Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

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

MeteoSwiss

# Data-Driven Quality Control for Surface Data

KNMI Data Science Symposium
December 9th, 2020

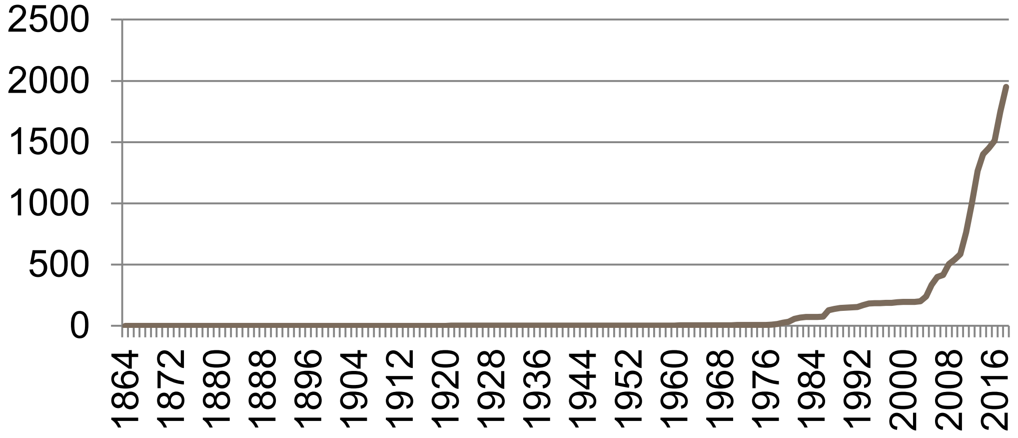
Christian Sigg, Valentin Knechtl and Deborah van Geijtenbeek
christian.sigg@meteoswiss.ch

# Outline

- A Challenge and an Opportunity
- From Rule-Based to Data-Driven QC
- Example: Consistency to Other Measurements
- Example: Learning from Expert Feedback

# A Challenge and an Opportunity

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

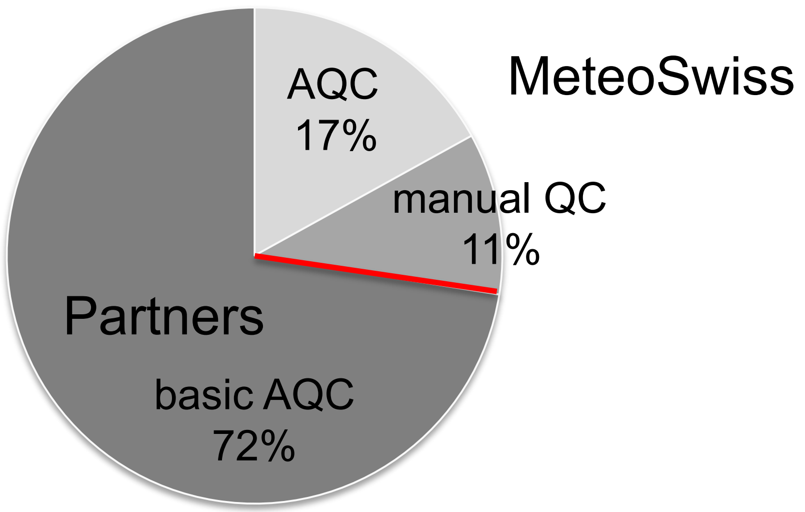



surface data records [millions]

# A Challenge and an Opportunity

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

- Only a tiny fraction can be inspected manually → automated QC must act as a powerful filter

