Generation of synthetic remote sensing images with ultrasimple but ultrafast approaches

Gregoire Mariethoz, University of Lausanne, Switzerland

Workshop Geolearning, Fréjus, 2 April 2025

GEOLEARNING
CHAIRE /// Data Science for the Environment

ANDRA

BNP PARIBAS

CCR

SCOR
FONDATION POUR LA SCIENCE

MINES PARIS | PSL

INRAE
la science pour la vie, l'humain, la terre

Fondation
MINES PARIS

Allez Denis!

A wealth of data – but it is never enough

We would like to measure everything, everywhere, all the time.

It is impossible.

Generating data: why?

We have lots of data, but need to generate even more!

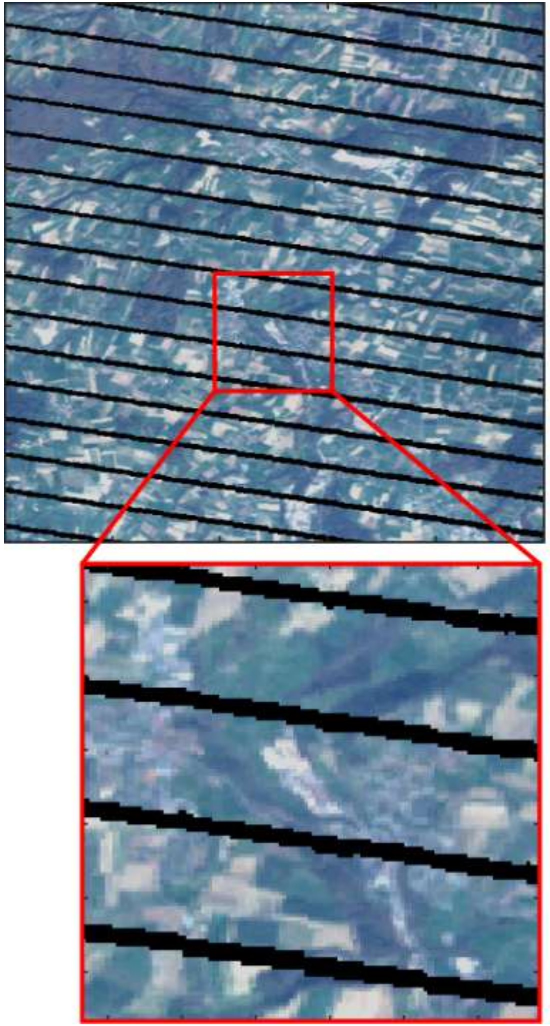
Specific type of problem suited to data-driven models:
machine learning, geostatistics (parametric/non-parametric).

We may need to generate data for:

- Filling spatial gaps: interpolation → Multiple-point geostatistics ♥
- Uninformed scales: downscaling
- Recovering missing colors: colorization (=multivariate)
- Generating uninformed epochs (past/future) (=spatio-temporal)

Generating spatial data (with MPS)

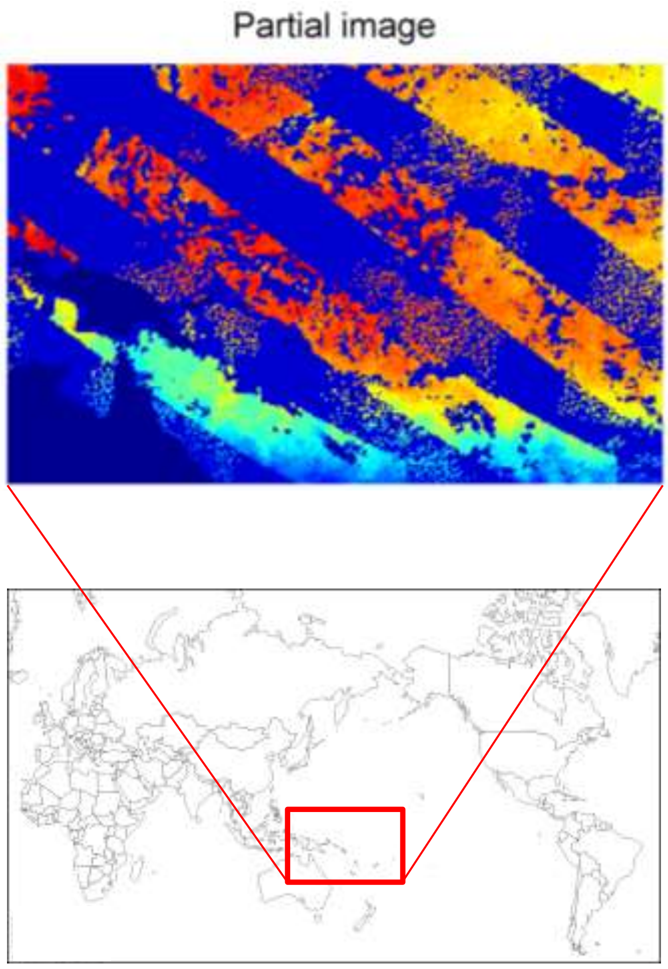
Application to Landsat 7 SLC-off images



Yin, G., et al. (2017). "Gap-filling of landsat 7 imagery using the direct sampling method." Remote Sensing **9**(1).

Gap-filling (pour Thomas)

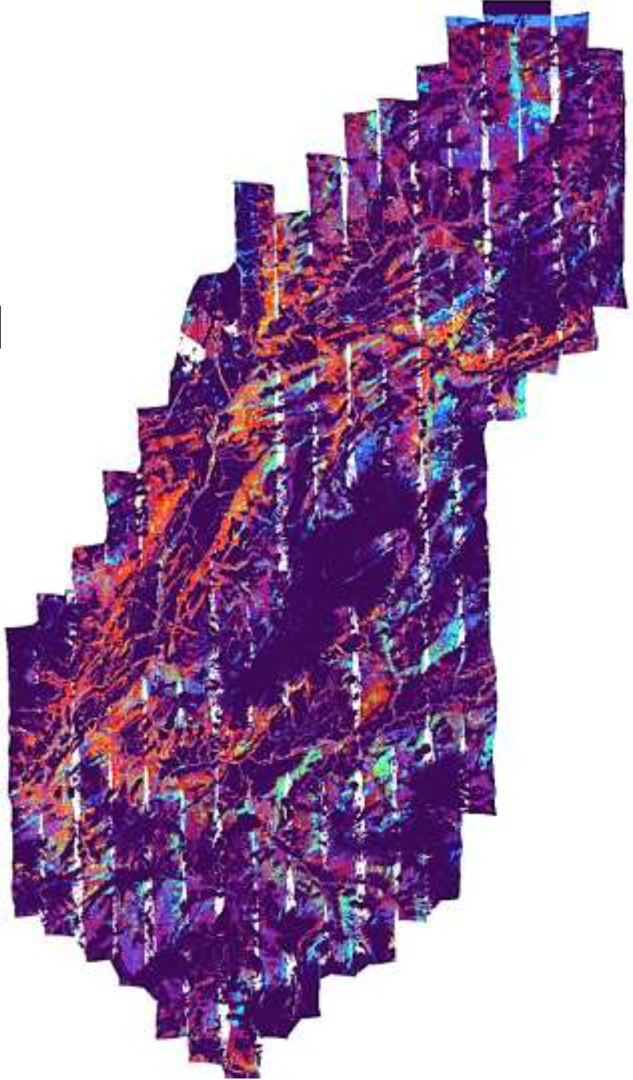
- Sea surface temperature
- Gaps due to orbital characteristics, clouds, etc
- The informed parts are sufficiently large to be used as training image.
- Reconstruction is non-unique

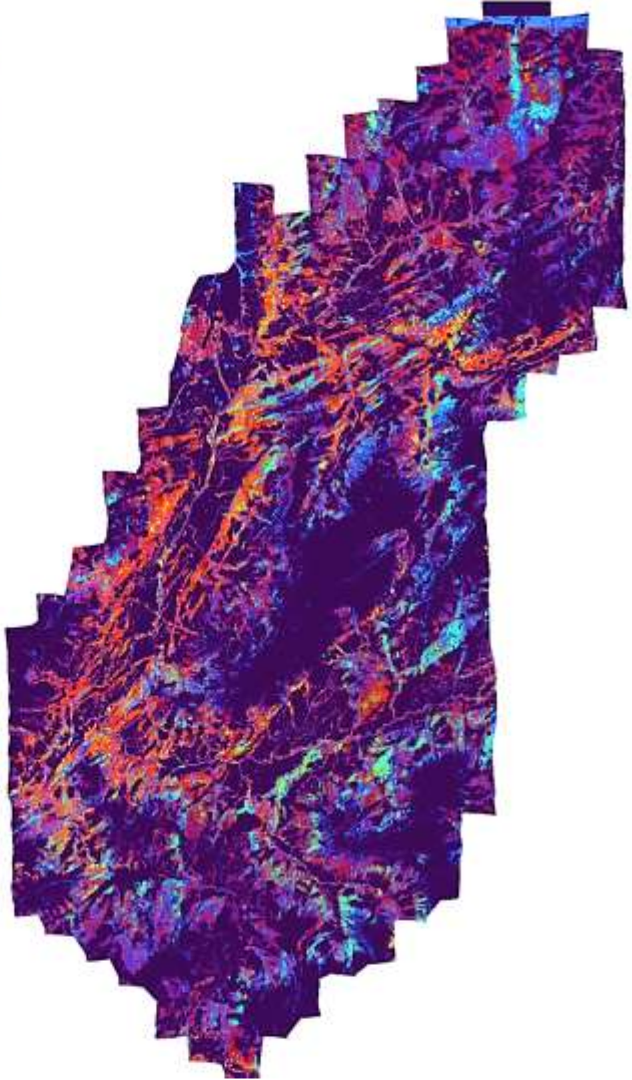


Mariethoz, G., M.F. McCabe, and P. Renard, *Spatiotemporal reconstruction of gaps in multivariate fields using the direct sampling approach.* WRR, 2012. **48**(10).

