Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Home Affairs FDHA
Federal Office of Meteorology and Climatology MeteoSwiss

MeteoSwiss

Using Machine Learning to Develop our Quality Control and Communicate its Results

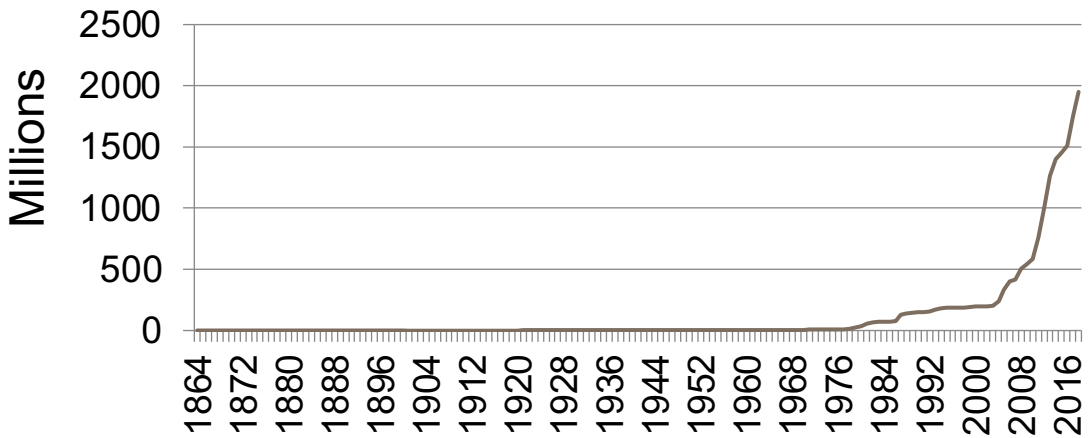
E-AI Summer Workshop: Products and Services – Offenbach, 10.07.2025

Christian Sigg, Francesco Pinto, Deborah van Geijtenbeek and Gian-Duri Lieberherr

christian.sigg@meteoswiss.ch

A Challenge and an Opportunity

The growing volume of surface data is both a challenge and an opportunity

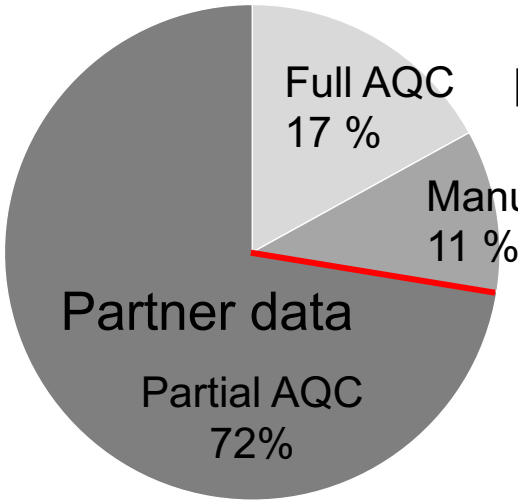


Number of surface data records in the MeteoSwiss DWH

A Challenge and an Opportunity

The growing volume of surface data is both a **challenge** and an opportunity:

- Only a tiny fraction of all surface data can be inspected manually
  → Automated QC (AQC) must act as a powerful filter



MeteoSwiss data

On average 580 values are treated manually each day

Surface data series in the MeteoSwiss DWH

A Challenge and an Opportunity

The growing volume of surface data is both a challenge and an **opportunity**:

- Only a tiny fraction of all surface data can be inspected manually
  → Automated QC (AQC) must act as a powerful filter
- ML-based quality control works better if more data is available

Daily precipitation sums are available for 510 sites in Switzerland

Overview

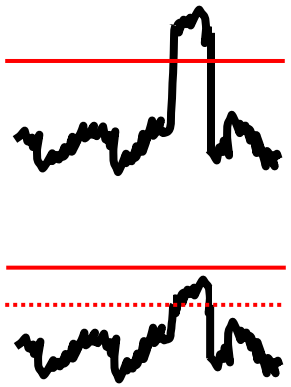
- Evaluation and Optimization of QC Tests
- Learning to Combine Rain Gauge and Radar Data
- Summarizing Quality Information with Naïve Bayes
- Conclusion

Overview

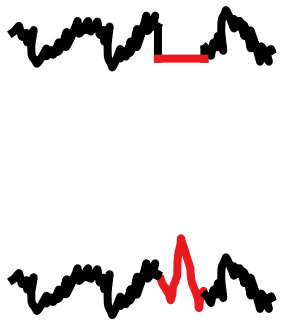
- Evaluation and Optimization of QC Tests
- Learning to Combine Rain Gauge and Radar Data
- Summarizing Quality Information with Naïve Bayes
- Conclusion

Rule-Based Quality Control

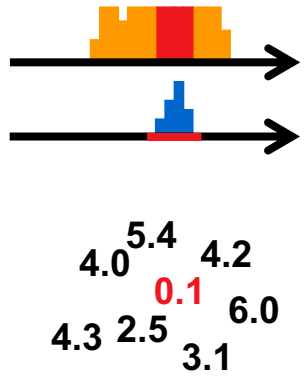
Before 2015, we employed a rule-based expert system (RBES), following WMO guidelines WMO (2012)



Hard and soft limits



Variability limits



Consistency

WMO (2012) "Guide to the Global Observing System"

How Well Does a QC Test Perform?

