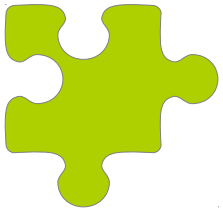
Empirical Statistical Downscaling

Diagnostics & Predicions



Diagnostics & Predictions

Statistics & Physics

Predictors

Analysis & Evaluation

Calibration

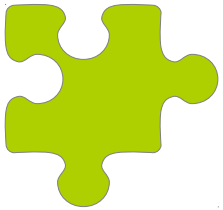
The two-component aspect

Model-observation bridging

Perfect Prog - 'PP'

Model Output Statistics - 'MOS'

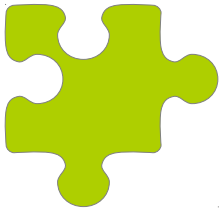
'PP-MOS' Hybrid approaches



Statistics & Physics

Physical dependencies must be reflected in the statistics.

The statistics is more robust when it reflect the physics.



Predictors

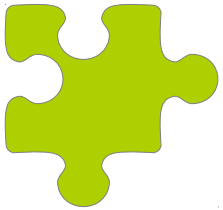
Stationarity - a common problem

Physical consistency - same physical units on both side of the equation (dimensional analysis & the Buckingham Pi Theorem).

[temperature]= [energy]/[M] = $[L]^2/[T]^2$

$$C = \frac{\Delta Q}{\Delta T}$$

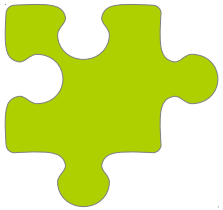
Predictors must 'carry the signal', be well-simulated, and involve a stationary dependency. Relation as strong as possible.



Analysis & Evaluation

Most important - evaluation & diagnostics

- Detrended calibration
- Cross-validation
- Predictor patterns
- Residuals
- ANOVA
- 'Comprehensive' evaluation - the entire chain & for ensembles.



Calibration: large-scales

Predictor pattern & spatial coherence.

Predictor set?

Predictor domain?

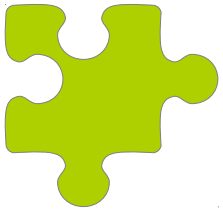
Calibration method?

Linear, non-linear.

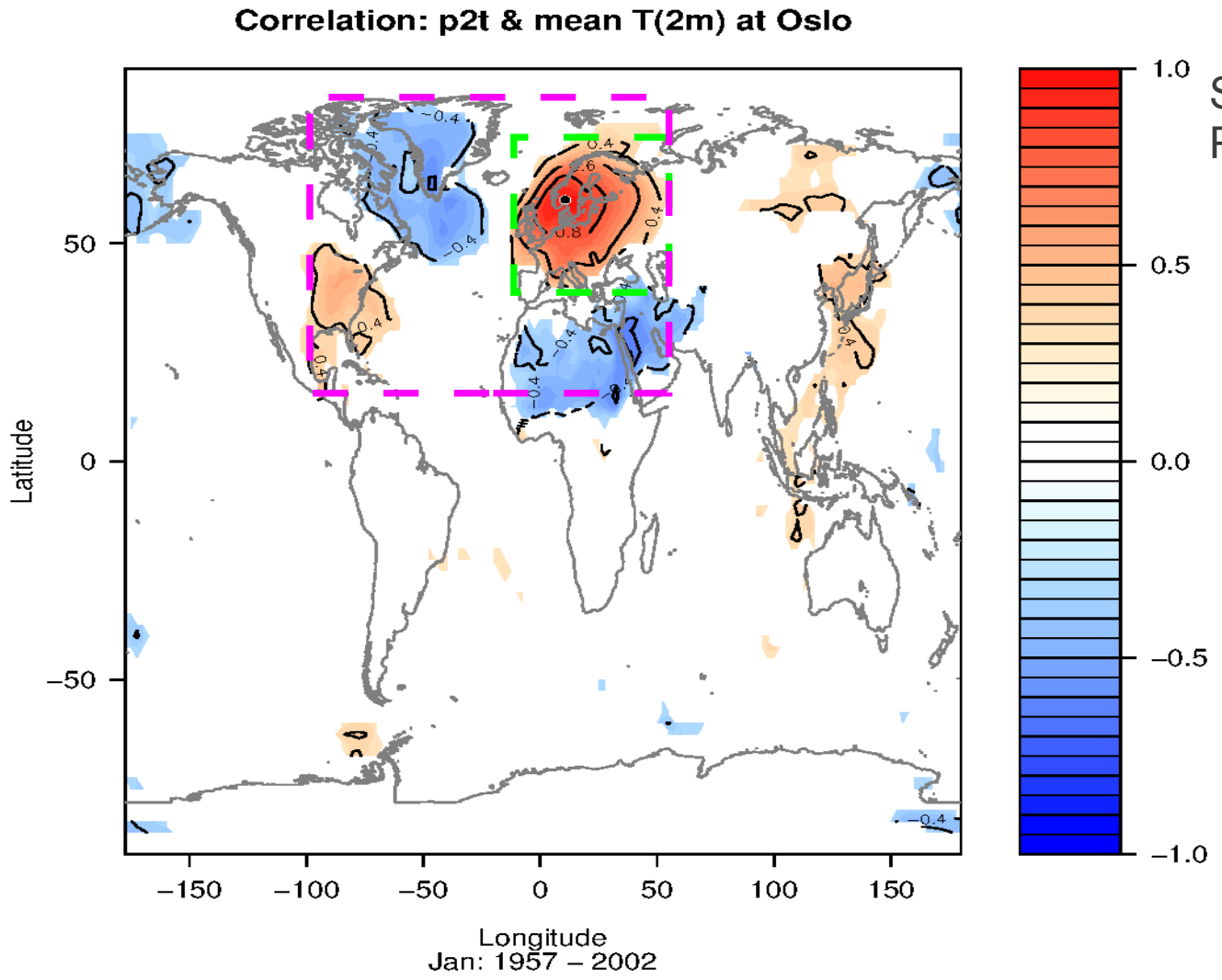
Examine the physical picture.

Testing & diagnostics:

- ANOVA
- Spatial predictor patterns



Calibration: Predictor domain



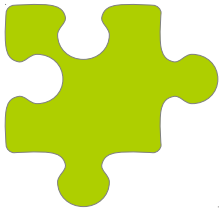
Spatial coherence
Positive correlation

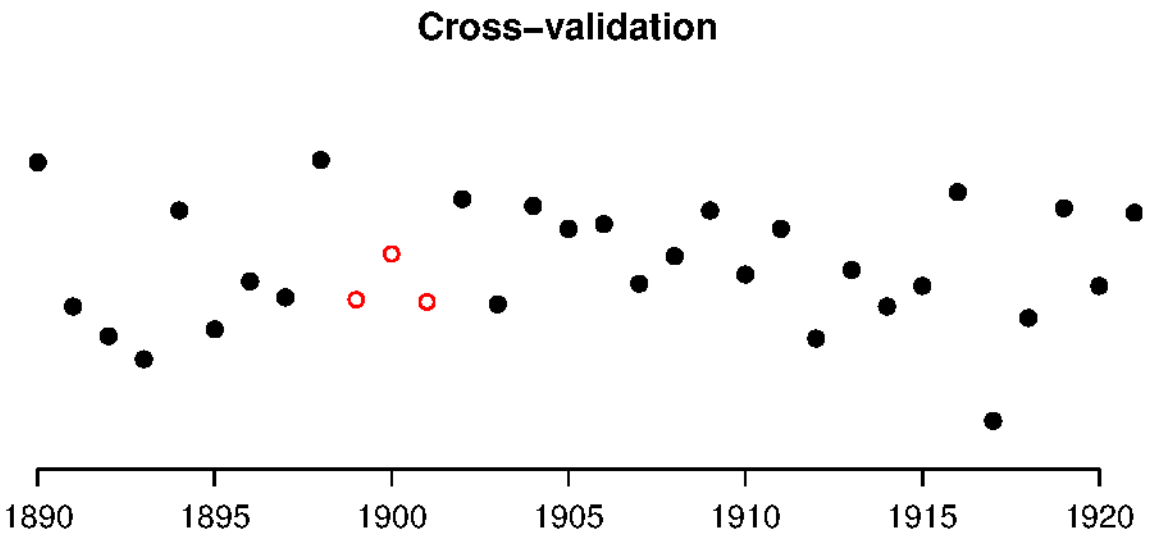
Calibration: Cross-validation

Potential problem: over-fit and fortuitous weighting giving accidental good match.

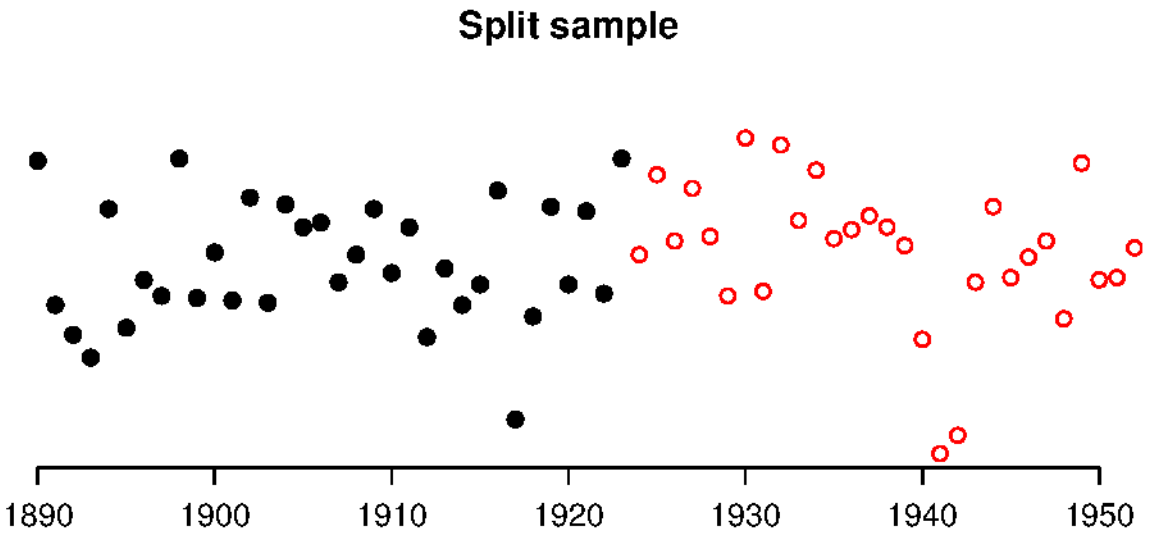
Solution: **Cross-validation** or **Split sample** if long series.

Alternatively: Stepwise screening (stepwise regression), or a combination.

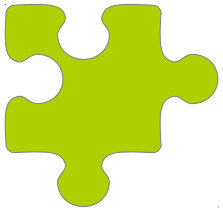




Short series
Auto-correlation?



