Obtain original web presentation here:

https://slides.com/odineidolon/chym2017-8/fullscreen#/

This PDF version is of lower quality

HOW TO DEAL WITH OBSERVATIONAL DATA FOR (HYDROLOGICAL) MODELLING PURPOSES

ADRIANO FANTINI

ICTP, Trieste, Italy

afantini@ictp.it

Online presentation: https://bit.ly/2MaOjrP

Which observations do you need for hydrology?

- Precipitation (possibly hourly, esp. for small basins)
- Temperature
- Snow
- Elevation data
- Land use data
- Discharge / water stage

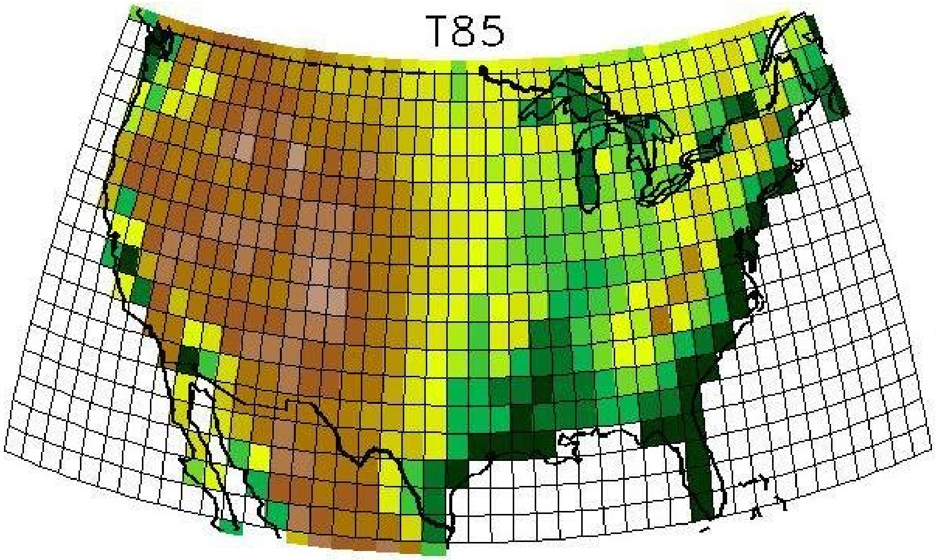
Dense or sparse?

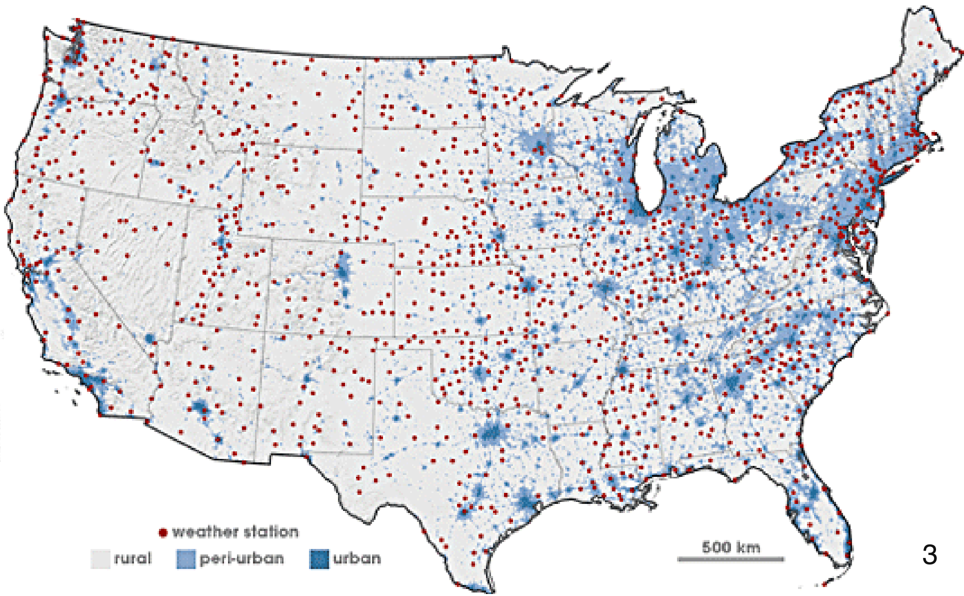
Gridded:

- Precipitation
- Temperature
- Elevation
- Land use

In-situ:

- Precipitation
- Temperature
- Discharge
- Water stage





Gridded

Advantages

- Uniform availability, often global
- Compare easily with models
- Generally straighforward formats (e.g. NetCDF)
- Efficient processing
- Different variables on the same grid
- Usually quality-controlled

Disadvantages

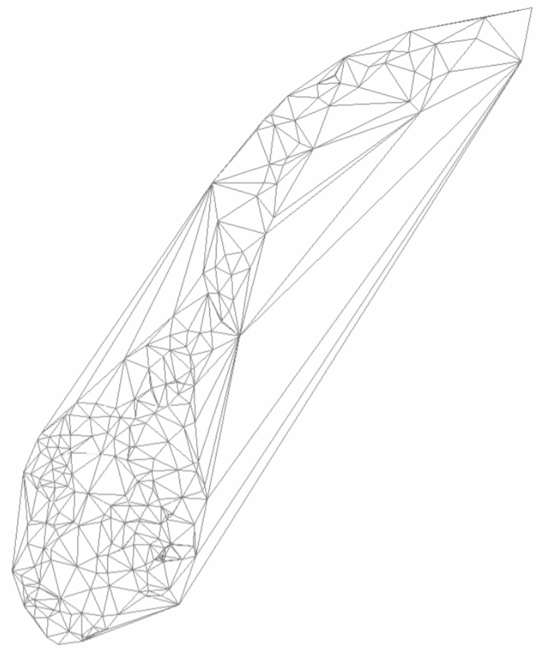
- Heavily dependent on gridding method
- Not suitable for comparison over specific points
- Usually derived from in-situ data
- Dataset resolution != actual resolution (!!!)

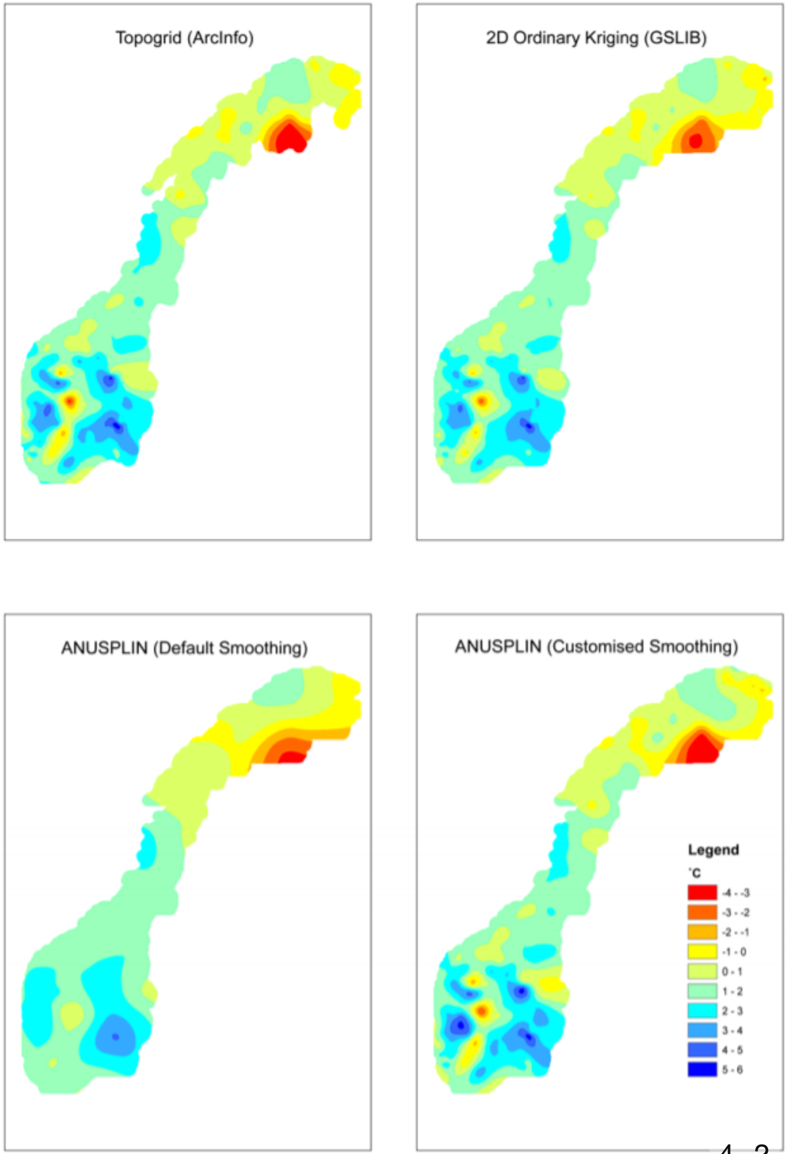
Gridding methods

Basic categories:

- Inverse Distance Weighting
- Kriging
- Spline Interpolation
- Surface polygons

CAN HAVE DIFFERENT RESULTS!





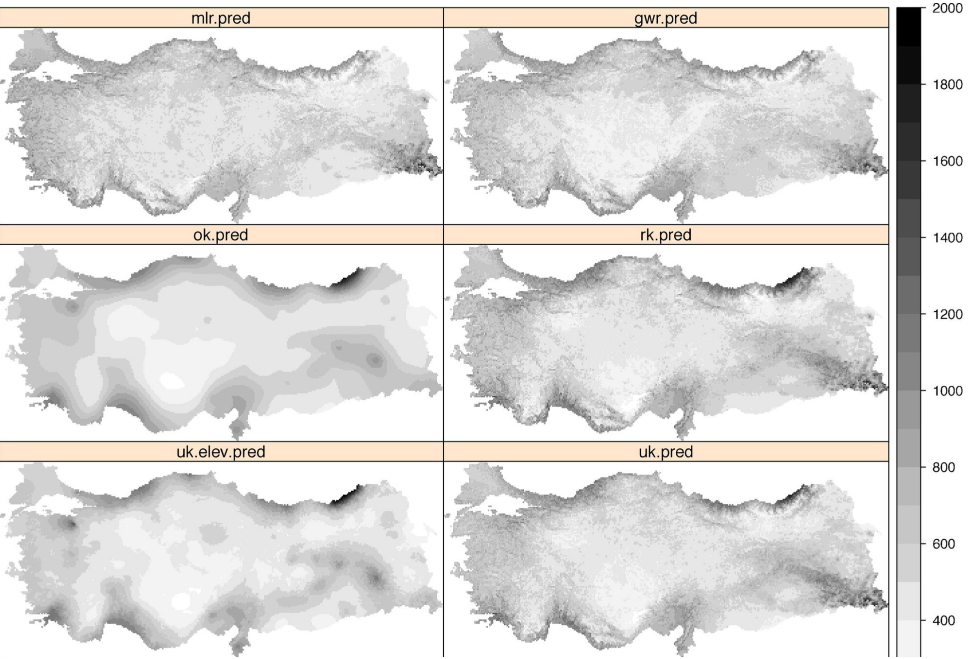
Mohr, 2008

Gridding methods

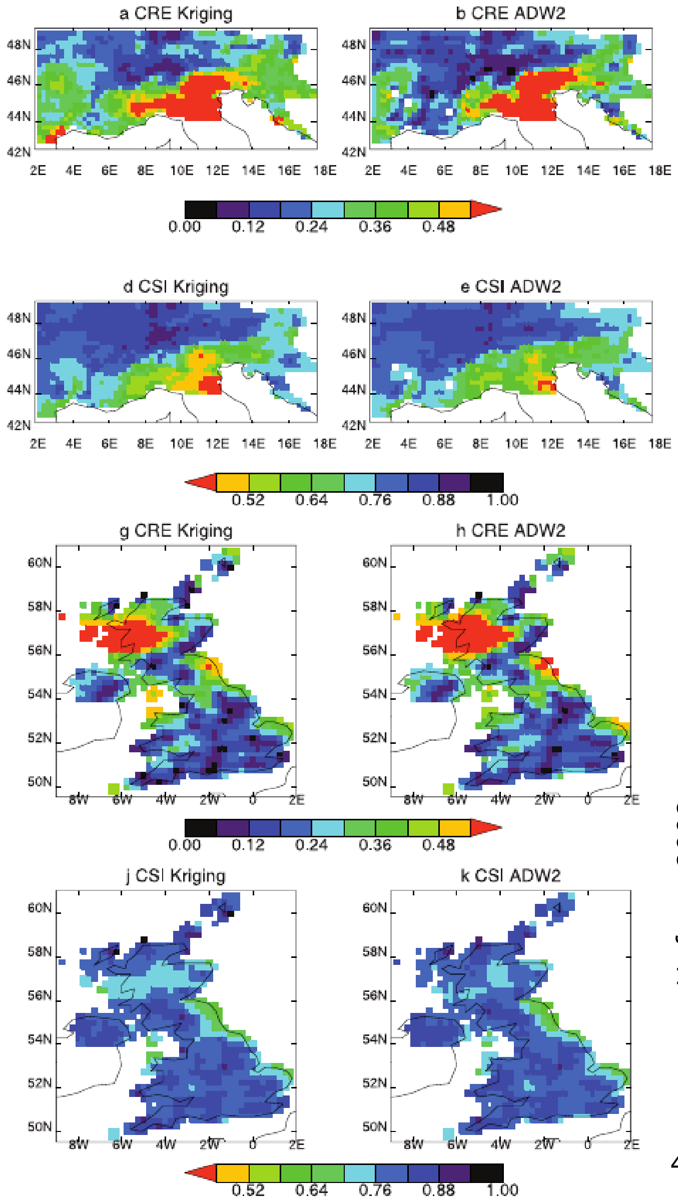
Basic categories:

- Inverse Distance Weighting
- Kriging
- Spline Interpolation
- Surface polygons

CAN HAVE DIFFERENT RESULTS!