Gap-filling (grand)

- AVIRIS hyperspectral imagery
- 262’500’000 pixels
- ~4h on a small cluster





Downscaling by pattern matching

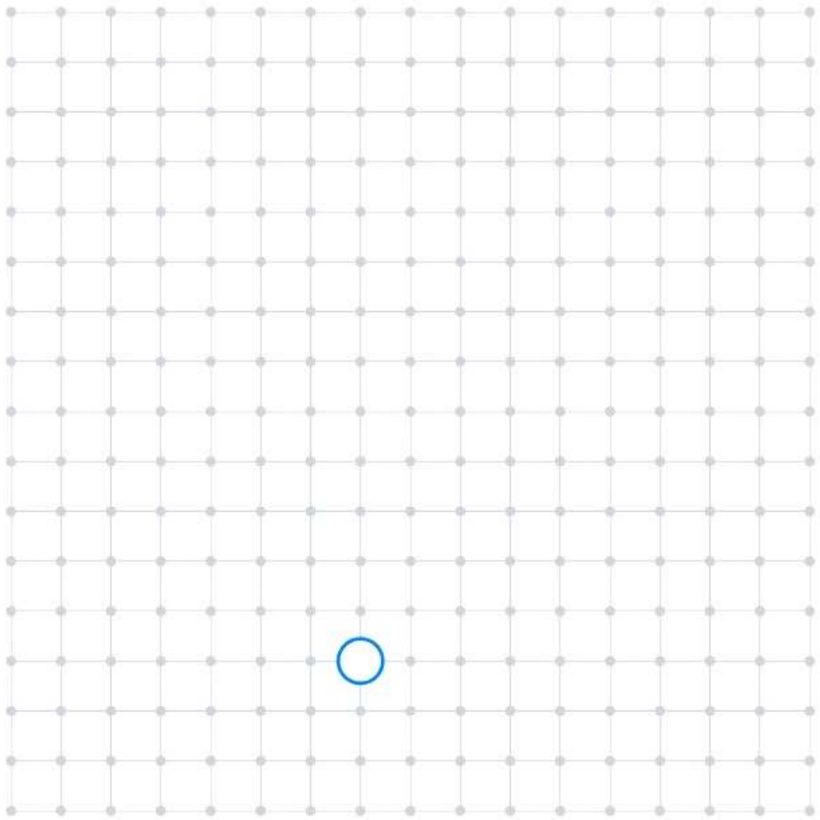
Lower portion of the image only known with low-quality sensor

Geostatistical models are often pixel-based

- e.g. sequential simulation.
- The generated patterns are based on a training image or a covariance model.

Works but…

Difficult to scale to XXL++ space-time domains.

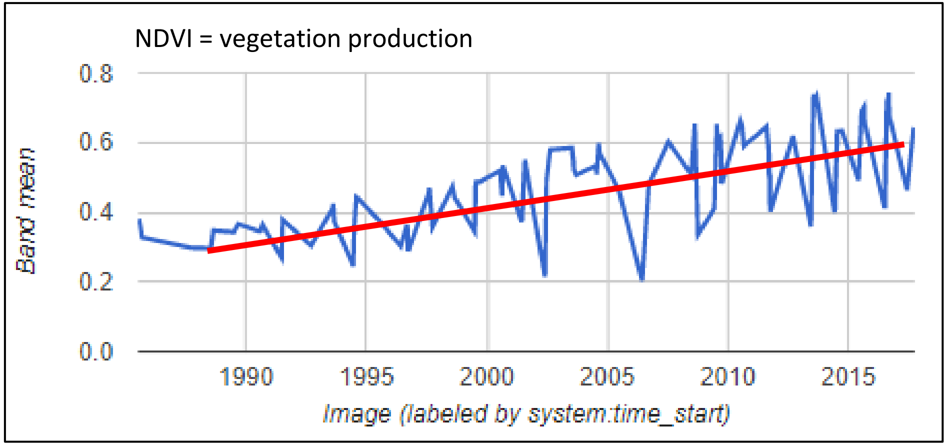


Video R. Nussbaumer

Biggest challenge and need: the temporal dimension

Dense time series of images

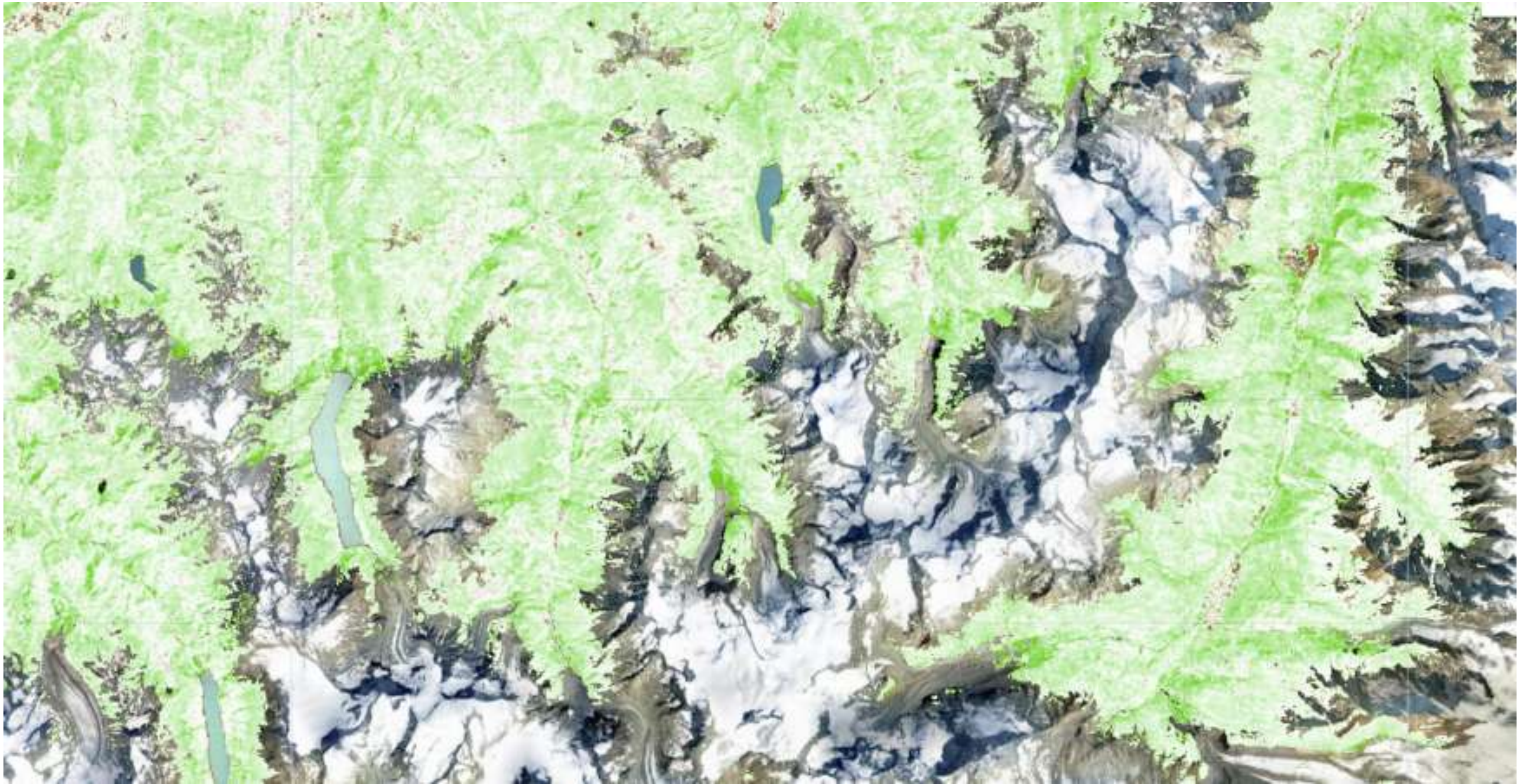
- Landsat: over 50 years of continuous data.
- For example, we study snow and vegetation processes in the Alps, based on a time series of all Landsat images.



