06 Error Modelling

Parameter Retrieval from Earth Observation

VO 120.034

Overview

Introduction to statistics

A-posteriori error characterization and validation

- Scaling
- Characterization of statistical dependency
- Characterization of absolute deviations
- Triple collocation
- Example

Definitions

Measurement – *"process of experimentally obtaining one or more quantity values that can reasonably be attributed to a quantity"*

Measurand – *"quantity intended to be measured"*

Quantity – *"property of a phenomenon, body or substance, where the property has a magnitude that can be expressed by a number or reference"*

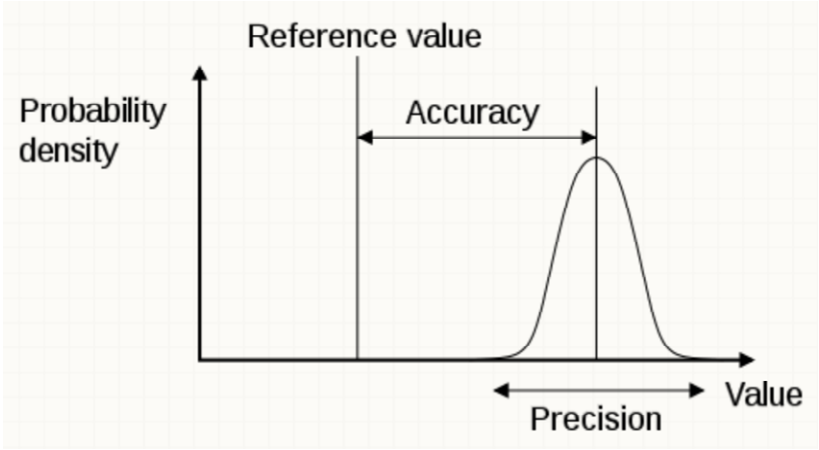
Measurement result – *"set of quantity values being attributed to a measurand together with any other available relevant information"*

Quantity value – *"number and reference together expressing magnitude of a quantity"*

Error vs. Uncertainty

Error

The measured value minus the true ***unknown*** value of the measurand. In practice, the error is unknowable, except when the measured value can be compared with a reference value of ***negligible*** uncertainty.



Uncertainty

A non-negative parameter characterising the spread of the quantity values attributed to a measurand, given the measured value and an understanding of the measurement.

Uncertainty types

Type-A:

uncertainty estimates using statistics i.e. by taking multiple readings and using that information

Type-B:

uncertainty estimates from any other information, e.g. past experience, calibration certificates, etc.

INTRODUCTION TO STATISTICS

Introduction to statistics

Random variables – elementary outcomes of a random experiment

- The probability of the random variable taking a particular value is determined by the probability of the outcome.

- If X is the sum of payoffs from 2 coin flips (heads is 1EUR, Tail 0EUR):

  P(X=0) = P({TT}) = ¼

  P(X=1) = P({HT})|P({TH}) = ½

  P(X=2) = P({HH}) = ¼

X takes on values {0,1,2} with p {¼, ½, ¼ }

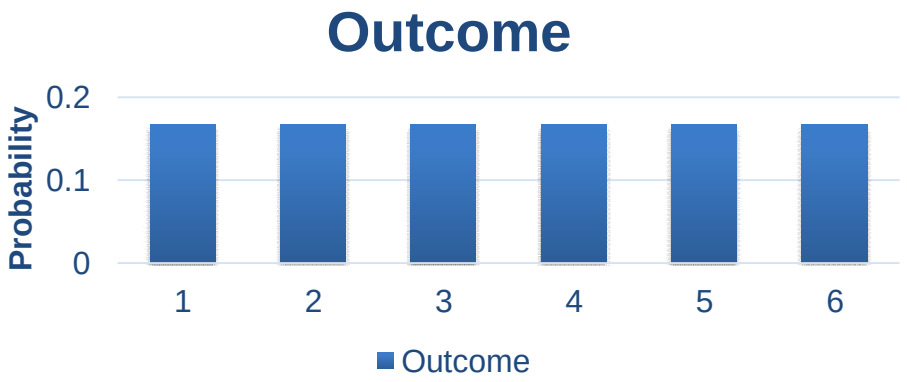
There are two types of random variables:

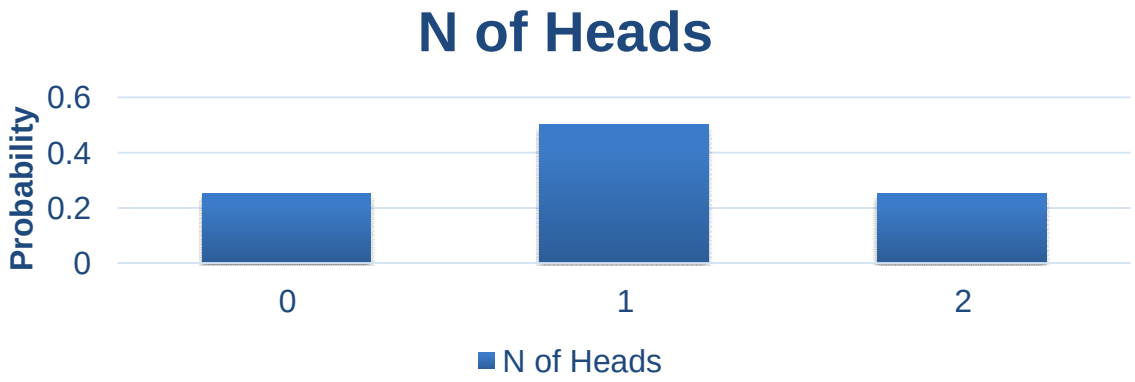
- Discrete random variables
- Continuous random variables

Discrete Random Variables

- Discrete – variables whose outcomes are separated by gaps – countable!
  - throwing dice → 1,2,3,4,5,6
  - Coin toss → Head/Tail

- Defined by probability mass function p
  - Sums to one
  1. Uniform distribution – equal probability for all outcomes
  2. Binomial distribution – yes/no, True/False

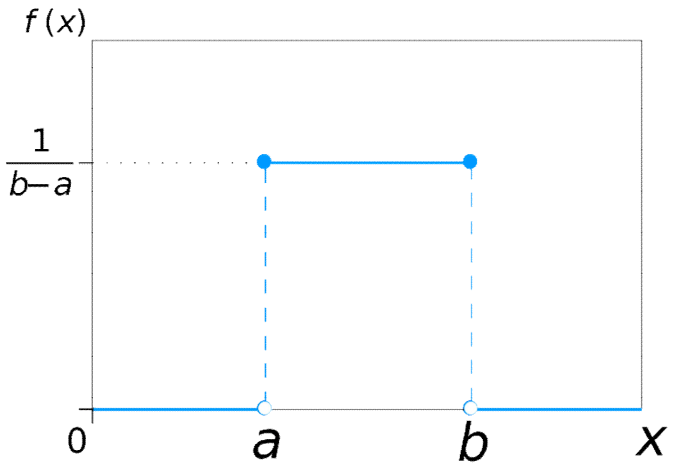


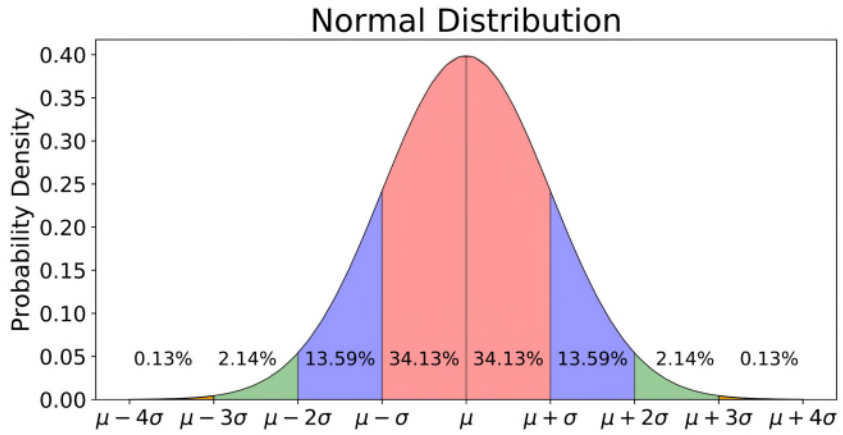


Continuous Random Variables

- Continuous – it can assume an infinite number of real values
  - Temperature
  - Height $? \geq a \geq 0$

- Probability of an event is expressed as the integral of a probability density function
  - Area under the pdf is equal to 1
  1. Continuous uniform distribution
  2. Normal (Gaussian) distribution





6.1 Statistical parameters

- If we know the distribution of a random variable we know all there is to know about the random variable.

- **But with real data we do not know the full distribution**

Moments: set of statistical parameters to measure a distribution.

- Geophysical parameters under validation are usually treated as continuous random variables
  - Focus on their distribution as given by their probability density function

Central tendency – first moment of distribution

Expected value - mean

- Discrete distribution – **X** discrete random variable (number on die) with list of probabilities p
  - Weighted sum of all values ***x*** of **X**, where weights are the probabilities p_i

$$\langle \mathrm{X} \rangle = E[(\mathrm{X})] = \sum x_i p_i$$

- Continuous distribution – **X** continuous random variable, probability density p as a function of *x*: p(*x*)
  - Integral of all values **x** of **X** weighted by their probability density function

$$\mu_{\mathrm{X}} = E[(\mathrm{X})] = \int_{-\infty}^{\infty} x p_{\mathrm{X}}(x) dx$$

Central tendency – first moment of distribution

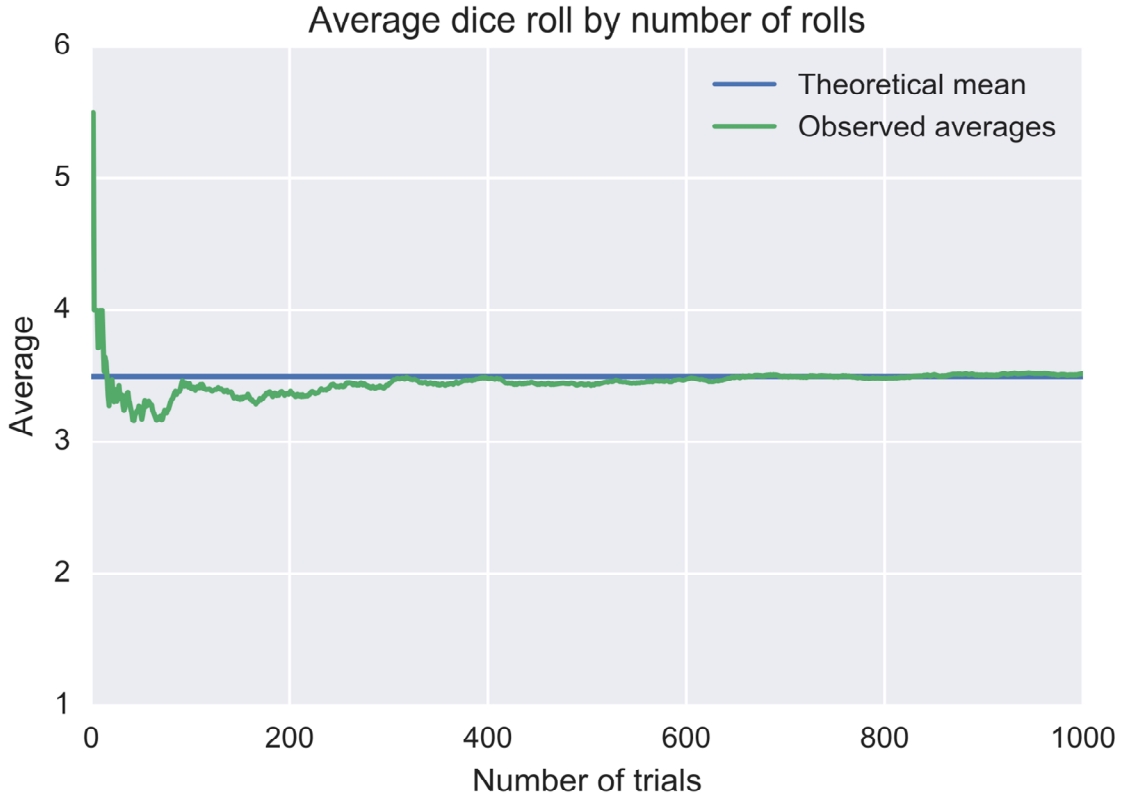
Total on dice	Pairs of dice	Probability	xi*pi
2	1+1	1/36 = 3%	0.06
3	1+2, 2+1	2/36 = 6%	0.17
4	1+3, 2+2, 3+1	3/36 = 8%	0.33
5	1+4, 2+3, 3+2, 4+1	4/36 = 11%	0.56
6	1+5, 2+4, 3+3, 4+2, 5+1	5/36 = 14%	0.83
7	1+6, 2+5, 3+4, 4+3, 5+2, 6+1	6/36 = 17%	1.17
8	2+6, 3+5, 4+4, 5+3, 6+2	5/36 = 14%	1.11
9	3+6, 4+5, 5+4, 6+3	4/36 = 11%	1.00
10	4+6, 5+5, 6+4	3/36 = 8%	0.83
11	5+6, 6+5	2/36 = 6%	0.61
12	6+6	1/36 = 3%	0.33
E[(X)]			**7.00**