Bostan, 2012



Hofstra, 2008

In-situ

Advantages

- No gridding/smoothing -> good for extremes
- Easy to compare with models (e.g. discharge at a given point)
- Do not hide anything from the user
- Dataset resolution == actual resolution
- Metadata!

Disadvantages

- Scarse data availability
- Often in very weird formats
- Often lacking quality control
- Hard to compare with gridded (e.g. climate) models (PR, T)

Common problems with in-situ measurements

Temporal and spatial problems:

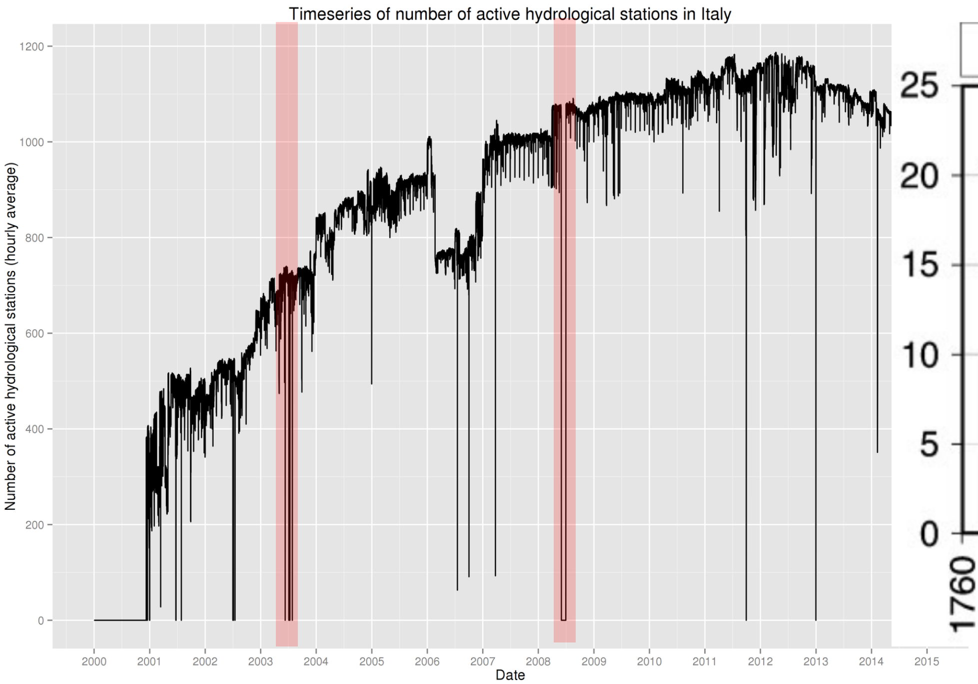
- Short timescale
- Missing periods
- Low station density
- Missing timesteps

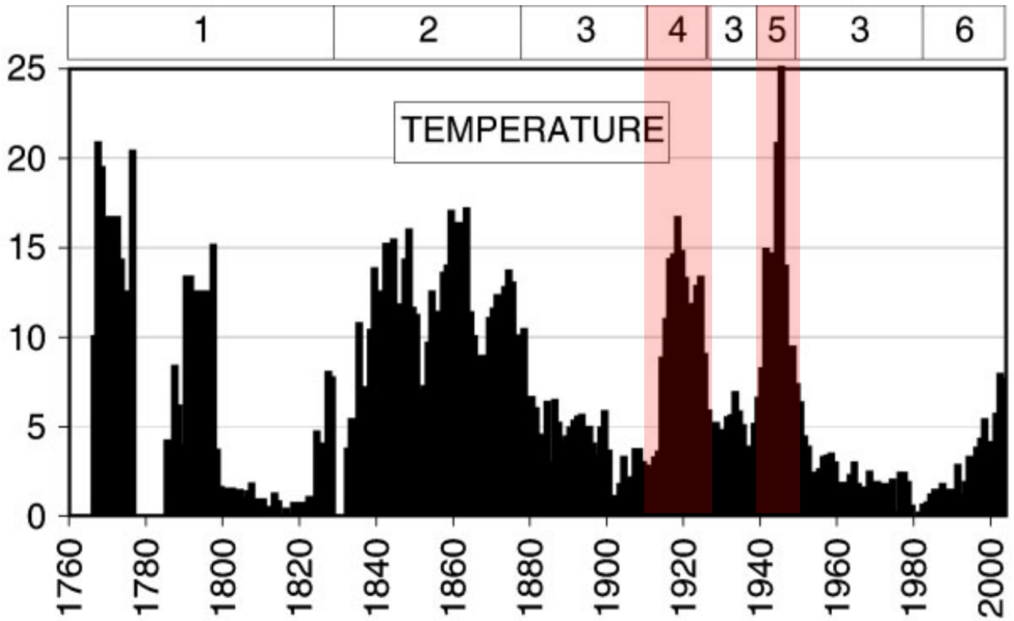
Data quality problems:

- Breaks and inhomogeneities
- Manual measurement errors
- Equipment errors and failures
- Weather-related measurement errors

Temporal and spatial problems

- Short timescale
- Low station density
- Missing timesteps
- Missing periods

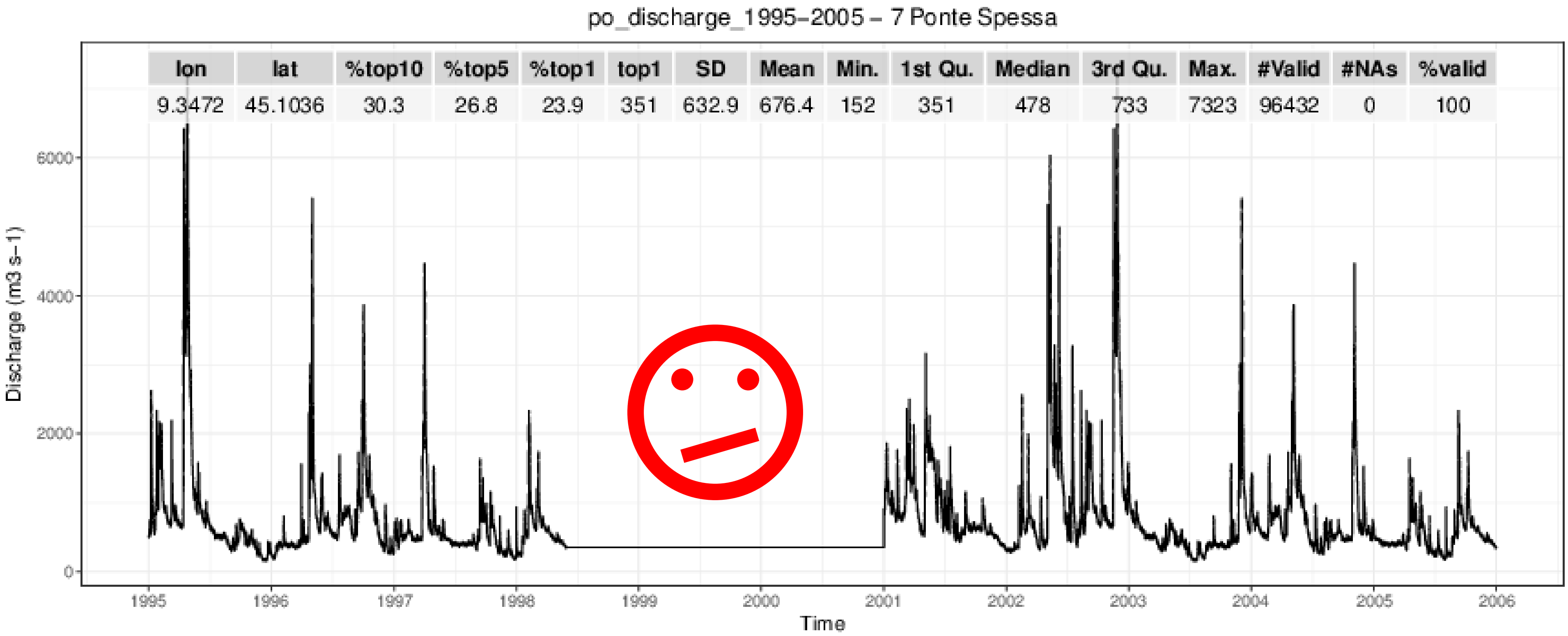




HISTALP database, Bohm et al., 2013

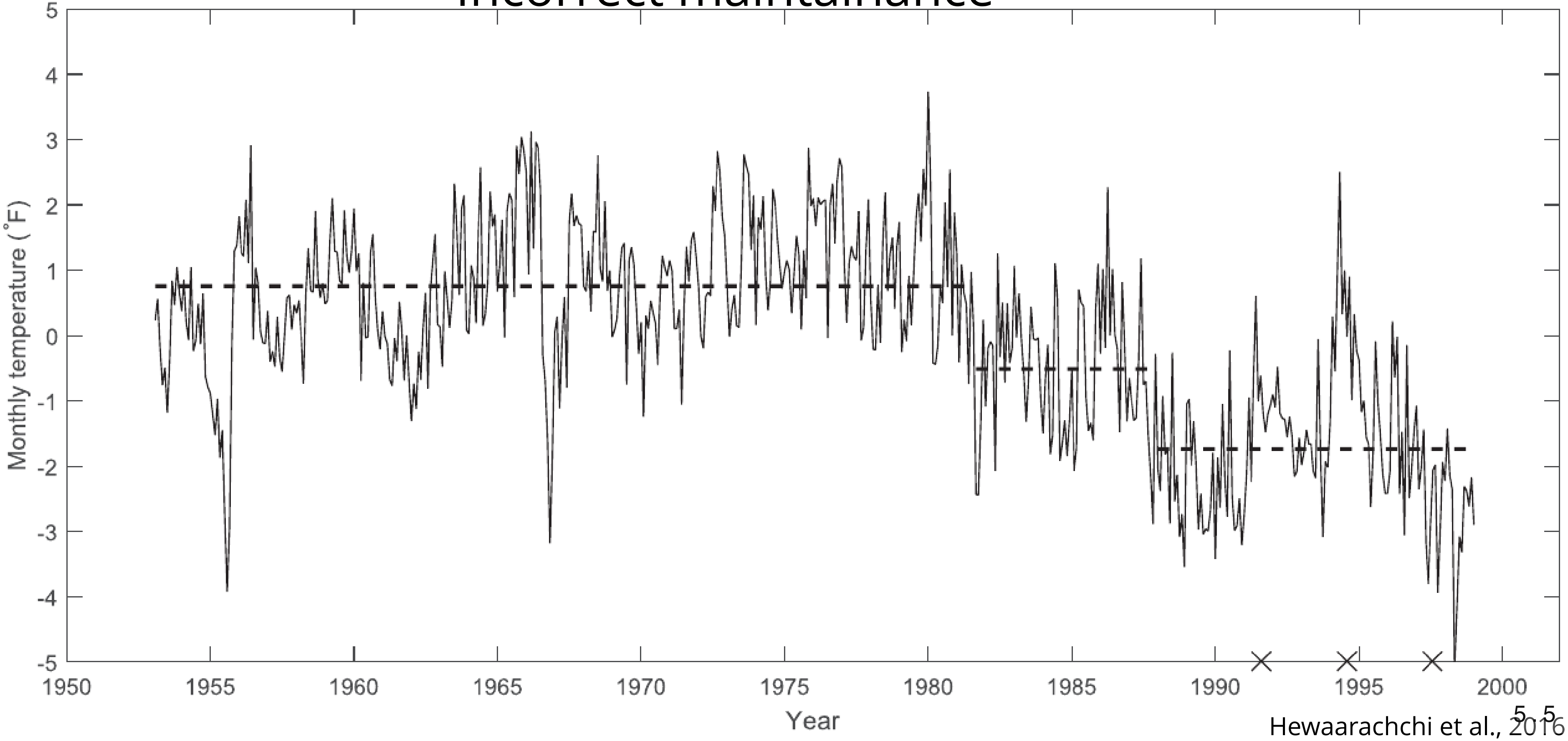
Data quality problems

- Manual measurement errors
- Equipment errors and failures



Inhomogeneities

- Changes in measurement time
- Station relocations
- Instrumentation upgrades
- Incorrect maintainance



