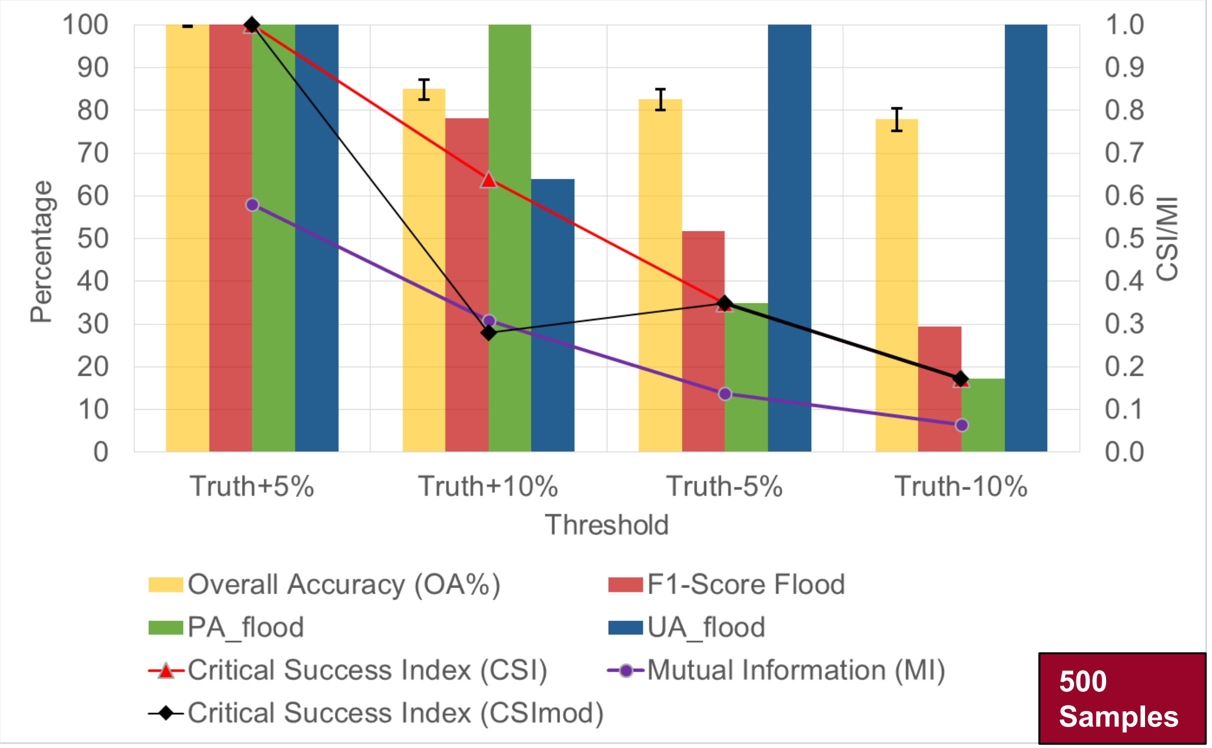


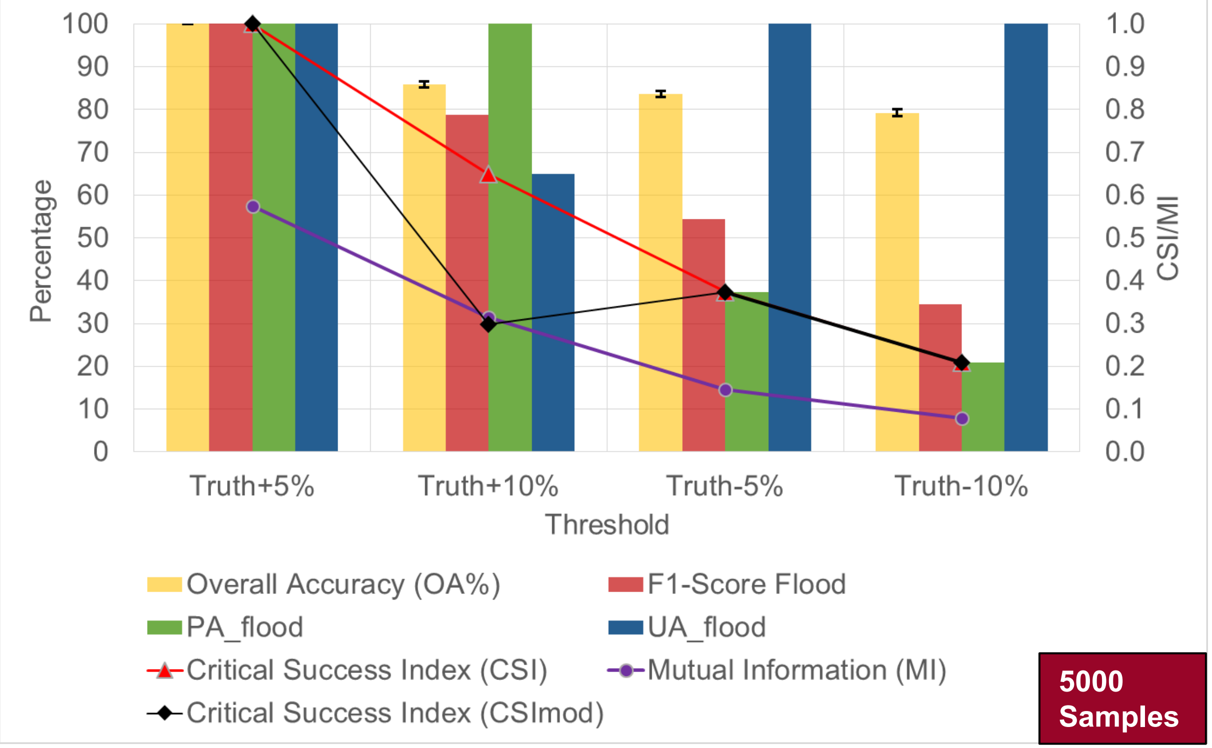
All metrics biased towards over prediction

Low sensitivity demonstrated by low variation in metric values as a function of threshold variations