Law of large numbers

If we have N independent observations of x_1, …, x_N which are independent and identically distributed the average of the results comes close to the expected value. The average is calculated as:

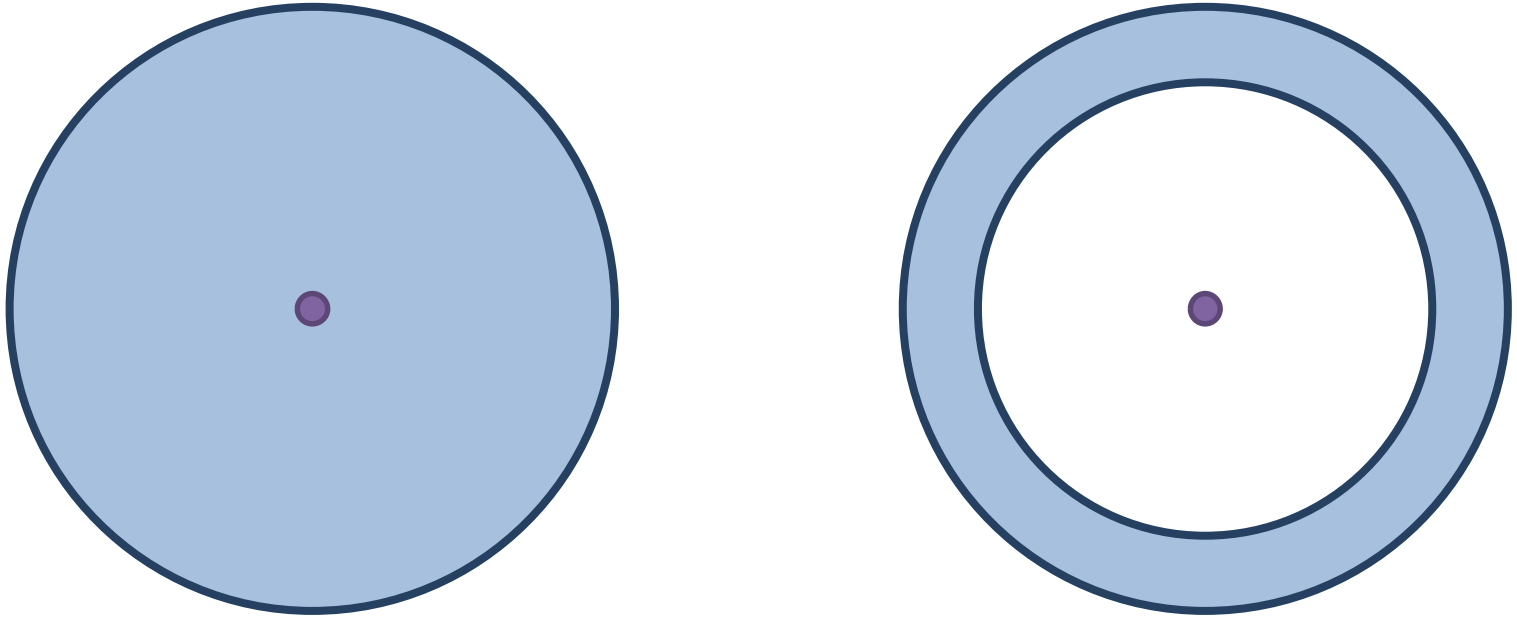
$$\hat{\mu}_X = \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

If $N \to \infty$,$\hat{\mu}_X$ comes close to the true parameter value μ_X.



Central tendency – first moment of distribution

Doesn’t tell us much yet about the distribution



Dispersion – second moment of distribution

Variance: How dispersed the distribution is **centered around the mean**:

$$\sigma_{\mathrm{X}}{}^2 = Var[\mathrm{X}] = E[(\mathrm{X} - E[\mathrm{X}])^2] = \sum (x_i - E[\mathrm{X}])^2 p_i$$

Dispersion – second moment of distribution

- Variance of continuous distribution

$$\sigma_{\mathrm{X}}{}^2 = E[(\mathrm{X} - E[\mathrm{X}])^2] = \int_{-\infty}^{\infty} (x - \mu_{\mathrm{X}})^2 p_{\mathrm{X}}(x)\, dx$$

For large numbers:

$$\hat{\sigma}_{\mathrm{X}}{}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \hat{\mu}_{\mathrm{X}})^2$$

Standard deviation

Unit of variance is the unit squared → inconvenient

To measure the dispersion in the same dimension we take the square root of the variance:

$$\sigma_X = \sqrt{\sigma_X^2}$$

$$\hat{\sigma}_X = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \hat{\mu}_X)^2}$$

Covariance

How well one variable can be predicted as a linear function of the other.

The covariance is a function of two random variables characterized by their joint pdf $p_{XY}(x,y)$:

$$\sigma_{XY} = E[(X - E[X])(Y - E[Y])] = \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) p_{XY}(x,y)\, dx dy$$

For large numbers this can be estimated as:

$$\hat{\sigma}_{XY} = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \hat{\mu}_X)(Y_i - \hat{\mu}_Y)$$

Covariance

Based on the **covariance of the data sets**. Not very informative since it depends on the magnitude of X and Y

$$\hat{\sigma}_{\mathrm{XY}} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \hat{\mu}_{\mathrm{X}})(y_i - \hat{\mu}_{\mathrm{Y}})$$

Consider the following conditions:

- xi> μX and yi> μY then (xi−μX)(yi−μY) will be **positive**.
- xi< μX and yi< μY then (xi−μX)(yi−μY) will be **positive**.
- xi> μX and yi< μY then (xi−μX)(yi−μY) will be **negative**.
- xi< μX and yi> μY then (xi−μX)(yi−μY) will be **negative**.

Covariance

Variances and covariances of two or more variables are often arranged in a so called **covariance matrix**.

For our 2 variable example, the covariance matrix **C** is given by:

$$\mathrm{C} = \begin{pmatrix} {\sigma_X}^2 & \sigma_{XY} \\ \sigma_{YX} & {\sigma_Y}^2 \end{pmatrix}$$

The diagonal contains the variances, while the off-diagonal contains the covariances of the variables.

A-POSTERIORI ERROR CHARACTERIZATION AND VALIDATION

Validation: A-posteriori error characterization

We aim to characterize the accuracy of the parameter estimates by comparing them with reference measurements.

- No reference data set will ever perfectly measure the true state of any geophysical parameter → relative error levels with respect to the "truth"

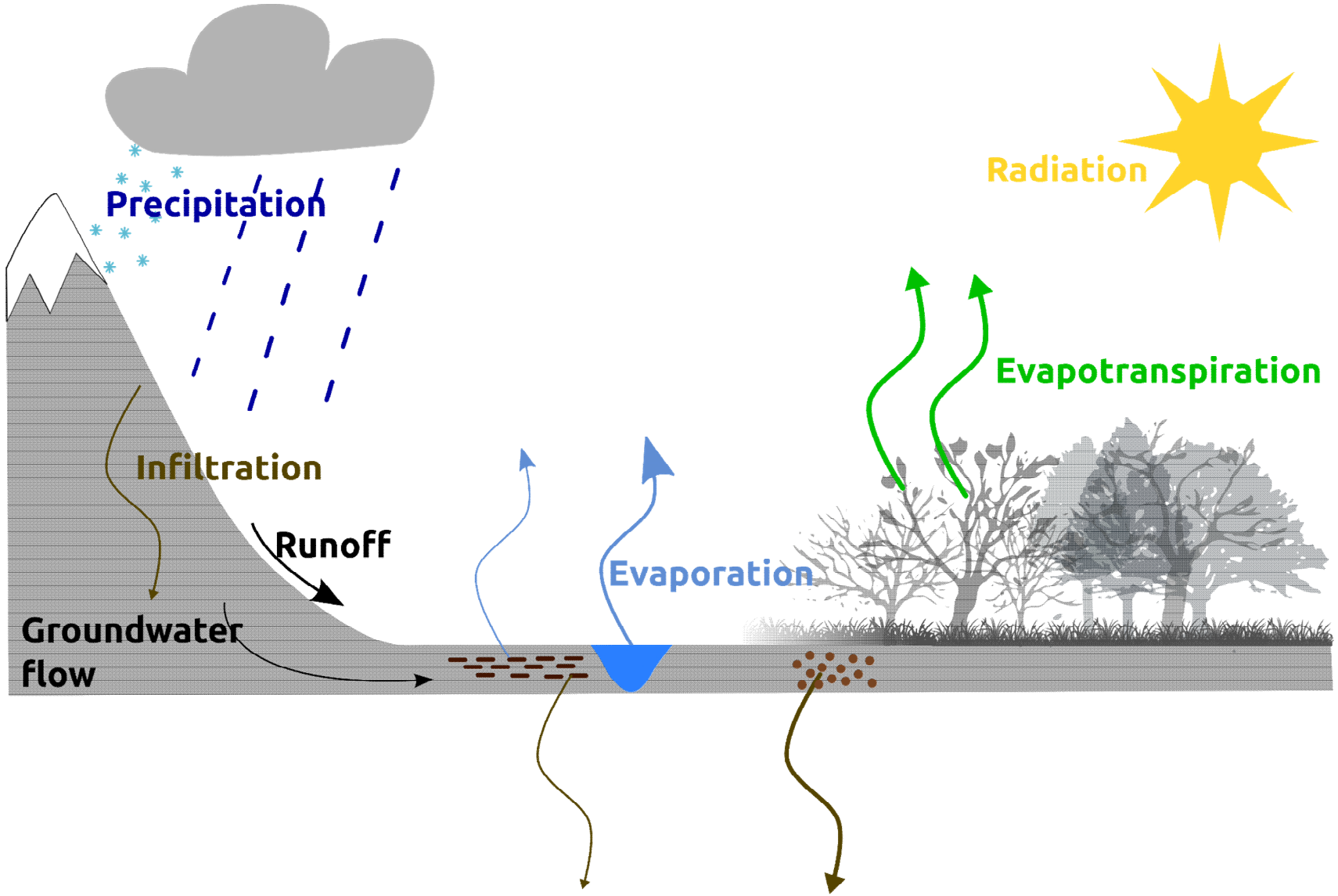
- Scaling differences can exist and need to be accounted for!