Estimate trend for each pixel

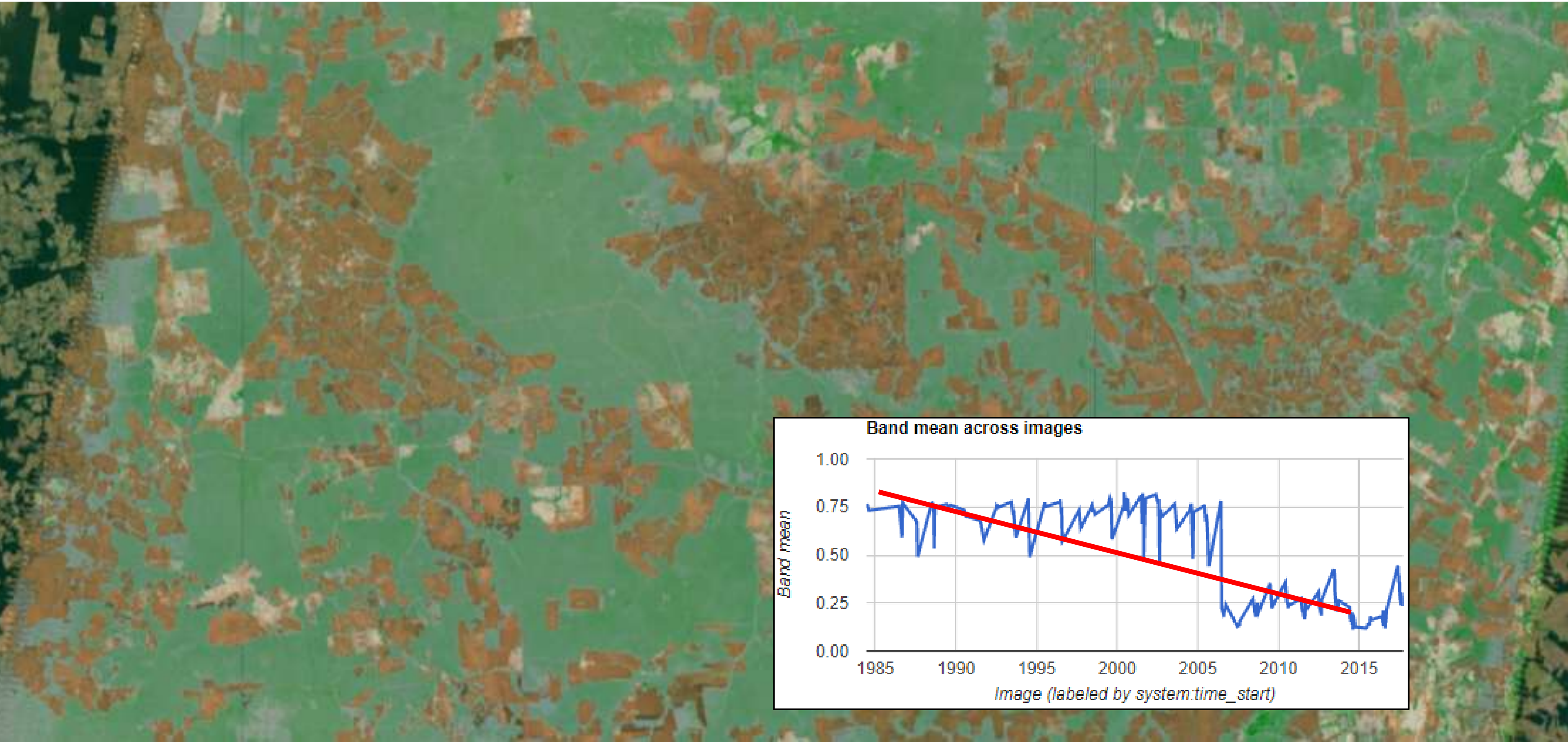
Progression of vegetation in the Alps (1980-2018, 30m resolution)

Slope of trend: Green=afforestation, brown=deforestation



Deforestation in Amazonia (1980-2018)

Slope of trend: Green=afforestation, brown=deforestation



The need for deep-time satellite data

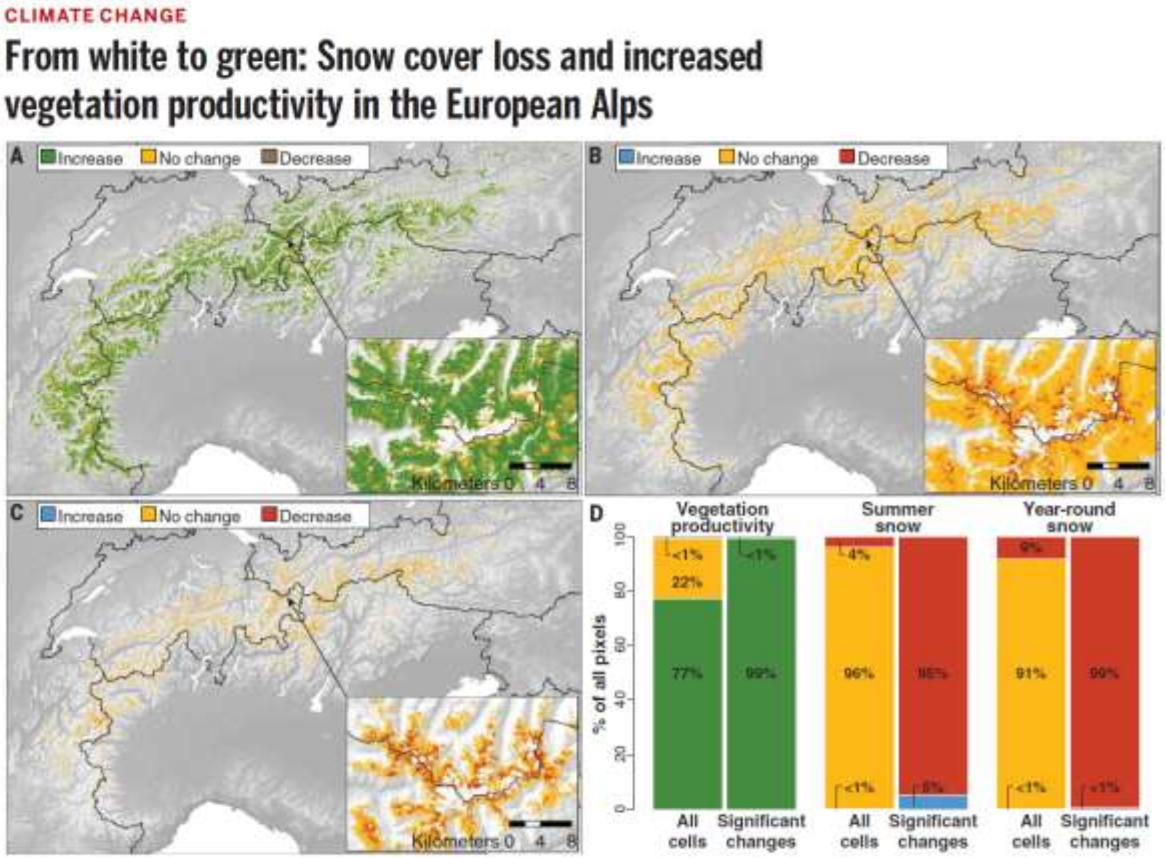
- Climate change in the Alps:

  Shorter snow season,
  Vegetation in higher altitudes,
  Implications for biodiversity,
  Hydrological resources,
  Tourism, etc.

- Quantifying such environmental change requires baselines.

- Entire Alps, 30m resolution
  - ~50 million pixels per image
  - Multivariate
  - At each pixel, a time series (1984-2023)

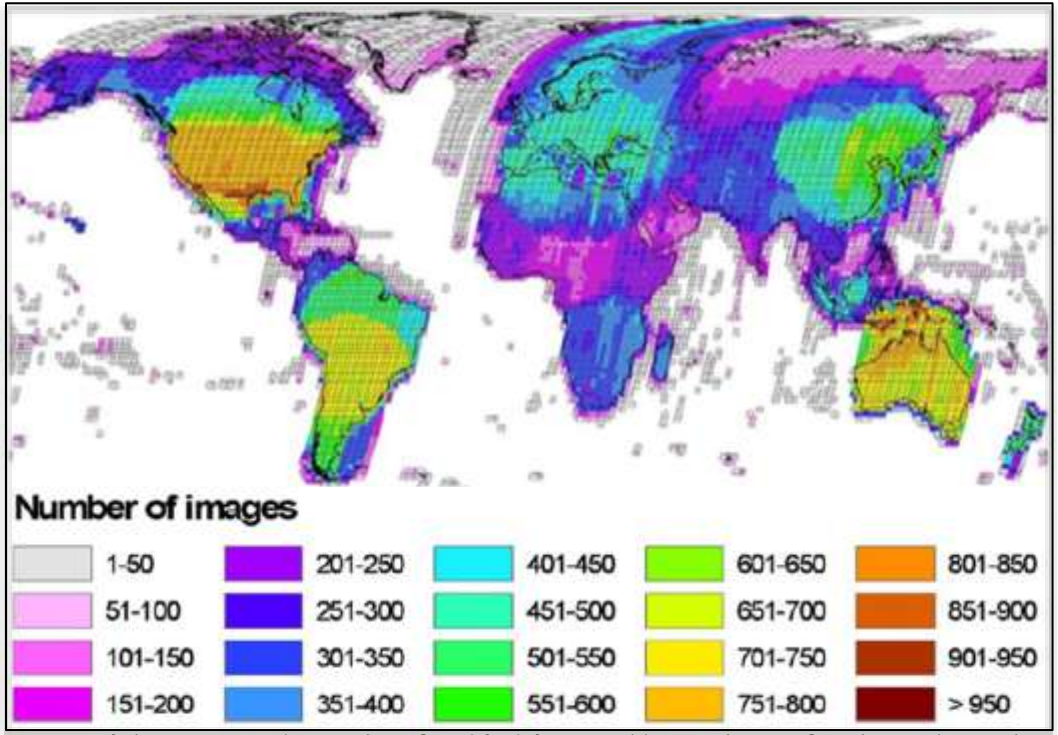
- **What about earlier than 1984?**



Rumpf *et al.*, From white to green: Snow cover loss and increased vegetation productivity in the European Alps. *Science* **376**,1119-1122(2022).DOI:10.1126/science.abn6697