Hewaarachchi et al., 2016

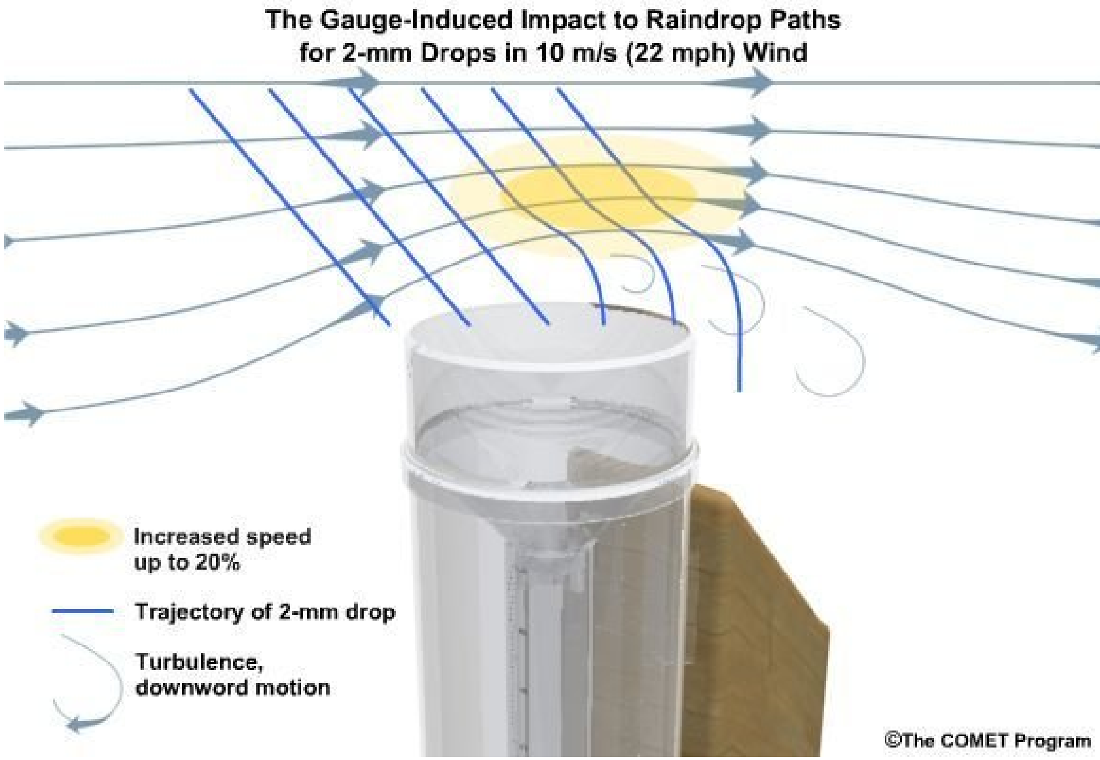
Data quality problems

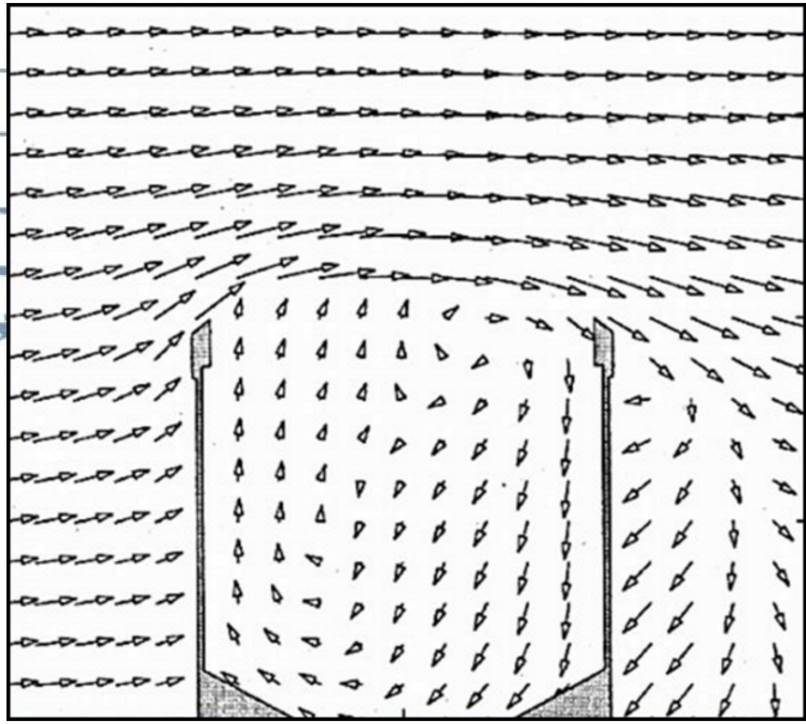
Measurement errors due to:

- Sensor icing
- Lack of power
- Vandalism
- Lack of maintenance
- Gauge undercatch

Data acquisition problems

EXAMPLE: PRECIPITATION GAUGE UNDERCATCH



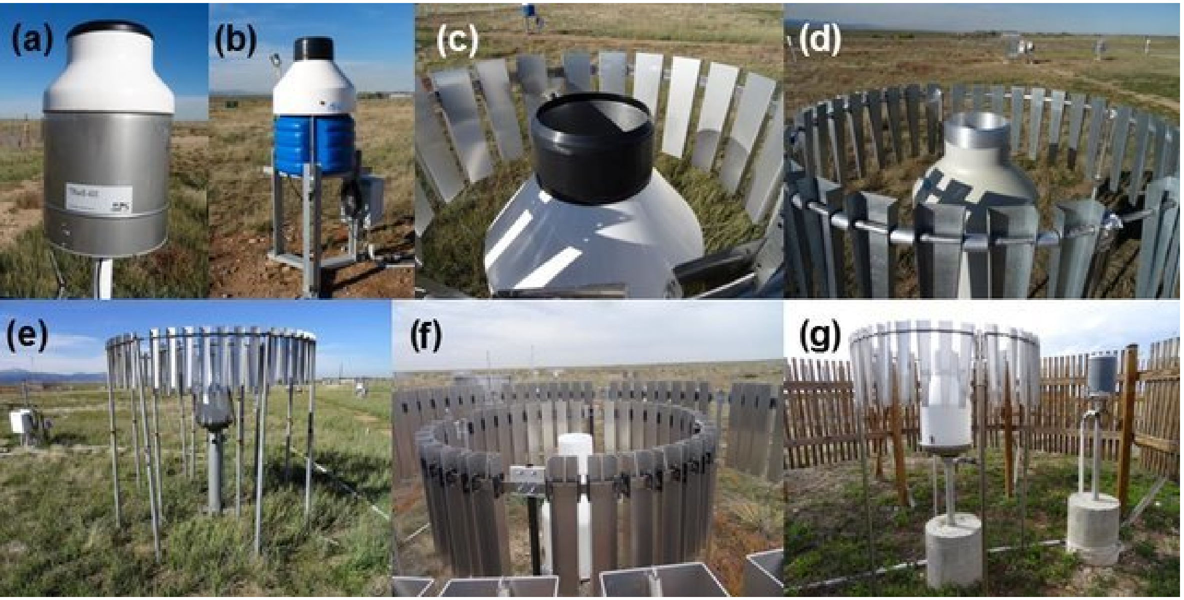


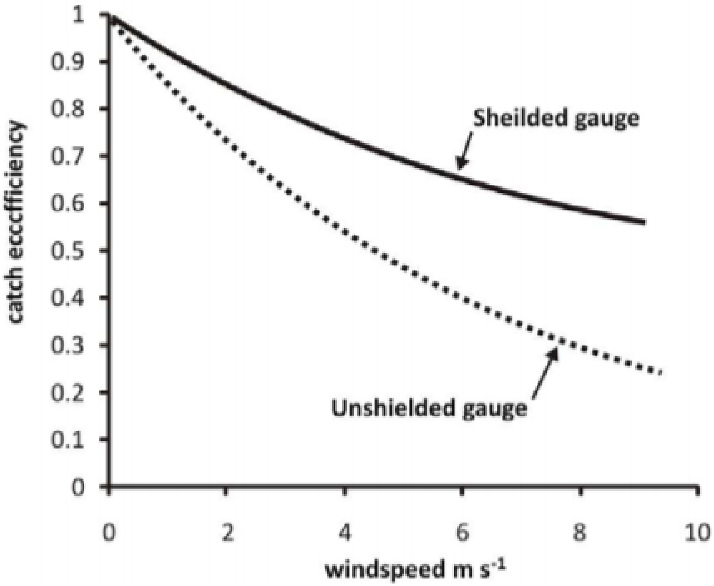
Nespor and Sevruk, 1999

- on average ~30% ?
- ~80% for extreme winter solid events?

Data acquisition problems

EXAMPLE: PRECIPITATION GAUGE UNDERCATCH





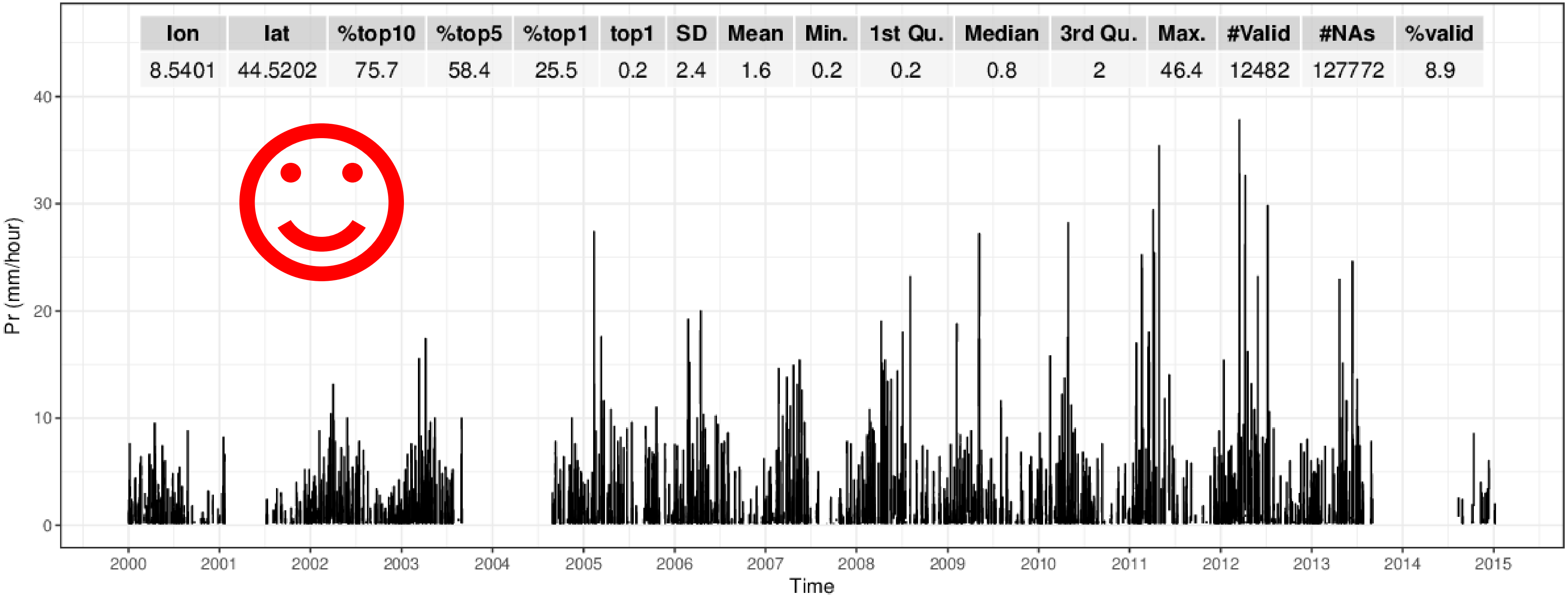
Macdonald and Pomeroy, 2008

An in-situ example

- Precipitation
- Hourly
- From different institutions
- ~2200 stations on average
- uneven spatial coverage
- 2000-2016