SCALING

Scales

Process	Measurement	Model
• Depends on drivers of geophysical variable	• Depends on instrument	• Depends on assumptions and input data

Process scale



Soil moisture:

- Soil type
- Terrain
- Vegetation cover
- Precipitation
- Radiation

Measurement and model scale

Soil moisture can be measured and modelled at different scales:

1. Remote sensing – **average over a footprint**
2. In situ – **point measurement**
3. Models – Any scale

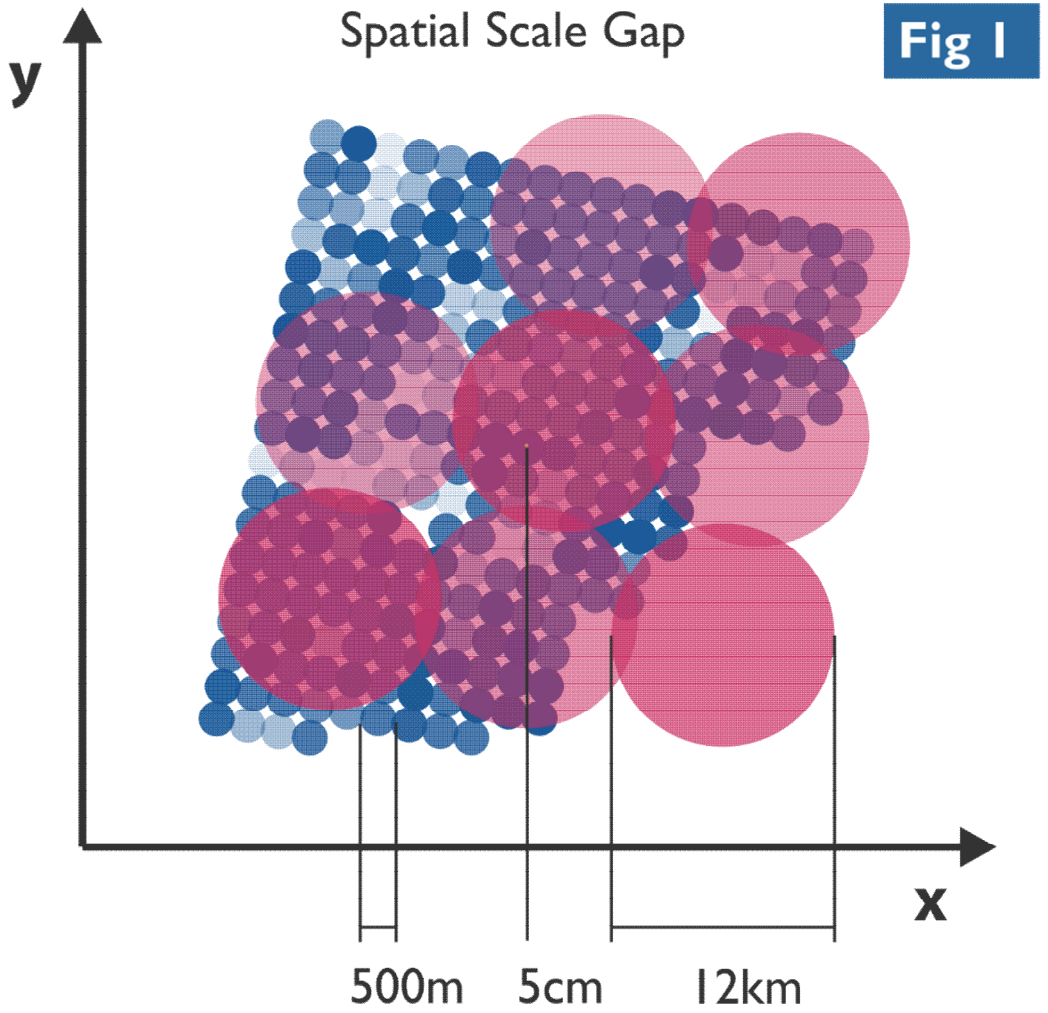
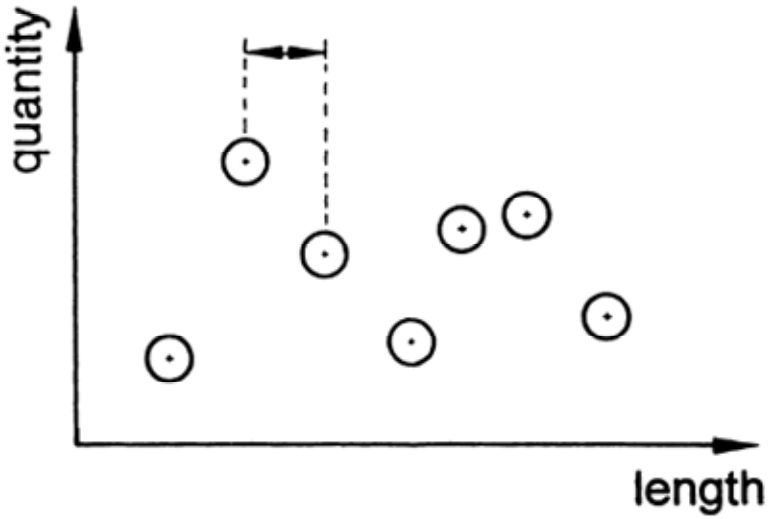


*Figure courtesy of Bernhard Bauer-Marschallinger et al. (2015).

Measurement and model scale

spacing

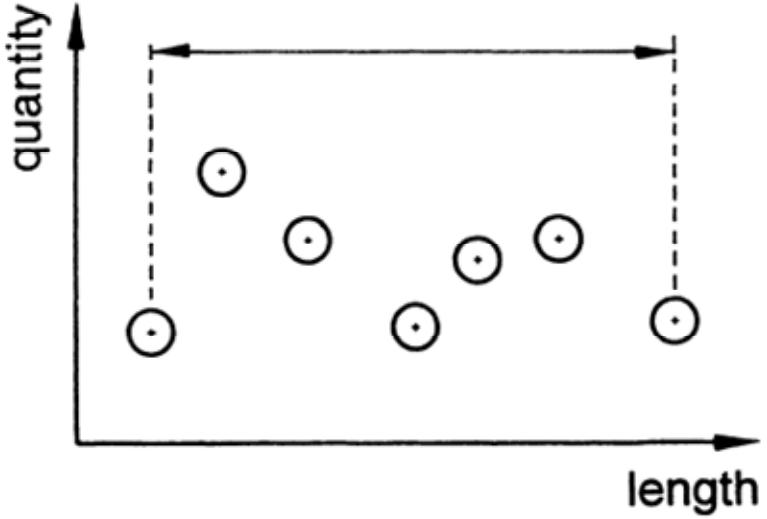


Distance between single measurements.

ASCAT: Resampled to a 12.5 km grid (spacing)

In situ: distance between sm stations

extent

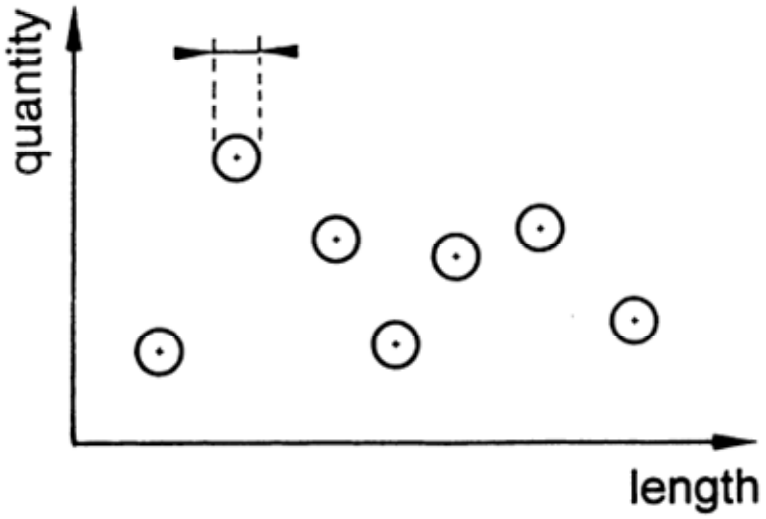


The overall sampled area.

ASCAT: 550 km wide swath (extent)

In situ: spatial coverage of the nework

support



The area over which the measurement integrates, i.e., the spatial resolution.

ASCAT: 25 km resolution (support)