- Each test emulates specific aspects of expert inspection
- But it can miss implausible measurements and create false alarms

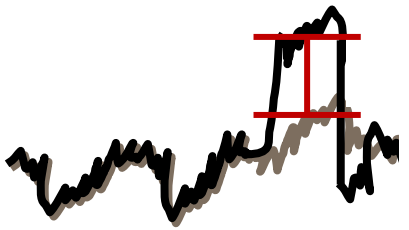
Statistical hypothesis testing:

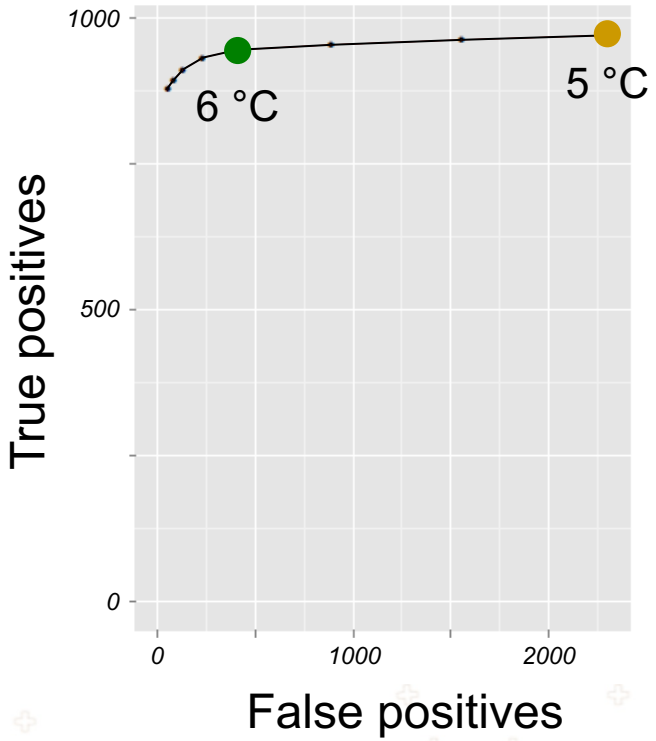
	Measurement plausible	**Measurement implausible**
Test passed	True negative (TN)	False negative (FN)
Test failed	False positive (FP)	True positive (TP)

→ Strong tests create few false positives and false negatives

Trade-Off between True and False Positives

Test based on the deviation from redundant 2 m temperature:





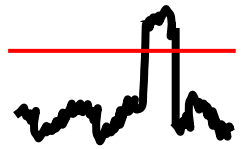
Increasing the allowable difference from 5 to 6 °C reduces false positives by an order of magnitude, while keeping almost all true positives

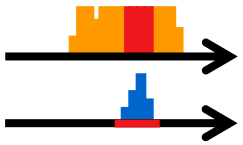
Strengths and Weaknesses of RBES Knechtl *et al.* (2015)

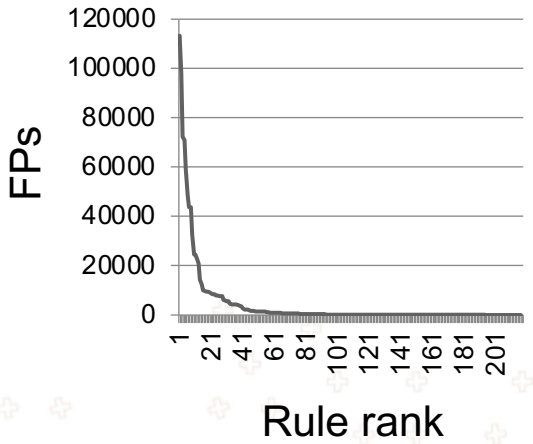
Evaluation of our rule-set in 2015:

- “Simple” rules create none or few FPs, but miss many implausible values
- Some consistency rules generate an inacceptable number of FPs, even though they seem sensible
- Most FPs are created by a small number of rules
- Redundancy: only 35 % of rules generated test failures

