Regional Climate Model Validation and its Pitfalls

Sven Kotlarski
Federal Office of Meteorology and Climatology MeteoSwiss, Zurich

4th VALUE Training School: **Validating Regional Climate Projections**
Trieste, October 2015

OUTLINE

1. The rationale of RCM evaluation
2. Techniques and measures
3. Potential pitfalls
4. Summary & conclusions

OUTLINE

1. The rationale of RCM evaluation
2. Techniques and measures
3. Potential pitfalls
4. Summary & conclusions

WHY SHOULD WE VALIDATE AN RCM?
(or a climate model, in general)

Why RCM Evaluation?

Does the model work for the purpose it has been built for?

Model = incomplete representation of the climate system

Structural and parametric uncertainties

Good evaluation = basic requirement for trust in regional climate scenarios

Model selection and weighting

If selection necessary: Evaluation can inform choice to some extent

Basis for excluding models with major deficiencies

Model setup and calibration

Choosing a specific setup

Calibration within a specific setup

Added value analysis

Is RCM application, or very high resolution really required?
Can SD deliver similar/better results? (-> **VALUE**!)

Identification of model deficiencies

Model development

OUTLINE

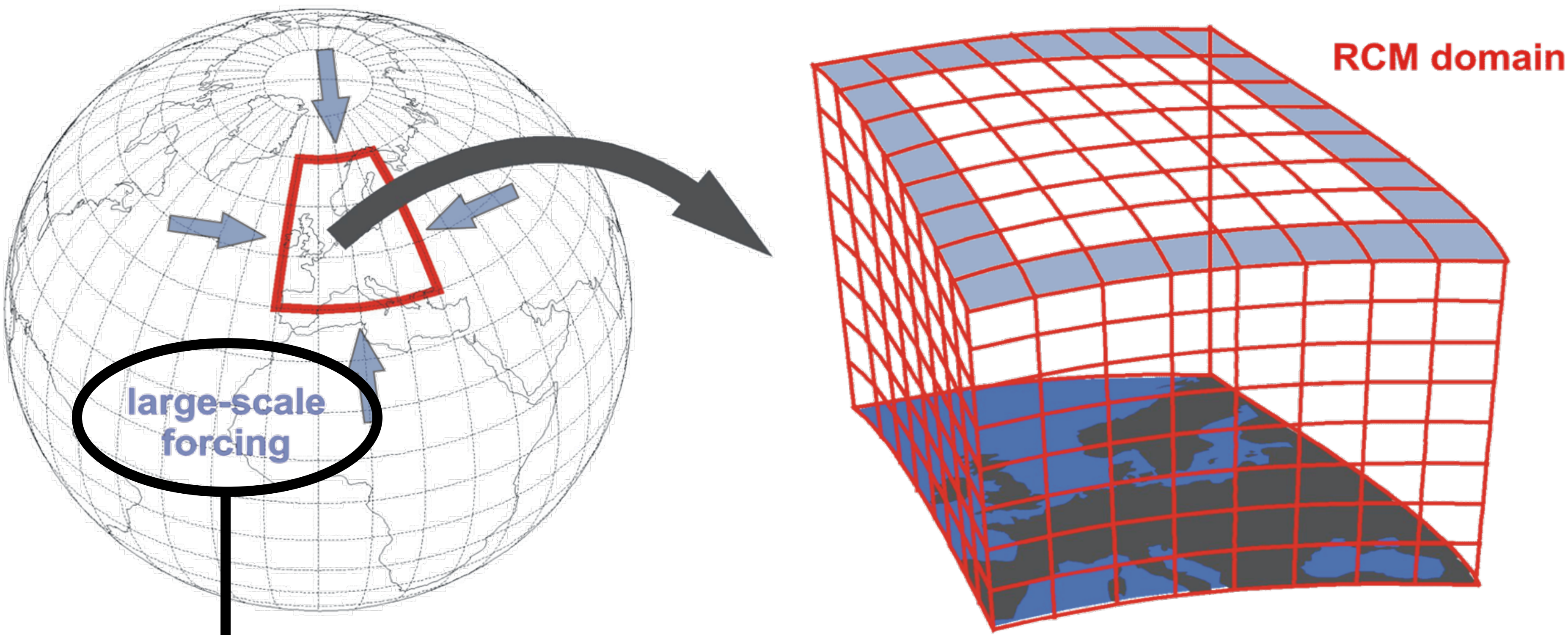
1. The rationale of RCM evaluation
2. Techniques and measures
3. Potential pitfalls
4. Summary & conclusions