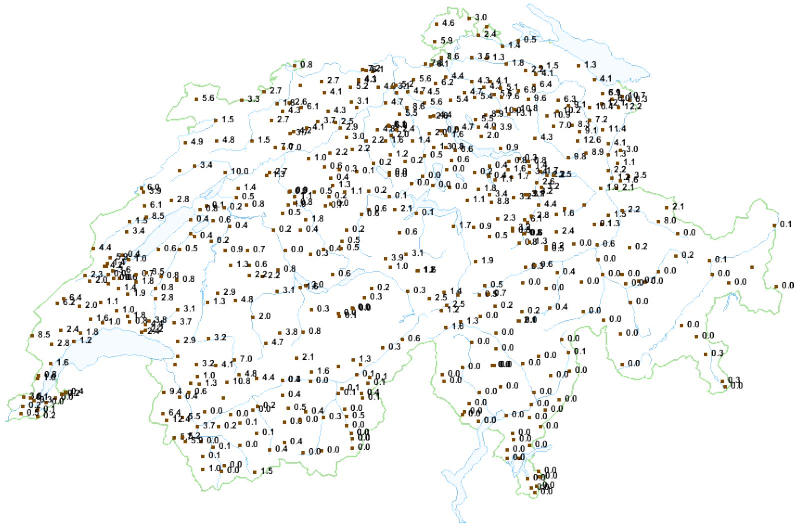



~ 400 suspect values receive daily manual inspection

# A Challenge and an Opportunity

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

- Only a tiny fraction can be inspected manually → automated QC must act as a powerful filter
- Data-driven quality control works better if more data is available



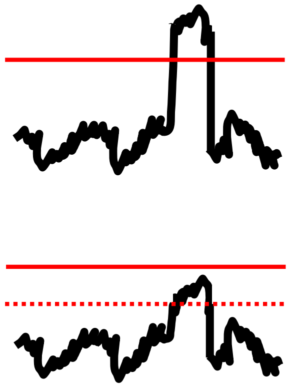
510 sites measuring daily precipitation

# Outline

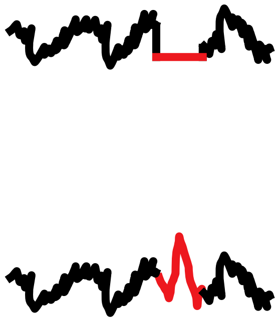
- A Challenge and an Opportunity
- From Rule-Based to Data-Driven QC
- Example: Consistency to Other Measurements
- Example: Learning from Expert Feedback

# Rule Design for Automated QC

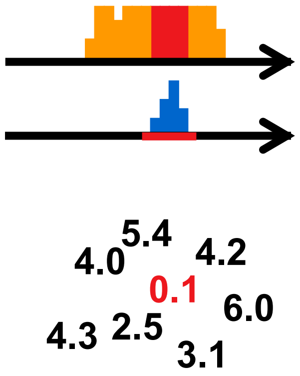
We employ a rule-based expert system, following WMO guidelines:
WMO (2012)



*Hard and soft limits*



*Variability limits*

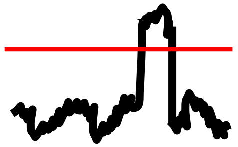



*Consistency*

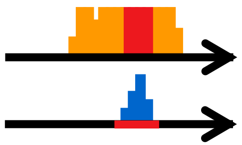
# Strengths and Weaknesses

Evaluation of our rule-set in 2015:

- “Simple” rules achieve a good TP to FP ratio, but miss many implausible values



- Consistency rules generate an inacceptable number of FPs, even though they look sensible on paper