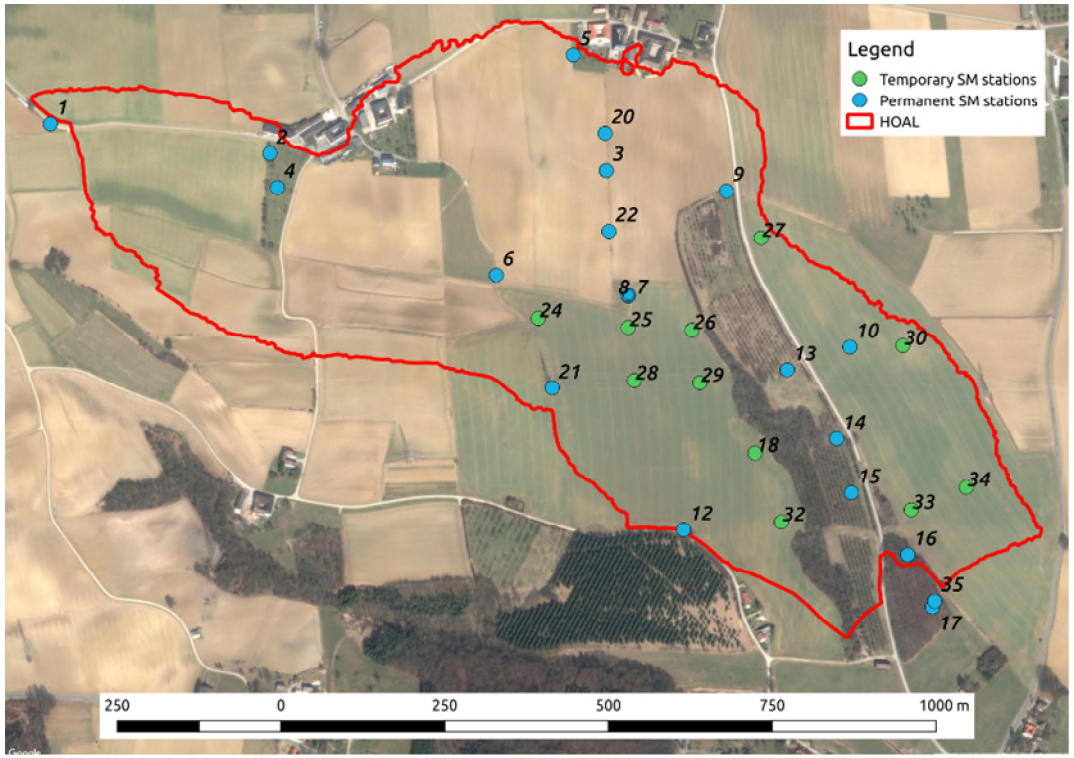
In situ: volume of the sample

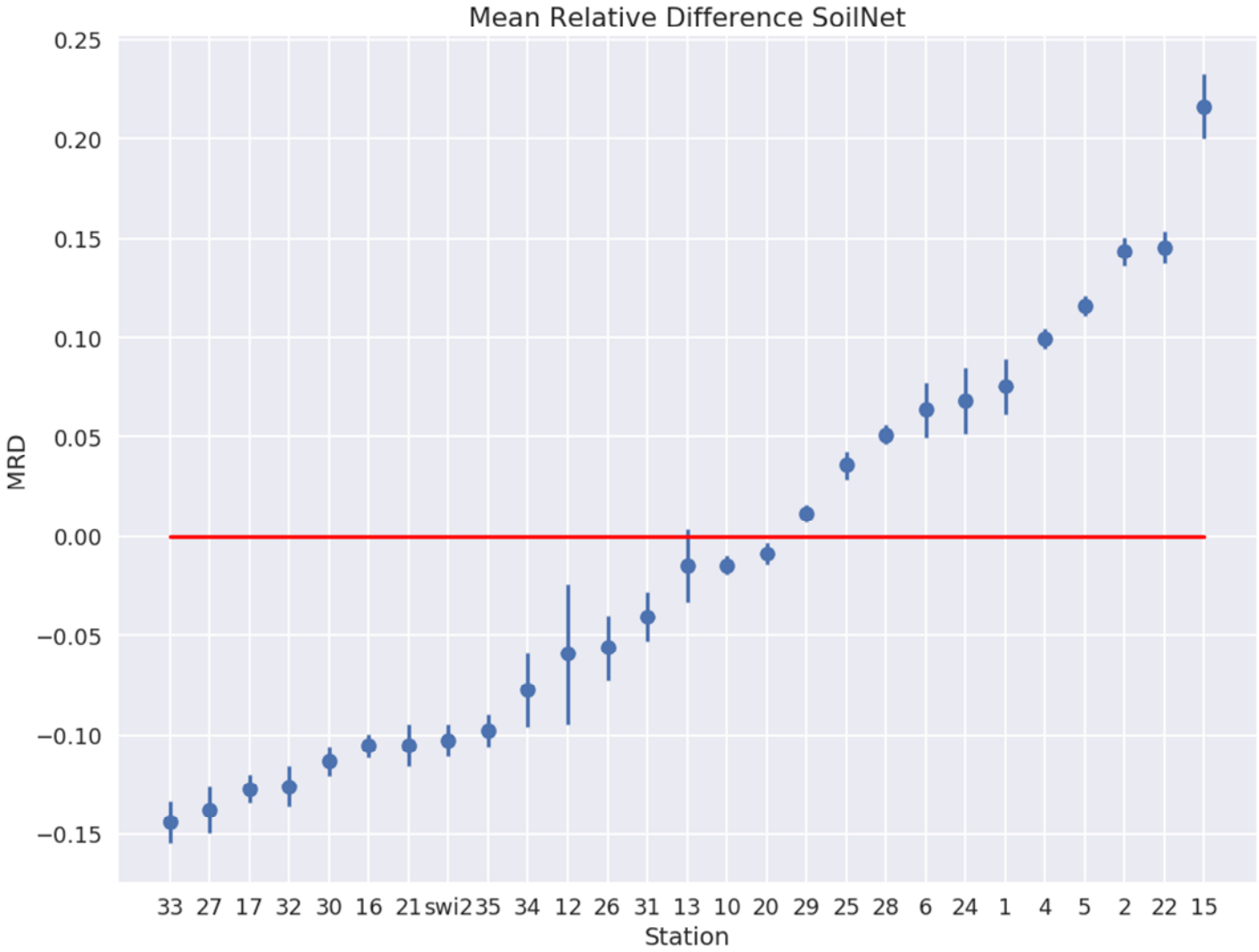
Scaling errors

- Systematic scaling errors – from stable properties
  - Topography
  - Soil texture
  - Land cover
- ➢ Can be corrected

- Random scaling errors – effect cannot be predicted at measurement scale
  - Precipitation
  - Radiation
- ➢ Can be characterized

Mean relative difference





- Which station is the most representative?

Rescaling methods

Two datasets X and Y

1. Linear relation: Y =a+bX
    1. Linear regression:
    2. Min-Max
    3. Mean Stadard deviation fit
2. Non linear relationship
    1. CDF-matching

Linear regression scaling

In the linear regression, the additive and multiplicative biases *a* and *b* are determined in a least-squares fashion, i.e., by minimizing the squared differences between the measurement.

Rescaled Y can then be calculated as:

$$Y^X = \frac{Y - a}{b}$$

Min Max scaling

- Another possible condition for the determination of the scaling coefficients is that the absolute minimum and maximum values should match

$$Y^X = \frac{Y - \min(Y)}{\max(Y) - \min(Y)} \cdot [\max(X) - \min(X)] + \max(X)$$

Mean standard deviation scaling

- Slightly more robust against outliers and small non-linearities than the minimum-maximum fit is the fitting of the mean and the standard deviation

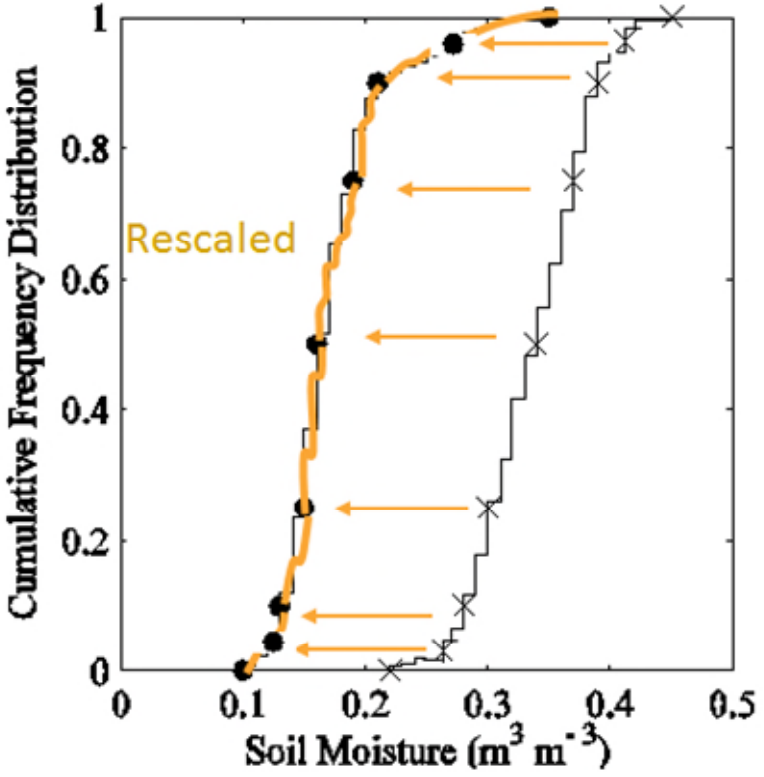
$$Y^X = \frac{Y - \bar{Y}}{\sigma_Y} \cdot \sigma_X + \bar{X}$$

CDF matching

- In some cases a non-linear relationship is observed between datasets X and Y

$$Y = a + bX + cX^2 + dX^3$$

- To account for this we match the cumulative distribution functions:
  - Fitting data through a higher order polynomial
  - Sorting data and splitting in percentiles and scale through linear regression



CHARACTERIZATION OF STATISTICAL DEPENDENCY

Correlation metrics

- ✓ Are largely independent of systematic bias as it is a measure of the covariance
- – Random errors will affect correlation

- Linear correlation (Pearson)
- Rank correlation (Spearman)

Linear correlation

Pearson product-moment correlation coefficient

- Measure of association that is not affected by changes in the scales of the variables
- Covariance between two datasets normalized with their respective standard deviation.
  - Scales between -1 and 1

$$\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma X \sigma Y} = \frac{\frac{1}{N-1}\sum_{i=1}^{N}[(x_i - \mu_X)(y_i - \mu_Y)]}{\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_X)^2}\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu_Y)^2}}$$

We use ρ for the population correlation, and R or r for the sample correlation

Rank correlation

Spearman rank correlation coefficient

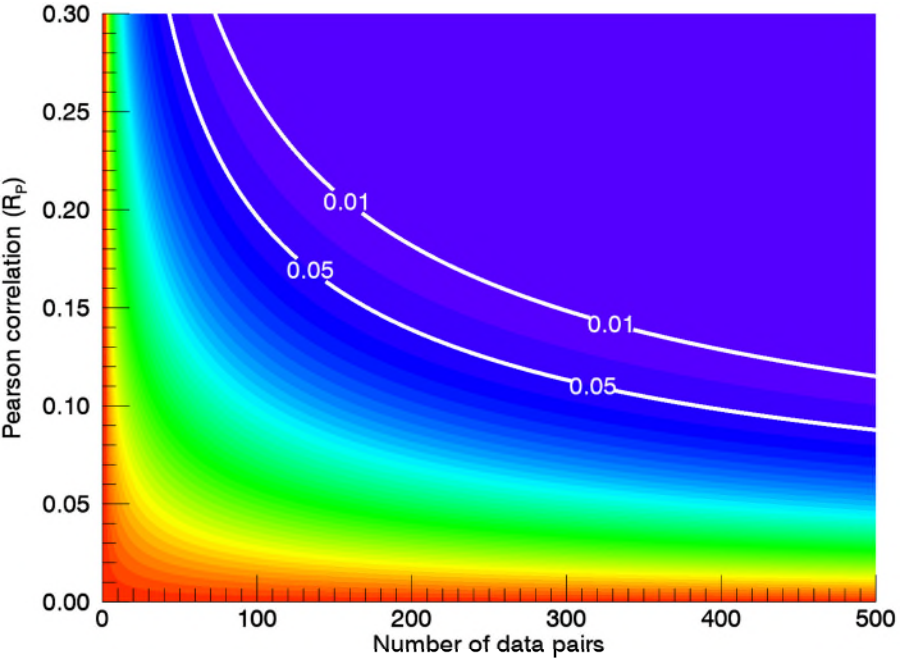
- relationship between two variables described using a monotonic function
- Same as Pearson between the rank values of two variables