Satellites are temporally short-sighted

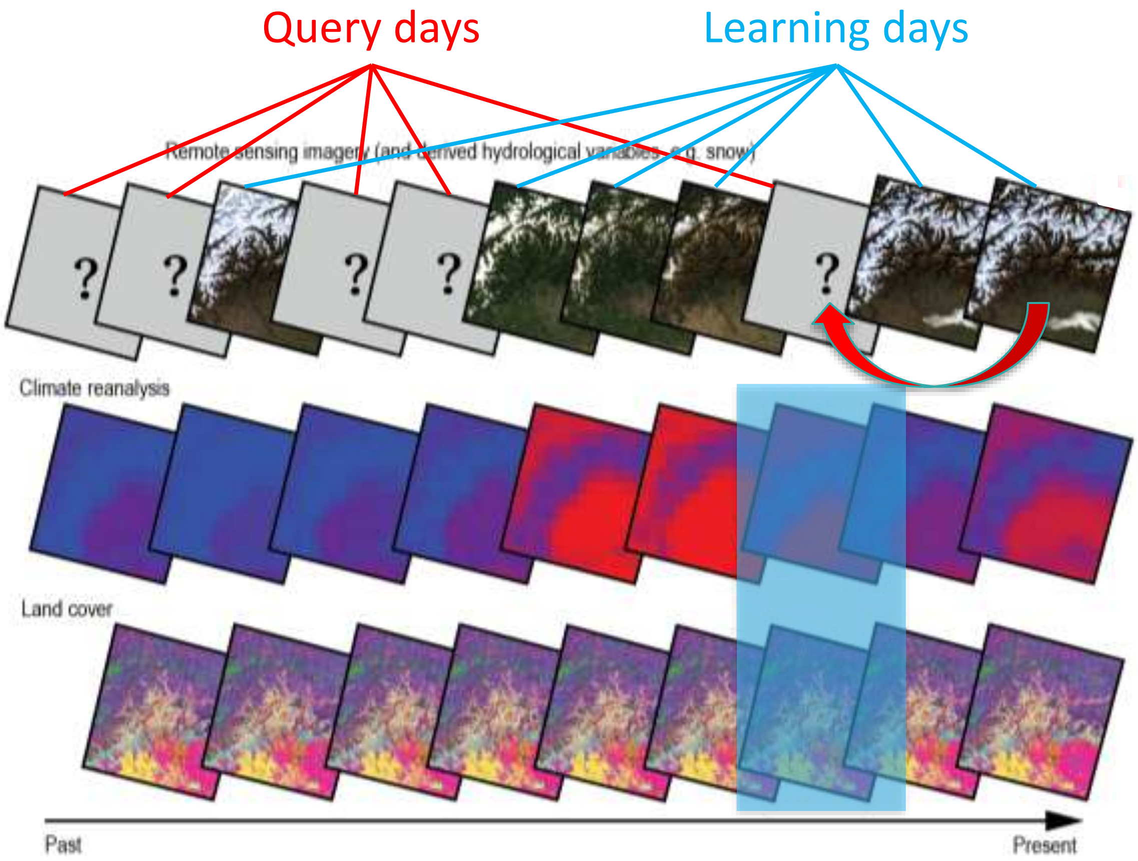
- Almost unlimited amounts of satellite data today (Sentinel, Planet Labs,…).
- Useful to see changes despite clouds.
- Much less before ~2008.
- Before 1999, on average only 1-2 cloudless images per year in central Europe.
- One image every 2-4 years in West Africa.



Status of the USGS Landsat archive [modified from Wulder et al., 2016]. Colors indicate the number of scenes available at each location for the period 2000-2009.

Generate missing epochs based on predictors

- Hypothesis: repetition of patterns under similar climatic conditions.
- Predictors are application-dependent
- To generate snow cover, it is temperature, precipitation, solar radiation, aspect.
- For ET, it is temperature (average, min, max), precipitation.
- Climate predictors informed from 1950, thanks to ERA5 reanalysis.
- Predictors not needed at high resolution!