→ Combine simple rules with data-driven ML models





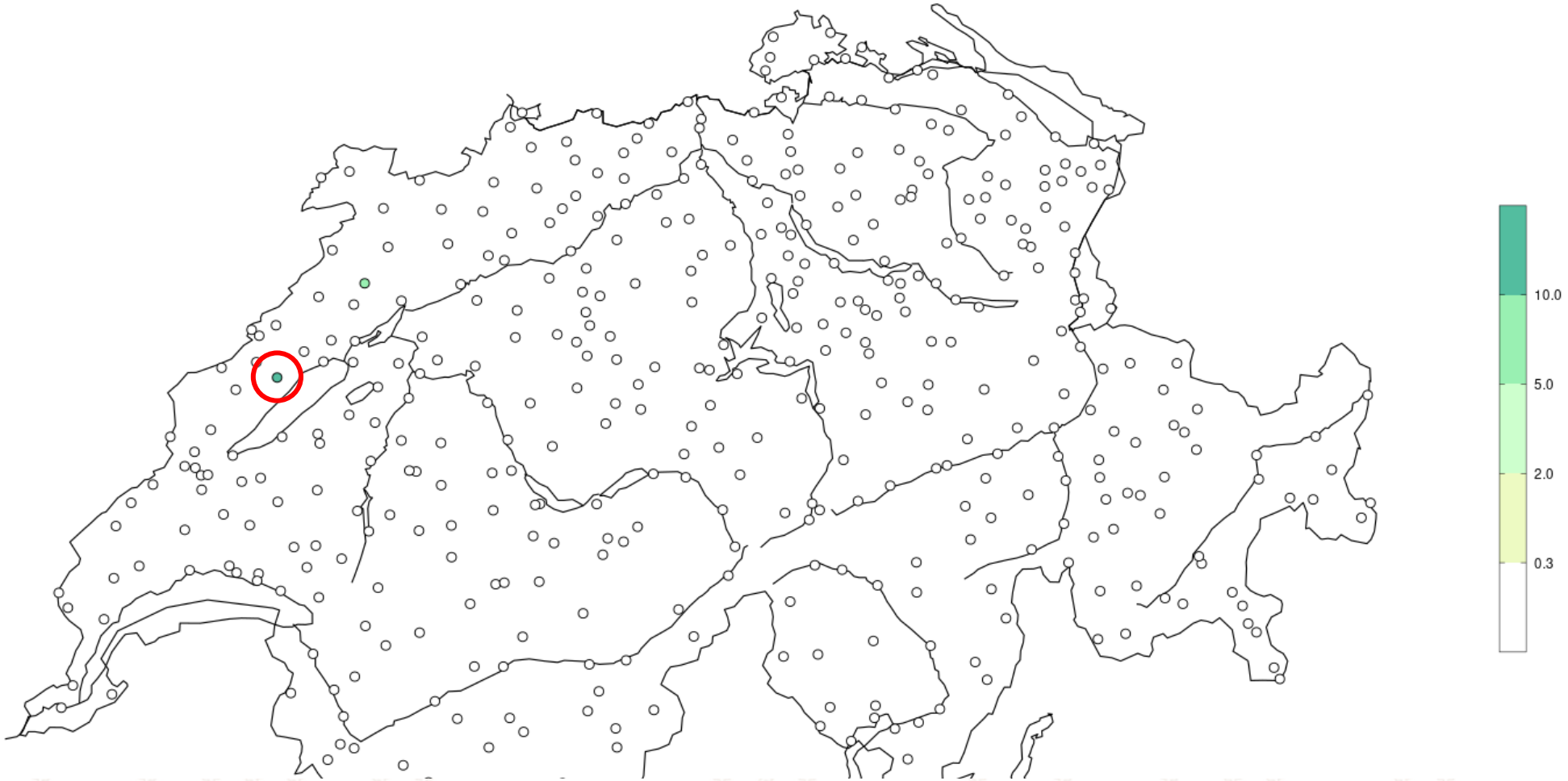


V. Knechtl, D. van Geijtenbeek and C. Sigg (2015) "A quantitative approach to optimise the quality control system for surface data at MeteoSwiss"

Overview

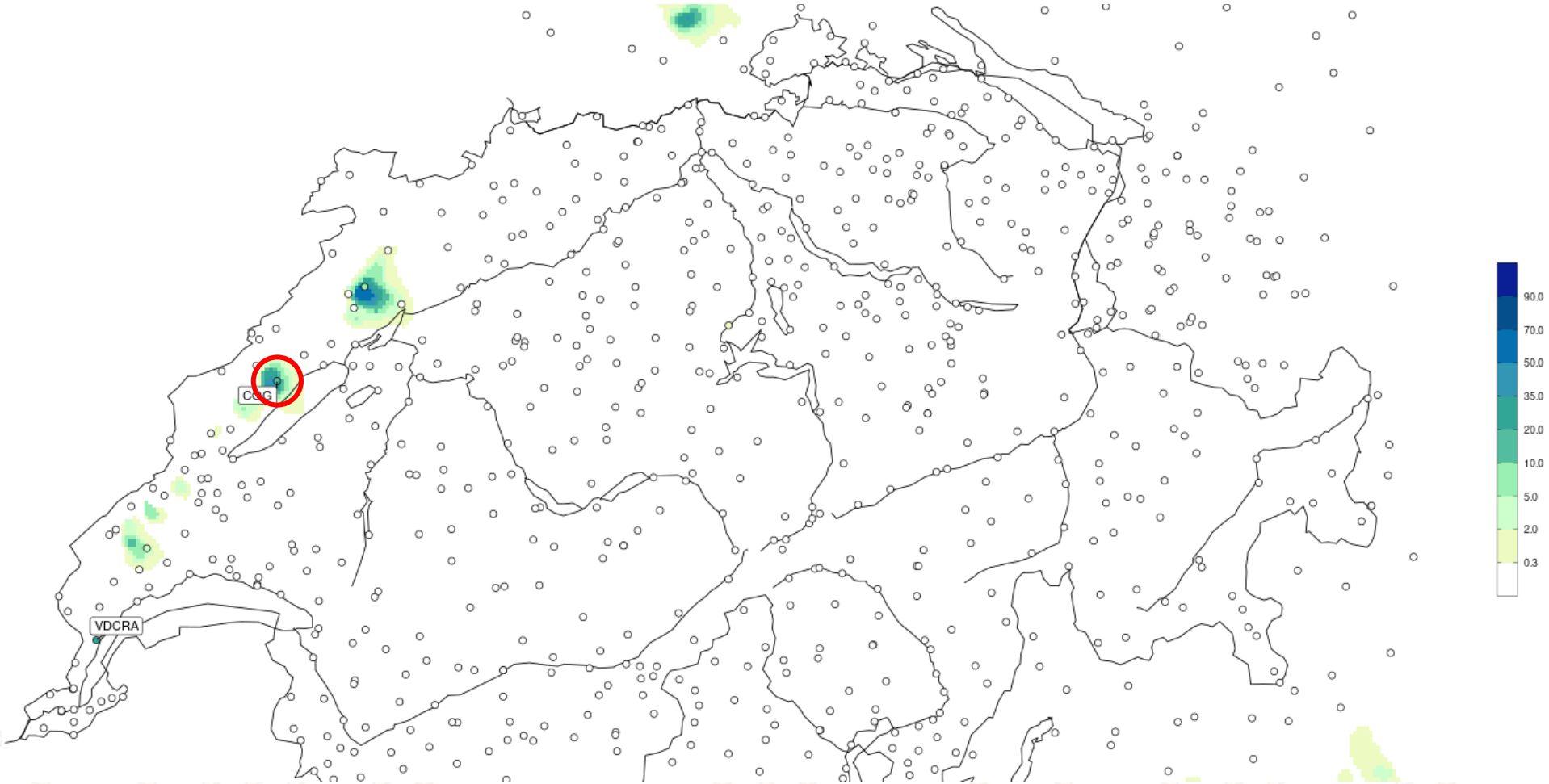
- Evaluation and Optimization of QC Tests
- Learning to Combine Rain Gauge and Radar Data
- Summarizing Quality Information with Naïve Bayes
- Conclusion