Long series
Long-term trends

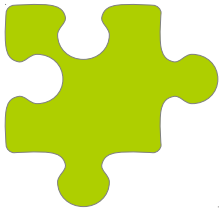


Calibration: detrended data

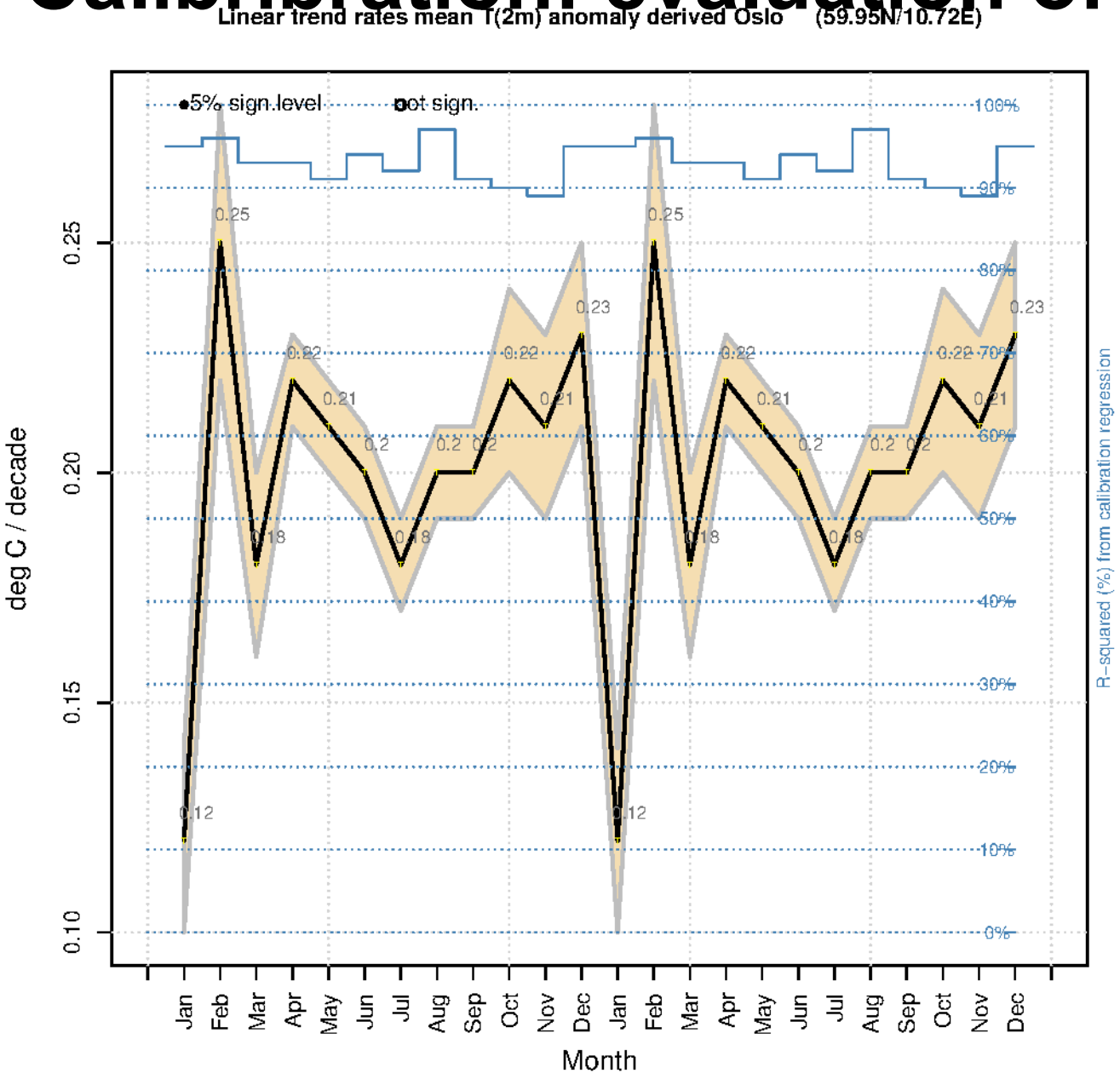
Detrended calibration - at least be able to predict long-term changes.

Test: - capture trend over calibration period? Past trends?

Expected trends – smooth function of season?



Calibribration: evaluation of trends



ANOVA: R^2 & p-values.
Trend = f(season)

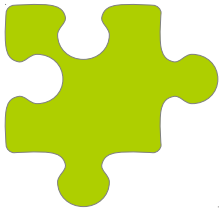
Physical plausible explanations

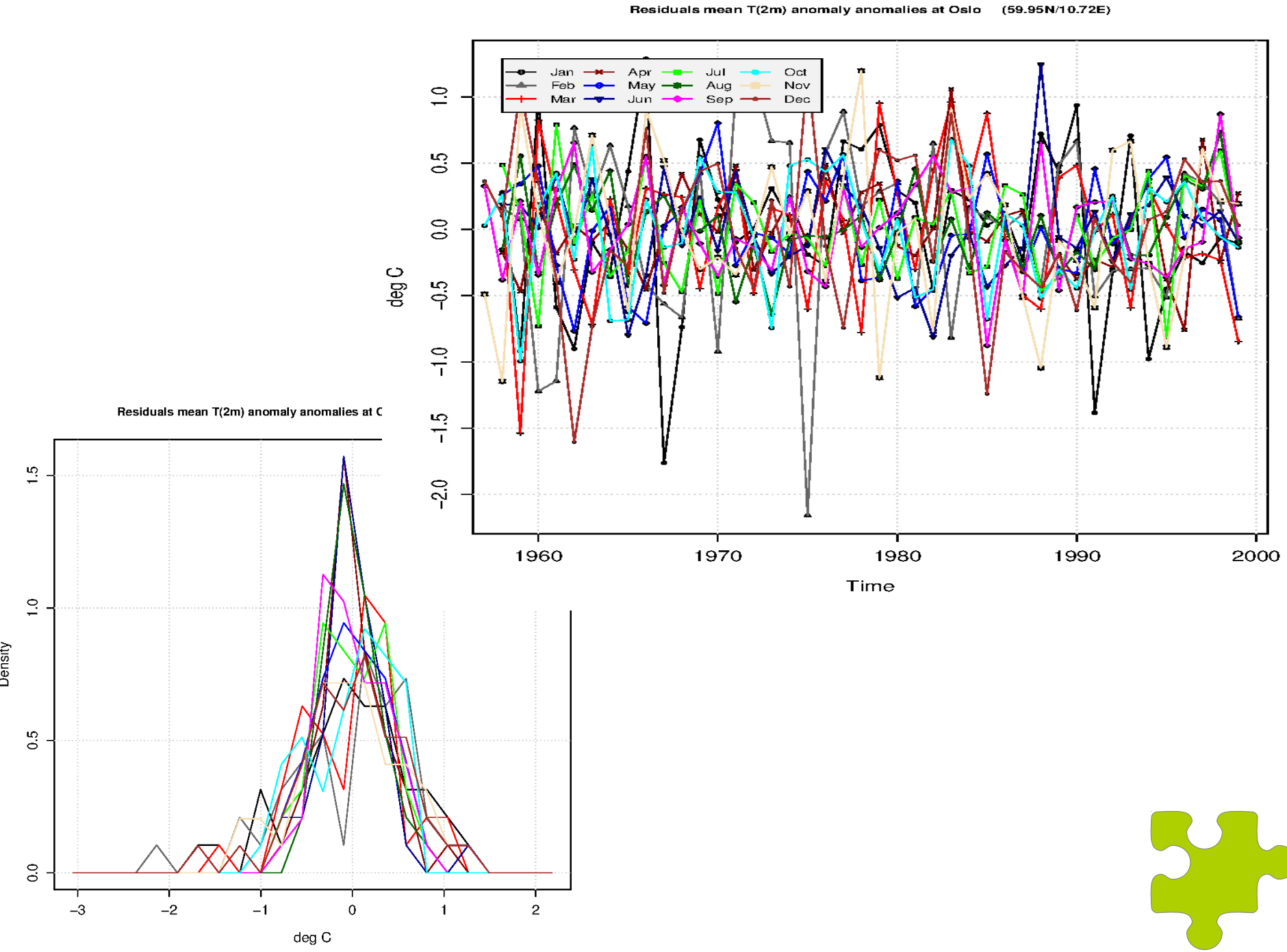
Calibration: residuals

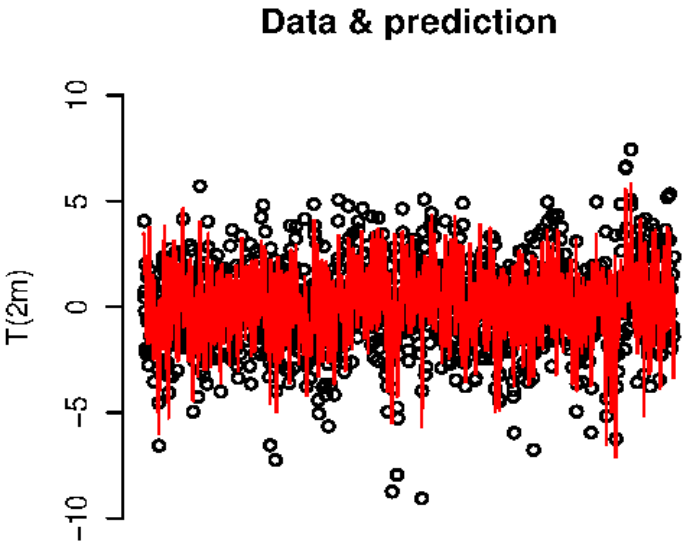
Residuals: structure? Trends? Distribution?

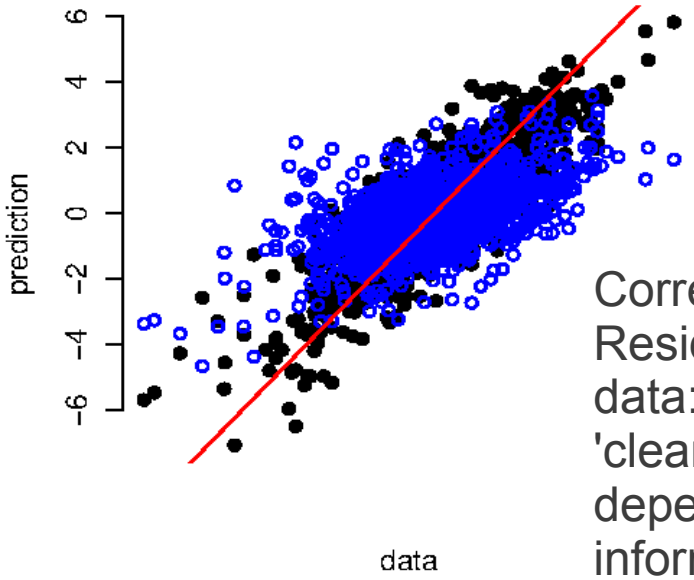
Weather generators can reproduce the statistics of the residual part (statistical models).