RCM Validation

Compare an RCM experiment against some **reference**

- «Observations» in historical periods
- A reconstruction of the historical climate (especially applies to paleoclimate studies)
- A different model that you trust in
  (could be, for instance, a re-analysis or a model based on first physical principles)
- A reference simulation of the same model

The Nesting Technique

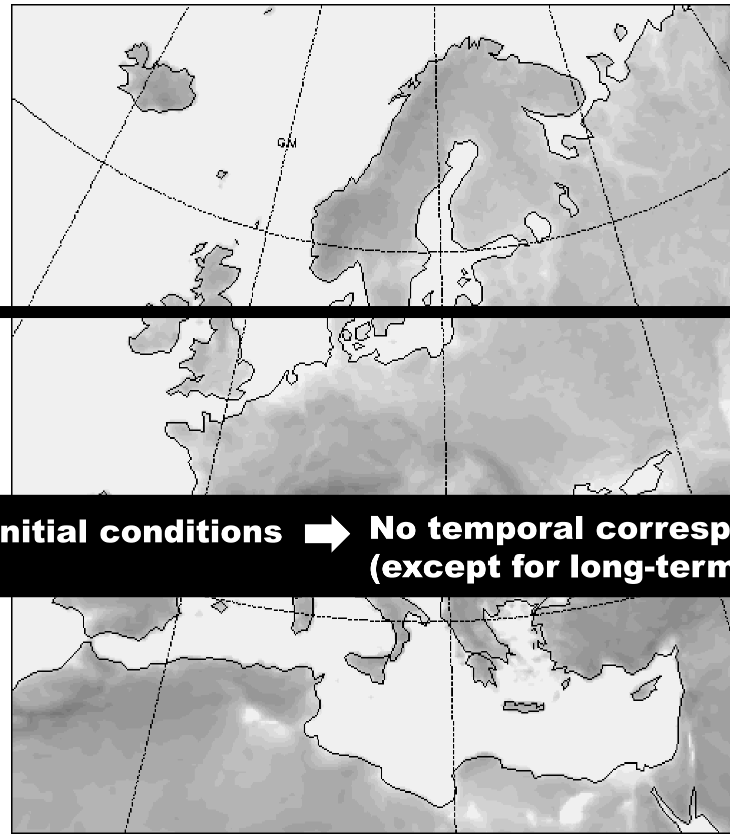


- Uncertainties / biases / differences in large-scale forcing will ultimately affect RCM results and, hence, evaluation
- «Garbage in – garbage out»

RCM Experiments for historical periods

boundary forcing
(global)

RCM



Re-analysis
(perfect boundaries)

Evaluation of
(pure)
downscaling

GCM
historical GHG

Evaluation of
GCM-RCM chain

Internal variability and uncertain initial conditions ➡ No temporal correspondence with «real-world» (except for long-term forced trends)

Re-analysis/GCM
Idealized setups

Sensitivities,
process
understanding

Types of Evaluation

EVALUATION RUN
(Re-analysis driven)

- Assumption of «**perfect boundaries**»
- Separation of downscaling performance from biases due to erroneous large-scale forcing
- Temporal correspondence on large temporal and spatial scales

SCENARIO RUN
(GCM-driven historical)

- Evaluation of combined GCM-RCM chain
- RCM results strongly influenced by errors in the boundary forcing («garbage in – garbage out»)
- **No temporal corresponden-ce!** (especially if driven by AOGCM)