Spatial Consistency of Rain Gauge Measurements



Combe-Garot (COG) on 2023-08-20: daily precipitation sum of 14.5 mm

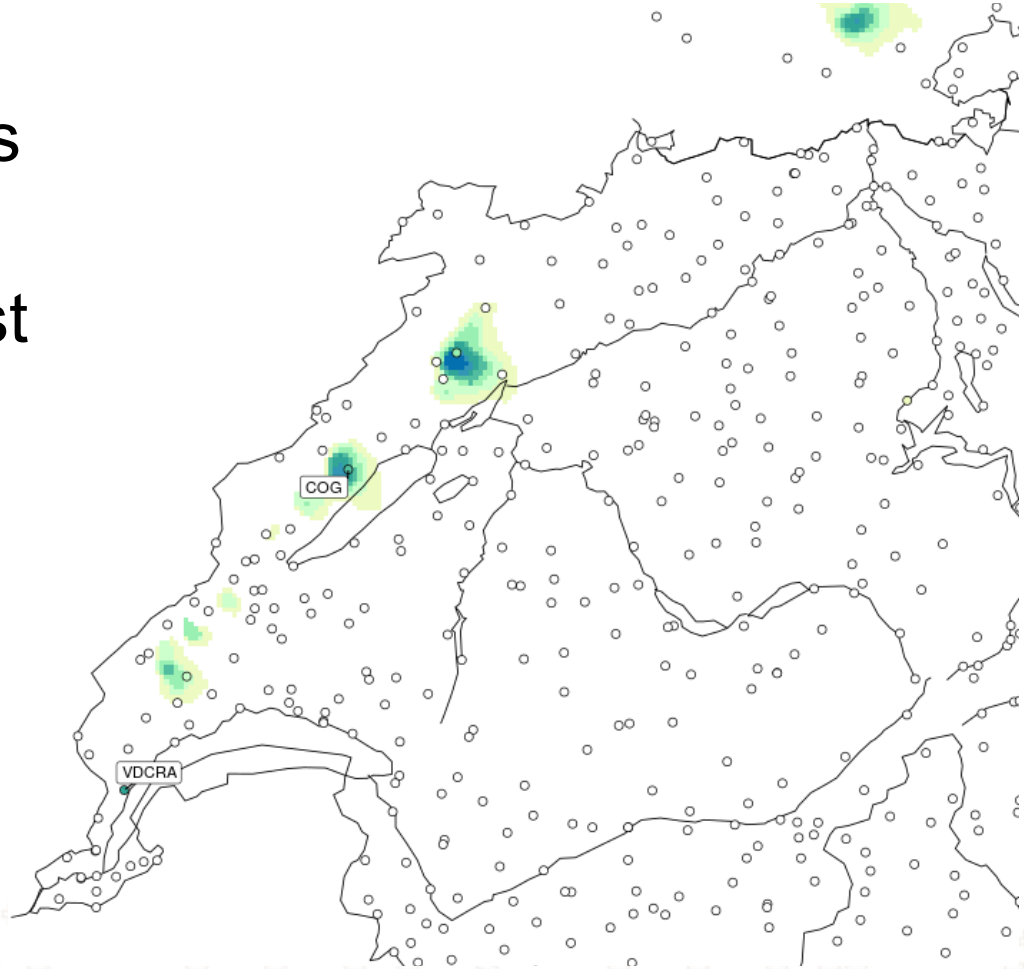
Including the 24 h Radar Accumulation



Combe-Garot (COG) on 2023-08-20: daily precipitation sum of 14.5 mm

QC by Cross-Validation

1. Withhold measurement under test
2. Predict using neighboring rain gauges and the radar field
3. Compare the measurement under test to the prediction
4. If the measurement is unlikely, given the prediction, create a QC case



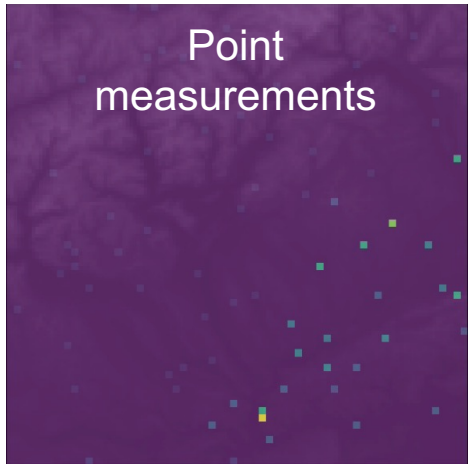
Challenge I: Different Data Modalities

Rain gauge network: Point measurements on the ground

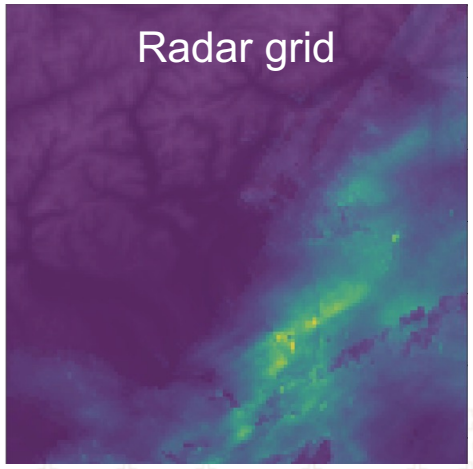
Radar: Remote sensing of the atmosphere, resulting in spatially dense grid