Main task is to capture all dependencies and regularities.

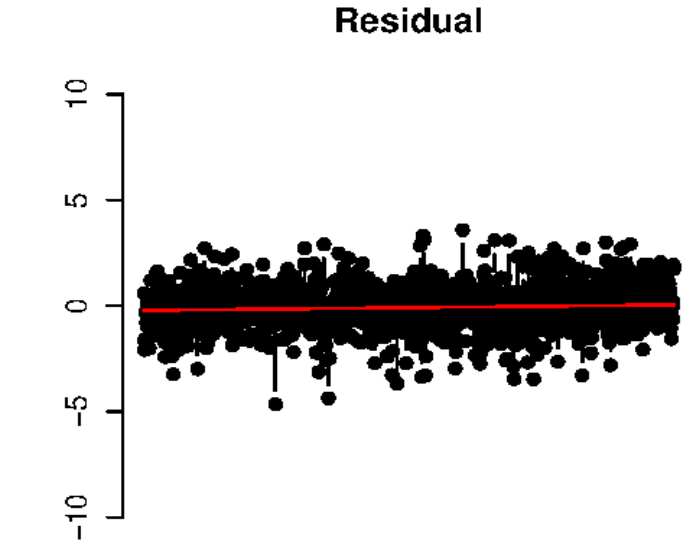


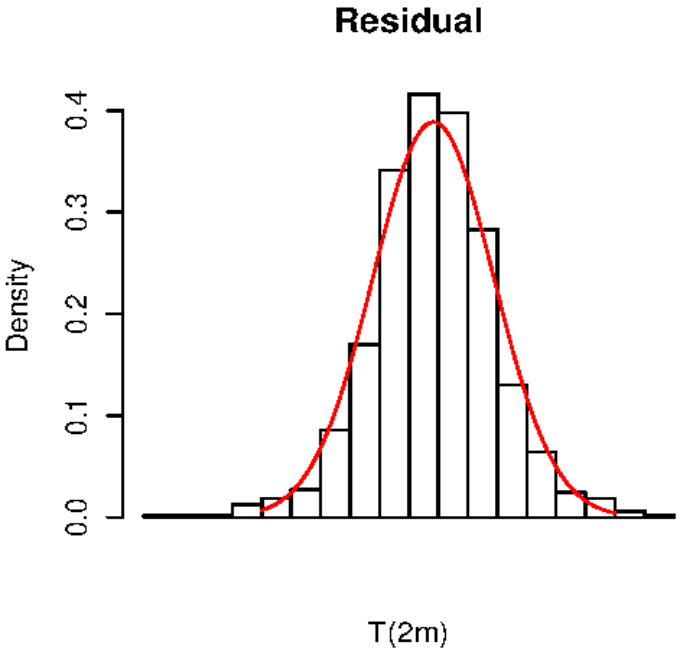

Residuals mean T(2m) anomaly anomalies at Oslo (59.95N/10.72E)
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
deg C
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
1960
1970
1980
1990
2000
Time
Residuals mean T(2m) anomaly anomalies at C
Density
1.5
1.0
0.5
0.0
−3
−2
−1
0
1
2
deg C

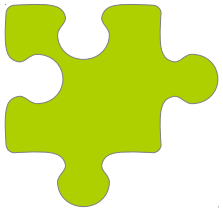




Correlation between Residual and original data: residual ont 'clean' – still contain dependent information.







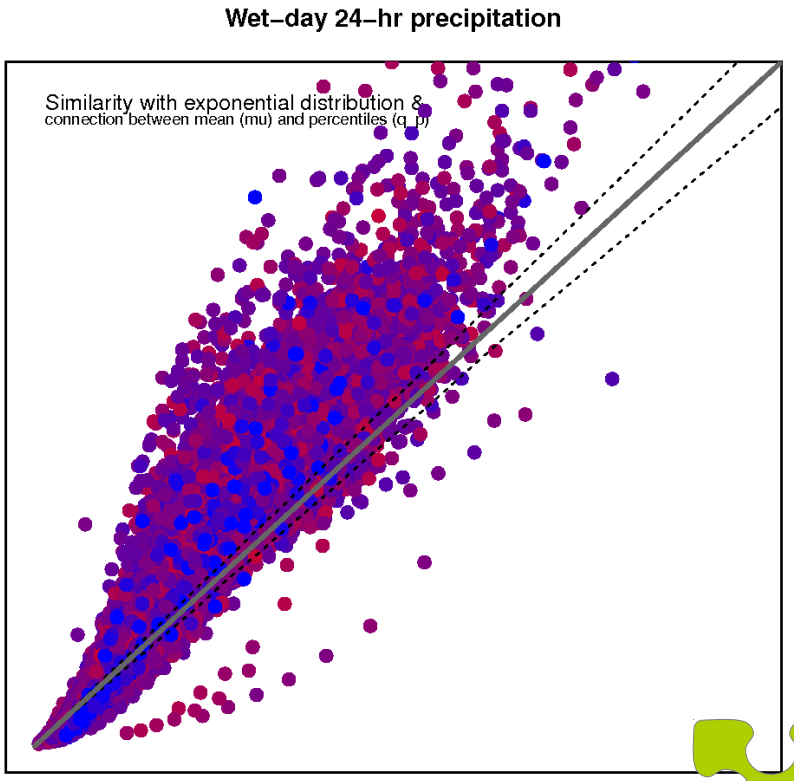
Virtual & real data space

Different structure:

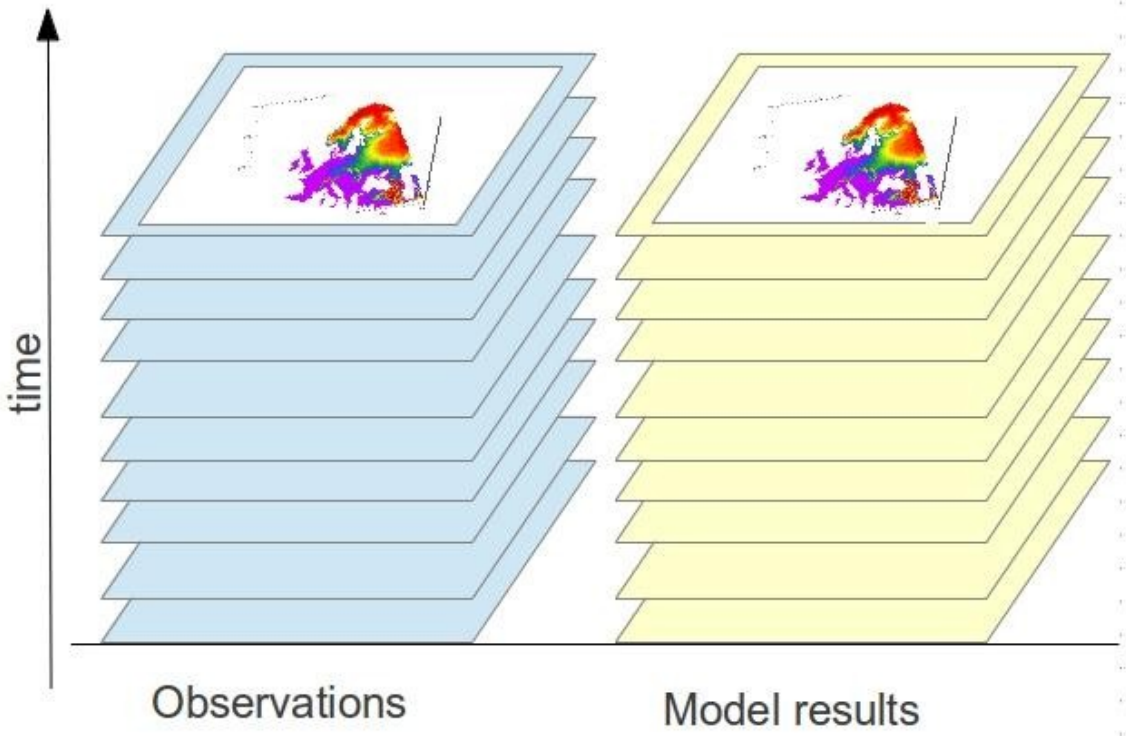
biases, level of detail.

Models do not provide an exact reproduction of the universe

ESD can bridge the two realms



How do I combine the information from the models and the observations?



Perfect Prog - 'PP'

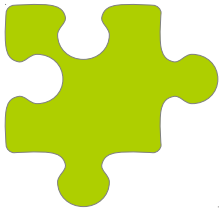
Calibration:

predictor = gridded observations (reanalyses).

predictand = point observations.

Match simulations corresponding to predictor with observations.

- Pattern-based, e.g. regression.
- EOFs - order or spatial structure?



Model Output Statistics - 'MOS'

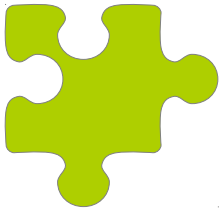
Calibration:

predictor = model results.

predictand = point observations.

No need to match simulations corresponding to predictor with observations. Coupled ocean-atmosphere models (GCMs) **not in synch** with observations.

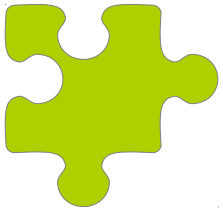
Weather forecast models.