*Stratified sampling over entire domain

Targeted Random Sampling*



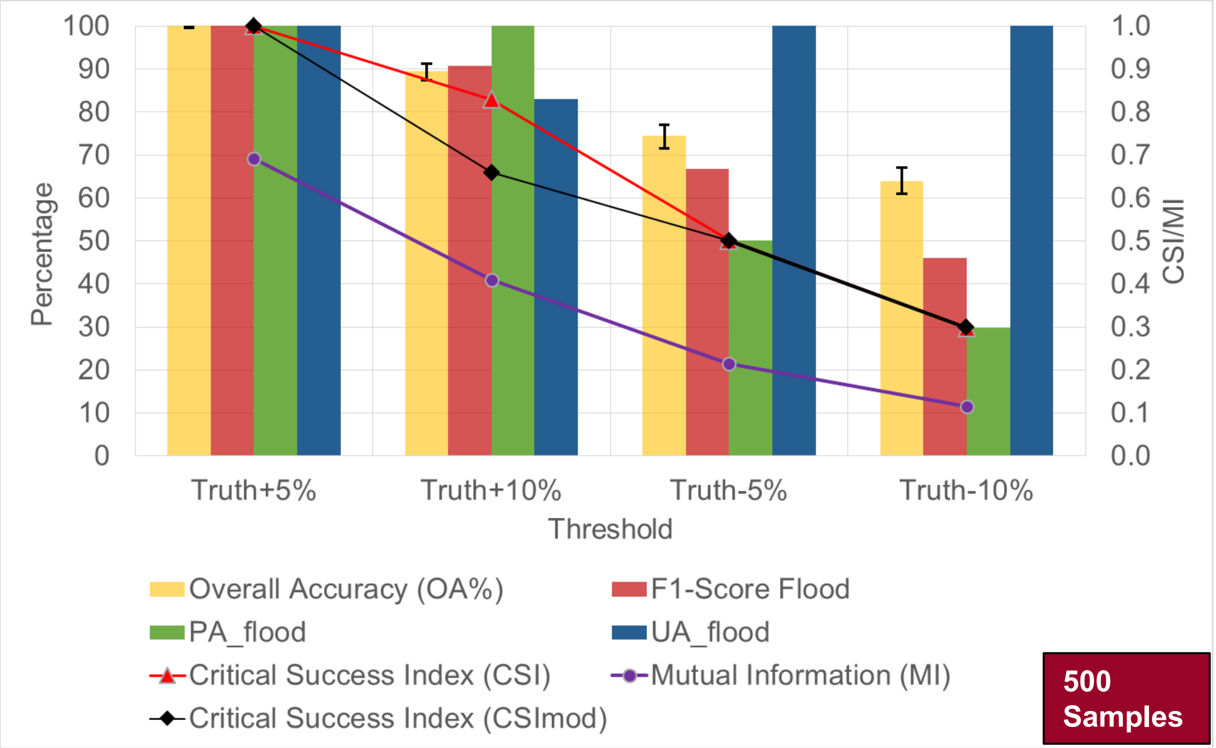


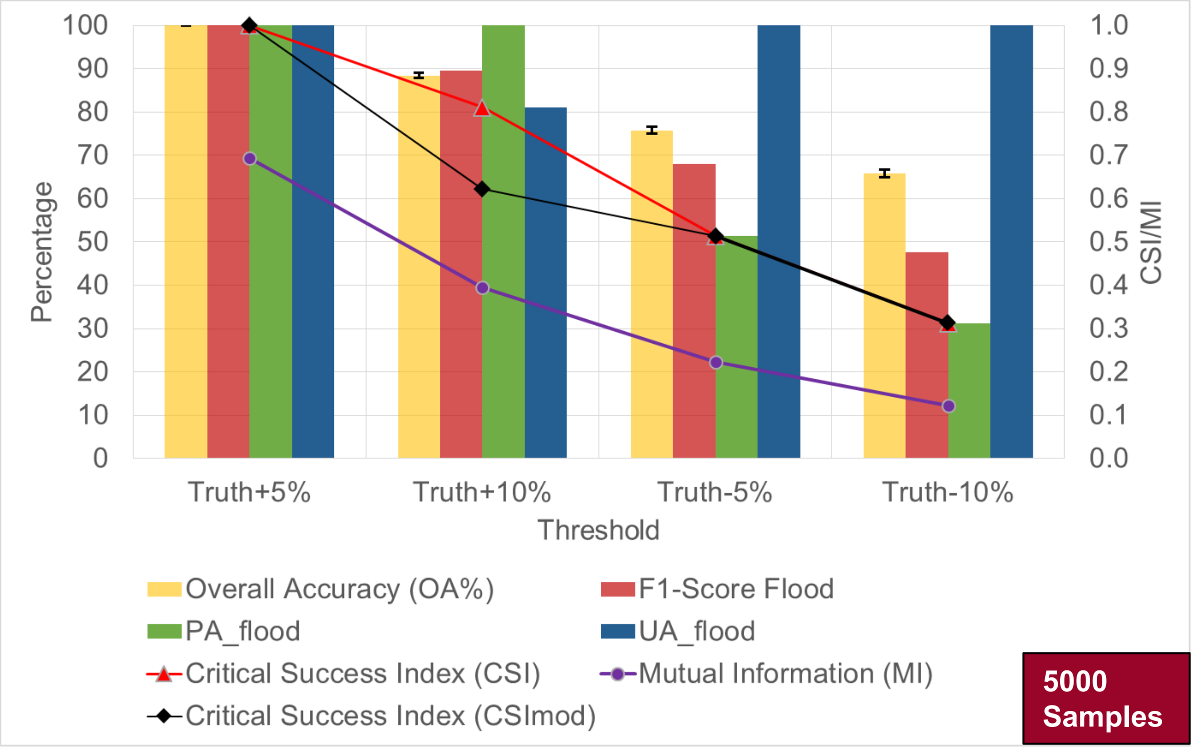
CSImod alters the CSI metric effectively such that similar over- and under- predictions yield similar values

Targeted sampling results in higher variation in metric values i.e. greater metric sensitivity

*Sampling in areas with RF classification confidence < 0.9

Targeted Stratified Random Sampling*





More flooded samples result in higher F1 and CSI scores for the same maps

Stratified sampling reduces variation in F1 and CSI and increases variation in MI and OA

