Empirical Statistical Downscaling

Diagnostics & Predictions
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).

\[ \text{temperature} = \text{energy}/[M] = [L]^2/[T]^2 \]

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

Correlation: p2t & mean T(2m) at Oslo

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.
Cross-validation

Short series
Auto-correlation?

Split sample

Long series
Long-term trends
Calibration: detrended data

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

Calibration: 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.
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.
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'
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:

\[ y = f(x) + n \]
Distribution and PDF

OSLO - BLINDERN

\[ \text{density} = \exp(-0.1741 \times \text{Precipitation (mm/day)}) \]
\[ \exp(-0.1514 \times \text{Precipitation (mm/day)}) \]

Low precip cut-off = 1 mm/day

Scenario: delta T=3C, delta P=0.1 mm/day
Distribution for 24-hr precipitation

Exp law: daily precipitation (1-order polynomial)

mm/day
Low precip cut-off: 1mm
Validation: 95$^{\text{th}}$ quantile

0.95 percentile: empirical v.s. theoretical

predicted $\ln[1-0.95]/\text{slope}$

empirical

Low precip cut-off: 1
Gamma & exponential distribution

Log-histogram

precipitation amount (mm/day)
Stochastic gamma distribution: scale= 5.86 shape= 1.17
PDF parameter prediction

Daily precipitation: two pieces of information – wet-day occurrence (frequency) and mean intensity ($\mu$).

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

Downscale $\mu$. 

thresh. = 1 mm/day; #stations= 11281 [qqplotter.R]
$r = 0.95\ (0.9\ -\ 0.98)\$

p-value = 0 %
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.

\[ C = XX^T \quad \rightarrow \quad C \, e_s = \lambda \, e_s \quad \rightarrow \quad X = E \, \beta \]

\[ X = U \, \Sigma \, V^T \]

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: \( Y^\top = X^\top \Psi - \zeta \)

\[ \Psi = (XX^\top)^{-1} XY^\top \]

EOFs for X give more robust results – not ill posed as \( E^\top E = I \):

\[ \Psi = (E_{(k)} \Sigma^2 E_{(k)}^\top)^{-1} XY^\top \]
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 mutual 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).

\[ g(y) = a_0 + a_1 x_1 + a_2 x_2 + \ldots \]

\( g(.) \) 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()

http://rcg.gvc.gu.se/edu/esd.pdf