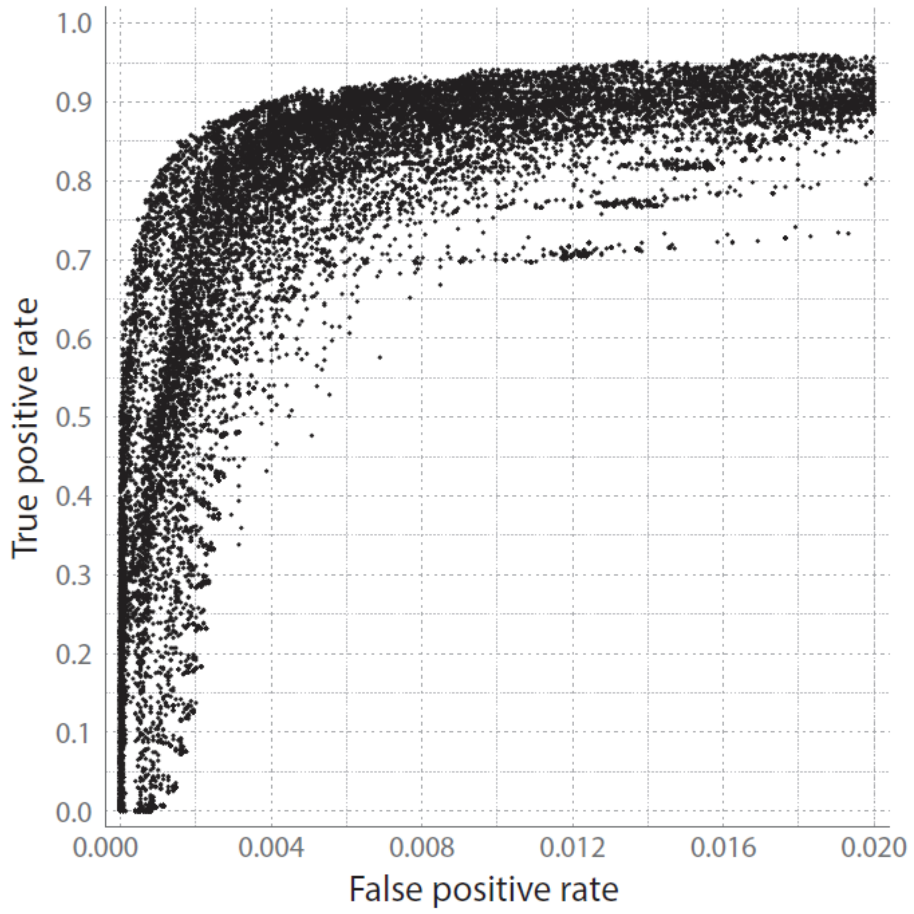
- Aggregate complexity: rule-set specification spans > 60,000 table rows
- Redundancy: only 35 % of rules generated test failures per year

→ Combine simple rules with complex but data-driven models

# Data-Driven QC

We use machine-learning techniques:

- To develop domain specific QC tests with an optimized cost-benefit ratio



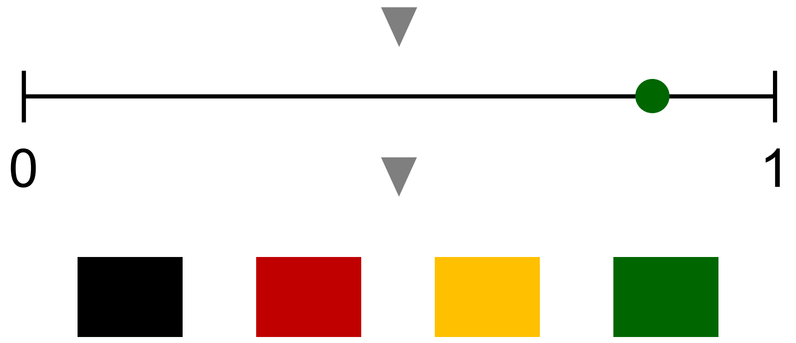

ROC for detecting spurious precipitation measurements in a weighing rain gauge

# Data-Driven QC

We use machine-learning techniques:

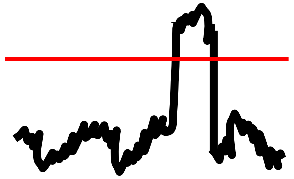
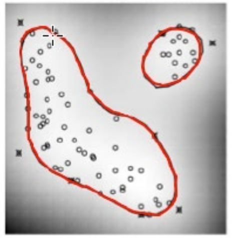
- To develop domain specific QC tests with an optimized cost-benefit ratio
- To provide a summary of all available quality information (QI) that is simple, well-defined and relevant to the user