$$\rho_{X,Y} = \frac{\sigma_{rXrY}}{\sigma_{rX}\sigma_{rY}}$$

Correlation significance

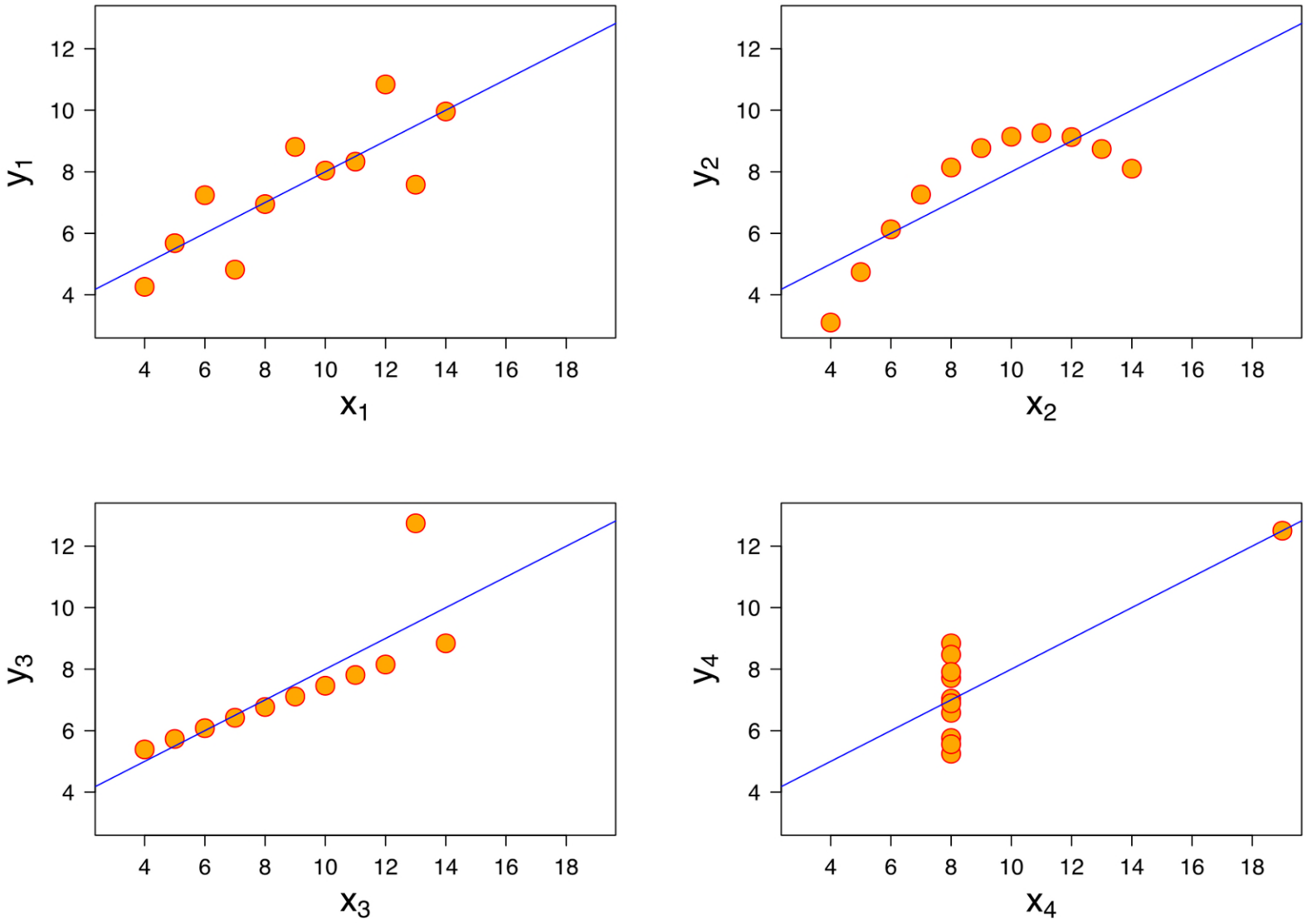
- How to test if correlation is a coincidence?
- Common test for correlation significance is the Student's t-test
  - links the observed correlation level to the number of available data pairs

$$t = R \cdot \sqrt{\frac{N-2}{1-R^2}}$$



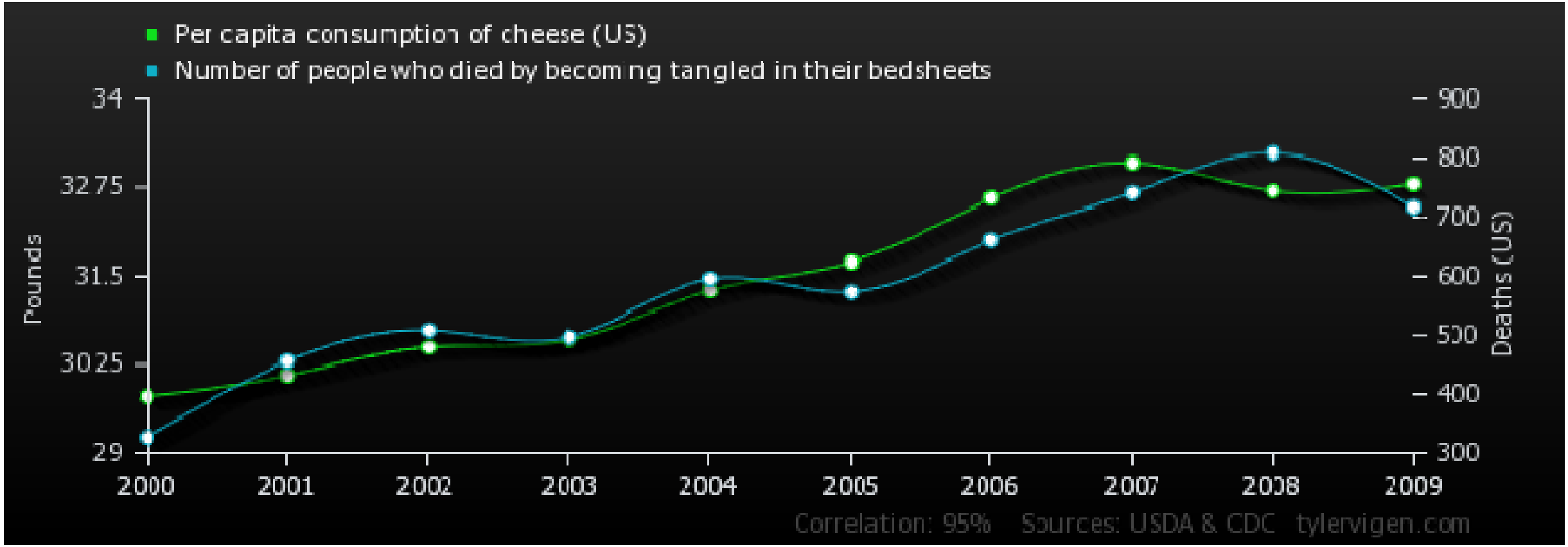
Relationship between the Pearson correlation, the number of data pairs and the correlation significance. White lines are the most common probability thresholds (0.05 and 0.01).

Be careful with correlations



- Four sets of data with the same correlation of 0.816

Be careful with correlations



R_p=**0.947091**

CHARACTERIZATION OF ABSOLUTE DEVIATIONS

Difference metrics

Statistical measures how well the actual measurement values agree.

Based on deviations from the mean and standard deviation

- Very sensitive to systematic biases
- Random errors affect the results

- Bias
- Standard deviation ratio
- Root Mean Squared Difference

Bias

Difference between the mean values of two data sets

- Used to characterize a systematic over- or underestimation of the estimated parameter states with respect to the reference

$$bias = \bar{X} - \bar{Y}$$

- spatial scale of the data sets matches
- no rescaling has been applied
- to particularly estimate systematic scaling errors using two data sets which are assumed to have negligible systematic measurement errors themselves

Standard deviation ratio

Measure for systematic differences between two data sets, but this time of second-order.

Compares the variability of the data sets instead of their mean values

$$SDR = \frac{\sigma_X}{\sigma_Y}$$

Root mean square difference

Square root of the averaged squared differences between two data set

Average absolute deviation between single measurement

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathrm{X}_i - \mathrm{Y}_i)^2}$$

RMSD consists of three error sources, correlation, bias and difference in variance as:

$$RMSD = \sqrt{MSD_{corr} + MSD_{bias} + MSD_{var}}$$

$$RMSD = \sqrt{2\sigma_X\sigma_Y(1 - R_P) + (\mu_X - \mu_Y)^2 + (\sigma_X - \sigma_Y)^2}$$

Root mean square difference

Characterize random and systematic errors of a data set individually:

1. Subtract the bias:

$$cRMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[\left(\mathbf{x}_i - \mu_{\mathbf{X}}\right) - \left(\mathbf{y}_i - \mu_{\mathbf{Y}}\right)\right]^2}$$

2. Scale datasets

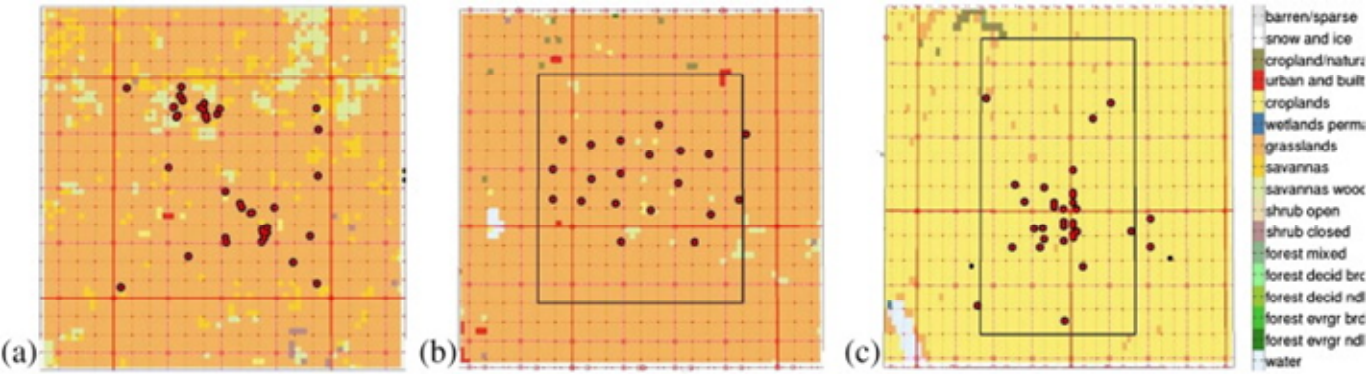
$$\mathrm{ub}RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{x}_i - \mathbf{y}_i^{\mathrm{X}}\right)^2}$$

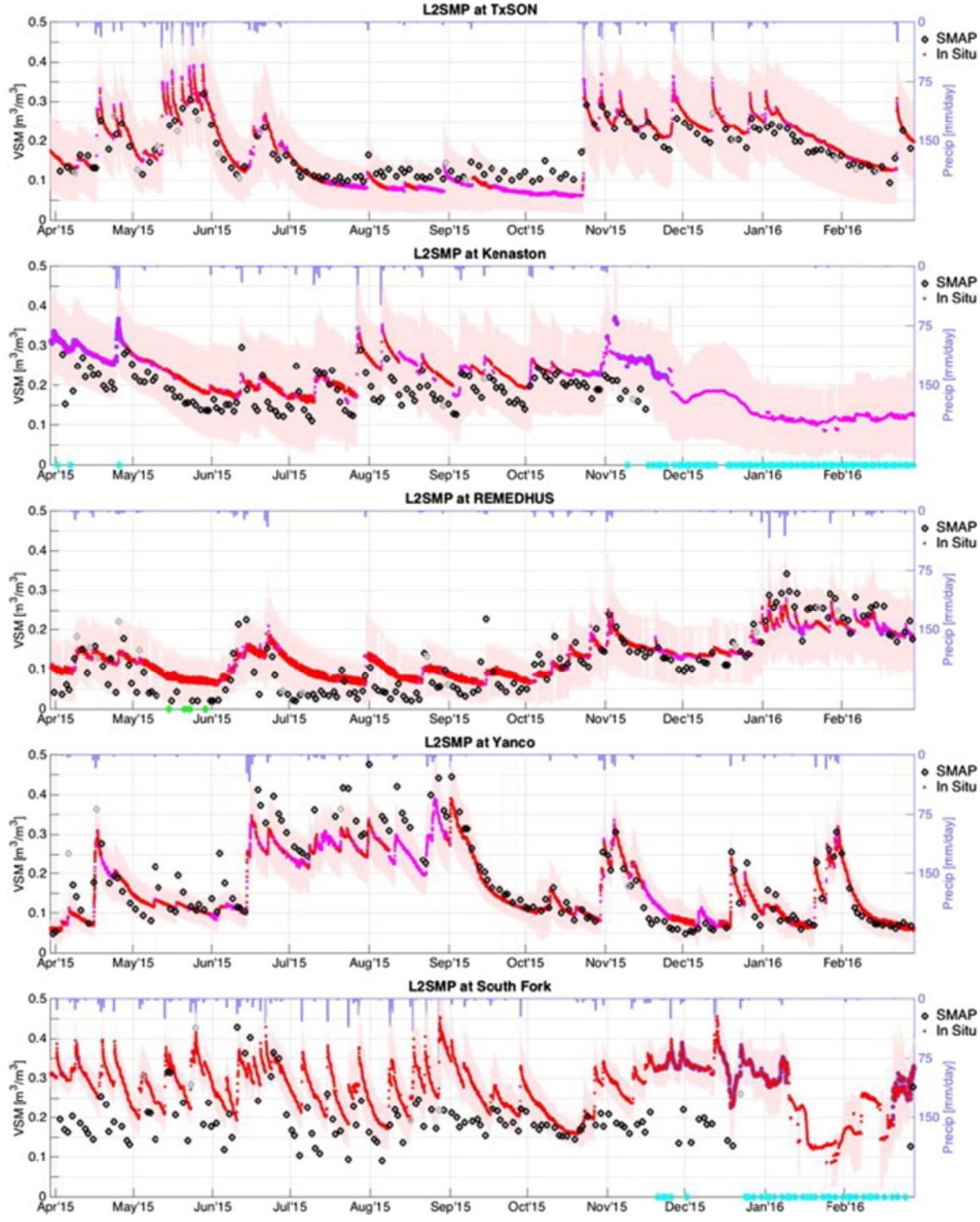
represent random errors of the data set under evaluation with respect to the reference data set, plus additional representativeness error