Goal: Fuse both modalities into common representation

Rain gauge



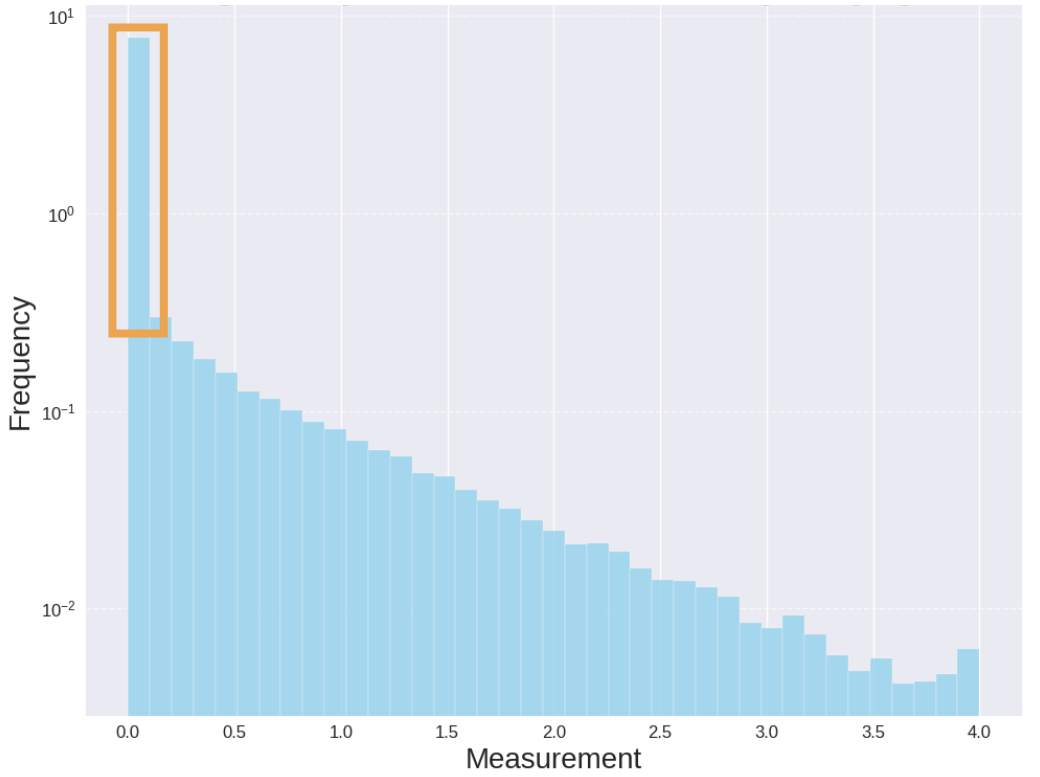
Radar



Challenge II: Two-Part Statistical Distribution

- Rain gauges are dry in 85 % of the measurements
- If wet, the amount can be any positive value with one decimal resolution

Goal: Jointly model presence and magnitude of precipitation



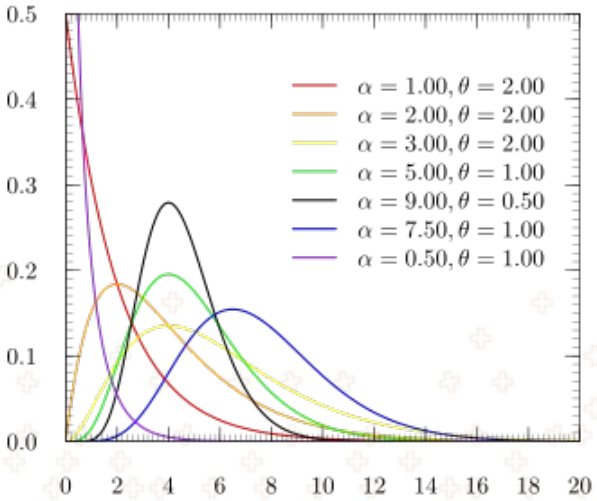
Gamma-Hurdle Distribution

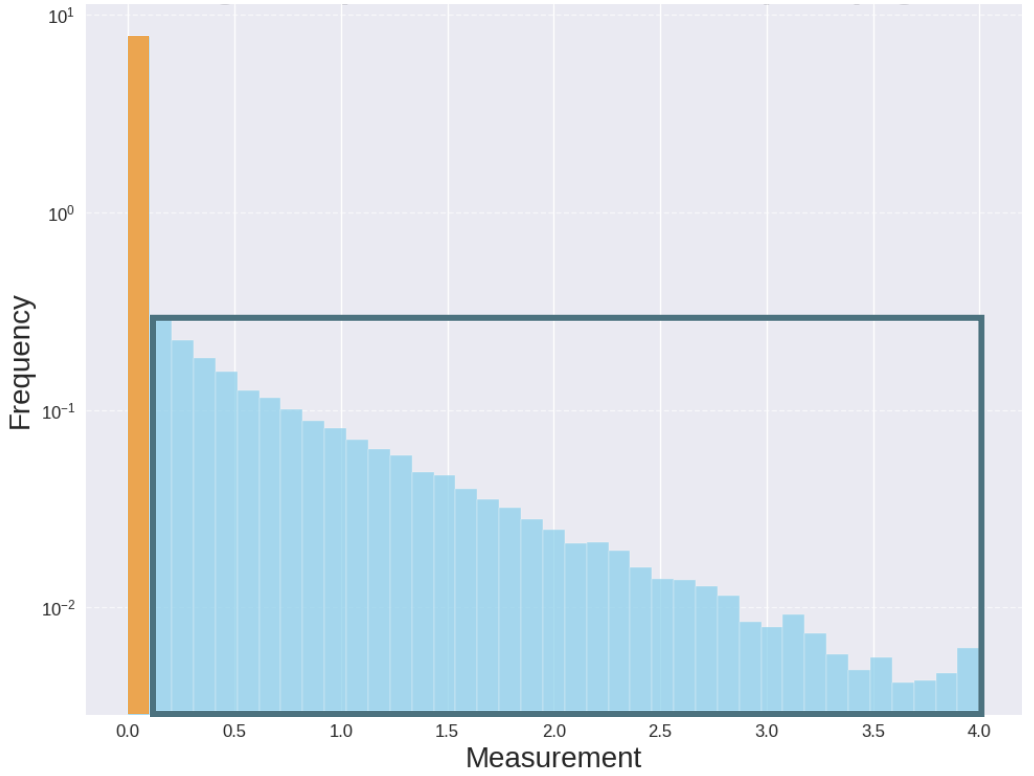
Probability of dry versus wet:

$$P(y > 0) \sim Bernoulli(\pi)$$

Distribution of positive magnitude:

$$P(y|y > 0) \sim Gamma(\alpha, \theta)$$

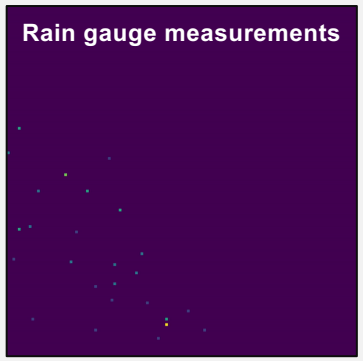




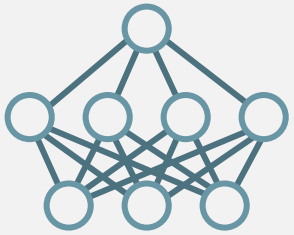
Sequential Model Overview

Step 1: Rain gauge interpolation

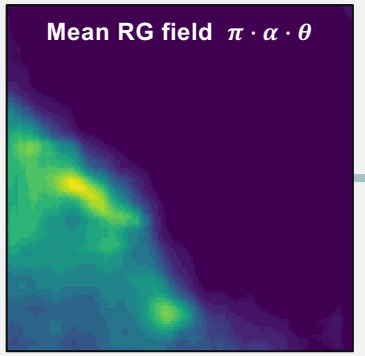
Input



ConvCNP
Gordon *et al.* (2020)

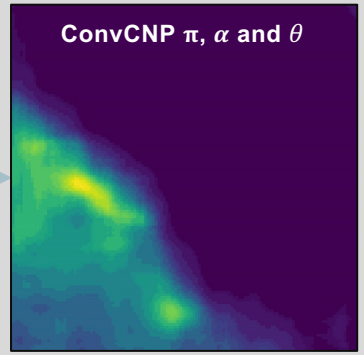


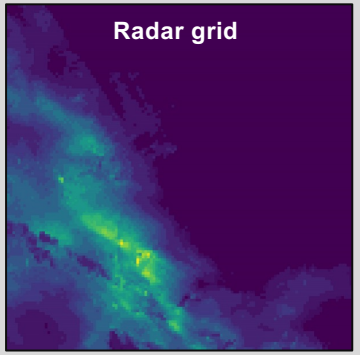
Output



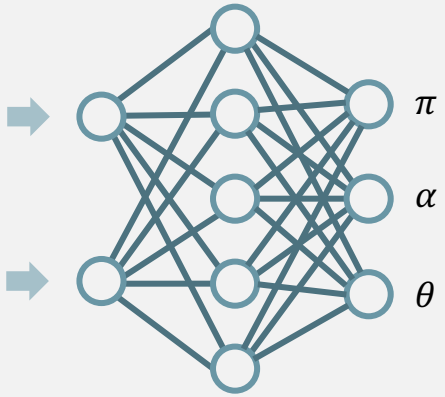
Step 2: Combination of Rain Gauge Field and Radar Grid

Input

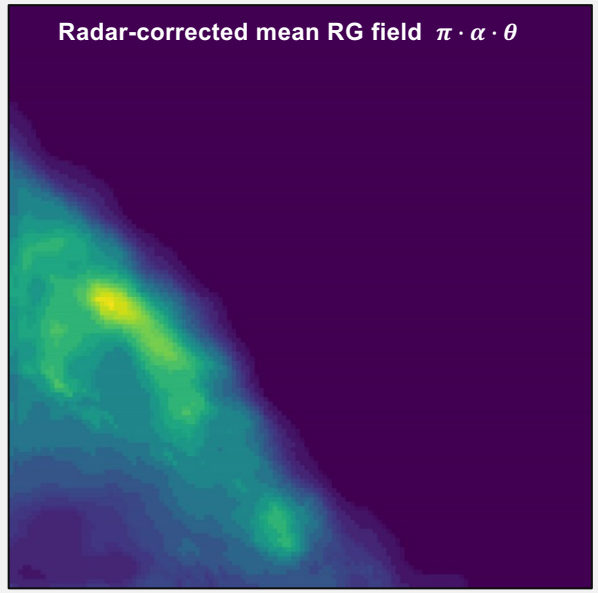




U-Net
Ronneberger et al. (2015)



Output



Gordon *et al.* (2020) "Convolutional Conditional Neural Process"
Ronneberger *et al.* (2015) "U-Net: Convolutional Networks for Biomedical Image Segmentation"

Comparison: RG-Only vs. RG-Radar vs. Baseline

Data: Hourly accumulations of 231 rain gauges for 2023

CNP: Learning a Conditional Neural Process on rain gauges only

CNP/U-Net: Learning to combine CNP with radar

CombiPrecip: Use Kriging to interpolate difference between rain gauge and radar, separately for every time step

Sideris *et al.* (2014)

Overall RMSE:

CNP/U-Net	Combi Precip	CNP
0.93	0.98	1.32

Overall scatter:

CNP/U-Net	Combi Precip	CNP
1.69	1.85	2.16

Overall log-bias:

CNP/U-Net	Combi Precip	CNP
-0.29	-0.32	0.10

Number of stations with lowest RMSE:

	CNP/U-Net	Combi Precip	CNP
January	140	54	41
February	121	54	53
March	122	57	56
April	127	83	25
May	133	92	10
June	118	106	11
July	128	91	16
August	112	108	15
September	116	103	16
October	145	62	28
November	156	56	23
December	150	52	33

Sideris *et al.* (2014) "Real-time radar-rain-gauge merging using spatio-temporal co-kriging with external drift in the alpine terrain of Switzerland"