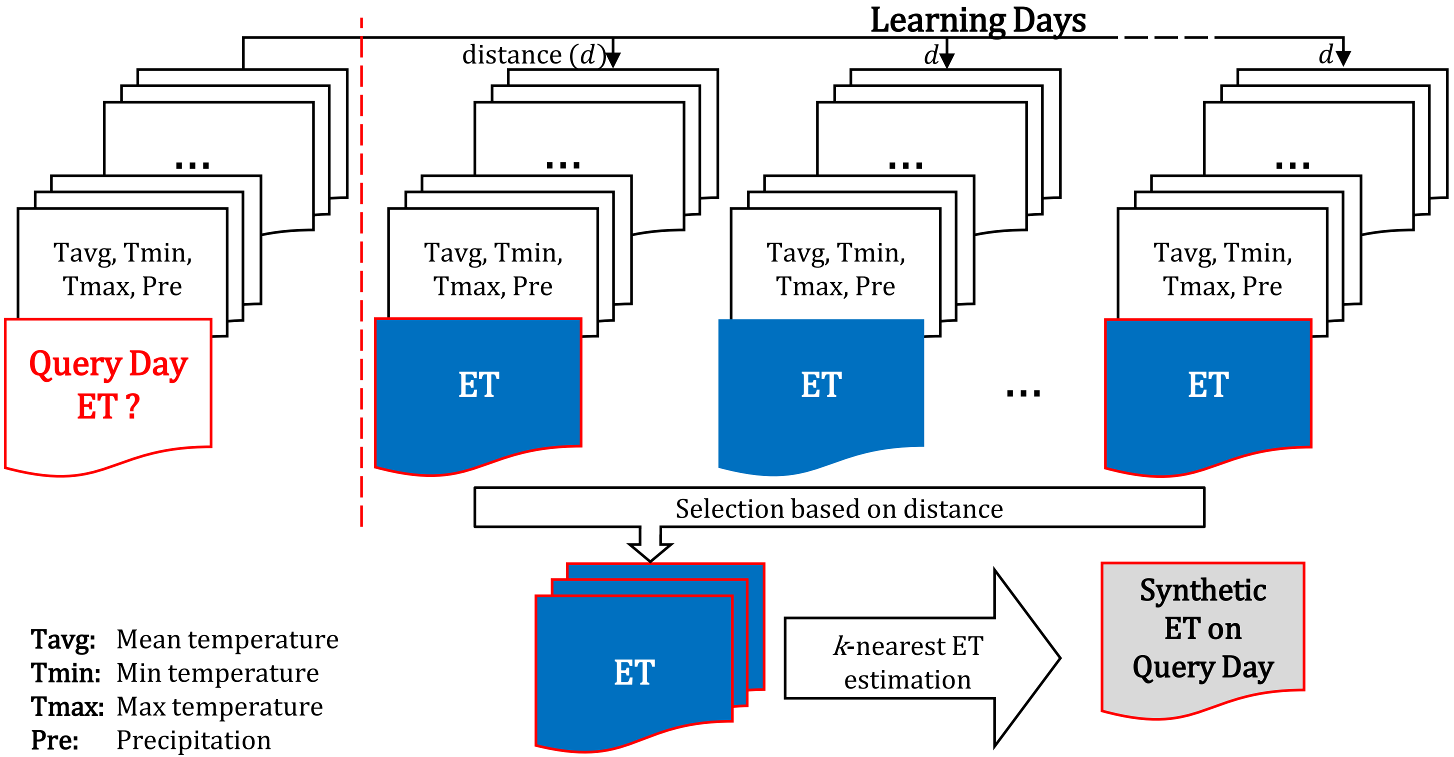
Guessing uninformed epochs

Loic Gerber

Fatemeh Zakeri

Said Obakrim

Estimation with a k-nearest neighbor approach



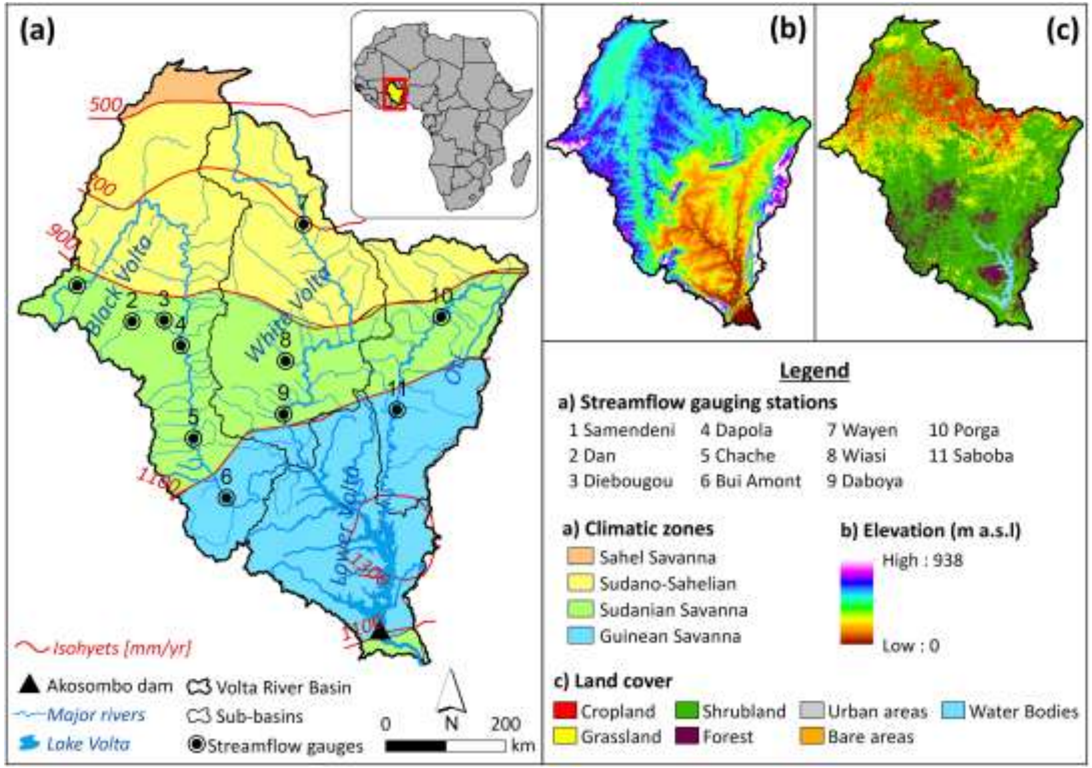
Tavg: Mean temperature
Tmin: Min temperature
Tmax: Max temperature
Pre: Precipitation
ET: Evaporation

Definition of a distance

- Distance between a given query day and all learning days.
- Computed including a number of preceding days.
- The k days with the lowest distance are then aggregated to obtain an estimate for the query day (mean, median, mode, etc).
- Parameters related to the distance (size of window, k, weights of variables) are optimized using cross-validation.

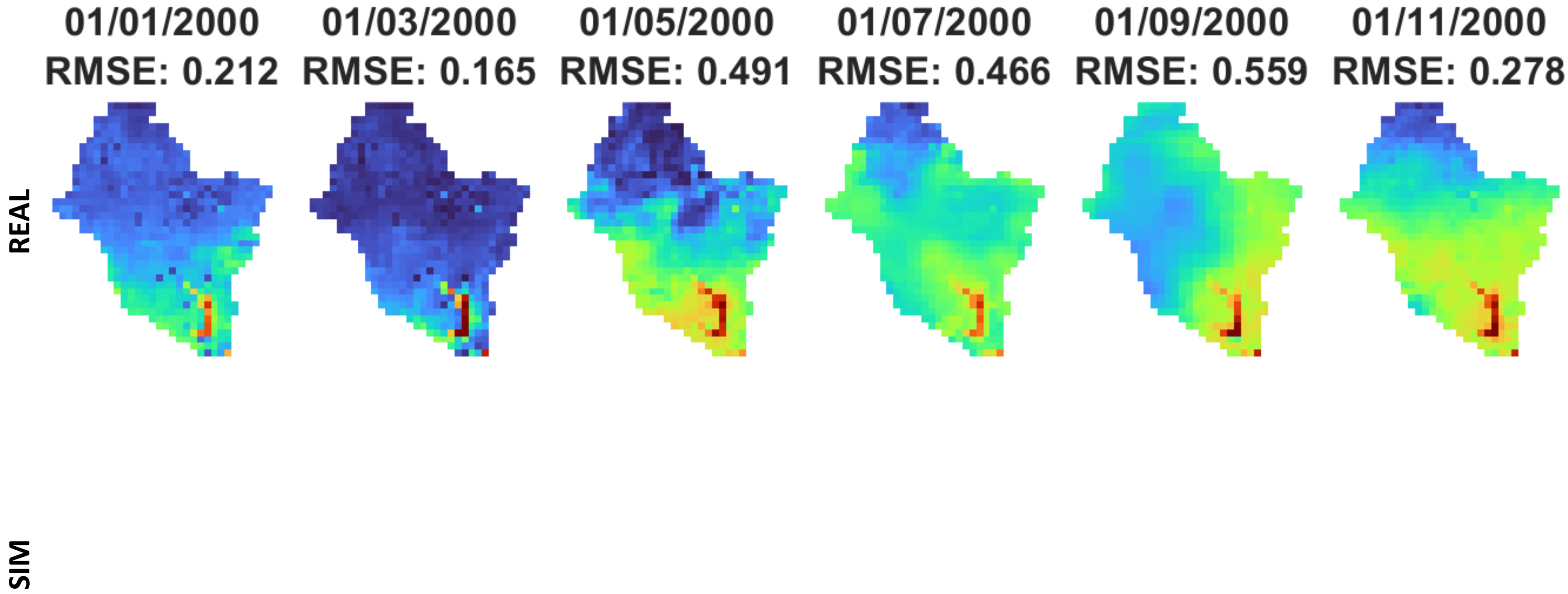
Application to ET in the Volta river basin

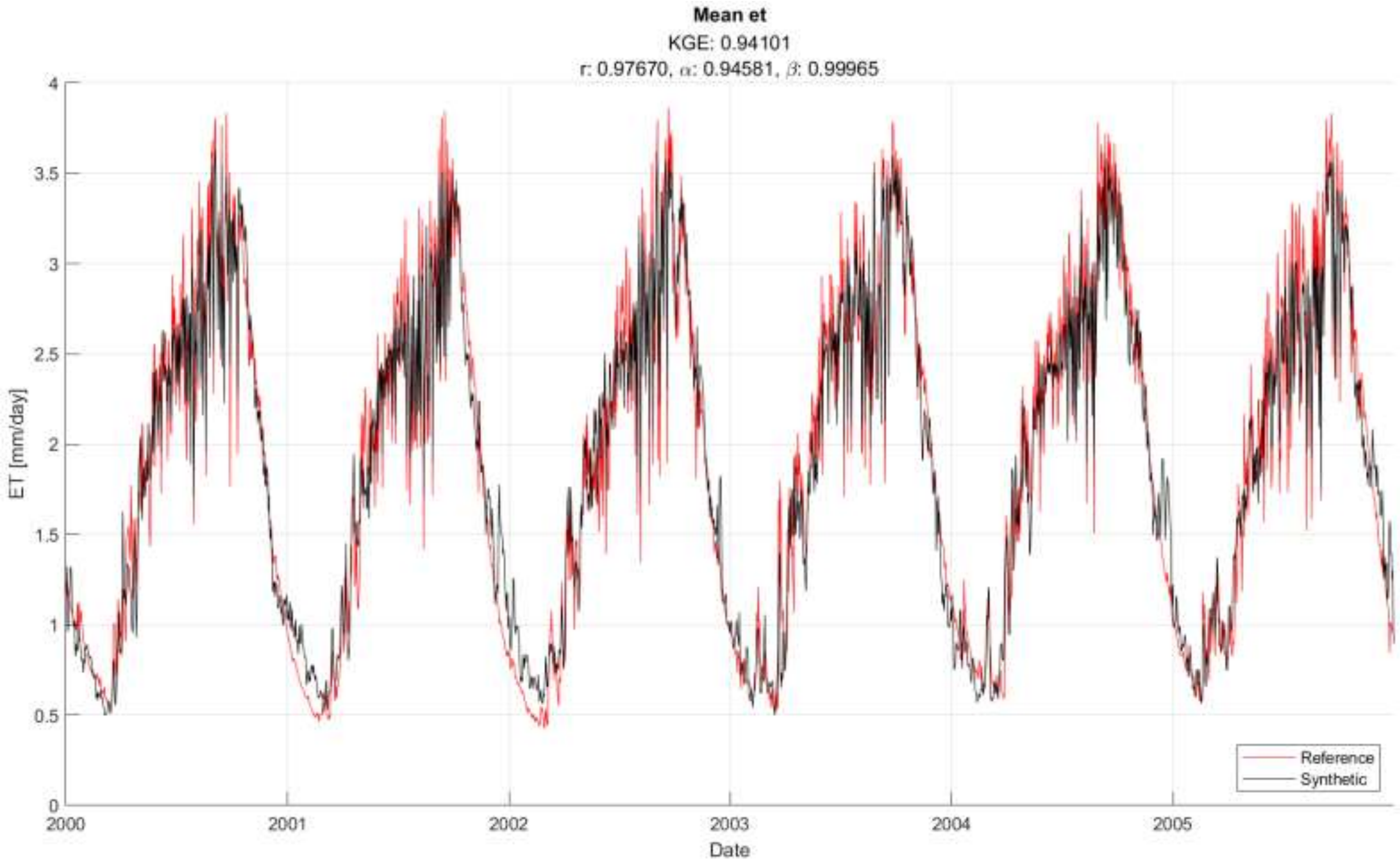
- Data scarce region
- Simulated:
  - ET (GLEAM)
  - Daily, 0.25°, 1980 – 2020
  - Split into training (1980-1999) and validation (2006-2020)
- Predictors
  - ERA5 Land: Precip, Tmin, Tmax, Tavg
- Multivariate tests ok (not shown…)



Dembélé et al., 2020

Simulated evapotranspiration




Mean et
KGE: 0.94101
r: 0.97670, α: 0.94581, β: 0.99965
ET [mm/day]
Date
Reference
Synthetic
0
0.5
1
1.5
2
2.5
3
3.5
4
2000
2001
2002
2003
2004
2005