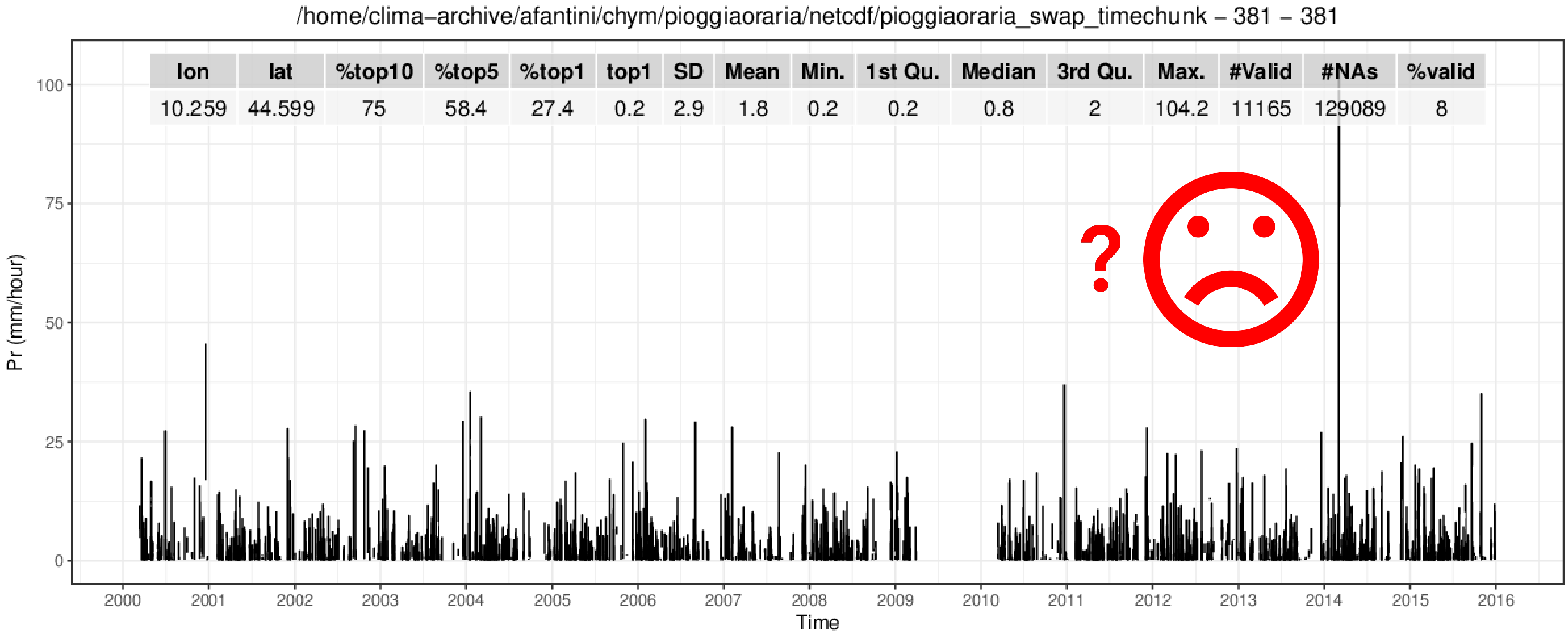
/home/clima-archive/afantini/chym/pioggiaoraria/netcdf/pioggiaoraria_swap_timechunk – 167 – 167

lon	lat	%top10	%top5	%top1	top1	SD	Mean	Min.	1st Qu.	Median	3rd Qu.	Max.	#Valid	#NAs	%valid
8.5401	44.5202	75.7	58.4	25.5	0.2	2.4	1.6	0.2	0.2	0.8	2	46.4	12482	127772	8.9

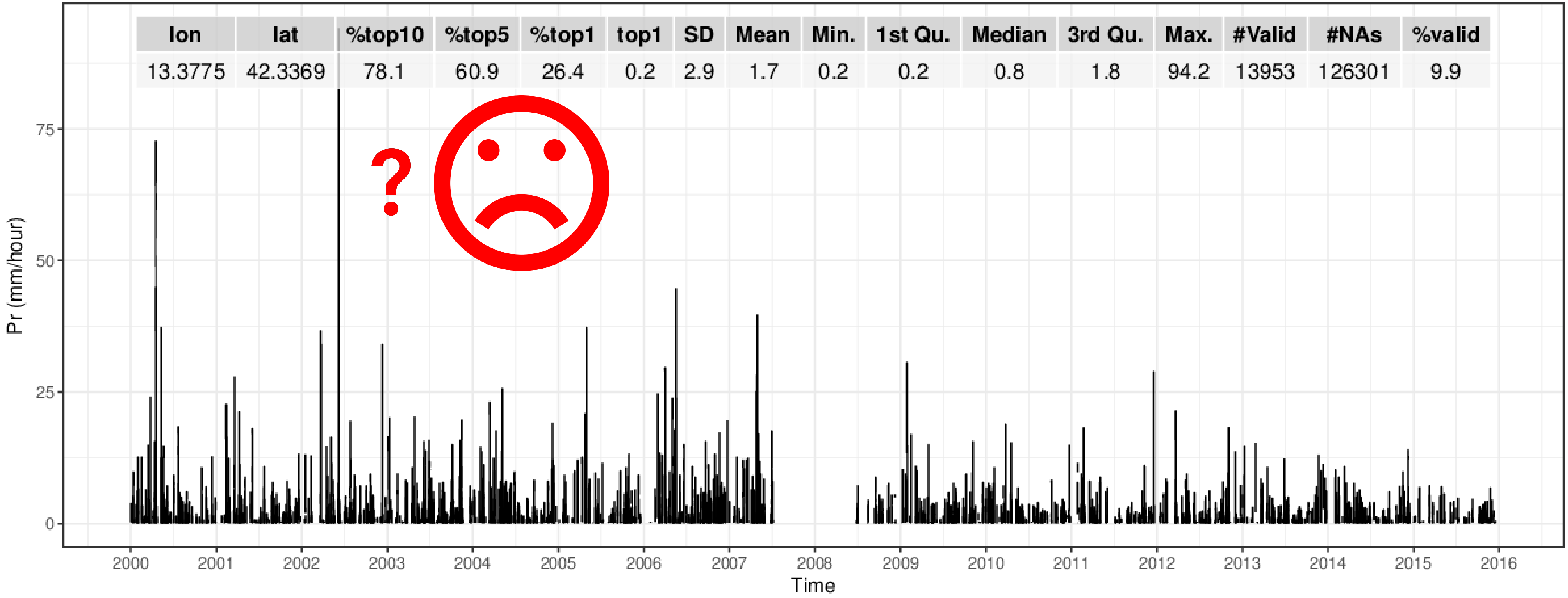


/home/clima-archive/afantini/chym/pioggiaoraria/netcdf/pioggiaoraria_swap_timechunk – 218 – 218

lon	lat	%top10	%top5	%top1	top1	SD	Mean	Min.	1st Qu.	Median	3rd Qu.	Max.	#Valid	#NAs	%valid
13.3133	42.3803	87	70.6	32.4	0.2	20	1.7	0.2	0.2	0.6	1.4	822	8022	132232	5.7


/home/clima-archive/afantini/chym/pioggiaoraria/netcdf/pioggiaoraria_swap_timechunk – 381 – 381
lon lat %top10 %top5 %top1 top1 SD Mean Min. 1st Qu. Median 3rd Qu. Max. #Valid #NAs %valid
10.259 44.599 75 58.4 27.4 0.2 2.9 1.8 0.2 0.2 0.8 2 104.2 11165 129089 8
Pr (mm/hour)
Time
?

/home/clima-archive/afantini/chym/pioggiaoraria/netcdf/pioggiaoraria_swap_timechunk – 248 – 248



lon	lat	%top10	%top5	%top1	top1	SD	Mean	Min.	1st Qu.	Median	3rd Qu.	Max.	#Valid	#NAs	%valid
13.3775	42.3369	78.1	60.9	26.4	0.2	2.9	1.7	0.2	0.2	0.8	1.8	94.2	13953	126301	9.9

What can we do based on this?

- Cut outliers over a given fixed threshold
- Variable threshold based on SD or IQR
- Remove consecutive suspicious values

What can we do based on this?

- Cut outliers over a given fixed threshold
- Variable threshold based on SD or IQR
- Remove consecutive suspicious values

Timeseries are usually not enough to identify inhomogeneities and errors

What can we do based on this?

- Cut outliers over a given fixed threshold
- Variable threshold based on SD or IQR
- Remove consecutive suspicious values

Timeseries are usually not enough to identify inhomogeneities and errors

Metadata

What can we do based on this?

- Cut outliers over a given fixed threshold
- Variable threshold based on SD or IQR
- Remove consecutive suspicious values

Timeseries are usually not enough to identify inhomogeneities and errors

Metadata **Spatial analysis**

Metadata

Metadata

all the information that is not data itself

- Gauge type and characteristics
- Station history (relocations, upgrades...)
- Recorded changes in station environment
- News about extreme events (hard to find for old data)

Metadata

all the information that is not data itself

- Gauge type and characteristics
- Station history (relocations, upgrades...)
- Recorded changes in station environment
- News about extreme events (hard to find for old data)

WE OFTEN DO NOT HAVE ACCESS TO THIS,
AND IT'S EXTREMELY TIME CONSUMING

Spatial analysis

Spatial analysis

Maps + comparison to neighbouring stations

- Can be automated, once a criterion is chosen
- Possibilities for choosing reference stations: nearest neighbours, distance radius, height range, high correlation...