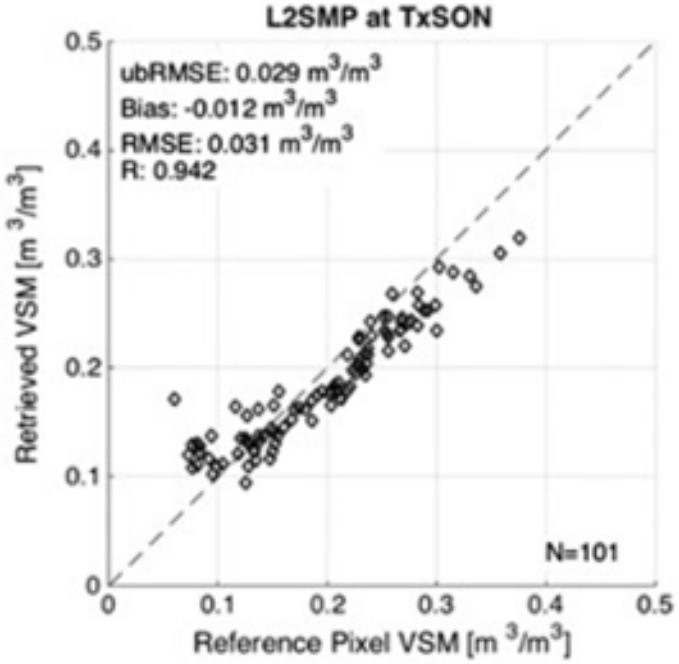
EXAMPLES

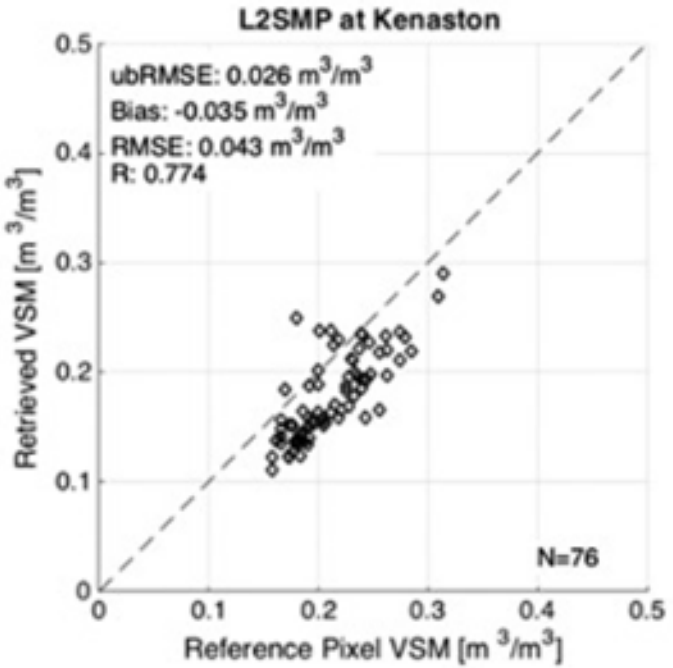
Example validation

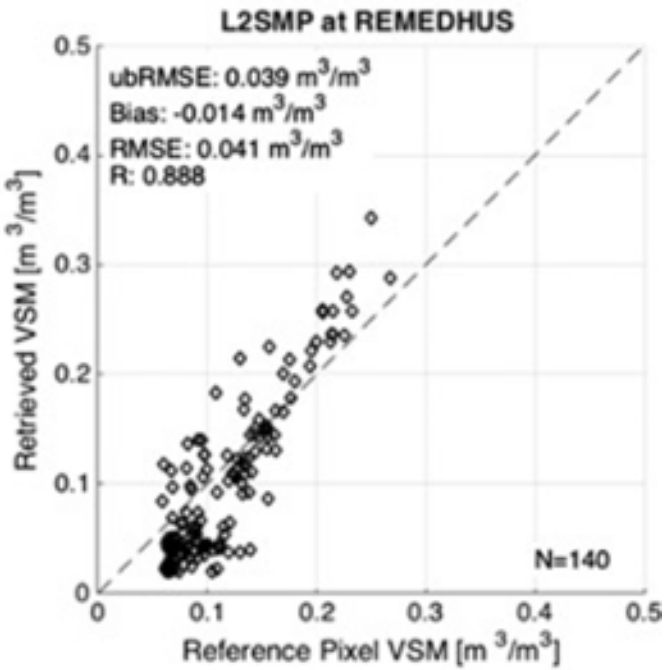
Validation of NASAs SMAP soil moisture product with different in situ data sets

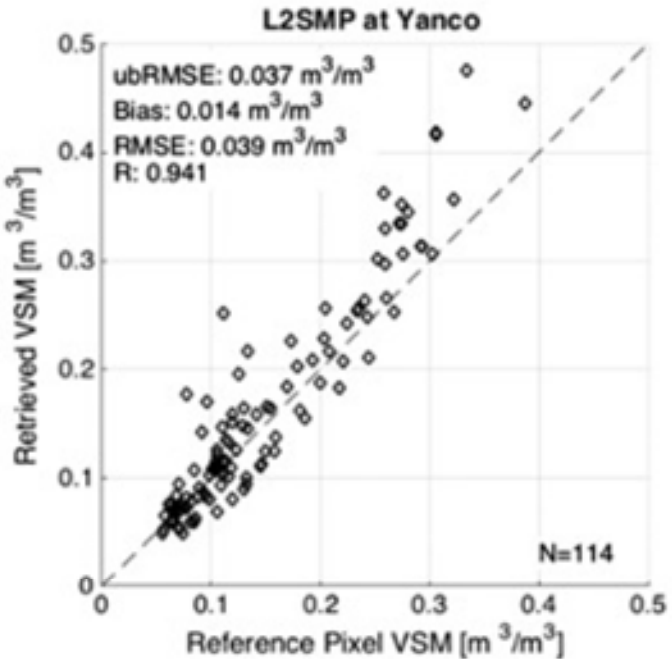


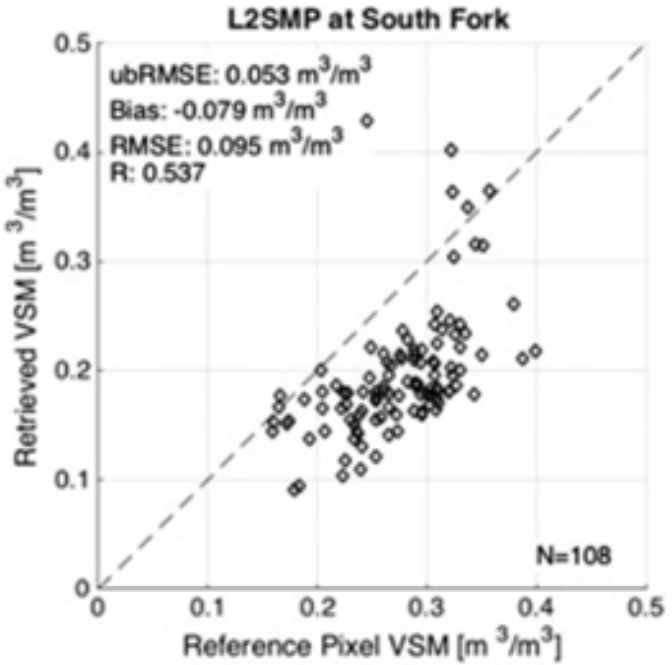



L2SMP at TxSON
ubRMSE: 0.029 m3/m3
Bias: -0.012 m3/m3
RMSE: 0.031 m3/m3
R: 0.942
N=101
Retrieved VSM [m 3/m3]
Reference Pixel VSM [m 3/m3]

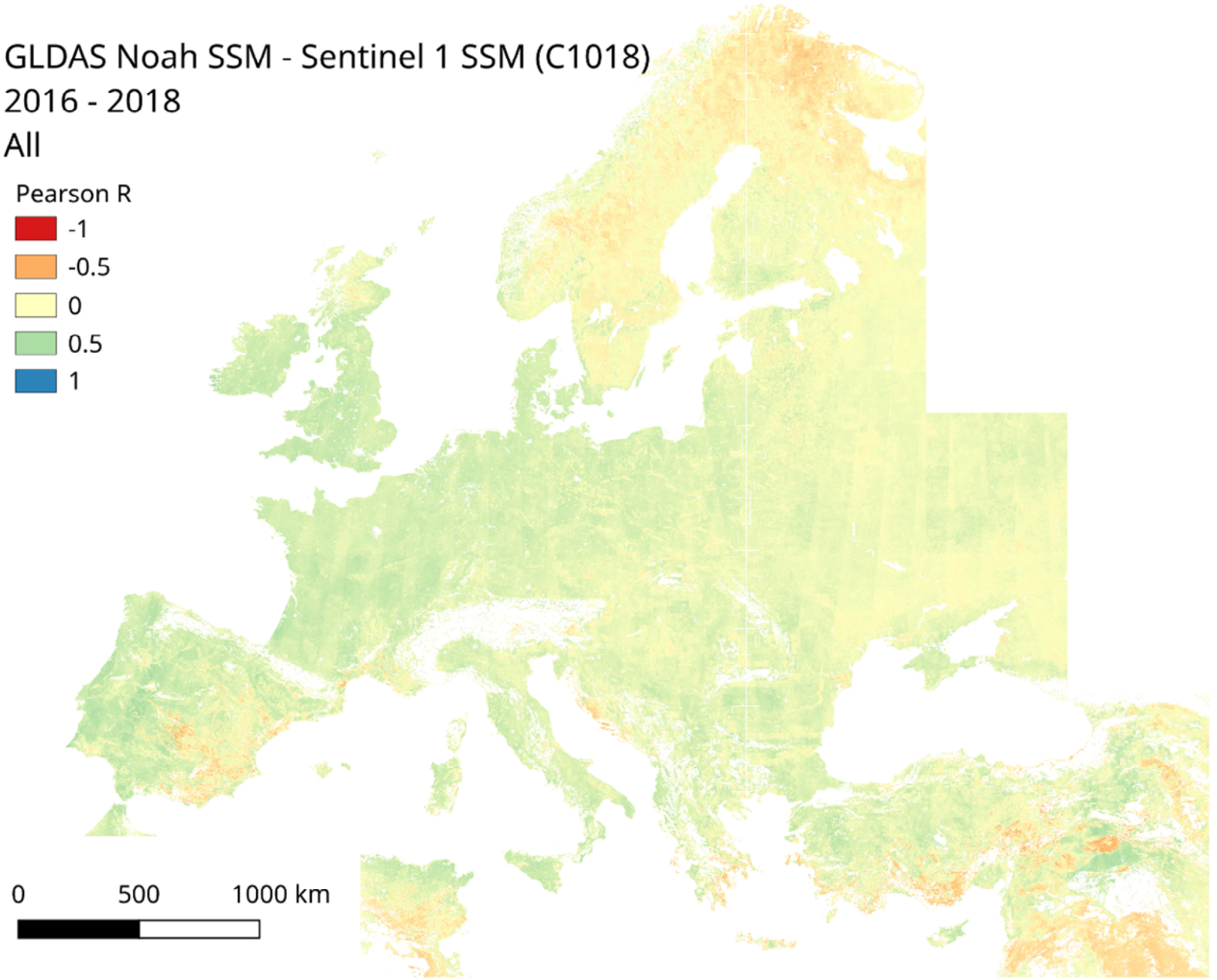

L2SMP at Kenaston
ubRMSE: 0.026 m3/m3
Bias: -0.035 m3/m3
RMSE: 0.043 m3/m3
R: 0.774
N=76
Retrieved VSM [m 3/m3]
Reference Pixel VSM [m 3/m3]