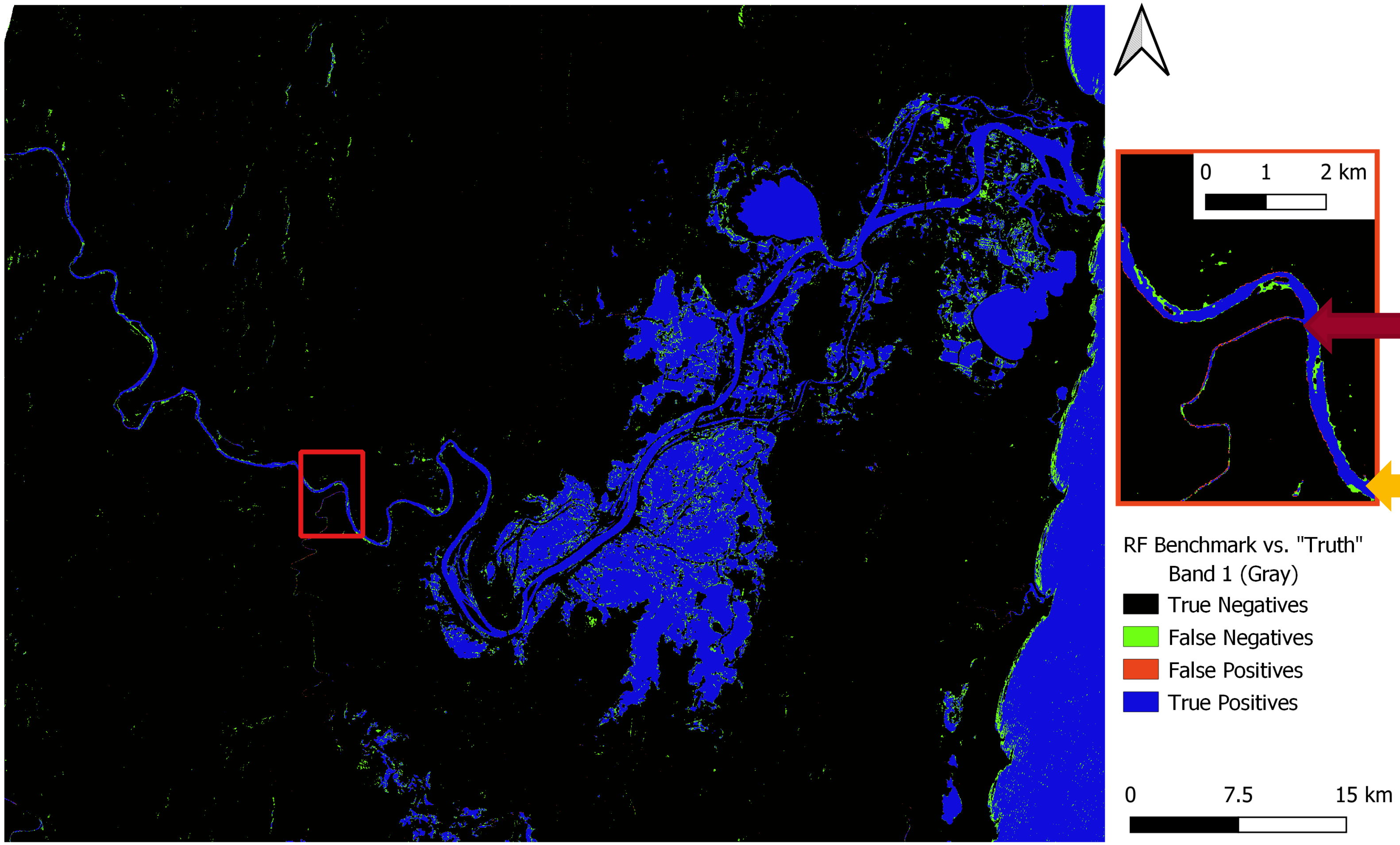
*Stratified sampling in areas with RF classification confidence < 0.9

Research Question 2: How does the validation data error influence accuracy assessment and flooded area calculations?

SYNTHETIC VV-BINARIZED “TRUTH” VS. RANDOM FOREST BENCHMARKS

Benchmark Sentinel-1 Classification

Random Forest Classification using **S1 VV, VH, elevation (CopDEM30), Otimized GLCM Texture PCs 1 and 2**