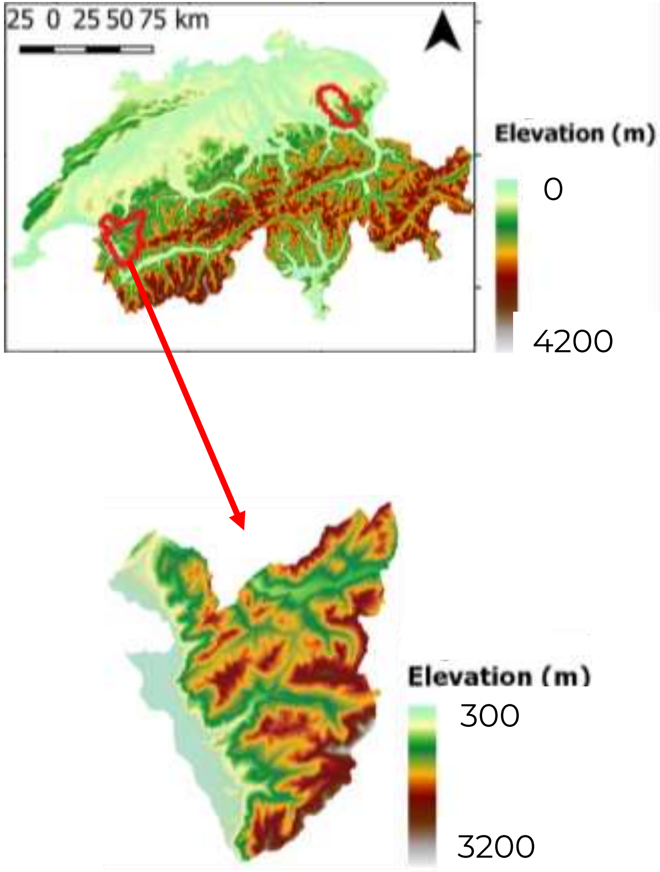
Application to snow cover

- Predicted:
  - <u>Daily</u>, 20y, Landsat 30m/Sentinel-2 snow cover - binary

  (real Landsats every <16 days only)

- Two Scenarios for the predictors:
  1. Satellite age (including MODIS-observed snow cover, 500m)
  2. Pre-satellite age

- Two resolutions for climate data:
  - ERA5 (temperature and precipitation, 11km)
  - Swiss national reanalysis product (1km)

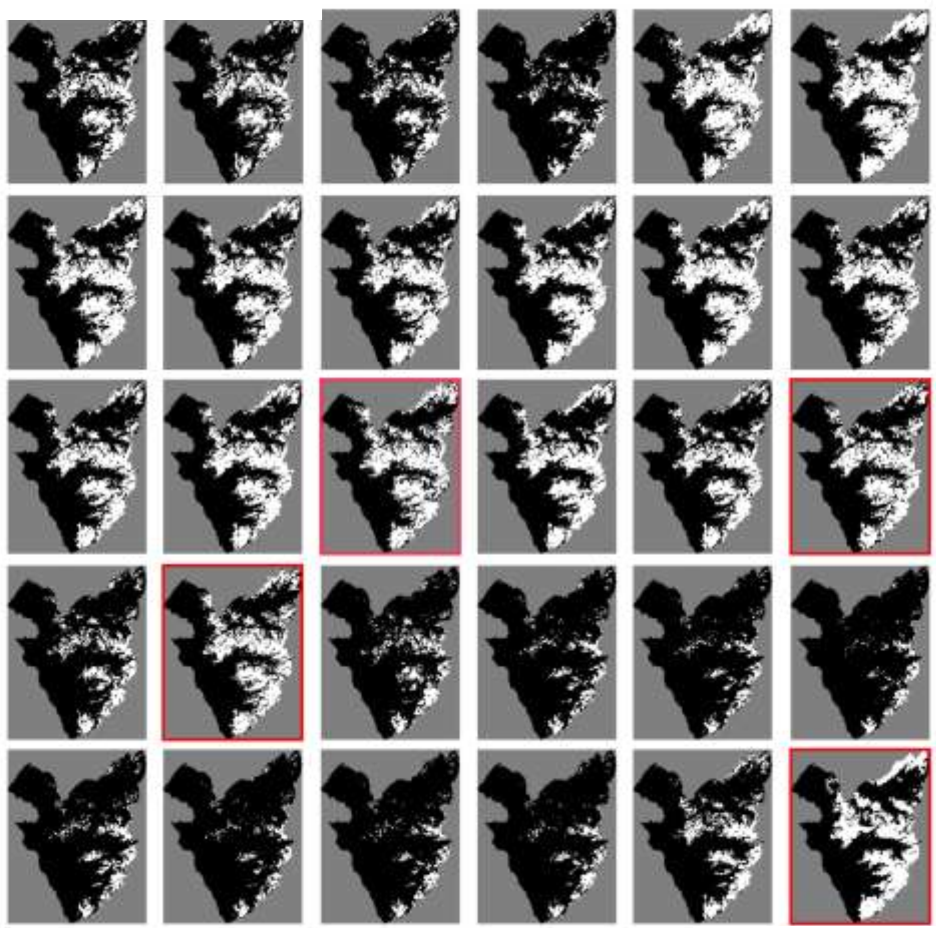


Western Swiss Alps

Results

- Optimal parameters:
  k=11
  window=60 days
  (interpretable)
- One month daily snow cover in the Western Swiss Alps.
- April: spring melt with occasional snowfall,
- Realistic transitions reproduced,
- Real Landsats in red.

2019/04/01



2019/04/30

Validation against ground stations

		Kappa					OA				
	Longitude/ Latitude	Satellite-Age		Pre-Satellite		Actual snow cover	Satellite-Age		Pre-Satellite		Actual snow cover
WSA		1 km	11 km	1 km	11 km		1 km	11 km	1 km	11 km	
La Comballaz	7.08/46.38	**0.82**	**0.82**	0.81	0.80	0.86	91.99	**92.22**	91.32	90.95	93.95
Les Diablerets	7.17/46.35	**0.74**	0.71	0.71	0.66	0.78	**89.47**	88.62	88.60	86.61	91.69
Leysin	7.02/46.35	**0.76**	0.73	0.72	0.66	0.79	**90.77**	90.12	89.43	87.72	92.7
Château-d'Oex	7.14/46.48	**0.78**	0.78	0.76	0.73	0.89	91.79	**92.16**	91.20	90.25	96.47
MeanAll		0.77	0.76	0.75	0.71	0.83	91.00	90.78	90.14	88.88	93.70
TJS											
Degersheim	9.19/47.36	**0.67**	0.63	0.59	0.59	0.86	**91.51**	89.91	89.54	88.75	96.69
Mogelsberg	9.14/47.36	0.43	**0.50**	0.34	0.33	0.73	90.21	**91.16**	89.43	89.47	95.03
St. Peterzell	9.17/47.32	**0.42**	0.41	0.39	0.39	0.37	**86.30**	85.52	85.86	85.33	88.95
Unterwasser Iltios	9.31/47.18	0.65	0.69	0.68	**0.71**	0.77	84.07	85.83	85.64	**86.75**	90.61
MeanAll		0.54	0.56	0.50	0.50	0.68	88.02	88.10	87.62	87.57	92.82

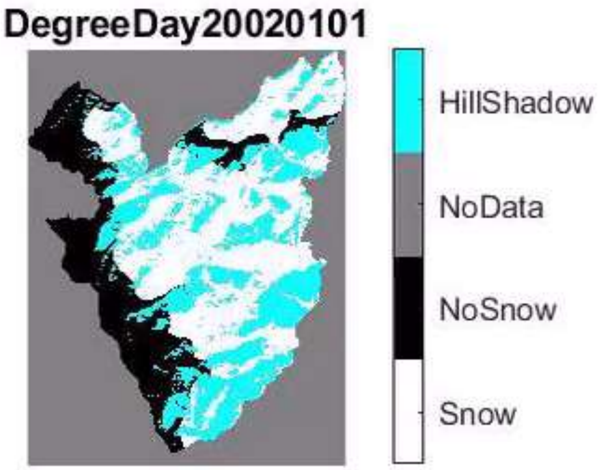
Resolution of predictors has little influence!