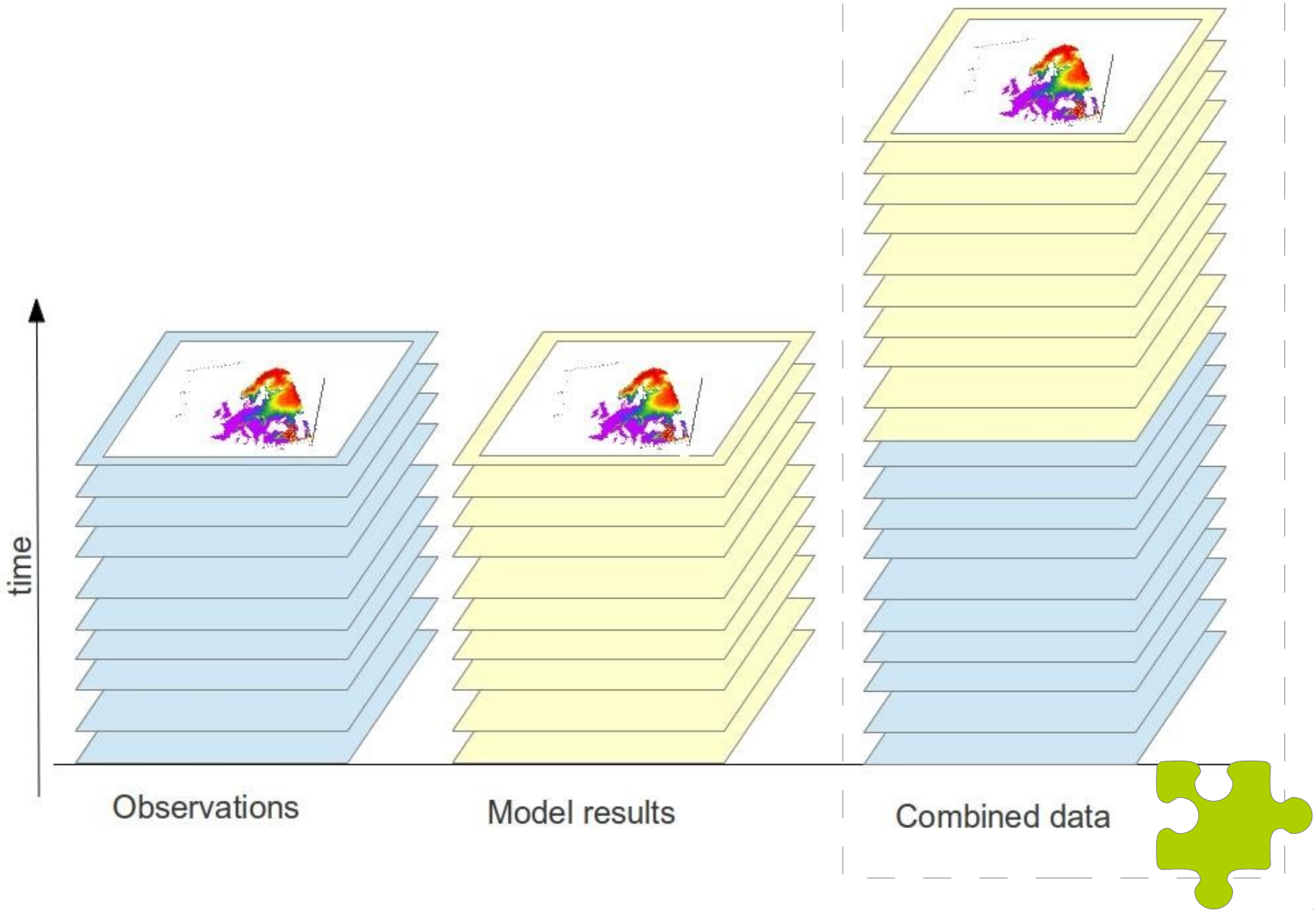
'PP-MOS' Hybrid approaches

common EOFs: a simple mathematical technique to ***ensure*** model-observation correspondence.

Mathematically simple, differs through processing of data: combine observations and model results on the same grid.



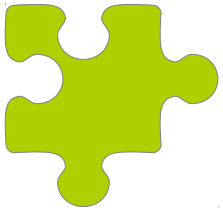
EOF analysis



Model-observation bridging

Ideal real is cleaner than the messy real world.

Degradation of variance.



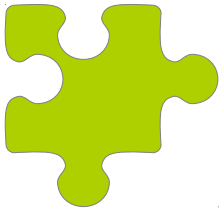
Calibration: 'inflation'

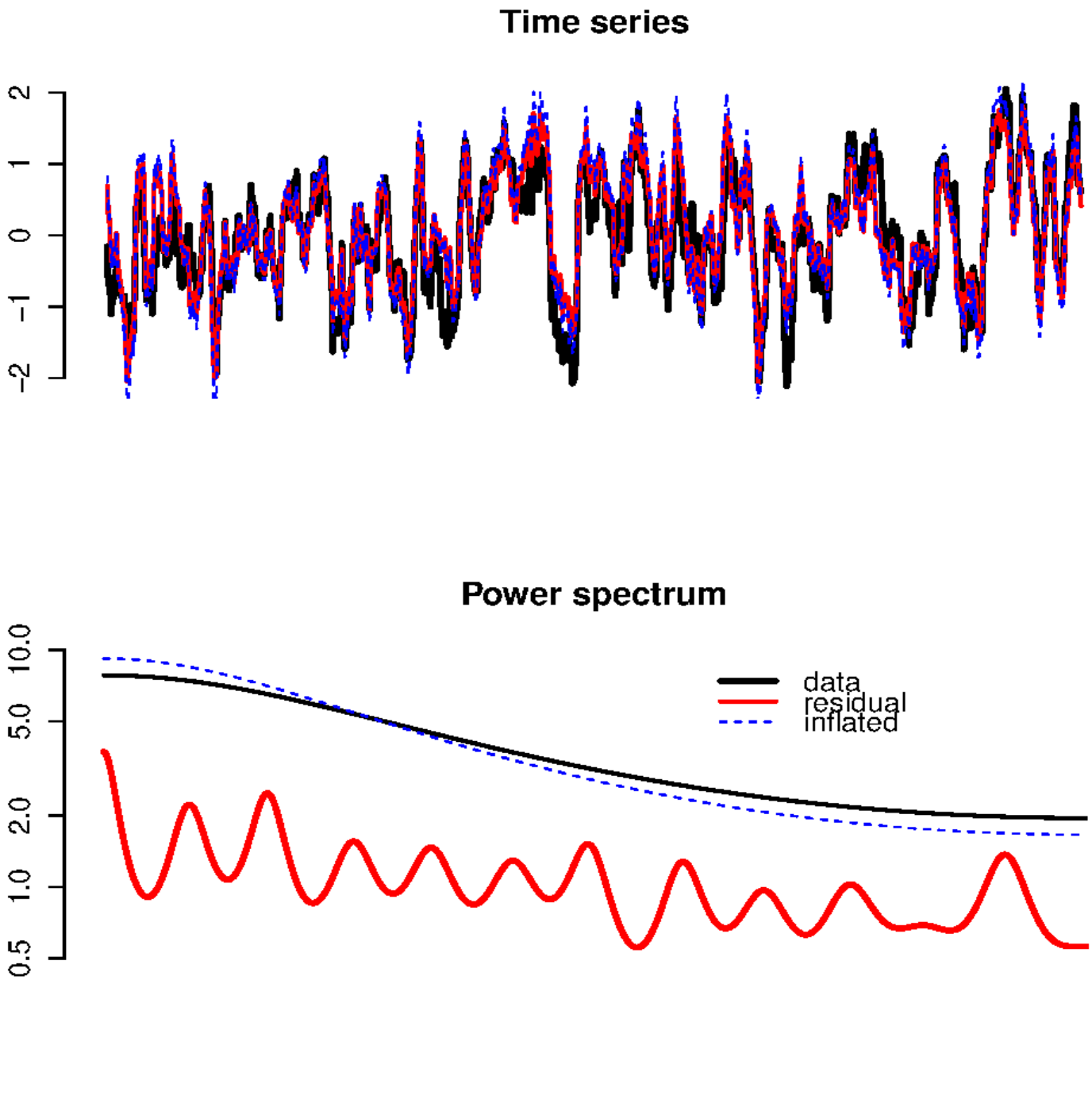
Scaling up the predicted values to have same variance as original data.

Changes the statistical character of the data – **bad idea**.

Artificial – we know for sure that the deficient fraction cannot be explained be explained by the predictor.

Solution: two-component approach where one part is the predictable 'signal' and the other the unpredictable 'noise'




Time series
2
1
0
−1
−2
Power spectrum
10.0
5.0
2.0
1.0
0.5
data
residual
inflated

The two-component aspect

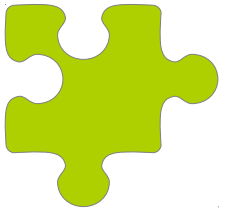
The local climate - predictand:

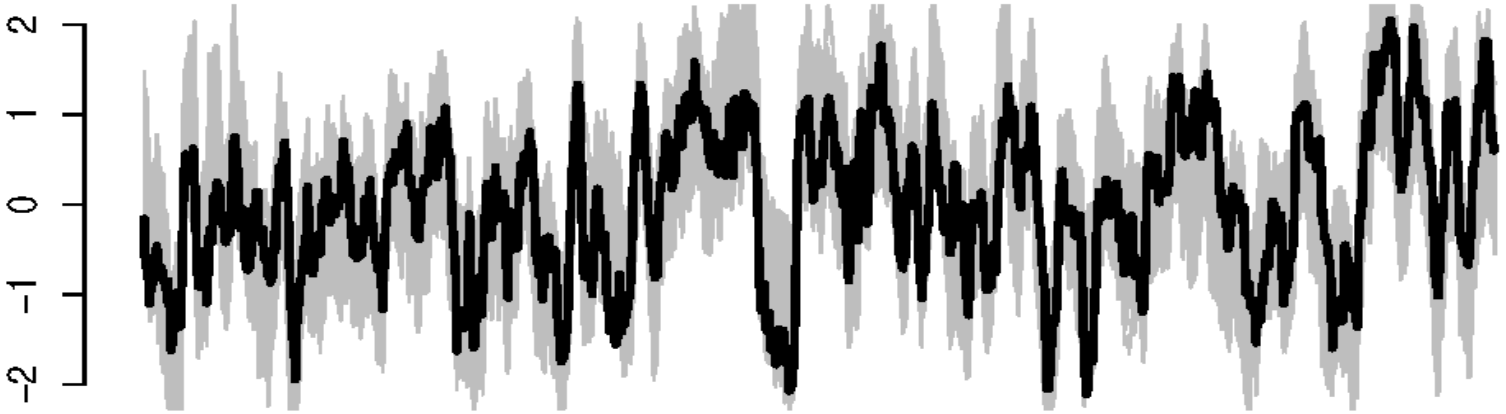
Two components: principally predictable + unpredictable.
Former important for climate change.