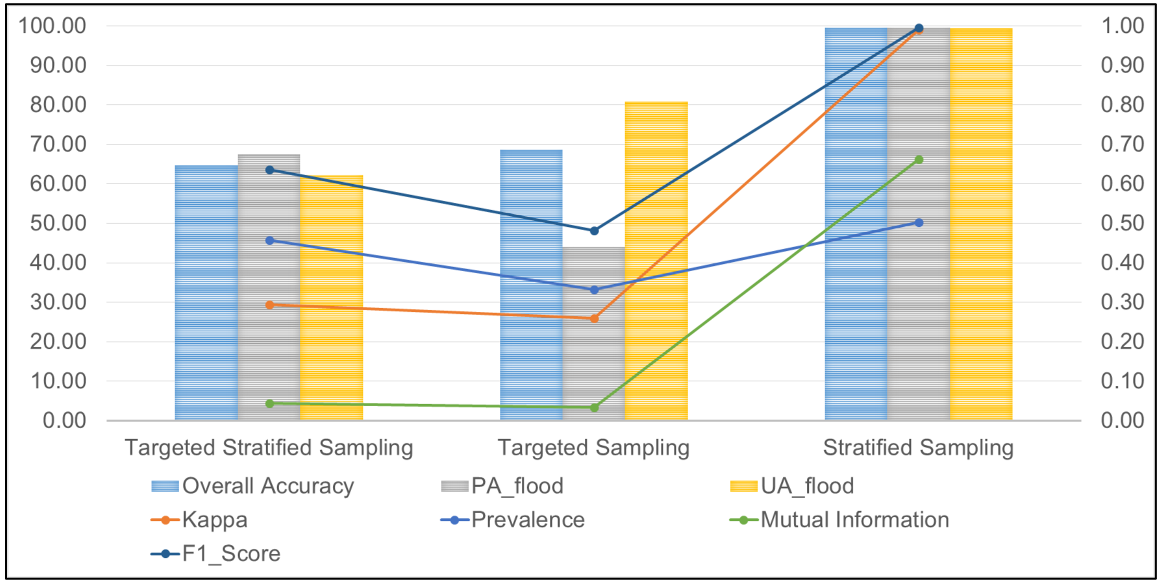


False positives at the edge of ephemeral streams

False negatives at channel banks

Comparison against S-1 RF Benchmark

Sampling Strategy (500-1000 pts)		Targeted Stratified Sampling		Targeted Sampling		Stratified Sampling	
Confusion Matrices				Reference			
		Flood	Non-flood	Flood	Non-flood	Flood	Non-flood
"Truth"	Flood	**308**	**204**	**73**	**64**	**501**	**3**
	Non-flood	**149**	**337**	**93**	**270**	**2**	**494**



Stratified sampling provides spuriously high accuracy values

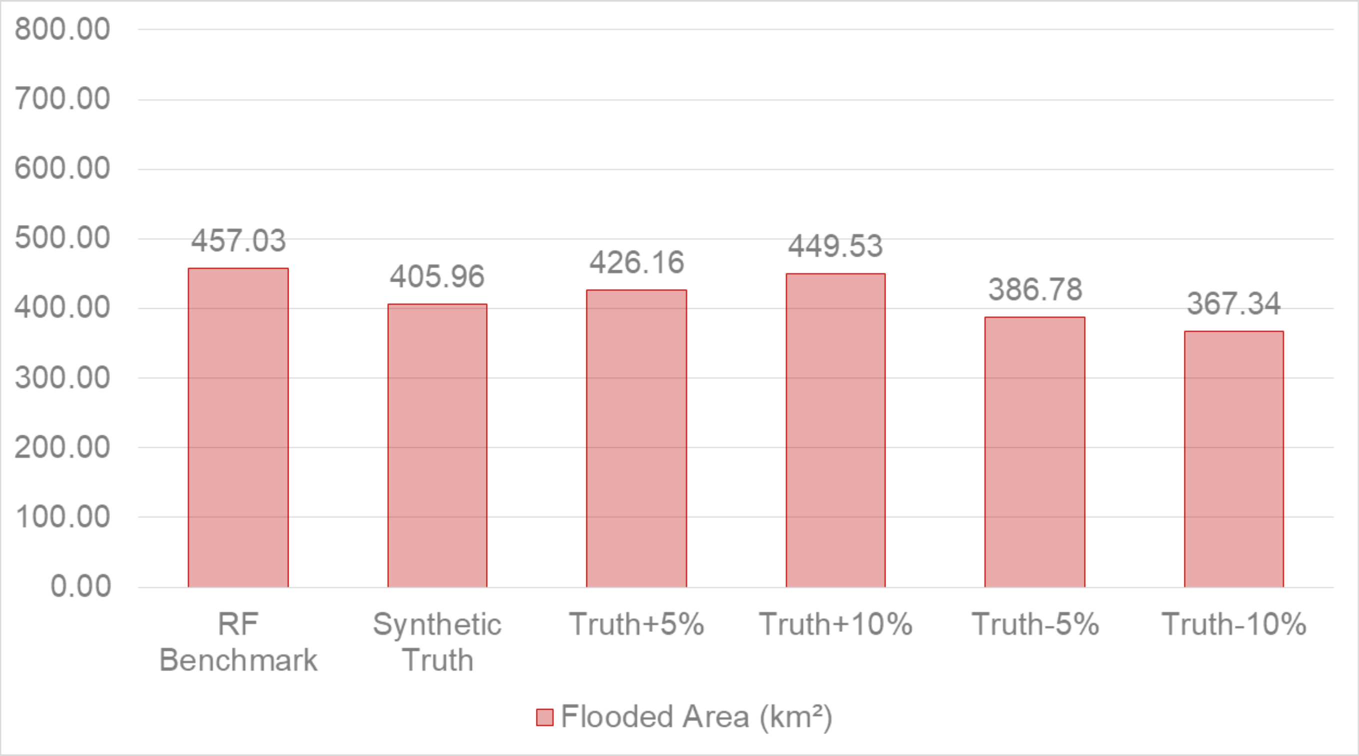
Targeted sampling proves better than other sampling designs