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



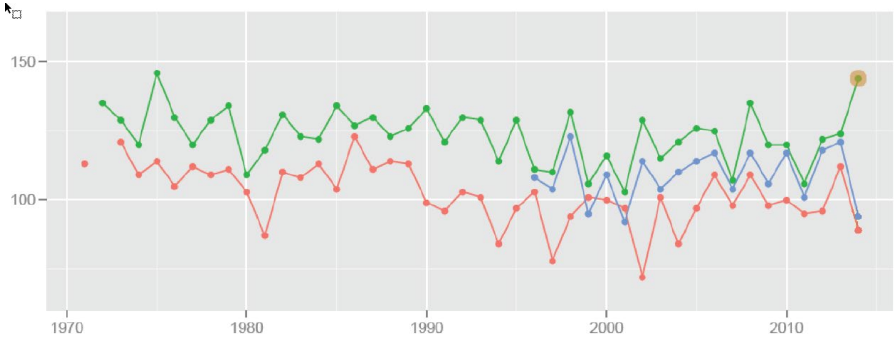
# 3 Information Sources for Data-Driven QC

**1. Relative frequency of occurrence**
Principle: “Rough errors are rare”
Model: Outlier detection

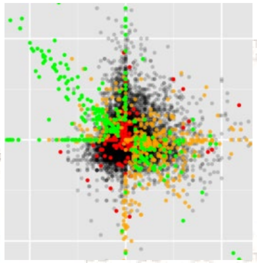



Schölkopf *et al.* (2001)

**2. Relationships to other measurements**
Principle: “Implausible values are inconsistent”
Model: Continuous regression

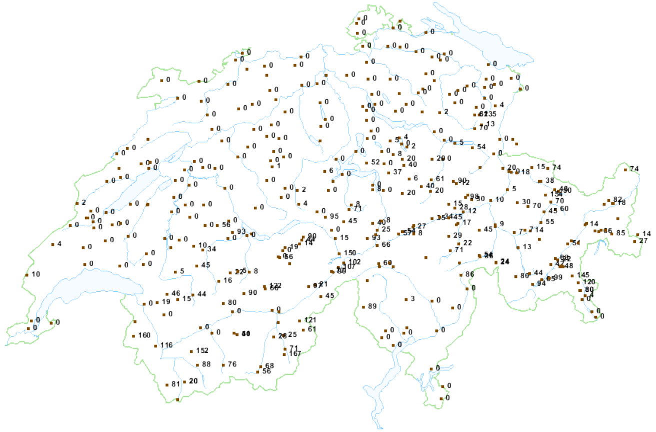



**3. Expert feedback**
Principle: “Model imitates expert”
Model: Discrete classification



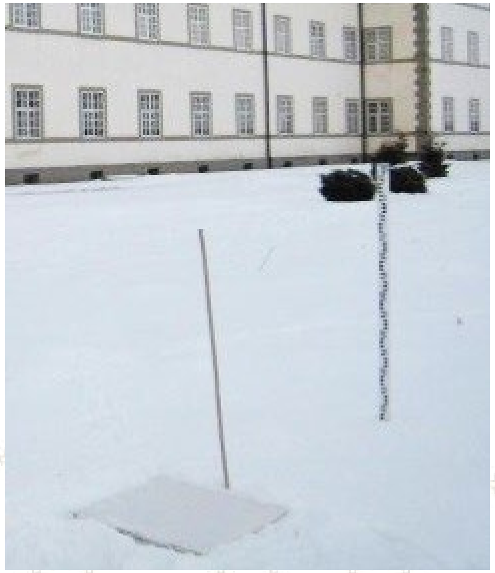
# Outline

# QC of Manual Snow Measurements



MCH and SLF together operate 372 manual snow measurement sites:

- For avalanche warnings, climatology and hydrology
- Daily measurements of total and new snow depth
- Reported per SMS text



Challenges for QC:

- Drifting snow
- Typos during manual entry
- Observers not following protocol
- Spatial correlation between sites can be low