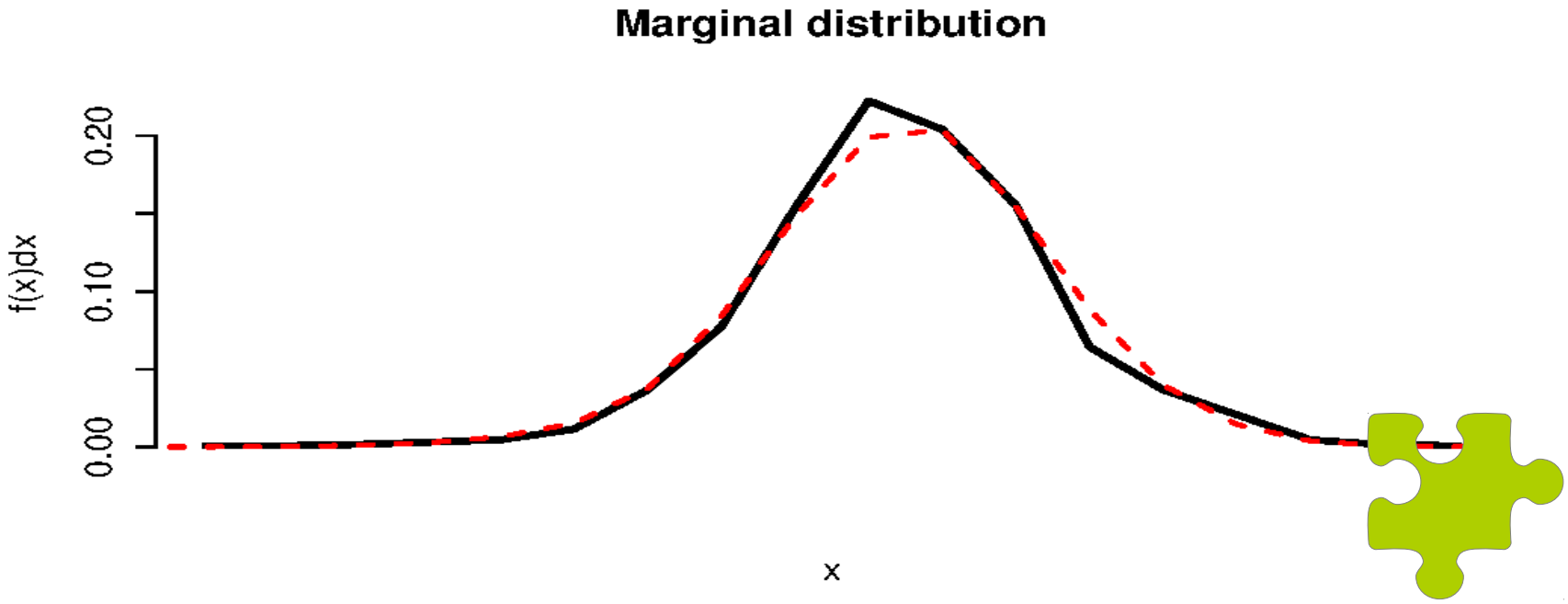
Know *a priori* that only the predictable part is predictable.

Combine the predictable part related to large scales with stochastic part unrelated to large scales:

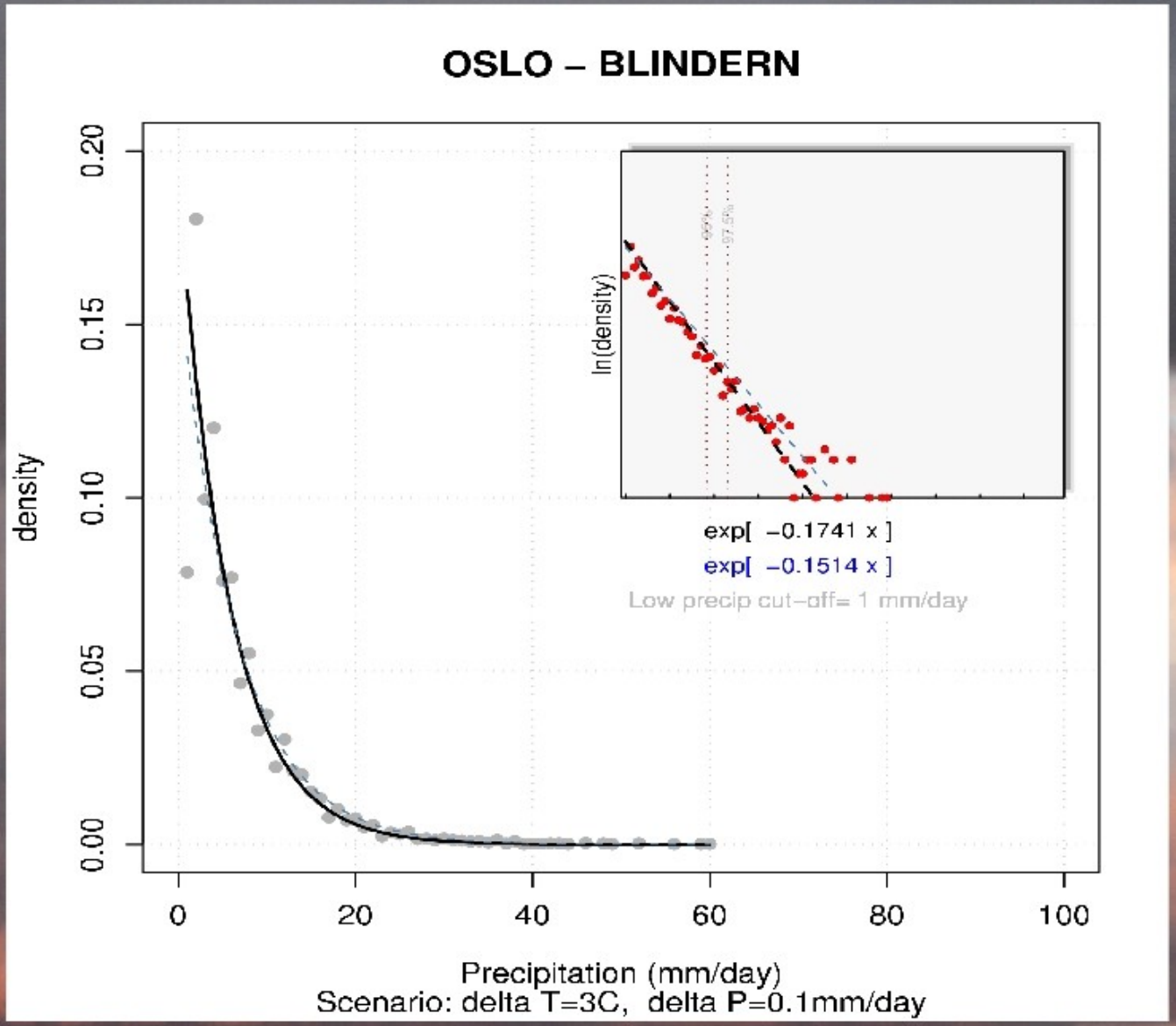
$$y = f(x) + n$$



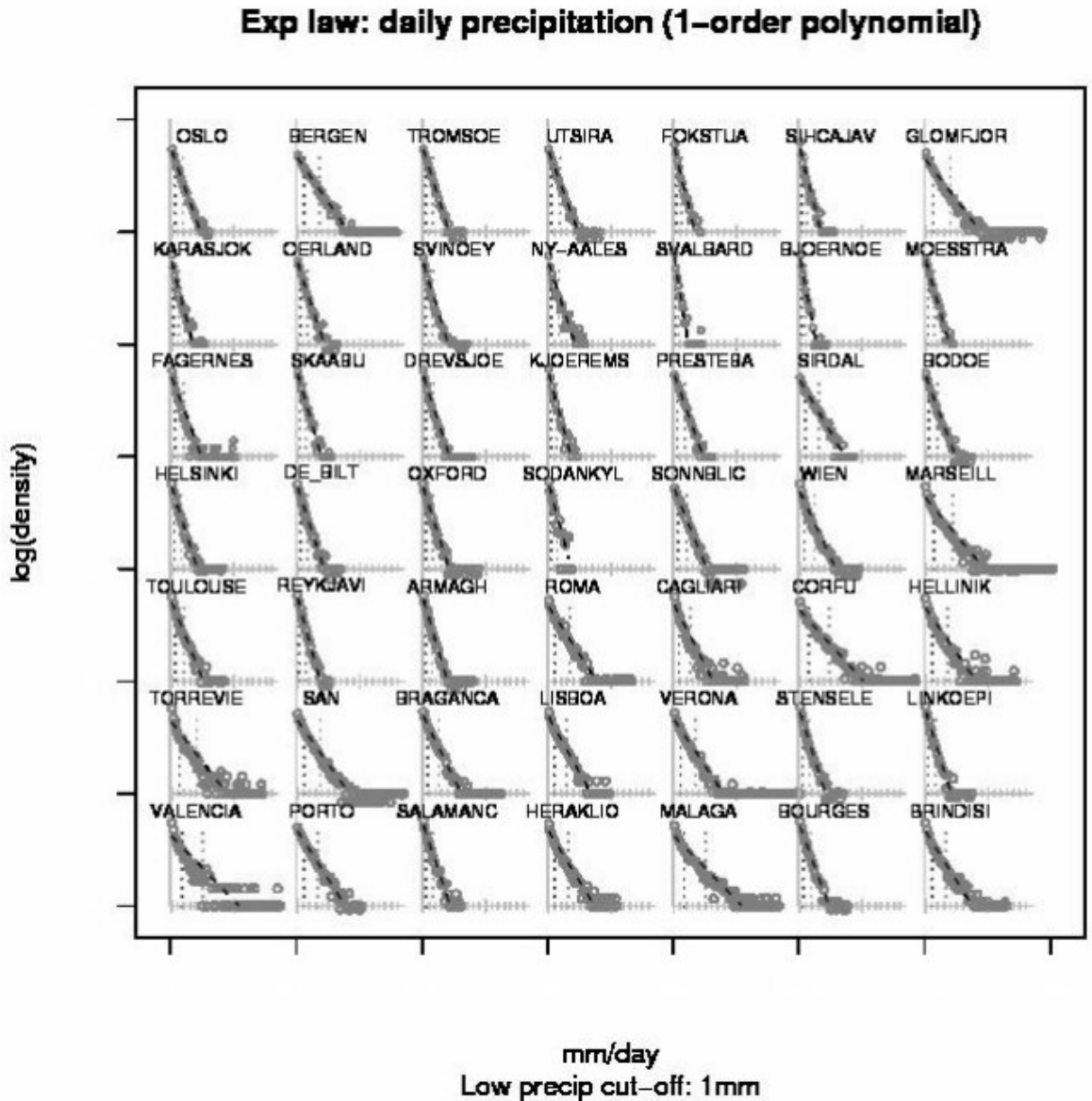

2
1
0
−1
−2


Marginal distribution
f(x)dx
0.00
0.10
0.20
x

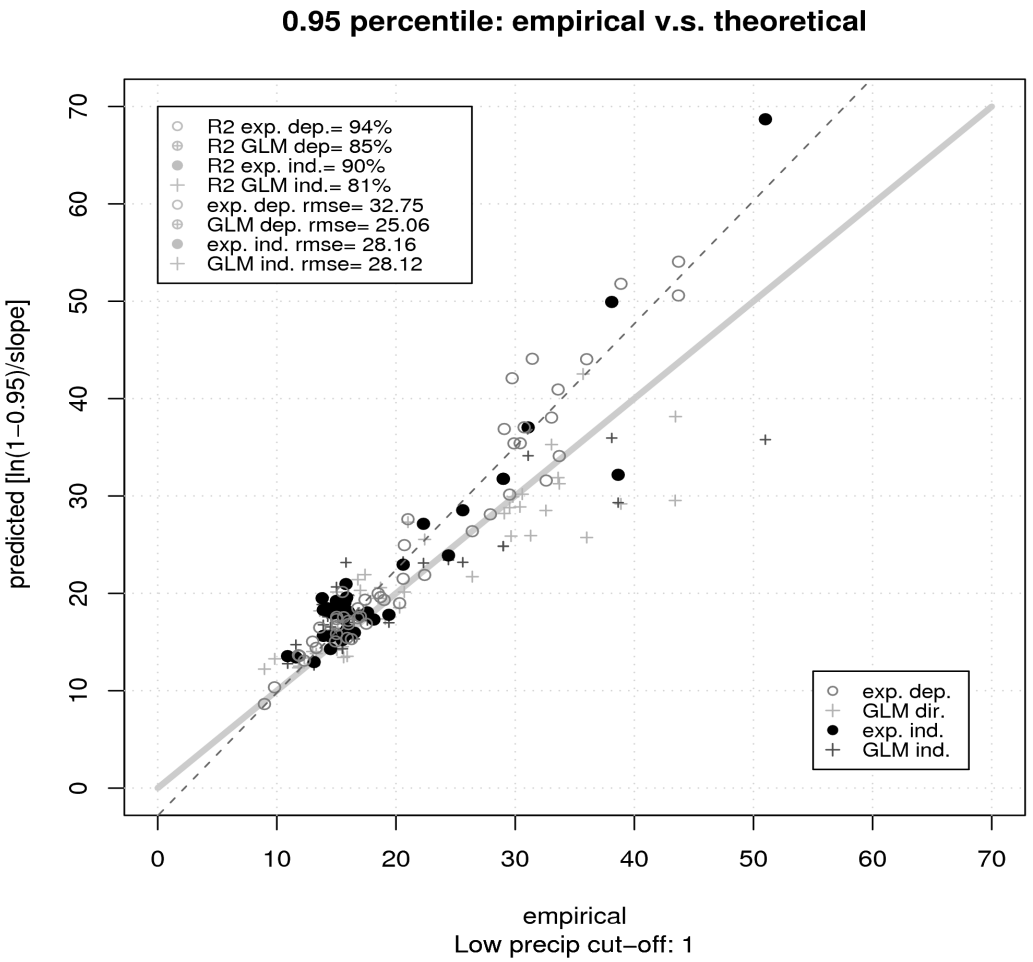
Distribution and PDF

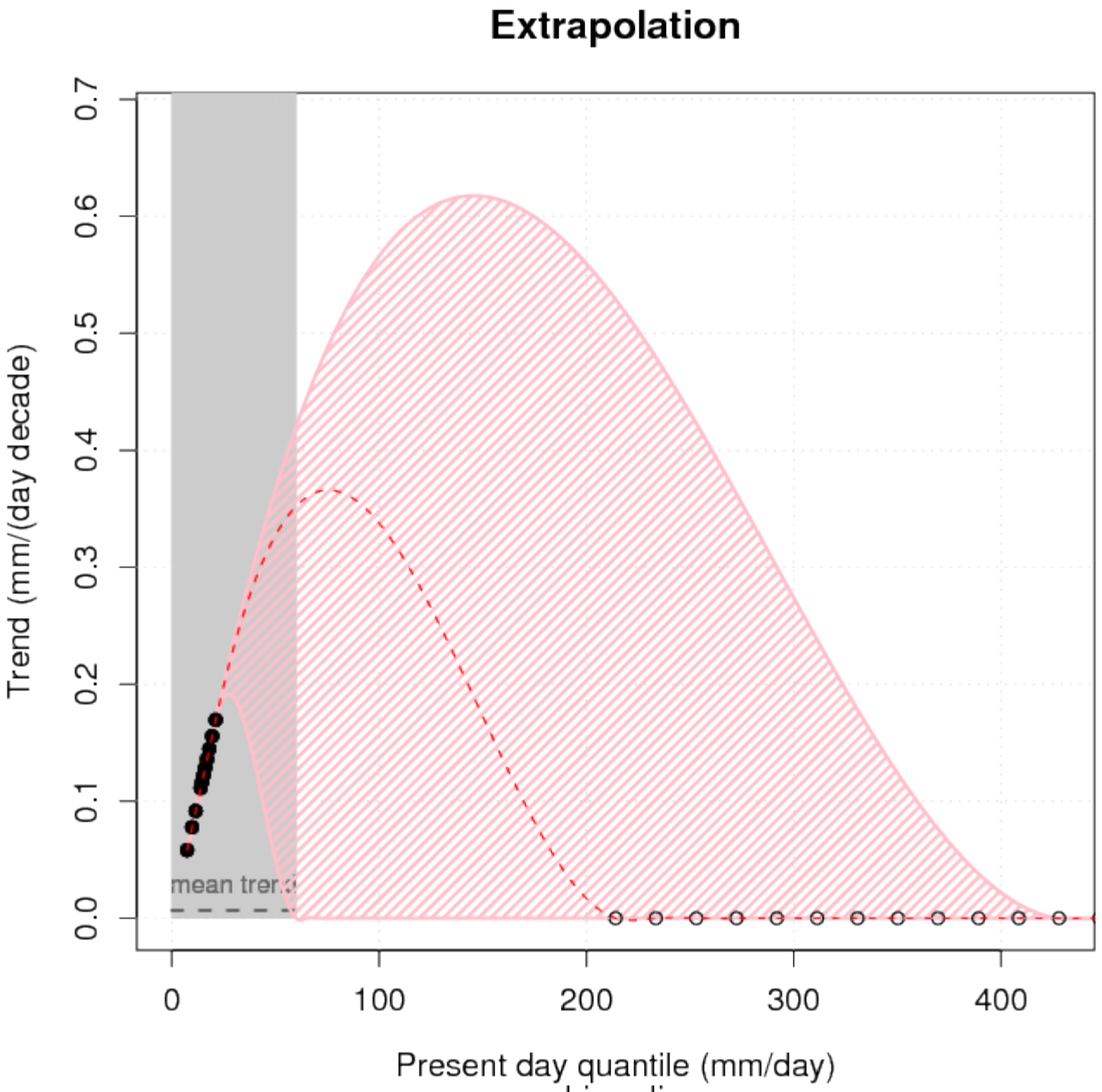


Distribution for 24-hr precipitation

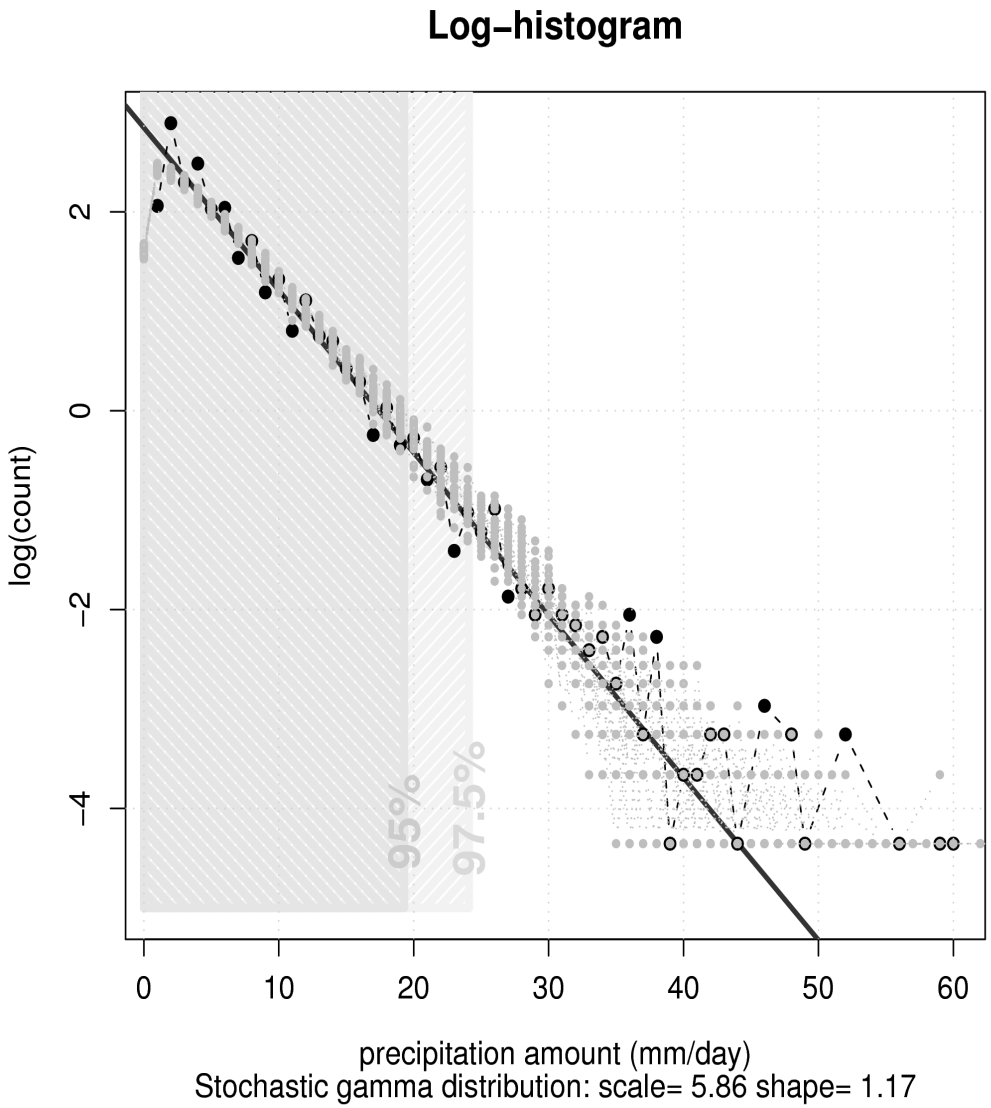


Validation: 95th quantile

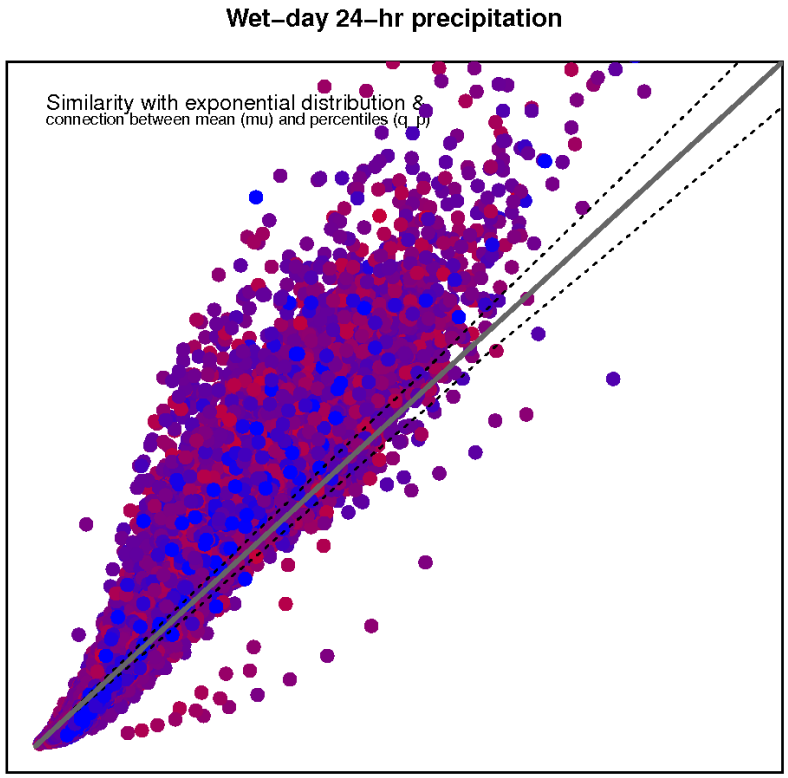



Extrapolation
Trend (mm/(day decade)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
100
200
300
400
Present day quantile (mm/day)