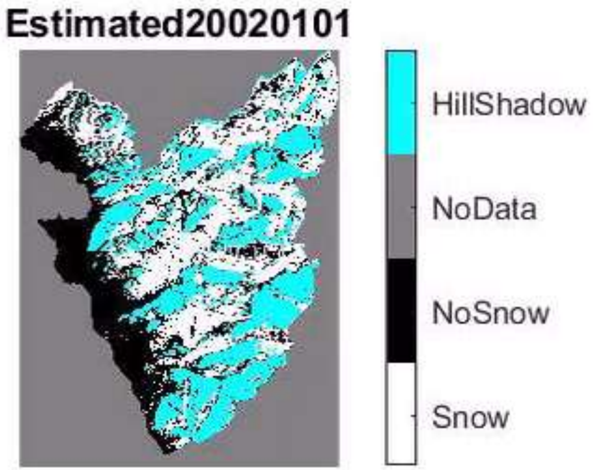
Comparison with a degree-day snow model

Method	*Overall Accuracy*
kNN vs actual snow cover	93.84%
degree-day vs actual snow cover	88.20%
kNN vs degree-day	88.63%

Also validated against high-resolution Planet Lab images, other snow reanalysis products

➔ Generally almost as good as real Landsat images



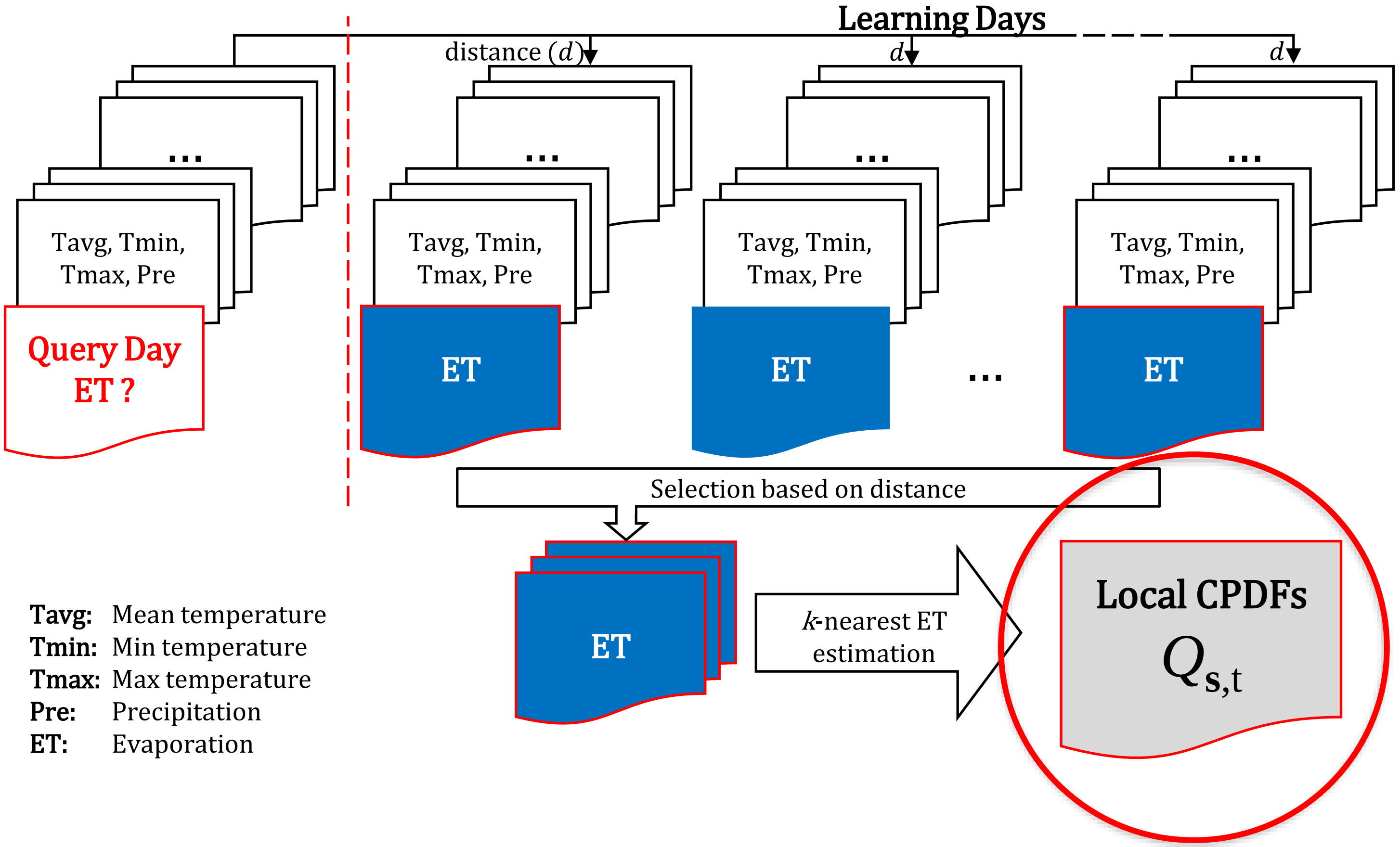


Nice but…

- Advantages:
  - Ultrafast because the distance is only computed on low-resolution predictors
  - kNN brings dependance to multiple predictors and complex temporal dependance
  - Naturally deals with non-stationarity
  - Non-parametric (no need to fit a model)
  - Resampling-based → cannot produce extreme values, but can generate succession of sub-extremes
- Drawbacks:
  - Over-smooth (high K → high smoothness)
  - The temporal covariance is solely driven by the predictors
  - No uncertainty quantification
  - Non-parametric (stuck with historical data, hard to represent extreme values)
  - Requires lots of data (but who doesn't?)

A semi-parametric kNN-based approach

Hybrid parametric / non-parametric approach



A hybrid approach

A generator whereby, for each day to simulate:

- The kNN approach is used to select k nearest candidates based on predictors.
- A conditional distribution $Q_{\mathbf{s},\mathrm{t}}$ is inferred locally based on the k candidates.
- A latent Gaussian field approach is then used to simulate the target variable $Z(\mathbf{s},\mathrm{t})$ based on $Q_{\mathbf{s},\mathrm{t}}$
- A Gneiting space-time covariance inferred on the entire training period.

In details

$\mathbf{s}$: space

t : time

Transformation function (space-time covariance + local cdf)

cumulative distribution function

Local covariate

$$Y(\mathbf{s},t) = \Psi_{\mathbf{s},t}\big(Z(\mathbf{s},t)\big) = Q_{\mathbf{s},t}\left(\Phi\left(Z\left(\mathbf{s},t\right)\right) \mid \mathbf{X}\left(\mathbf{s},t\right) = \mathbf{x}\right)$$

Target variable

Gaussian space-time RF (geniting)

conditional quantile function

Simulation

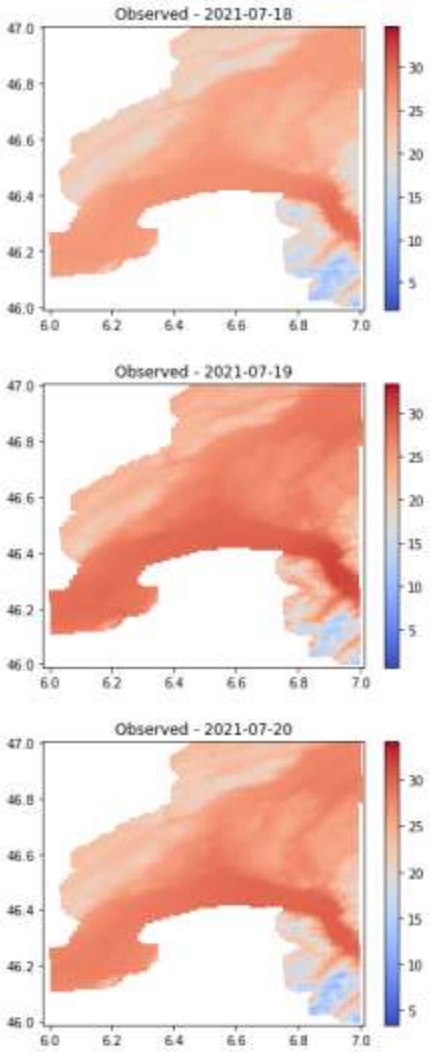
Estimation of the Gneiting covariance parameters by maximum likelihood. Pairwise likelihood is used.

Simulation in 2 parts:

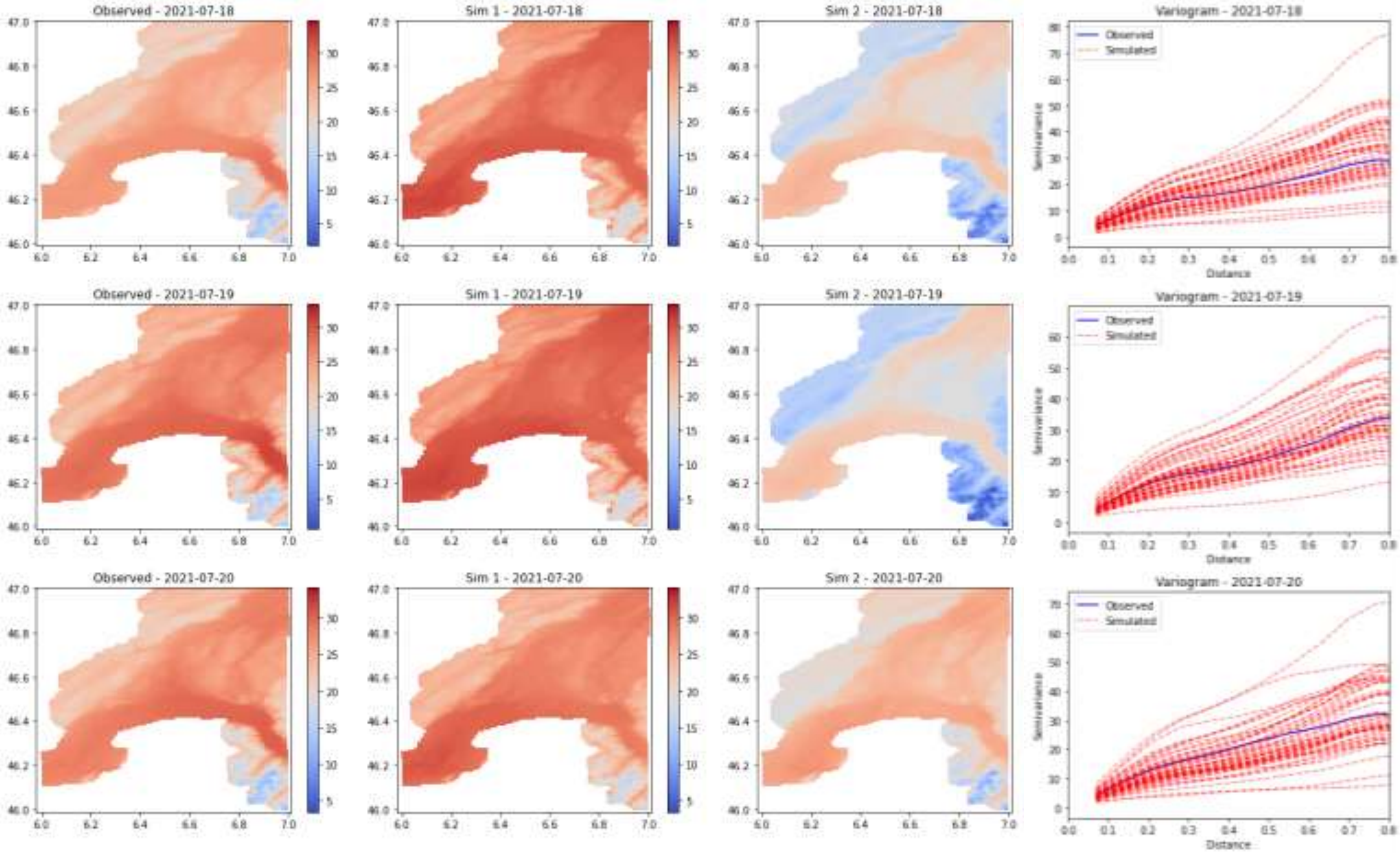
1. Simulation of the space-time GRF $Z(\mathbf{s},\mathrm{t})$
2. Local transformation $Z(\mathbf{s},\mathrm{t})$ of with the estimated $\Psi_{\boldsymbol{s},t}$