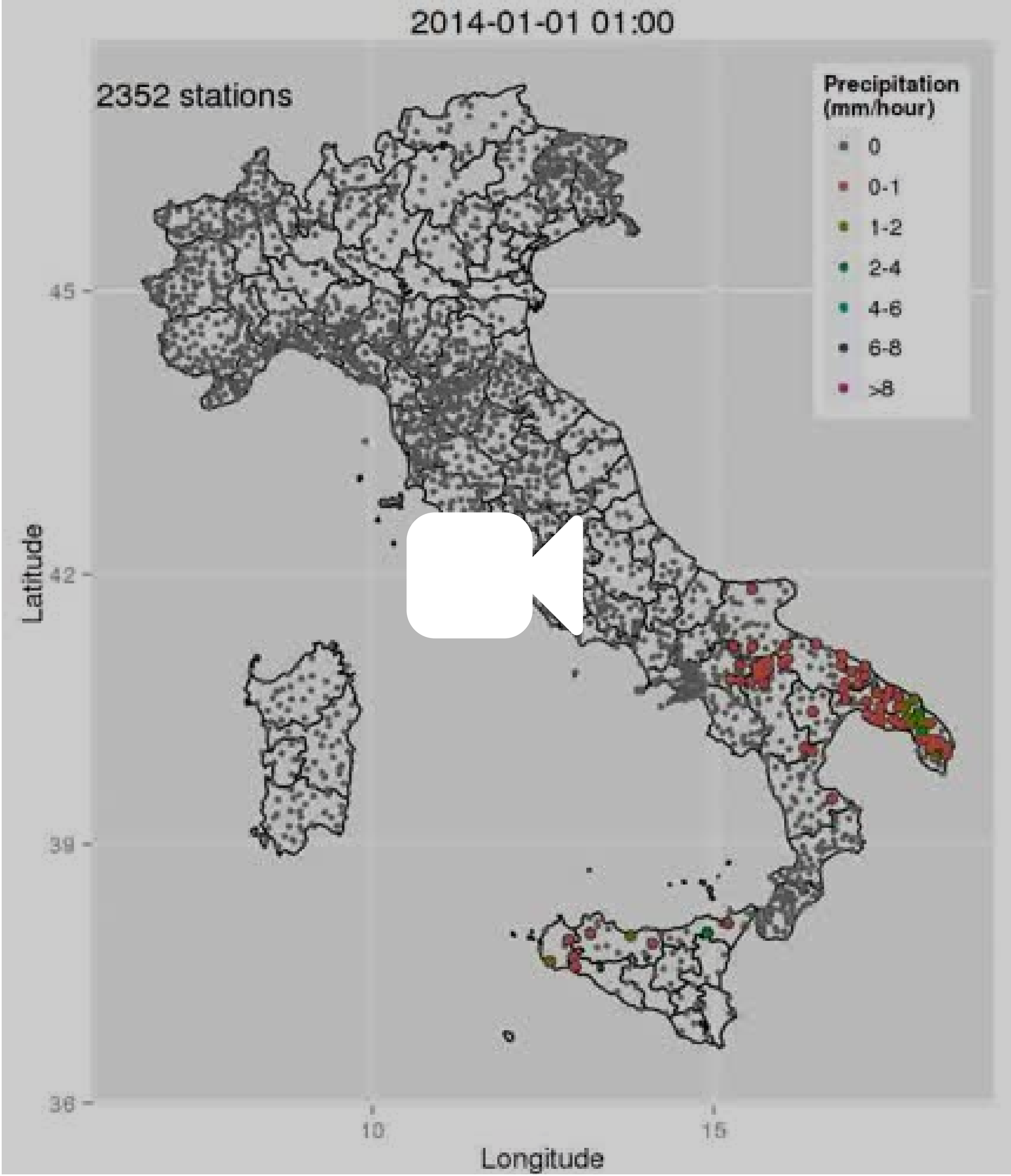
Spatial analysis

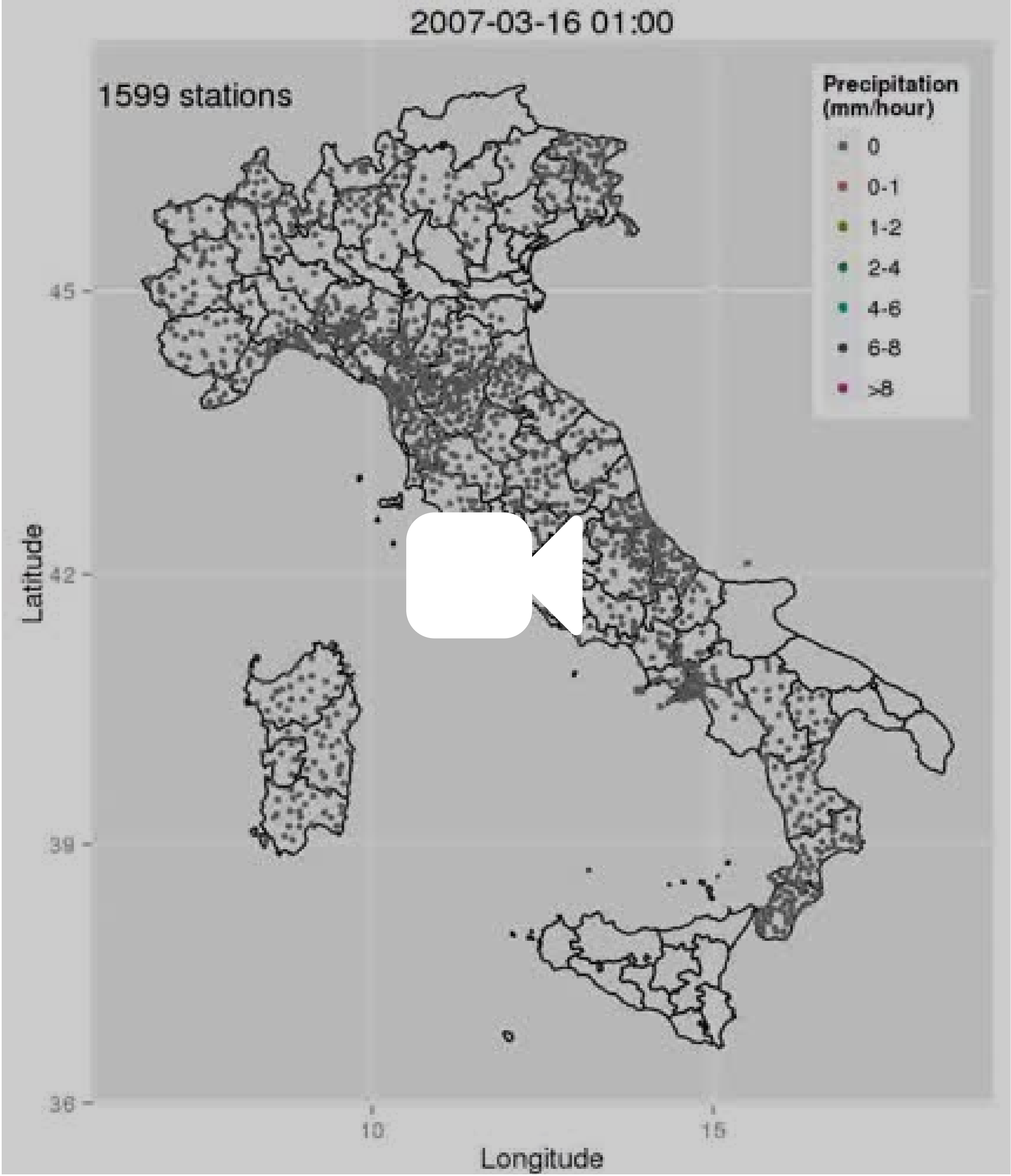
Maps + comparison to neighbouring stations

- Can be automated, once a criterion is chosen
- Possibilities for choosing reference stations: nearest neighbours, distance radius, height range, high correlation...

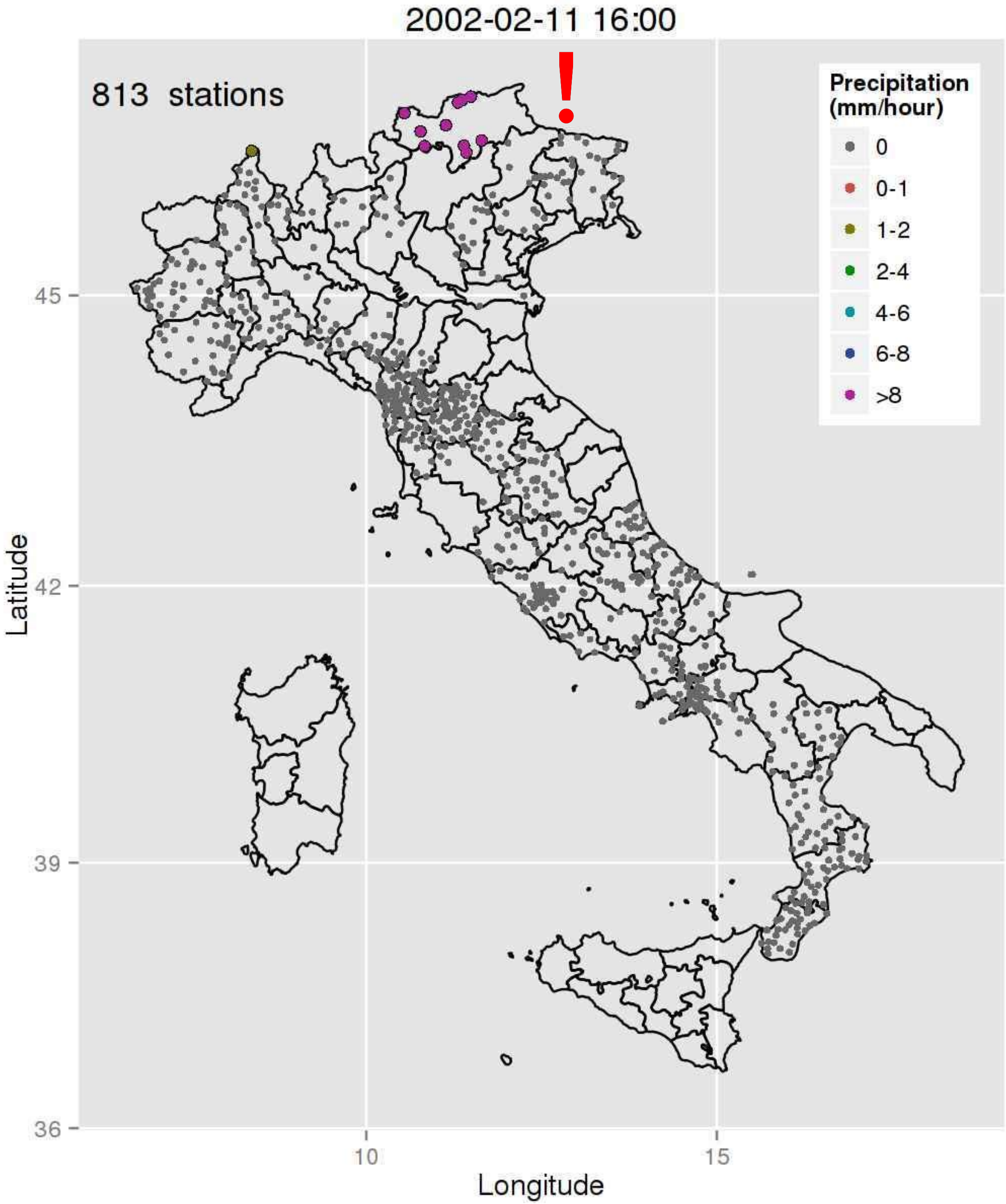
REQUIRES HIGH ENOUGH STATION DENSITY

HARD TO DO ON HIGHLY SPATIALLY VARIABLE FIELDS (e.g. PRECIPITATION) OR REGIONS (e.g. MOUNTAINS)

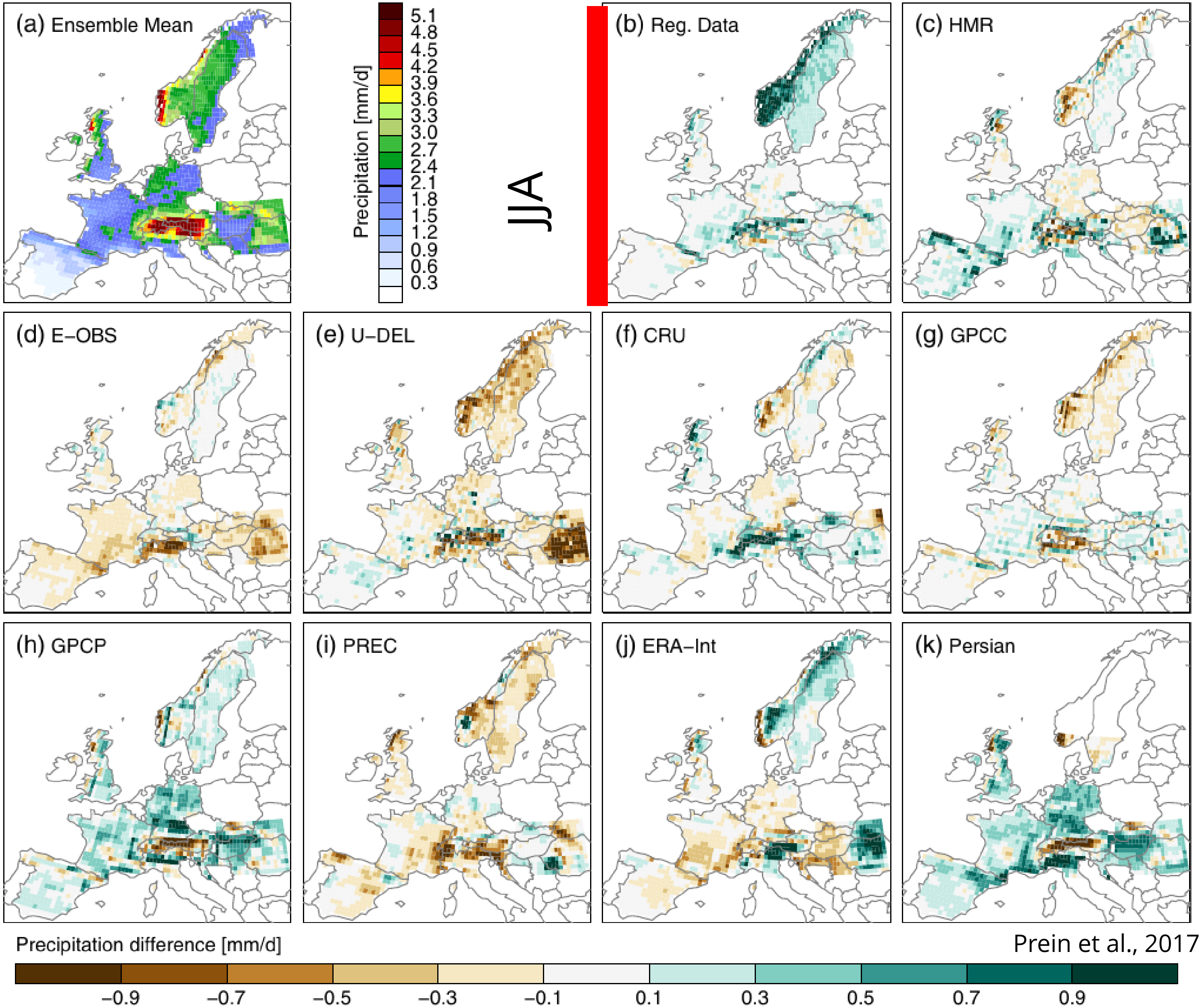

2014-01-01 01:00
2352 stations
Precipitation (mm/hour)
0
0-1
1-2
2-4
4-6
6-8
>8
Latitude
45
42
39
36
Longitude
10
15