Gamma & exponential distribution



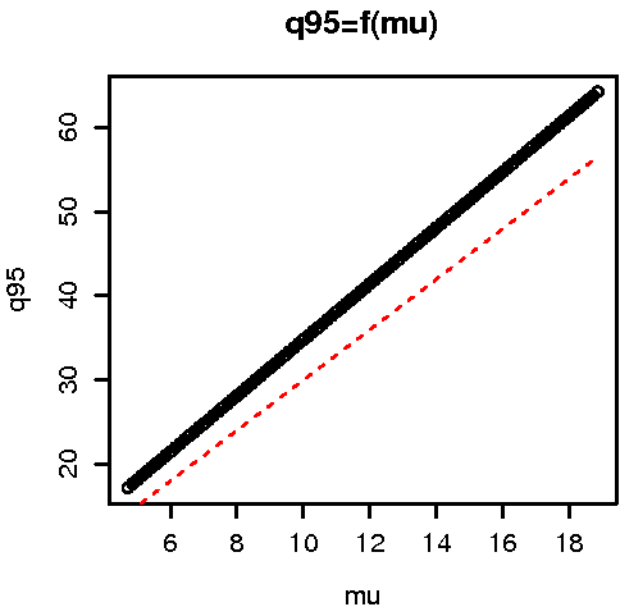
PDF parameter prediction

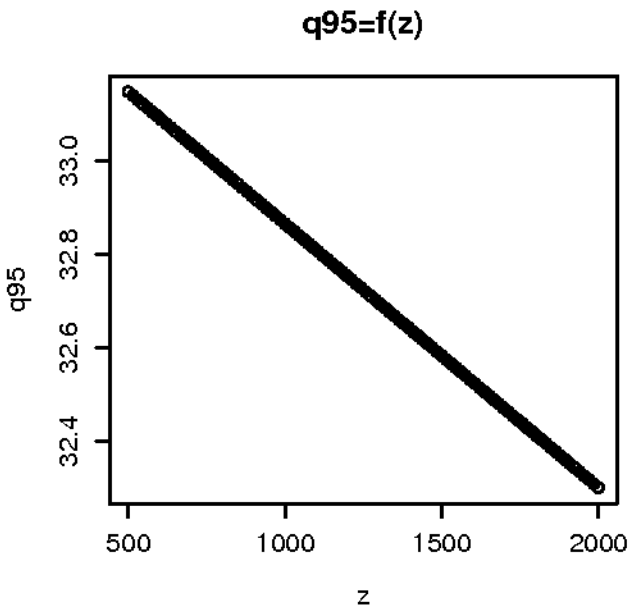


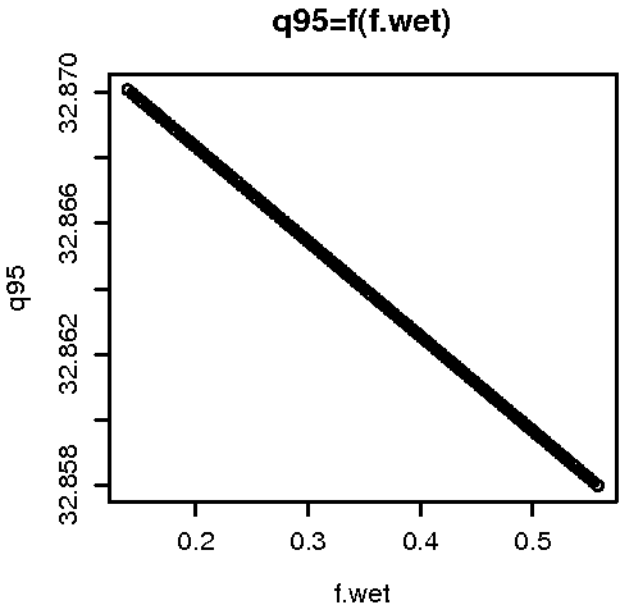
Daily precipitation: two pieces of information – **wet-day occurrence** (frequency) and **mean intensity** (μ).

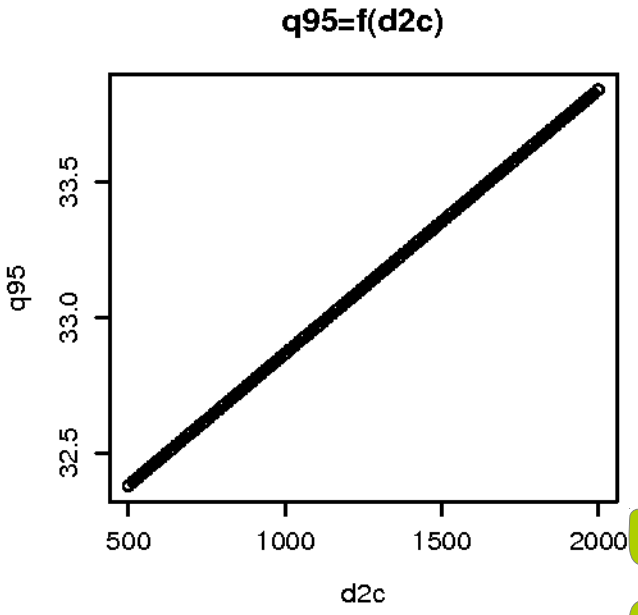
Exponential distribution:
$q_p = -\ln(1-p)\,\mu$

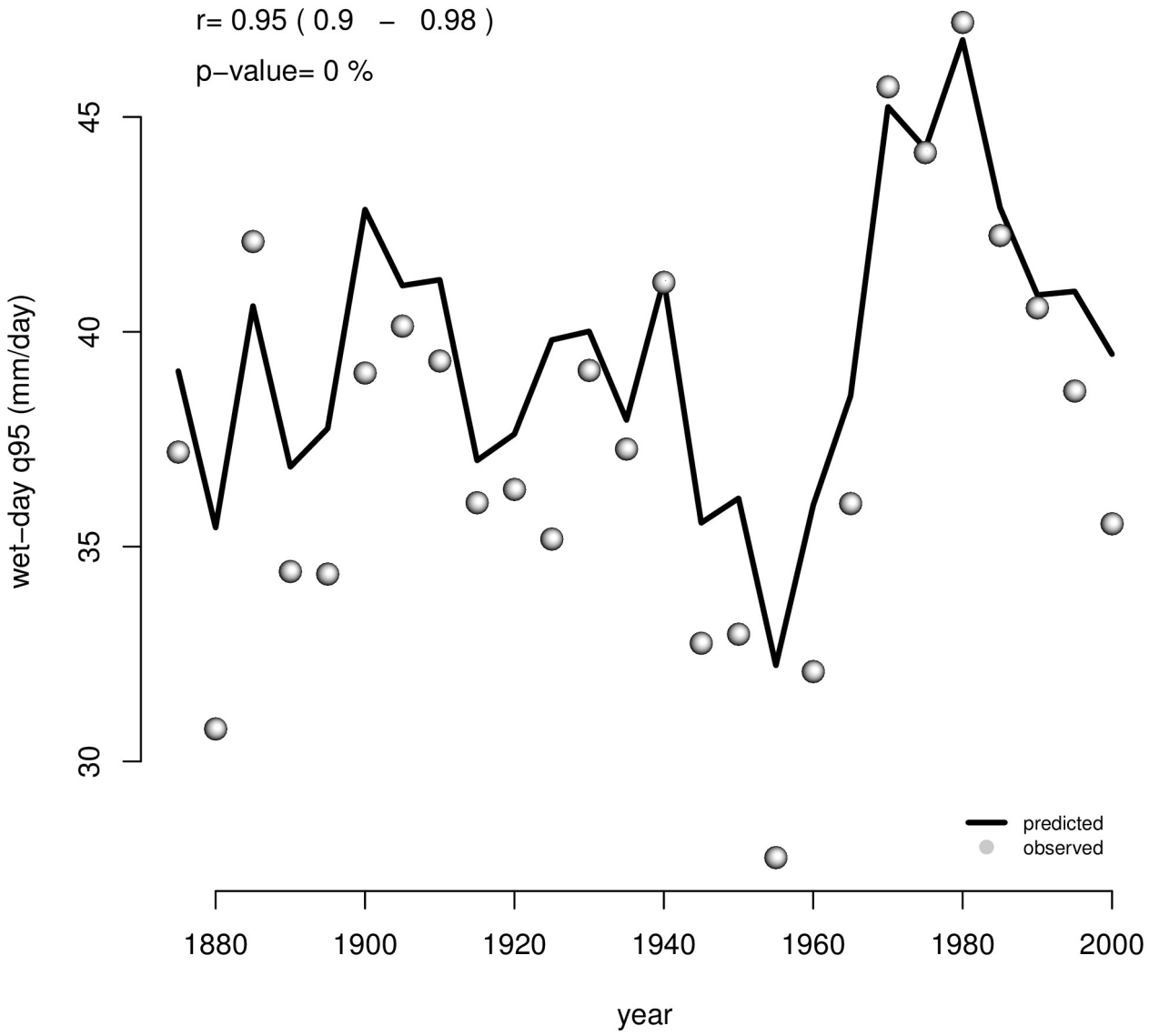
Downscale μ.


q95=f(mu)
q95
mu
20
30
40
50
60
6
8
10
12
14
16
18


q95=f(z)
q95
z
32.4
32.6
32.8
33.0
500
1000
1500
2000

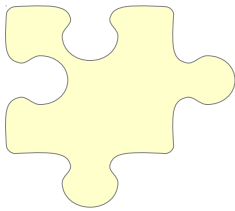

q95=f(f.wet)
q95
f.wet
32.858
32.862
32.866
32.870
0.2
0.3
0.4
0.5


q95=f(d2c)
q95
d2c
32.5
33.0
33.5
500
1000
1500
2000


r= 0.95 (0.9 – 0.98)
p–value= 0 %
wet–day q95 (mm/day)
30
35
40
45
1880
1900
1920
1940
1960
1980
2000
year
predicted
observed

Empirical Statistical Downscaling

Linear methods



Linear methods

Quantification of y & X

What to optimize?