*Validation data errors "correlated" in this case as same input data source i.e. Sentinel-1 SAR

*In the absence of RF confidence or map uncertainty, a buffer along the flood boundary might be considered as the area to "target"

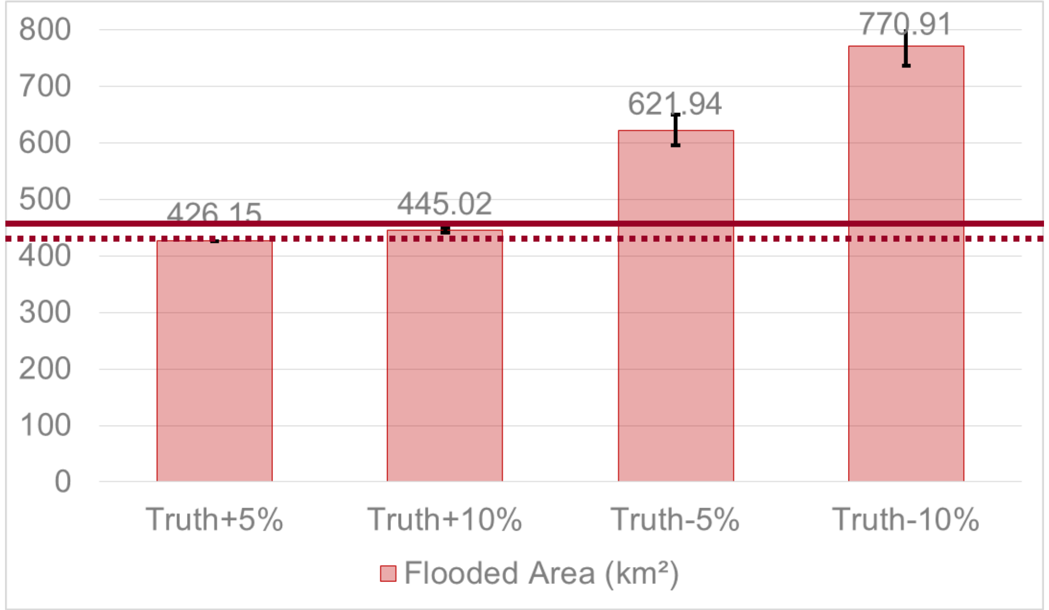
OSNABRÜCK UNIVERSITY

Flooded Area Estimation



Direct map-based flooded area estimates

Bias-corrected area estimates after Olofsson et al. (2014)



RF Benchmark Flooded Area

Synthetic Truth Flooded Area

***Standard bias-correction techniques for change area estimation NOT directly applicable to flood mapping**

***Bias correction only applicable to random, systematic and stratified sampling designs**

Olofsson, P., Foody, G. M., Herold, M., Stehman, S. V., Woodcock, C. E., & Wulder, M. A. (2014). Good practices for estimating area and assessing accuracy of land change. *Remote Sensing of Environment*, *148*, 42–57. https://doi.org/10.1016/j.rse.2014.02.015

Conclusions and Outlook

- Flood map accuracy assessments **strongly depend** on the choice of metrics, sampling strategy, and validation data quality.
- Increasing sample size **reduces the sensitivity** of accuracy estimation metrics.
- Confidence intervals could provide a clearer overview for decision-makers.
- Future work will focus on developing bias correction methods for flooded area calculations from satellite data.
- An assessment of impacts of land-use and elevation categories on accuracy assessments will also be considered.