L2SMP at REMEDHUS
ubRMSE: 0.039 m3/m3
Bias: -0.014 m3/m3
RMSE: 0.041 m3/m3
R: 0.888
N=140
Retrieved VSM [m 3/m3]
Reference Pixel VSM [m 3/m3]


L2SMP at Yanco
ubRMSE: 0.037 m3/m3
Bias: 0.014 m3/m3
RMSE: 0.039 m3/m3
R: 0.941
N=114
Retrieved VSM [m 3/m3]
Reference Pixel VSM [m 3/m3]


L2SMP at South Fork
ubRMSE: 0.053 m3/m3
Bias: -0.079 m3/m3
RMSE: 0.095 m3/m3
R: 0.537
N=108
Retrieved VSM [m 3/m3]
Reference Pixel VSM [m 3/m3]

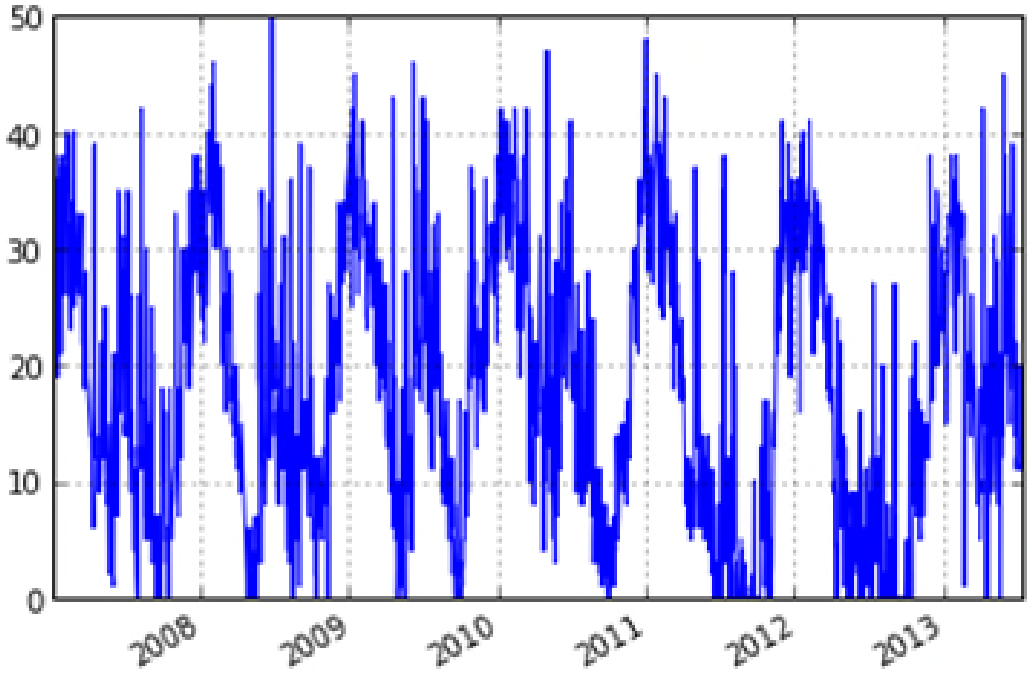
GLDAS Noah SSM - Sentinel 1 SSM (C1018)
2016 - 2018
All



Absolute vs Anomalies

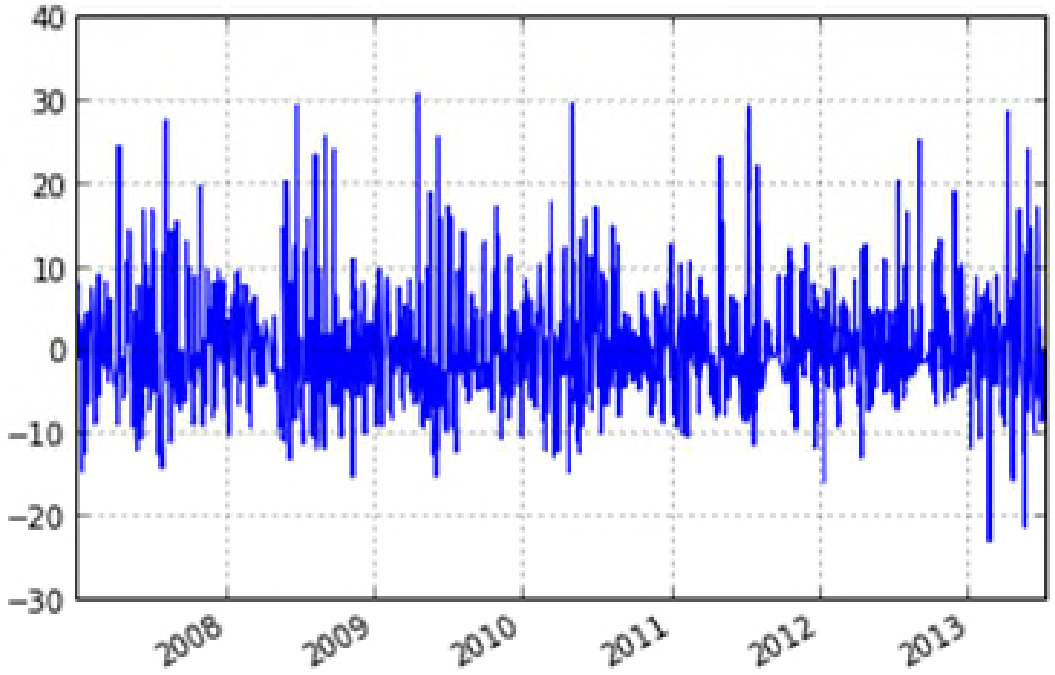
Absolute values

- Metrics are strongly related to seasonal signal



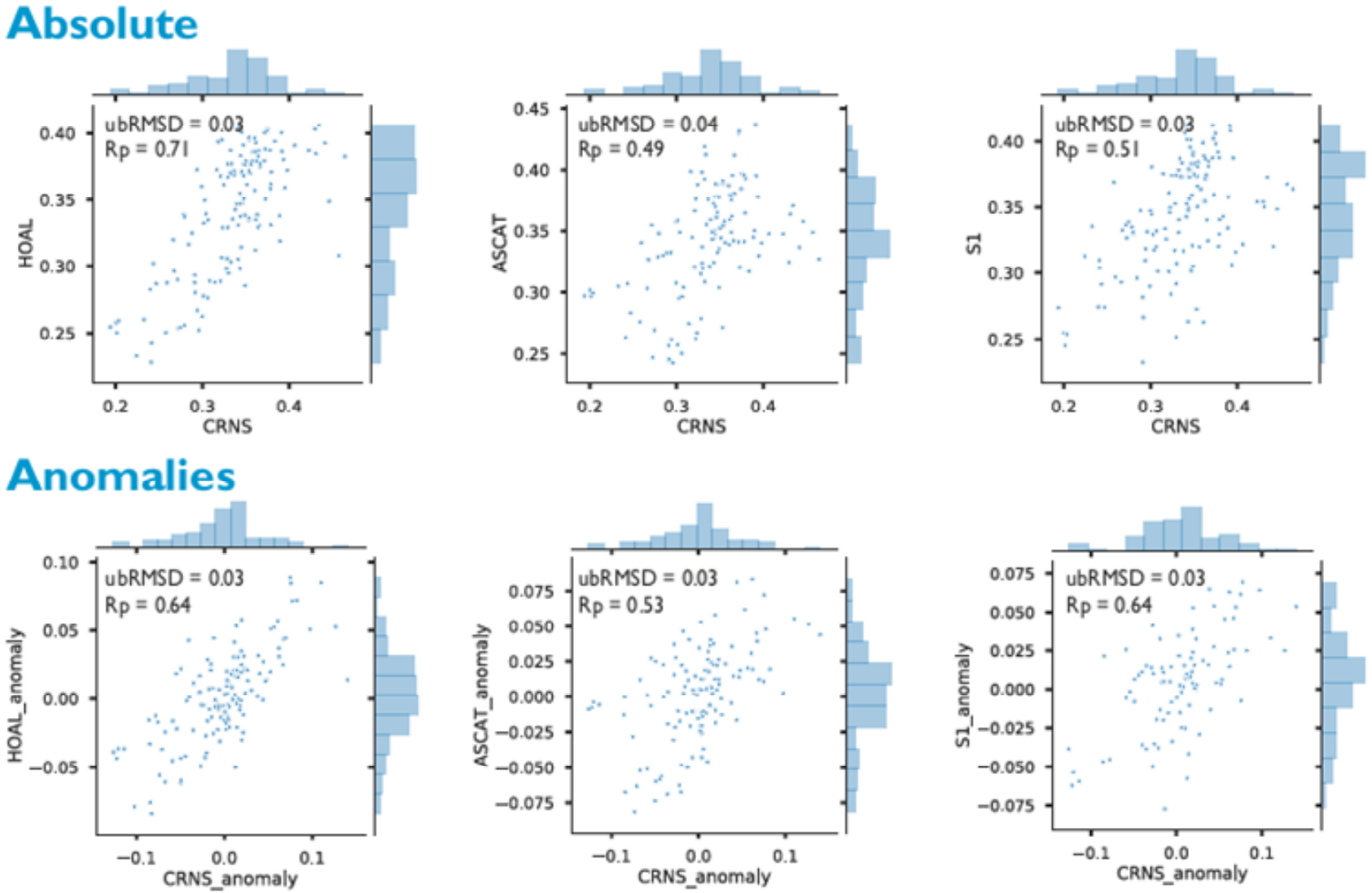
Anomalies

- Metrics are related to single events



Example

Correlation (R_p) and ubRMSD between soil moisture from in situ data, Sentinel-1 and ASCAT with a Cosmic Ray Sensor



Triple Collocation

All previous methods assume the reference dataset to be the “truth”

- strongly dependent on quality of reference dataset

If there is no true reference measurement:

Estimate the random error variance in three collocated datasets of the same geophysical variable (Stoffelen, 1998).

$$RMSE_{\mathrm{X}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathrm{X}_i - \mathrm{Y}_i)(\mathrm{X}_i - \mathrm{Z}_i)}$$

- We assume all three datasets are related to the same geophysical parameter
- All contain a systematic and random error term

$$X = \alpha_X + \beta_X\theta + \boldsymbol{\varepsilon}_X$$

Θ true value, α_i is the additive and β_i multiplicative systematic error terms (biases) and **ε the zero mean random error**.