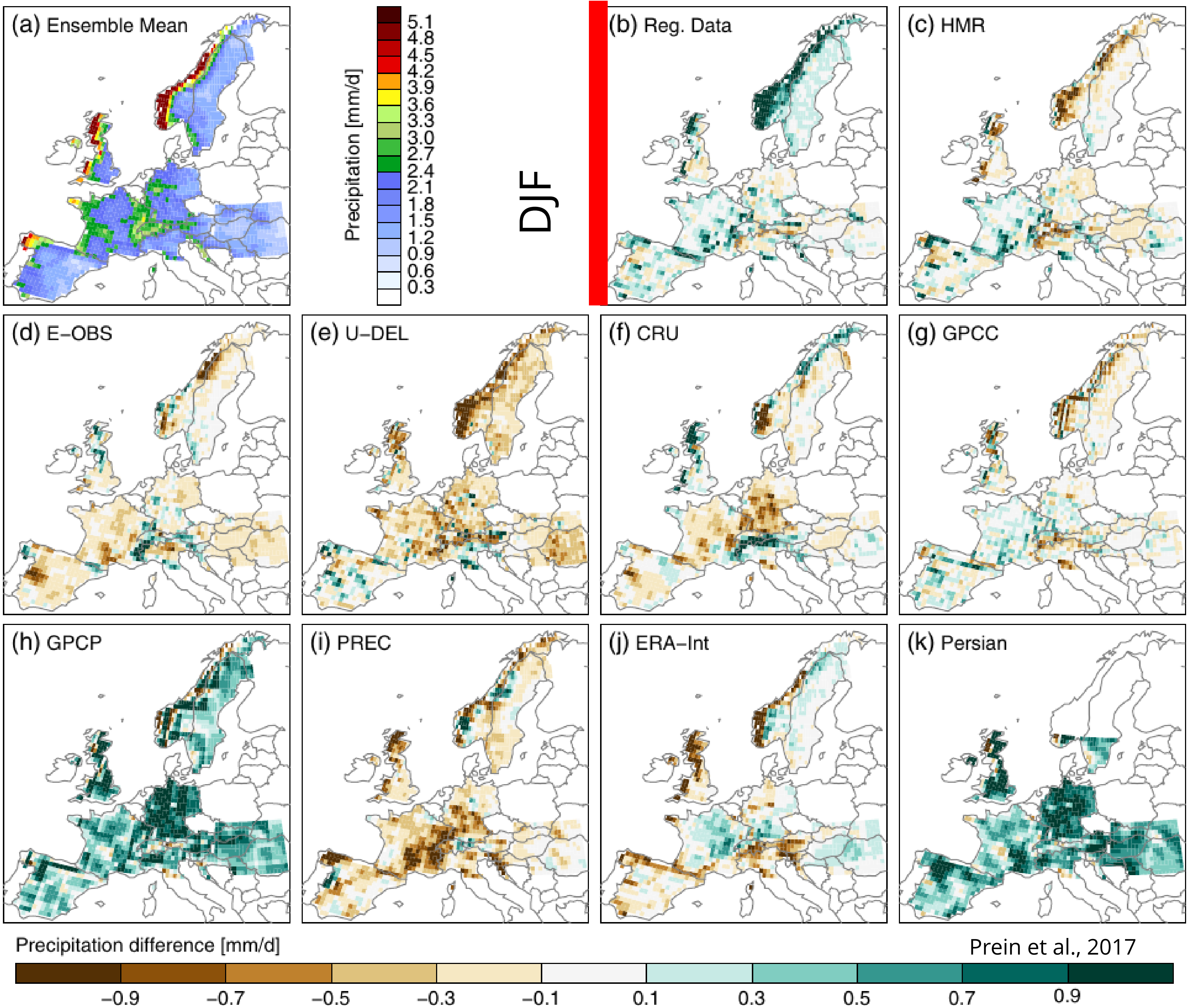
2007-03-16 01:00
1599 stations
Precipitation (mm/hour)
0
0-1
1-2
2-4
4-6
6-8
>8
Latitude
45
42
39
36
Longitude
10
15


2002-02-11 16:00
813 stations
!
Precipitation (mm/hour)
0
0-1
1-2
2-4
4-6
6-8
>8
45
42
39
36
10
15
Latitude
Longitude

Even after correction...