# Predicting Presence and Magnitude

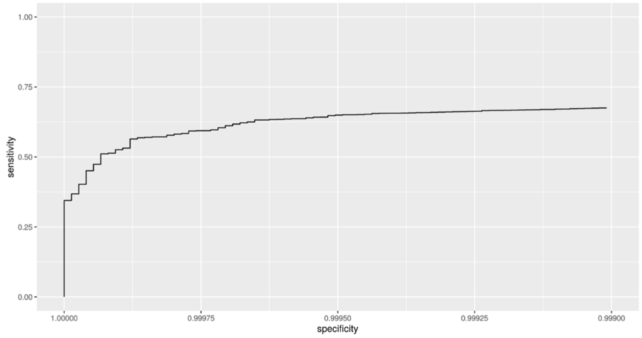
XGBoost classification and regression models for:

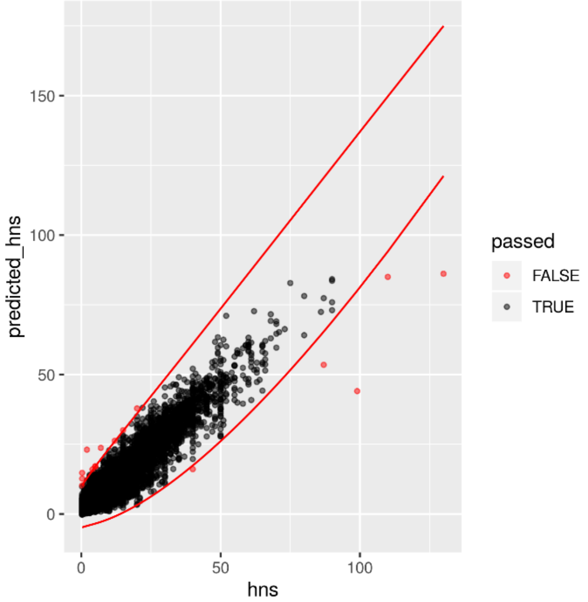
- Presence / absence of new snow and snow cover
- Depth of total and new snow

Features used for prediction:

- Past and future snow measurements
- Temperature, precipitation, global radiation
- Substitution of missing values with grid product estimates

QC pass / fail decision based on ROC curves and quantile regression



# Outline

- A Challenge and an Opportunity
- From Rule-Based to Data-Driven QC
- Example: Consistency to Other Measurements
- Example: Learning from Expert Feedback

# Detecting Spurious Precipitation

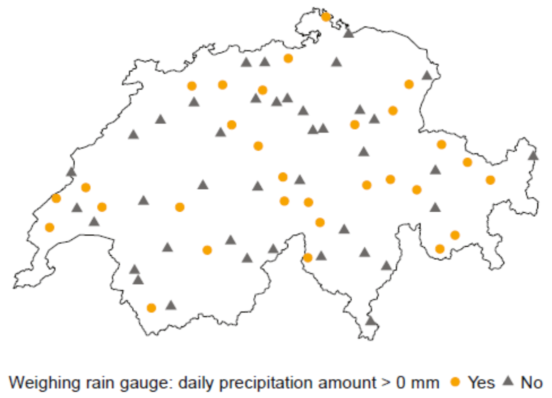
Automated precipitation network:

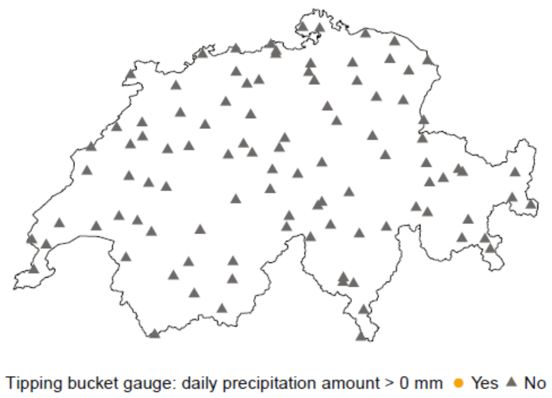
- 100 Lambrecht 15188/1518H3 tipping bucket
- 131 OTT Pluvio$^2$ weighing gauges

Spurious precipitation at weighing gauges:

- Isolated measurements of 0.1 to 0.5 mm / 10 min
- Have negative impact on climatological indices, gridded products and NWP verification

We established a QC regime to manually correct offending measurements to zero and performed a systematic review of all events in 2015.