Overview

- Evaluation and Optimization of QC Tests
- Learning to Combine Rain Gauge and Radar Data
- Summarizing Quality Information with Naïve Bayes
- Conclusion

Our Former Quality Information

Binary flag per test category:

Variable	Instrument	Reference time	Value	P	C	I	S
rre150z0	11356	05.01.2019 23:20	-23	1	0	0	0
rre150d0	9838	03.02.2019	5.5	0	0	0	1
tre200s0	20324	01.02.2019 08:10	11.5	0	1	1	0

- **P**hysical limit exceeded
- **C**limatological limit exceeded
- **I**nconsistent to another variable
- **S**patially inconsistent

Pros and Cons

+ Data model is straightforward:

1. Summarize test outcomes into categories
2. Store in binary attributes right along measurement

+ Categories are self-explanatory

– Categories combine tests with greatly varying sensitivity and specificity

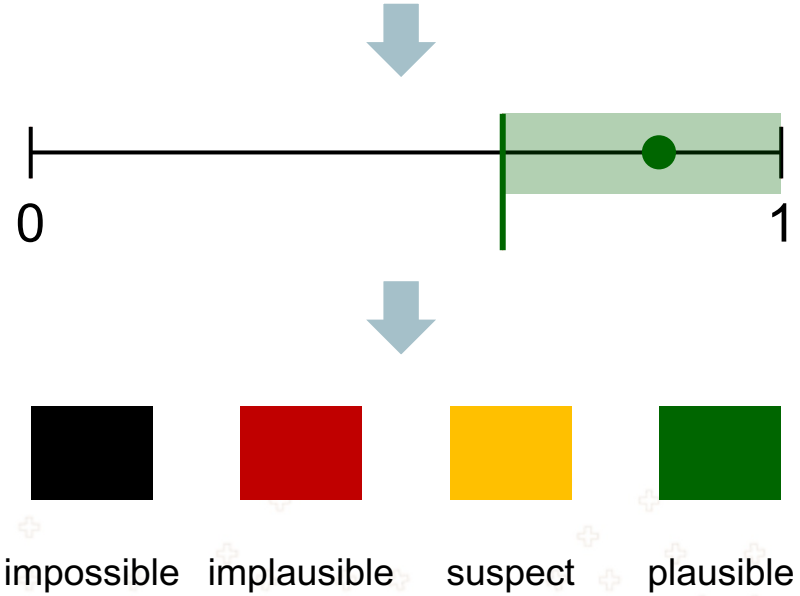
– Customers typically cannot integrate flags into their own databases

Internal and external users could not make effective use of flags beyond **P**

Probabilistic Plausibility Sigg *et al.* (2017)

- Store outcome of every test
- Use Naïve Bayes to compute probabilistic plausibility of measurements from test outcomes and likelihoods
- Users specify threshold for their application, either numerically or with a "traffic light" code

Measurement	Test	Passed
4614406274	8	N
4614406274	112	Y
4614406274	236	Y



Definition of Plausibility

Expert treatment is the reference:

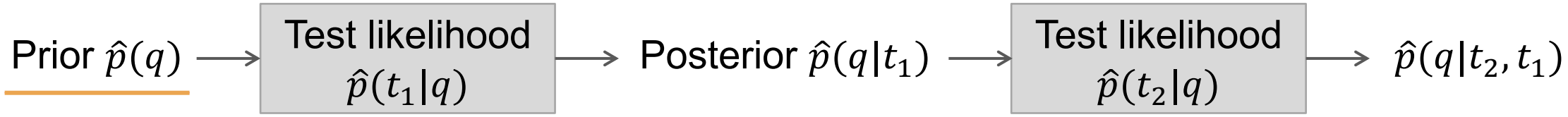
> A measurement is *plausible* if it is **confirmed** during expert inspection.
>
> A measurement is *implausible* if it is **corrected or suppressed** during expert inspection.

Expert inspection is incomplete: measurements are assumed to be plausible unless they are explicitly implausible

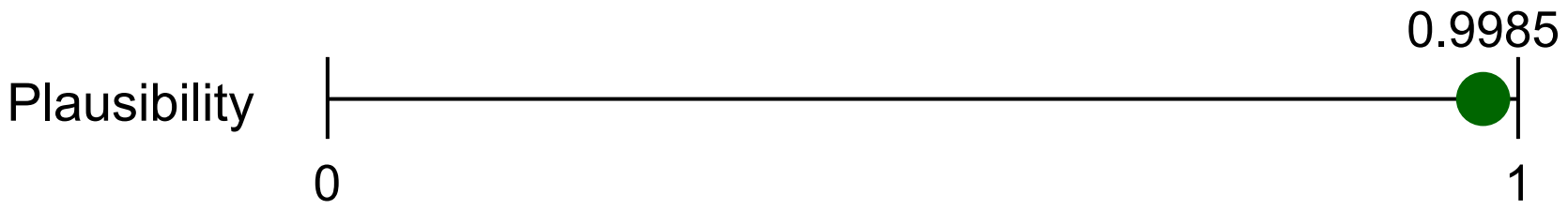
Calculation Example

- Plausibility of automated 2 m air temperature measurements
- 10 min granularity → 144 measurements per day
- Probabilities estimated from one year of data (values rounded)

Estimated Prior

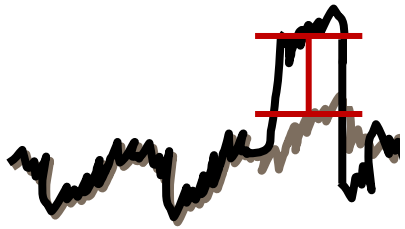


About 1 in 670 measurements is implausible:



Consistency Test

Prior $\hat{p}(q)$ → Test likelihood $\hat{p}(t_1|q)$ → Posterior $\hat{p}(q|t_1)$ → Test likelihood $\hat{p}(t_2|q)$ → $\hat{p}(q|t_2, t_1)$

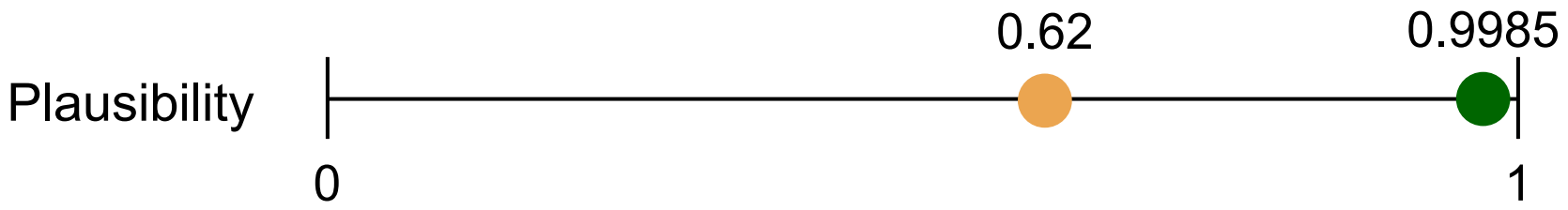


Limit of absolute difference to redundant measurement:

- 0.014 % false positive rate
- 5.7 % of implausible measurements fail consistency test

Estimated Posterior

Measurement fails consistency test, $t_1 = 0$:



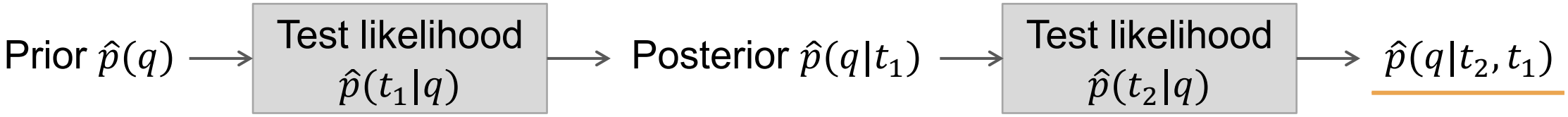
Combining Test Outcomes

Prior $\hat{p}(q)$ → Test likelihood $\hat{p}(t_1|q)$ → Posterior $\hat{p}(q|t_1)$ → Test likelihood $\hat{p}(t_2|q)$ → $\hat{p}(q|t_2, t_1)$

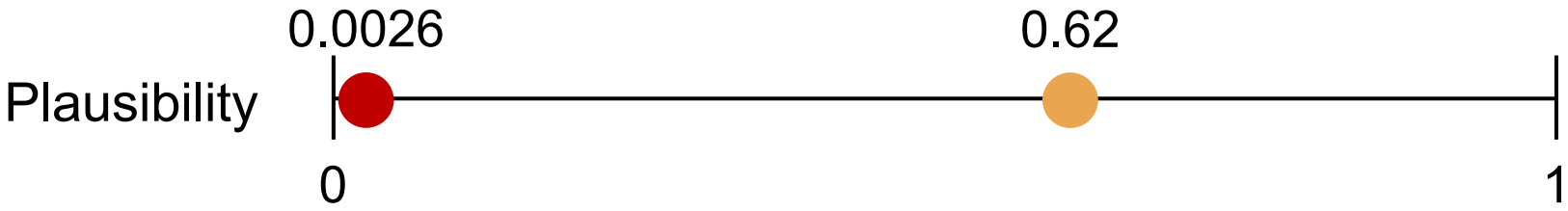
Minimum variability test:

- 0.0015 % false positive rate
- 1.2 % of implausible measurements fail minimum variability test

Updated Posterior



Measurement also fails the minimum variability test, $t_2 = 0$:



Discussion

- – Cannot store test outcomes in fixed-length bitmask
- – Computation necessary to obtain plausibility
- \+ Outcomes contribute according to their evidence:
  - Test outcomes increase or decrease plausibility
  - Accumulate weak evidence of several test outcomes into strong evidence
- \+ Combine outcomes from independent QC systems:
  - Incorporate new test outcomes whenever they arrive
  - Re-calculate plausibility using Naïve Bayes

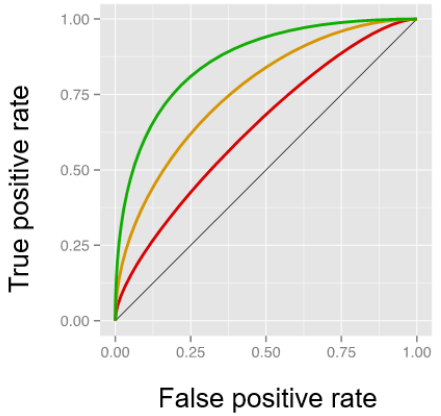
Overview

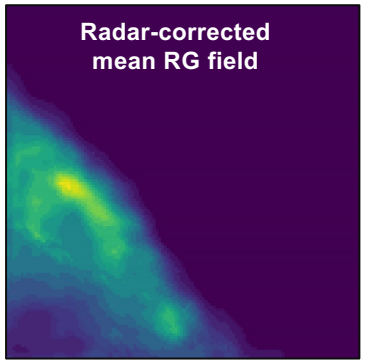
- Evaluation and Optimization of QC Tests
- Learning to Combine Rain Gauge and Radar Data
- Summarizing Quality Information with Naïve Bayes
- Conclusion

Conclusion

Three benefits of machine learning for QC:

1. Statistical evaluation provides the basis for further development and optimization of QC tests
2. ML models can capture complex relationships between different data modalities
3. Computing the probabilistic plausibility summarizea all available test outcomes into a simple but well defined score





Measurement	Test	Passed
4614406274	8	N
4614406274	112	Y
4614406274	236	Y