Triple Collocation

1. In order to get rid of relative systematic differences between the data sets, we scale all to one dataset:

$$Y^X = \hat{\Theta}^X + \varepsilon_Y^X$$
$$Z^X = \hat{\Theta}^X + \varepsilon_Z^X$$

2. Subtracting:

$$X^X - Y^X = \hat{\Theta}^X + \varepsilon_X^X - \hat{\Theta}^X + \varepsilon_Y^X = \varepsilon_X^X - \varepsilon_Y^X$$

3. Cross multiply differences:

$$(X^X - Y^X)(X^X - Z^X) = {\varepsilon_X^X}^2 - \varepsilon_X^X \varepsilon_Y^X - \varepsilon_X^X \varepsilon_Z^X - \varepsilon_Y^X \varepsilon_Z^X$$
$$(Y^X - X^X)(Y^X - Z^X) = {\varepsilon_Y^X}^2 - \varepsilon_Y^X \varepsilon_X^X - \varepsilon_Y^X \varepsilon_Z^X - \varepsilon_X^X \varepsilon_Z^X$$
$$(Z^X - X^X)(Z^X - Y^X) = {\varepsilon_{XZ}^X}^2 - \varepsilon_Z^X \varepsilon_X^X - \varepsilon_Z^X \varepsilon_Y^X - \varepsilon_X^X \varepsilon_Y^X$$

Triple collocation

4. If we have a large enough sample:

$$\left\langle {\varepsilon_i^X}^2 \right\rangle = Variance\left(\varepsilon_i^X\right) = MSE_i^X$$

$$\left\langle \varepsilon_i^X \varepsilon_j^X \right\rangle = Cov(\varepsilon_i^X \varepsilon_j^X)$$

5. Assuming mutually independent errors → $\left\langle \varepsilon_i^X \varepsilon_j^X \right\rangle = Cov(\varepsilon_i^X \varepsilon_j^X) = 0$

$$(X^X - Y^X)(X^X - Z^X) = {\varepsilon_X^X}^2 - \cancel{\varepsilon_X^X \varepsilon_Y^X} - \cancel{\varepsilon_X^X \varepsilon_Z^X} - \cancel{\varepsilon_Y^X \varepsilon_Z^X}$$

$$(Y^X - X^X)(Y^X - Z^X) = {\varepsilon_Y^X}^2 - \cancel{\varepsilon_Y^X \varepsilon_X^X} - \cancel{\varepsilon_Y^X \varepsilon_Z^X} - \cancel{\varepsilon_Y^X \varepsilon_Z^X}$$

$$(Z^X - X^X)(Z^X - Y^X) = {\varepsilon_{XZ}^X}^2 - \cancel{\varepsilon_Z^X \varepsilon_X^X} - \cancel{\varepsilon_Z^X \varepsilon_Y^X} - \cancel{\varepsilon_Y^X \varepsilon_Y^X}$$

6. Calculating the final error:

$$RMSE_X^X = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\langle\left[\left(X_i^X - Y_i^X\right)\left(X_i^X - Z_i^X\right)\right]\rangle} = \sqrt{Var(\varepsilon_X^X)} = StdDev(\varepsilon_X^X)$$

Triple Collocation

$RMSE_i$ are thus estimates of the standard deviations StdDev() of the random errors of the individual data sets.

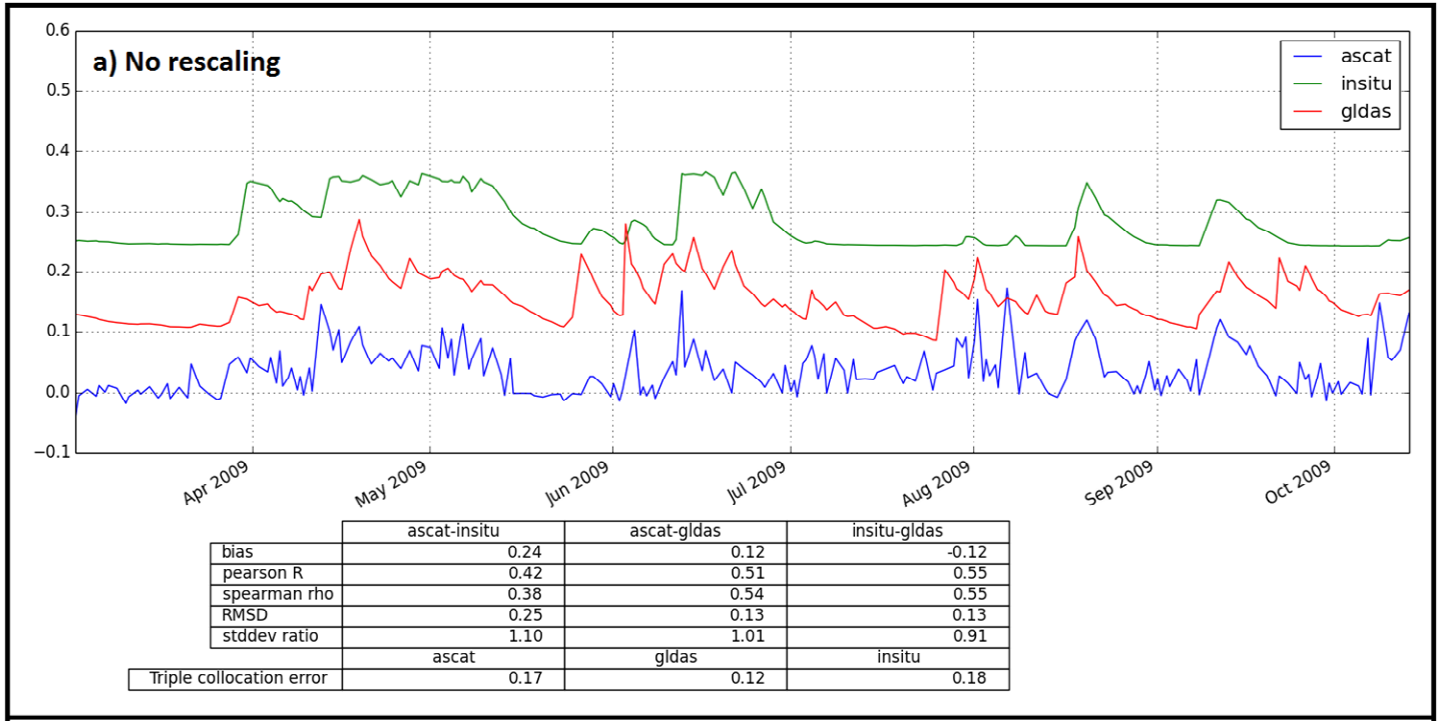
Assumptions:

- **errors of the individual data sets are to be mutually independent**
- triple collocation requires three data sets that are collocated in space and time
  - error estimates of the data sets will contain their individual measurement noise plus the representativeness error
- requires theoretically an infinite number of triplets for the covariances to converge to zero and for the total result to converge to the error variance
- error estimates will still contain the systematic error of the chosen reference with respect to the "common truth"

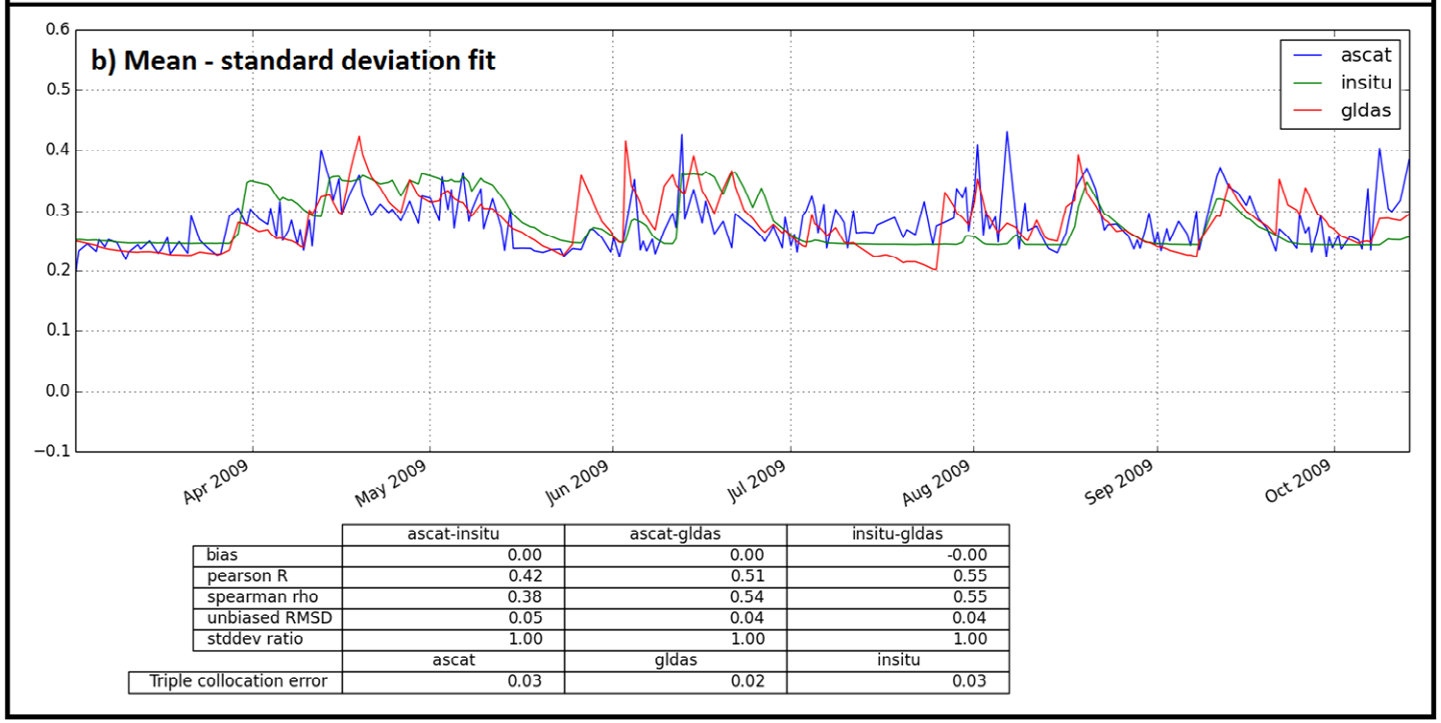
Example

With rescaling:

- Absolute metrics change
- Relative metrics stay the same



	ascat-insitu	ascat-gldas	insitu-gldas
bias	0.24	0.12	-0.12
pearson R	0.42	0.51	0.55
spearman rho	0.38	0.54	0.55
RMSD	0.25	0.13	0.13
stddev ratio	1.10	1.01	0.91
	ascat	gldas	insitu
Triple collocation error	0.17	0.12	0.18



	ascat-insitu	ascat-gldas	insitu-gldas
bias	0.00	0.00	-0.00
pearson R	0.42	0.51	0.55
spearman rho	0.38	0.54	0.55
unbiased RMSD	0.05	0.04	0.04
stddev ratio	1.00	1.00	1.00
	ascat	gldas	insitu
Triple collocation error	0.03	0.02	0.03

MULTIPLE CHOICE EXAM
15 JUNE