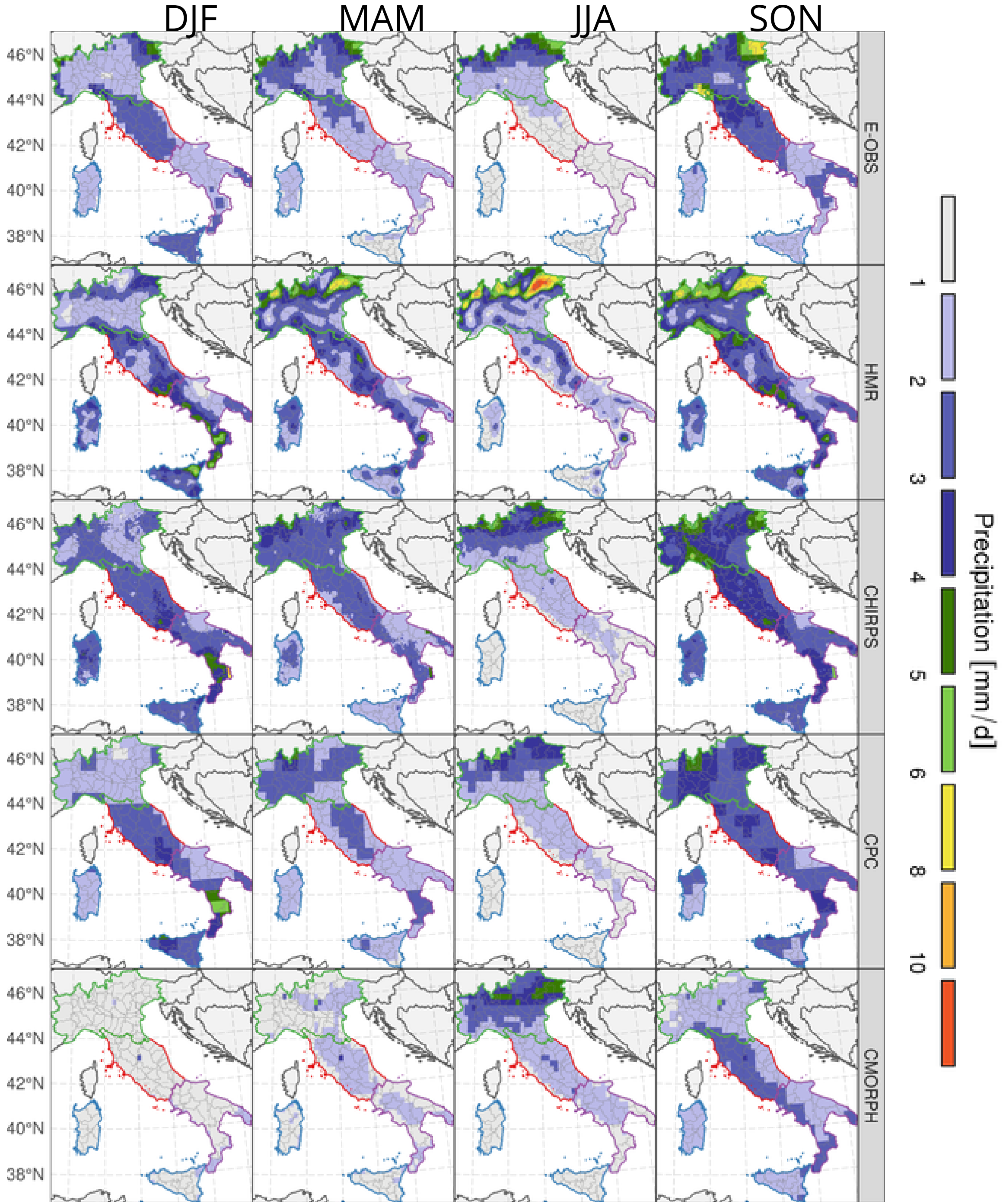


Prein et al., 2017

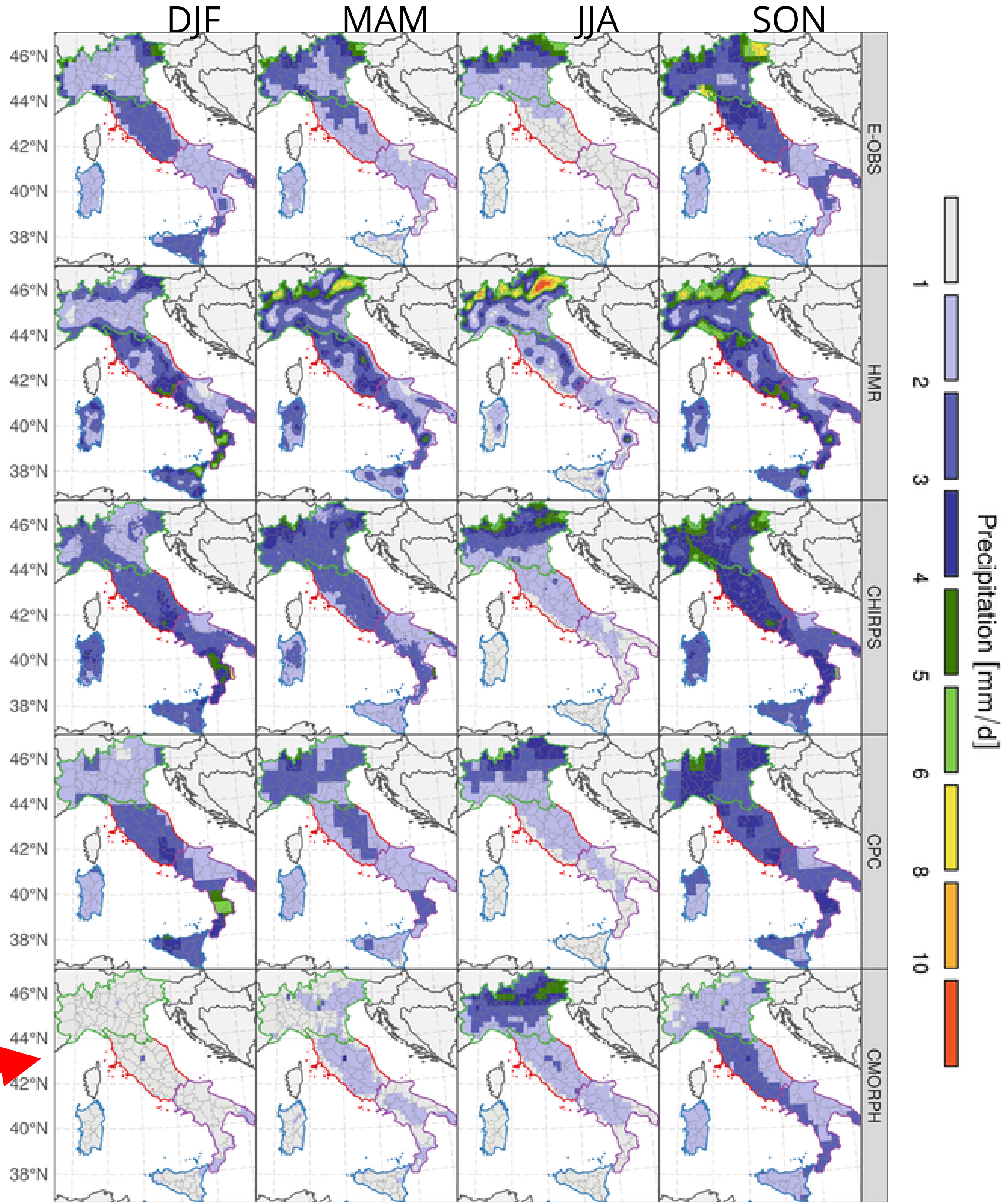


Prein et al., 2017

2001-2016
mean precip

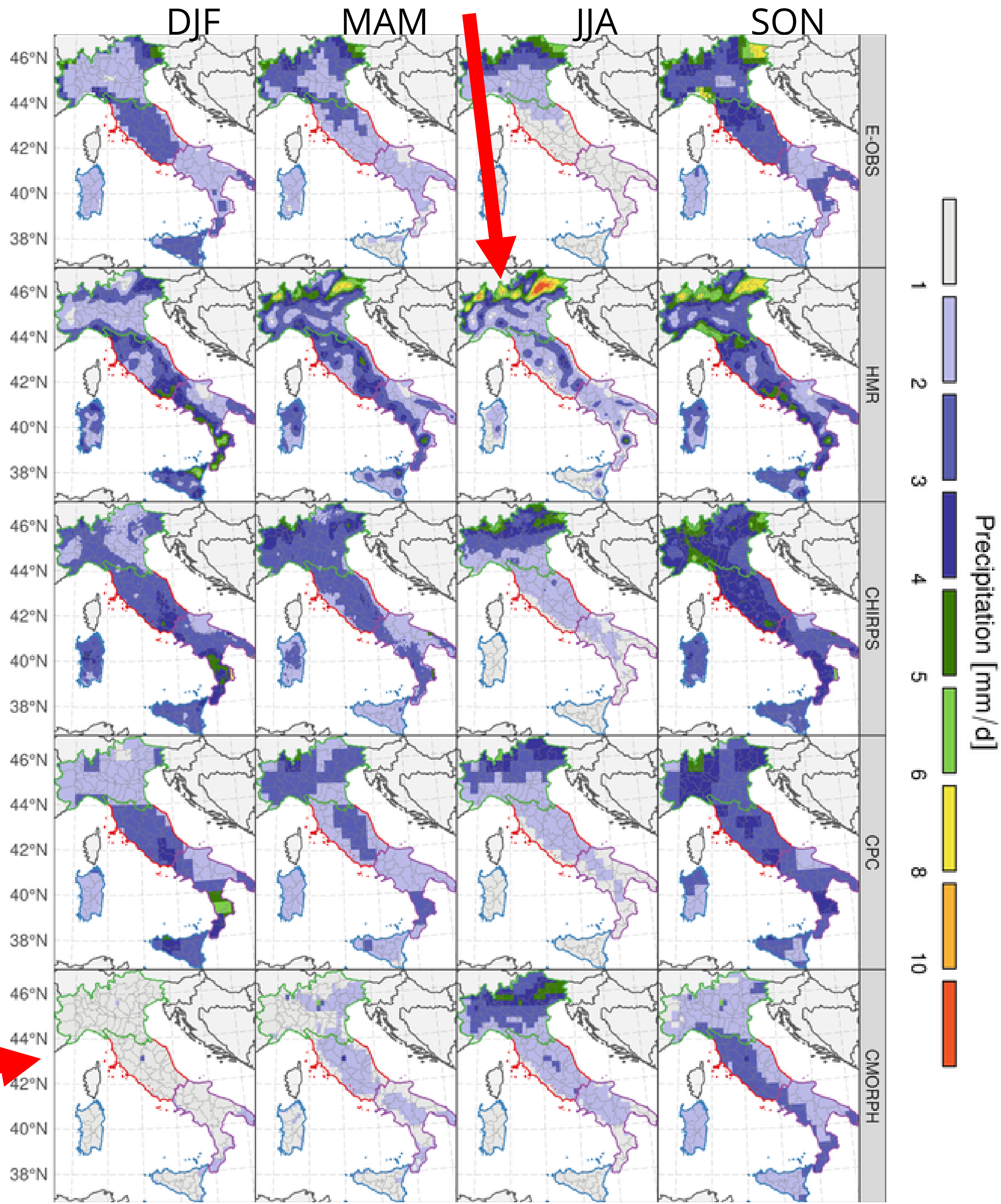


2001-2016
mean precip



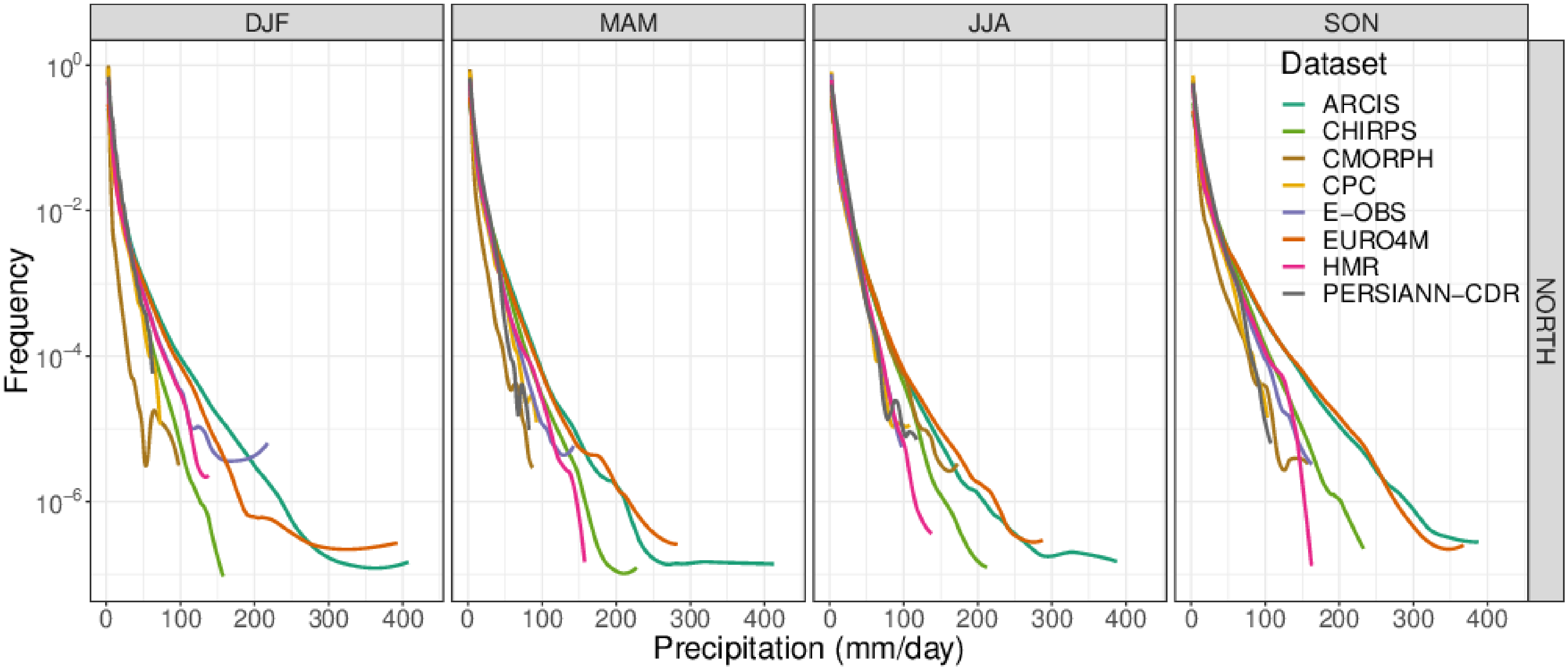
7.3

2001-2016
mean precip



2001-2016

Precipitation probability density function (Northern Italy only)



A few remarks

The best approach to correct data is heavily dependent on:

- Application
- Variable (e.g. precipitation > discharge > temperature)
- Availability of metadata
- Station density
- Length of the records
- Manual resources available

A few remarks

The best approach to correct data is heavily dependent on:

- Application
- Variable (e.g. precipitation > discharge > temperature)
- Availability of metadata
- Station density
- Length of the records
- Manual resources available

A CORRECTION WILL OFTEN NOT BE POSSIBLE → **OBSERVATIONAL DATA WITH VERY HIGH UNCERTAINTY**

But... what about other data sources?

RADAR

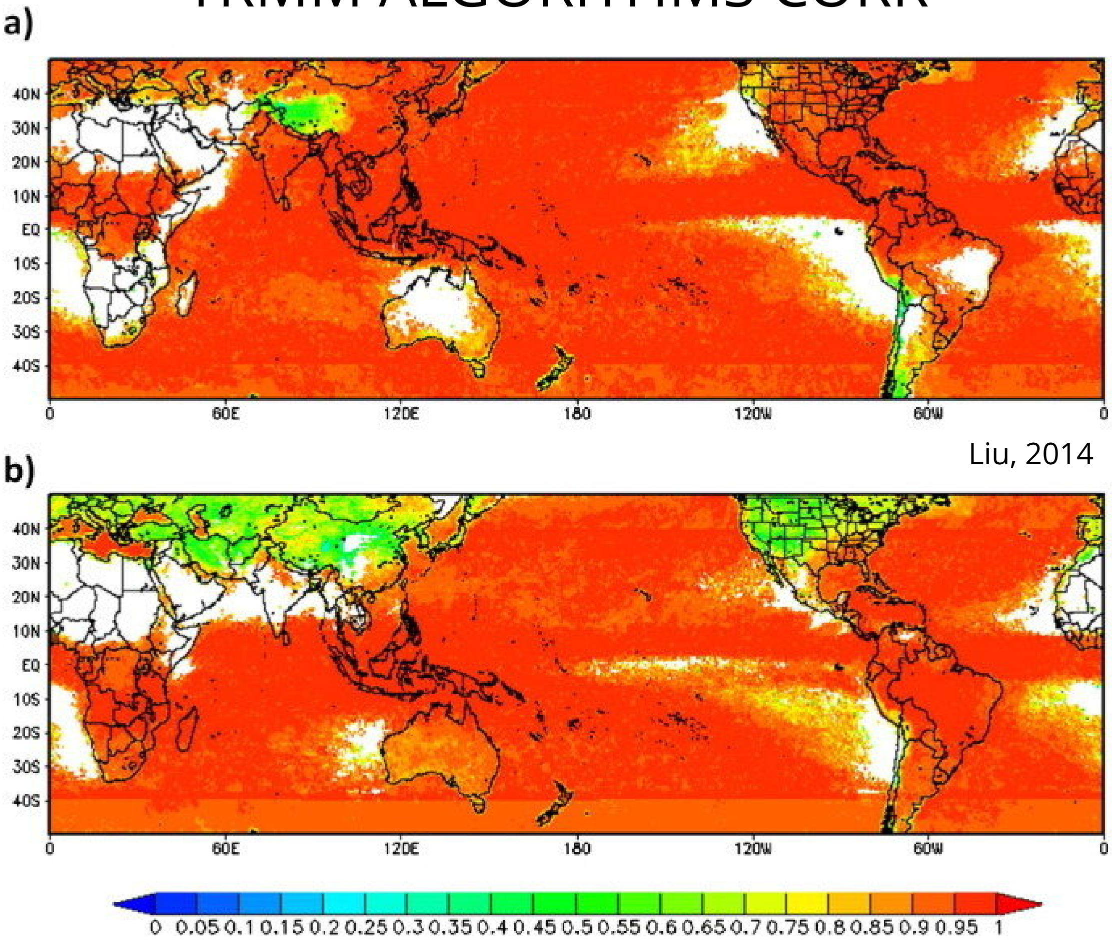
- Only for precipitation
- Depends heavily on location
- Can be shielded by topography
- Can be shielded by intense rain
- Frequent downtime

SATELLITE

- Precipitation, temperature
- ~ Worldwide
- The same algorithm is not necessarily good everywhere
- Resolution is generally poor (0.25° max)