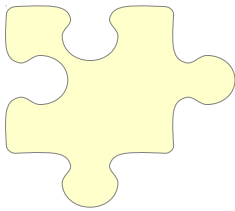
Least squares (two types)

Canonical correlation analysis (CCA)

Singular vector analysis (SVD)

Maximum likelihood methods

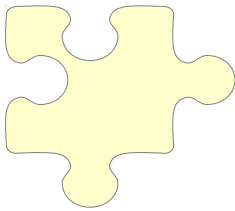
    Generalised linear models (GLM)



Quantification of y & X

The predictand y is a single or set of time series.

Predictor X represent an extensive (large-scale) area. Can be several grid boxes, indices, or EOFs.



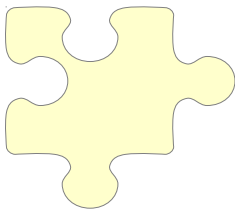
EOFs

Two types: eigenvalue-based and singular vector decomposition (SVD). Linear algebra & matrix notation.

$$C = XX^T \rightarrow C\, e_s = \lambda\, e_s \rightarrow X = E\, \beta$$

$$X = U\, \Sigma\, V^T$$

Here E and U are EOFs from different technique



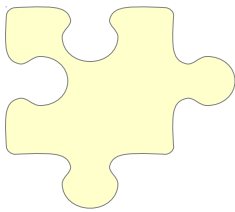
Common EOFs

Mathematically identical to ordinary EOFs, but applied to a combined data set.

Quality control on the GCMs.

Ensures time indices describe exactly the same spatial pattern.

One regression step – minimize loss of R^2.



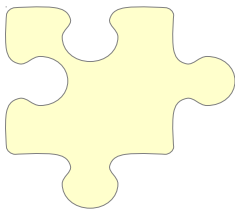
Rotated EOFs

The EOFs provide a reference frame in data space for analysis.

Weighting of these is equivalent to a rotation and a transform in data space.

Weights can be found from regression analysis.

Doubtful if varimax will benefit this type of analyses.



The equation

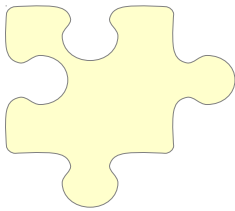
All equations are linear:

$y = f(X)$

The function f(X) may be non-linear.

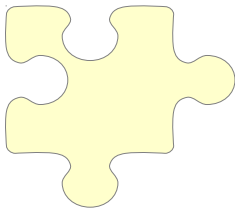
Seek transforms of the left and right hand sides so that they are represented by one term or a sum of terms: y & f(X).

Base the equation on known physics as far as possible.



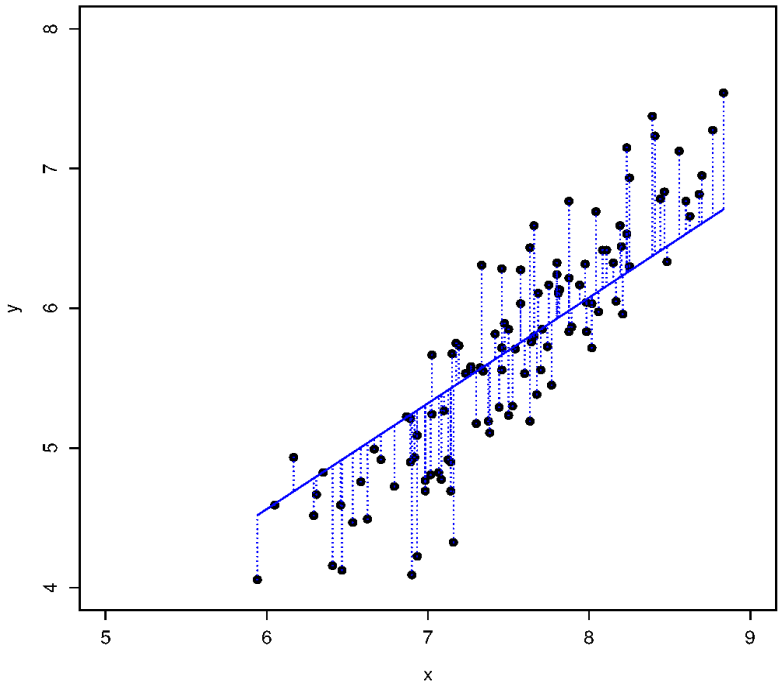
Training the model

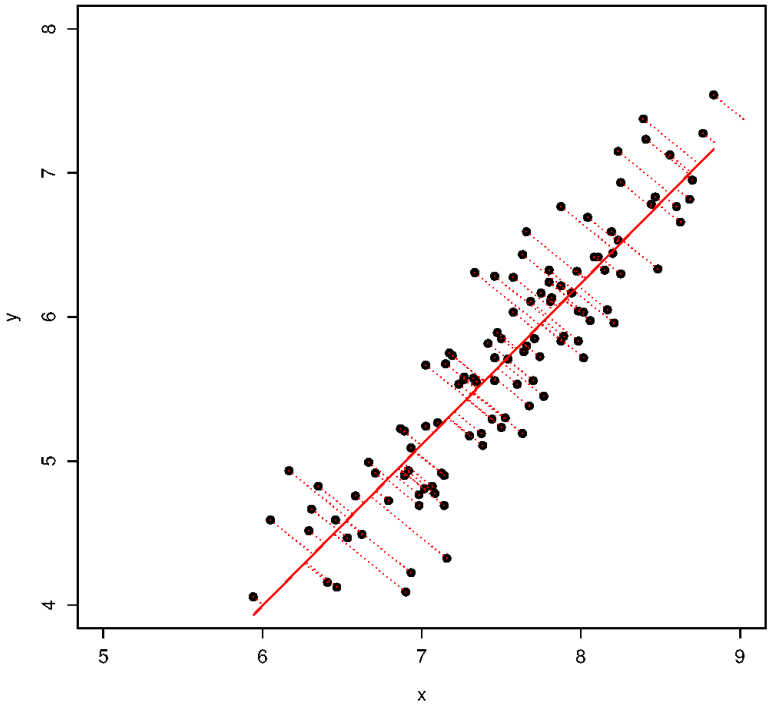
Estimate the value for the coefficients in the equation which give the best-fit – maximize or minimize some kind of cost function. OLR minimizes the distance between the points and the fitted line. LSF minimises RMSE.

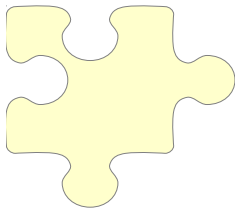


Least squares

2 types: minimizing the perpendicular distance to a line-fit and the errors in y:







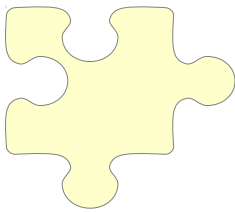
Multi-variate regression:

Projection – minimize perpendicular distance.
Linear algebra equation: $Y^T = X^T \Psi - \zeta$

$$\Psi = (XX^T)^{-1} XY^T$$

EOFs for X give more robust results – not ill posed as $E^T E = I$:

$$\Psi = (E_{(k)} \Sigma^2 E_{(k)}^T)^{-1} XY^T$$

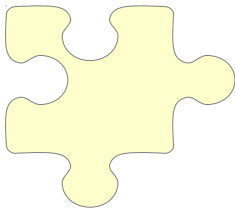


CCA

Two types: BP-CCA based on EOF products and CCA based on the gridded fields.

Find pairs of patterns which optimise mutual correlations.

Group of stations and fields.

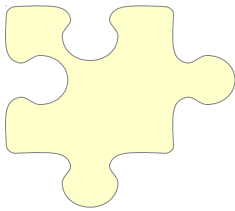


SVD

Not the same as SVD. Maximize the covariance between two fields. Similar to CCA.

Find pairs of patterns which optimise mututal variance.

Group of stations and fields.



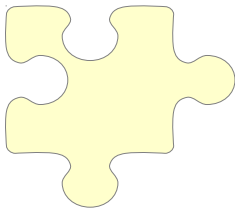
Maximum likelihood

If the variables are not normally distributed. Y may be an integer (days) or a probability [0,1]. Often more complicated than OLR minimizing distance or RMSE.

Generalised linear models (GLM).

$g(y) = a_0 + a_1 x_1 + a_2 x_2 + ...$

g(.) is the link function

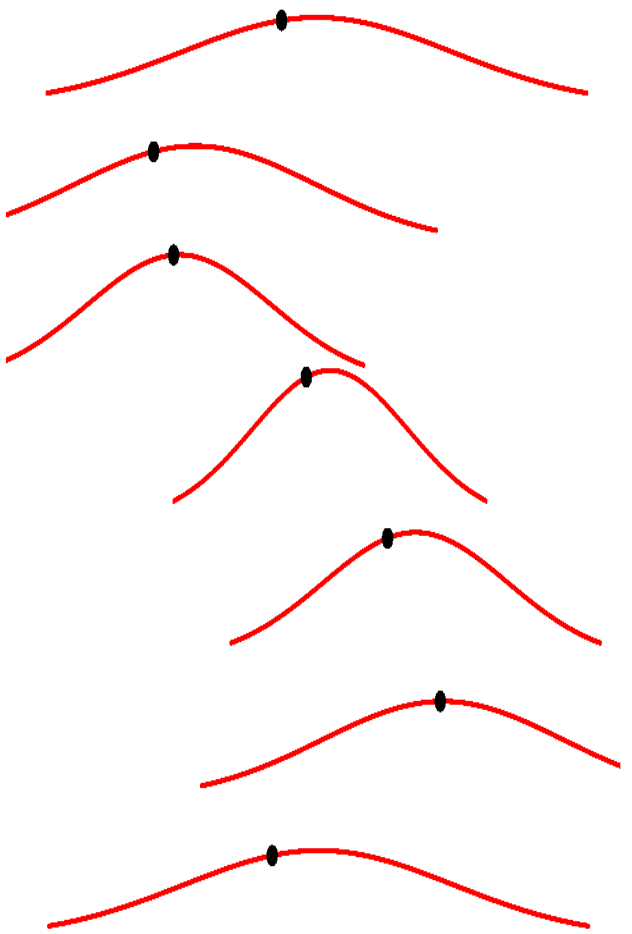


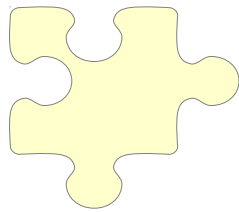
ML fitting

Different perspective.
Expectation value.

Optimisation by minimizing a cost function

Optimisation by finding parameters for PDFs.





Further reading

R:
> install.packages("clim.pact")
> library(clim.pact)
> ABC4ESD()

Benestad, R.E., Hanssen-Bauer, I. And Chen, D.(2008) "Empirical-Statistical Downscaling", World Scientific Publishers, ISBN 978-981-281-912-3; free Compendium on empirical-statistical downscaling (2007): http://rcg.gvc.gu.se/edu/esd.pdf