Spurious precipitation events on February 11, 2015

# From Analysis to Quality Control Knechtl *et al.* (2019)

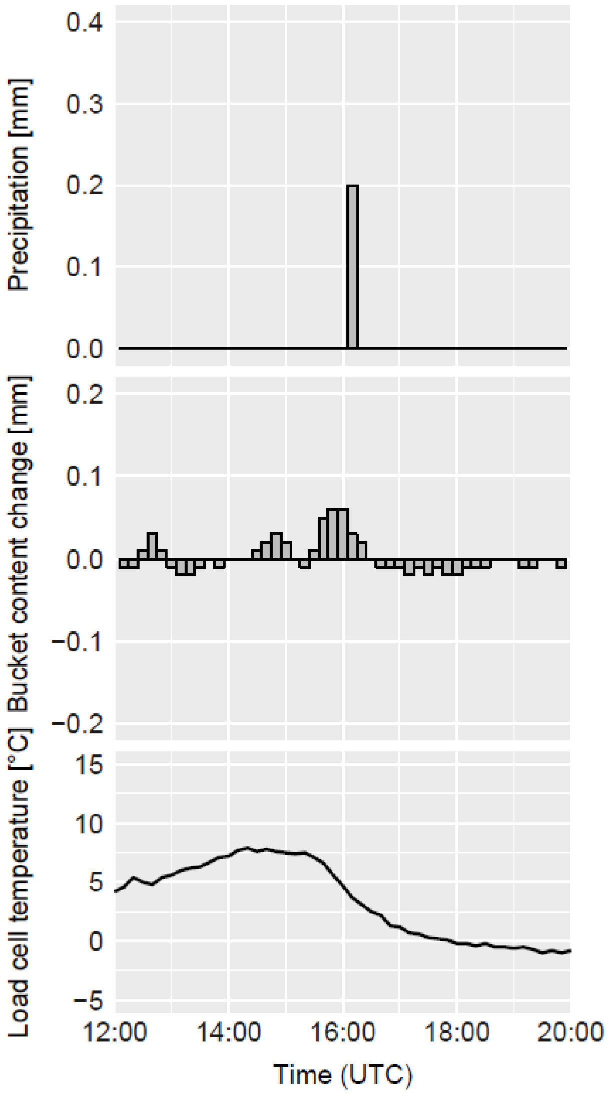
**Hypothesis:** Spurious precipitation is induced by rapid temperature changes of the load cell

**Analysis:** Training of SVM classifier on primary and auxiliary measurements and expert corrections

Classifier achieves high specificity and sensitivity

**Use for Quality Control:**

- Classifier is readily deployed as an AQC test
- Site-independent, near real-time
- Classifier relies only on instrument data
  → Collaboration with manufacturer to improve instrument



# Summary

- The growing volume of surface data is both a challenge and an opportunity
- Data-driven quality control works better if more data is available
- Machine-learning models enable a smooth transition from understanding the problem to implementing an automated QC algorithm

# Bibliography


Knechtl, V., M. Caseri, F. Lumpert, C. Hotz and C. Sigg (2019). *Detecting Temperature Induced Spurious Precipitation in a Weighing Rain Gauge.* Meteorologische Zeitschrift, 28(3), 215-225.

Schölkopf, B., J. C. Platt, J. Shawe-Taylor, A. Smola, and R. C. Williamson (2001). *Estimating the Support of a High-Dimensional Distribution.* Neural Computation, 13(7), 1443-1471.

WMO (2012). *Guide to the Global Observing System*. WMO No. 488.