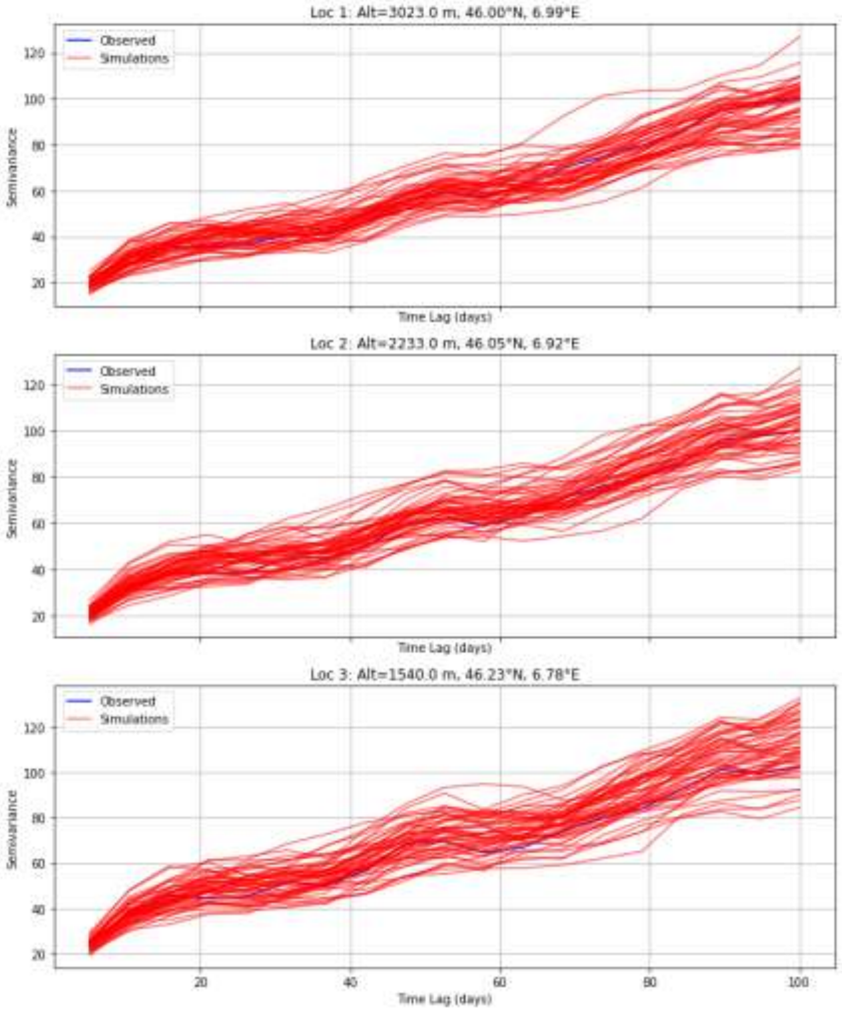
Test setting

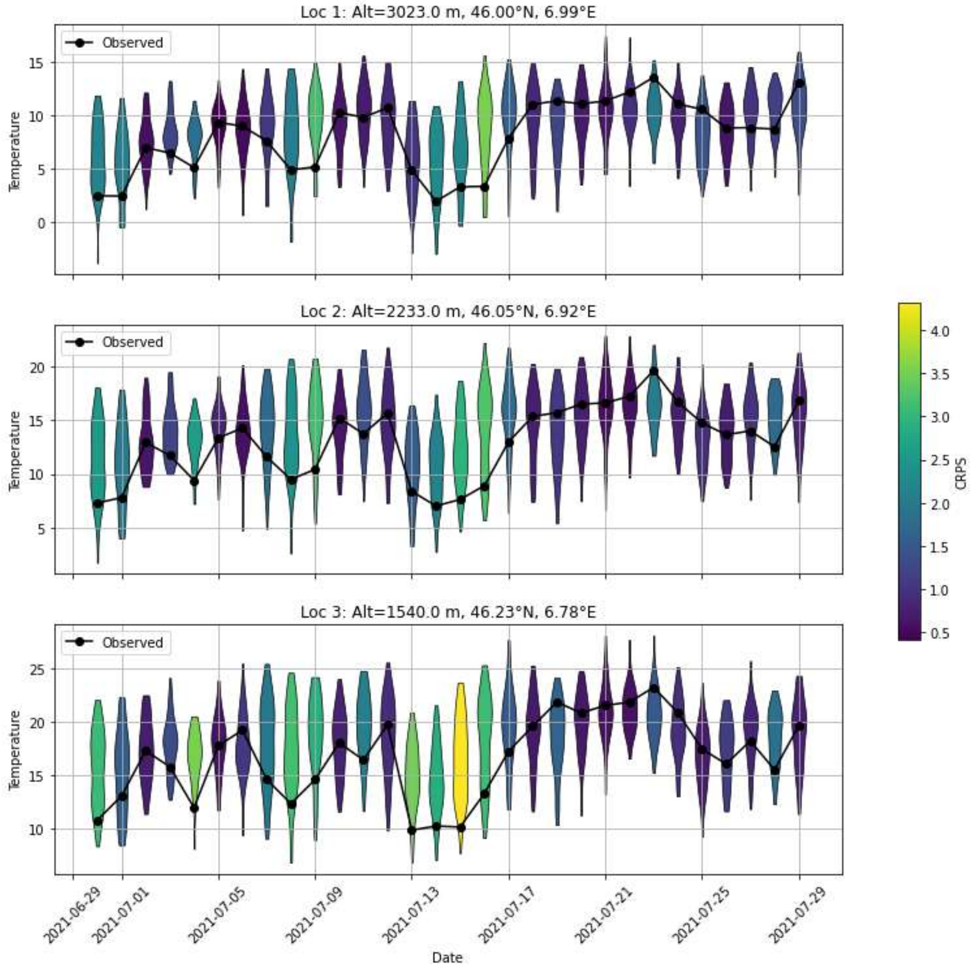
- Simulation of daily maximum temperature over western Switzerland
- Training: 1971-2019, daily, 1km.
- Based on Meteoswiss reanalysis
- Simulation: 2020-2022, daily, 1km.
- Single predictor: pressure (isopotential 500 hHa)



Some simulations




Loc 1: Alt=3023.0 m, 46.00°N, 6.99°E
Loc 2: Alt=2233.0 m, 46.05°N, 6.92°E
Loc 3: Alt=1540.0 m, 46.23°N, 6.78°E
Observed
Simulations
Semivariance
Time Lag (days)


Loc 1: Alt=3023.0 m, 46.00°N, 6.99°E
Loc 2: Alt=2233.0 m, 46.05°N, 6.92°E
Loc 3: Alt=1540.0 m, 46.23°N, 6.78°E
Observed
Temperature
Date
CRPS

Conclusions / takeaways

- Synthesizing data to overcome the limitations of remote sensing data regarding spatial coverage, spatial coverage, or for data fusion.
- Possible for the past, present, and possibly future.
- Requirement: having large amounts of data to resample (=learning period).
- Low-resolution predictors perform well, because the temporal patterns allow selecting candidates.
- Potentially global use. GEE implementation to come.
- Potential to generate entire synthetic multispectral images rather than derivatives, however the distance chosen should be application-specific.
- Climate predictors do not allow accounting for human-induced effects.
- kNN is not the only way: e.g. generative models.


Thank you
Questions?