They are just proxies! → Requirement to choose an algorithm

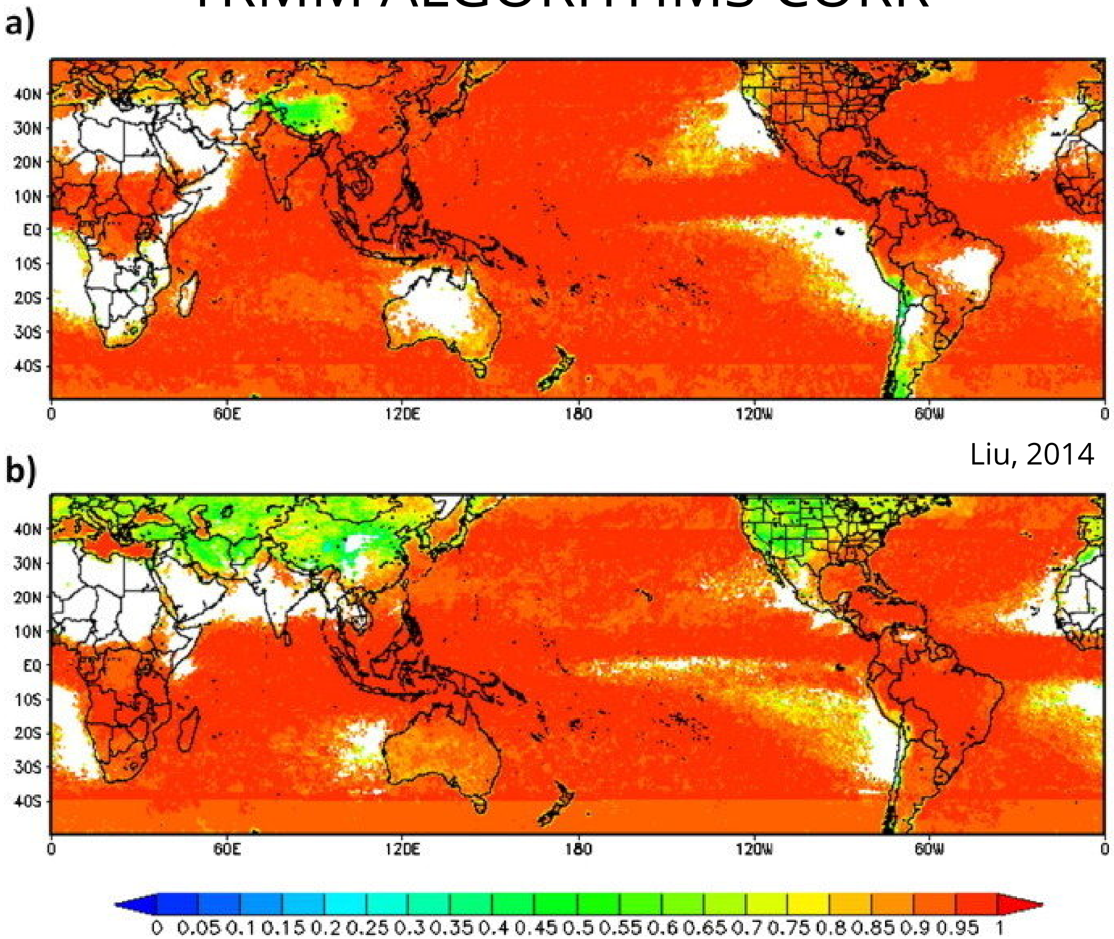
TRMM ALGORITHMS CORR



Liu, 2014

They are just proxies! → Requirement to choose an algorithm

TRMM ALGORITHMS CORR



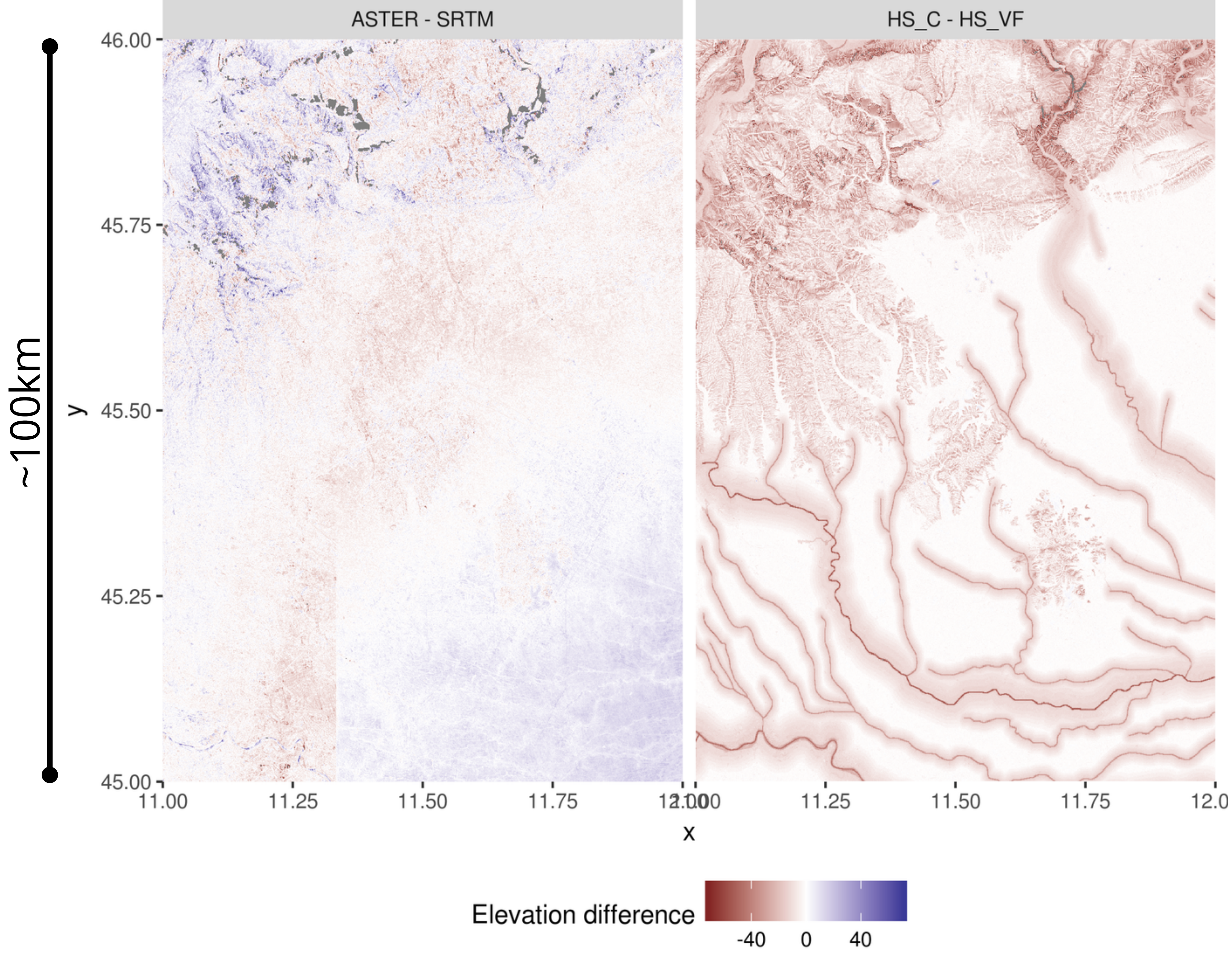
But they are getting better and better!

DEMs

Digital Elevation Models

- ASTER (30m)
- SRTM (30/90m)
- HydroSHEDS (90m)
- JAXA ALOS (30m)
- GTOPO (1km)
- WorldDEM (12m)
- Local, national DEMs
- ...

Usually satellite based, sometimes LIDAR



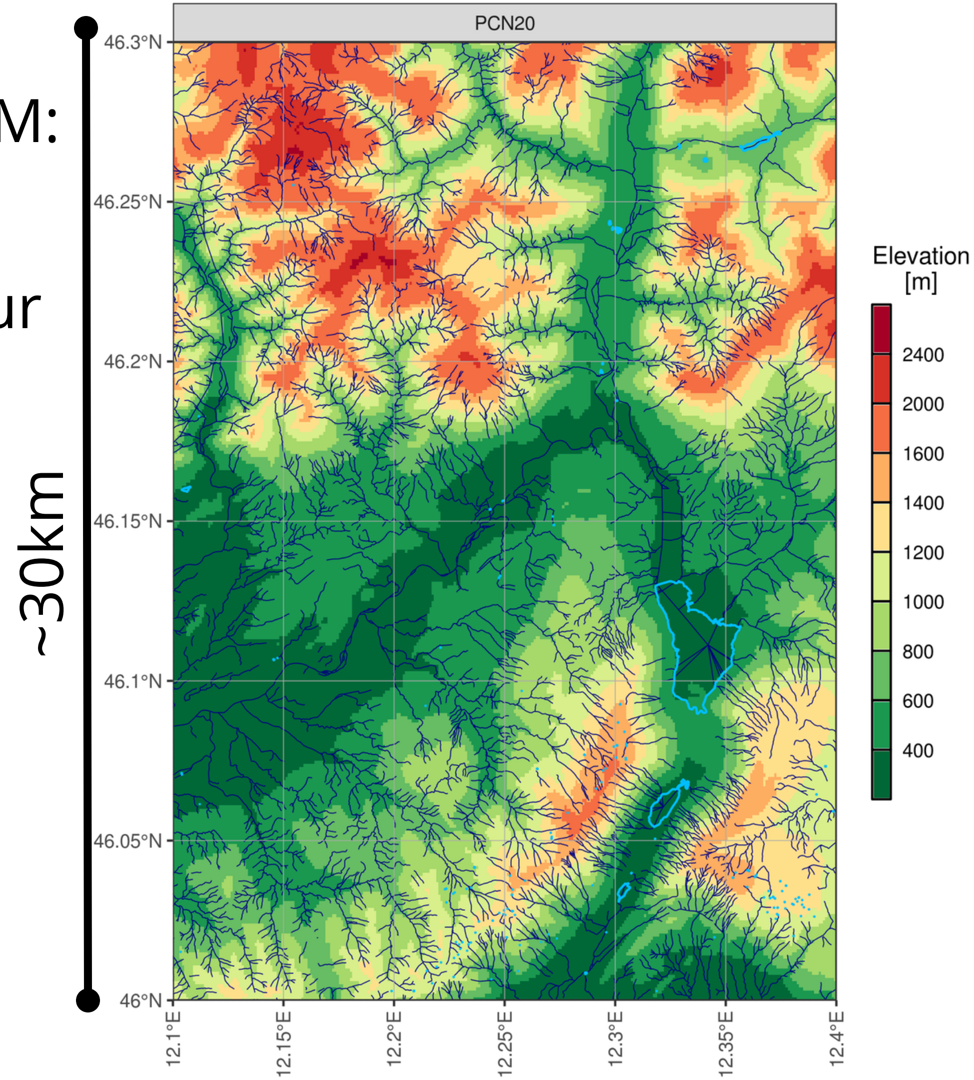
~100km
ASTER - SRTM
HS_C - HS_VF
46.00
45.75
45.50
45.25
45.00
y
11.00
11.25
11.50
11.75
12.00
x
Elevation difference
-40
0
40

9.2

Another example: comparison over a small area

High resolution Italian official DEM:

- 20m resolution
- Obtained from military contour maps
- Comparison with:
  - ASTER
  - HS-c
  - HS-vf
  - JAXA
  - SRTM
  - TINITALY01
- All remapped on a 100m grid



ASTER

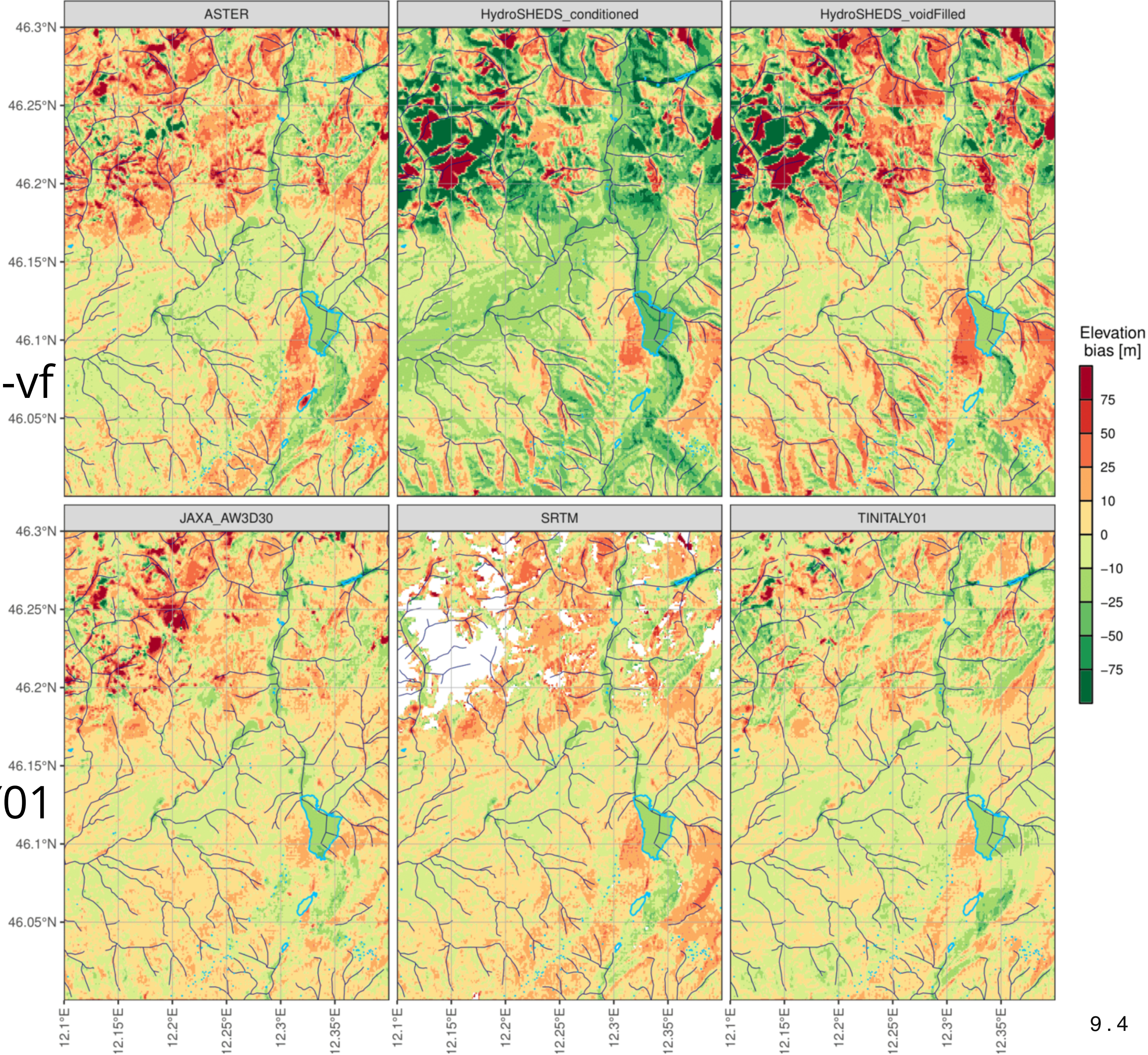
HS-c

HS-vf

JAXA

SRTM

TINITALY01



9.4

DEM	Mean	StdDev	Q05	Median	Q95
ASTER	3.9	30.1	-21	1	36
HydroSHEDS Void Filled	0.4	57.8	-52	0	51
HydroSHEDS VF + Cond.	-16.4	59.7	-86	-11	29
JAXA AW3D30	5.0	19.6	-14	3	26
SRTM	4.7	14.1	-14	3	25
TINITALY/01	-0.6	14.1	-19.3	-0.6	17.7

- Average height differences of several meters
- 5th-95th bias percentile range up to 115 meters
- Standard deviation of bias up to 60m
- Even the best alternative DEM has a 5th-95th bias percentile range of 35 meters, more than enough to introduce issues in river routing!
- CHOICE OF DEM MATTERS!

Take home message

- Do not underestimate observational uncertainty
- Choose your data source based on *your* application
- Never-ever blindly trust un-checked obs data!!!