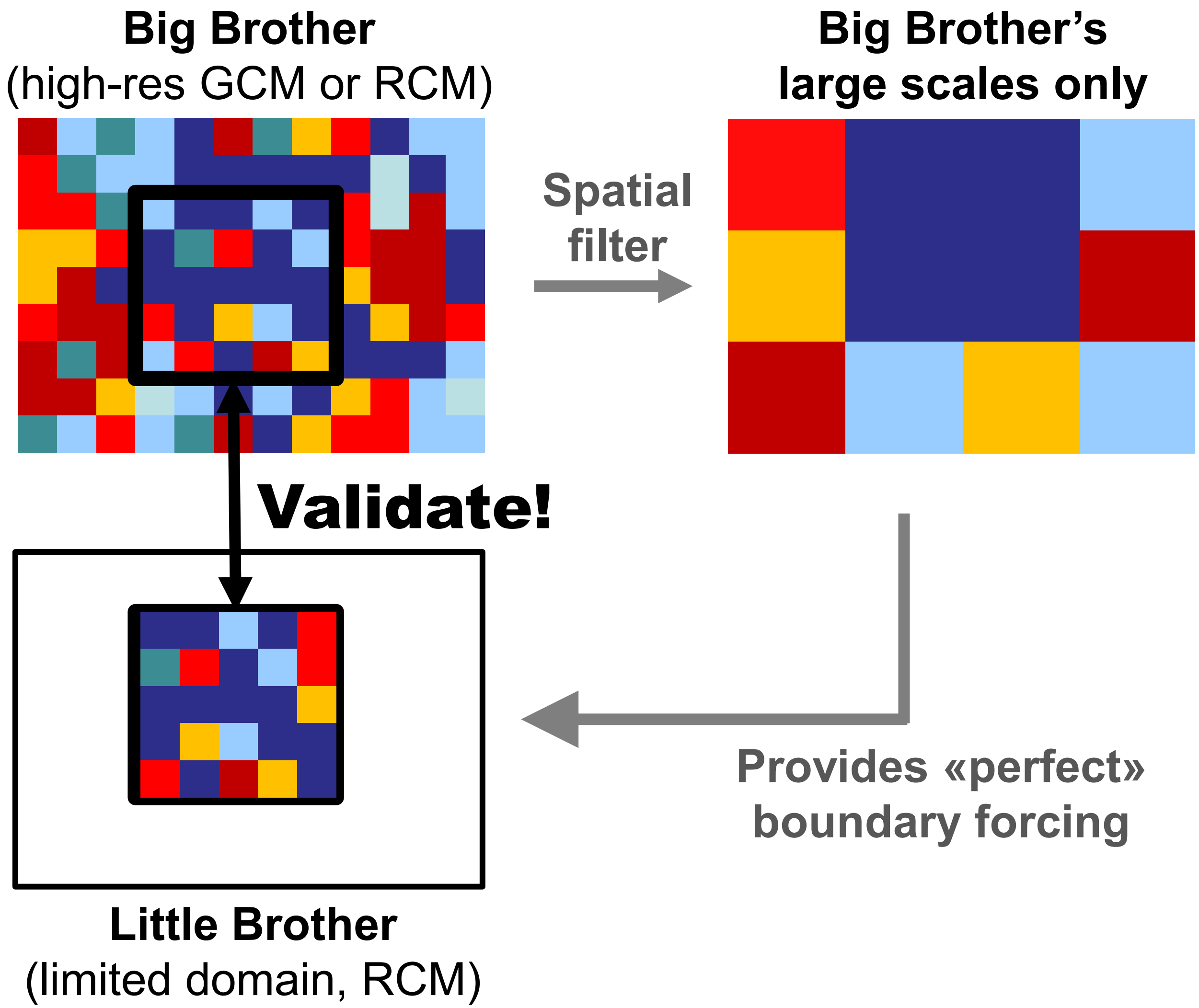
SENSITIVITY RUN

- Scope of evaluation strongly depends on specific setup
- Typically physical-based evaluation
- Reference: often another simulation of the same model

REFERENCE

The Big Brother Protocol (Denis et al. 2002)

Isolates the errors of the nesting strategy



Performance Metrics (1)

SIMULATION → **Comparison** ← **REFERENCE**

↓

Performance metric

- Metrics should **measure/quantify** the model performance against a **given reference dataset** for a **specific aspect**: «Is the model able to simulate things we have observed?»
- Combined scores (accounting for several aspects / variables) possible
- Ideally, a metric should allow a **comparison of the performance of different models** («good performance» -> «bad performance»): scalar quantity
- Usually not desgined to diagnose **reasons** for model errors
- Assessment of **temporal and spatial variability** of performance of a given model

Performance Metrics (2)

APPLICATION-DRIVEN

«I'm only interested in mean annual tempe-rature, therefore my metric should only consider performance wrt. mean annual temperature.»

«I'm only interested in the Alps, therefore my metric only needs to consider model performance in this region»

Often easy to carry out.

But potentially dangerous: Compensating errors might indicate good model performance.

Provides little evidence whether or not the physics are well represented.

PHYSICS- AND PROCESS-RELATED

Assess model performance with respect to the representation of physical processes.

Typically requires to include more than one variable.

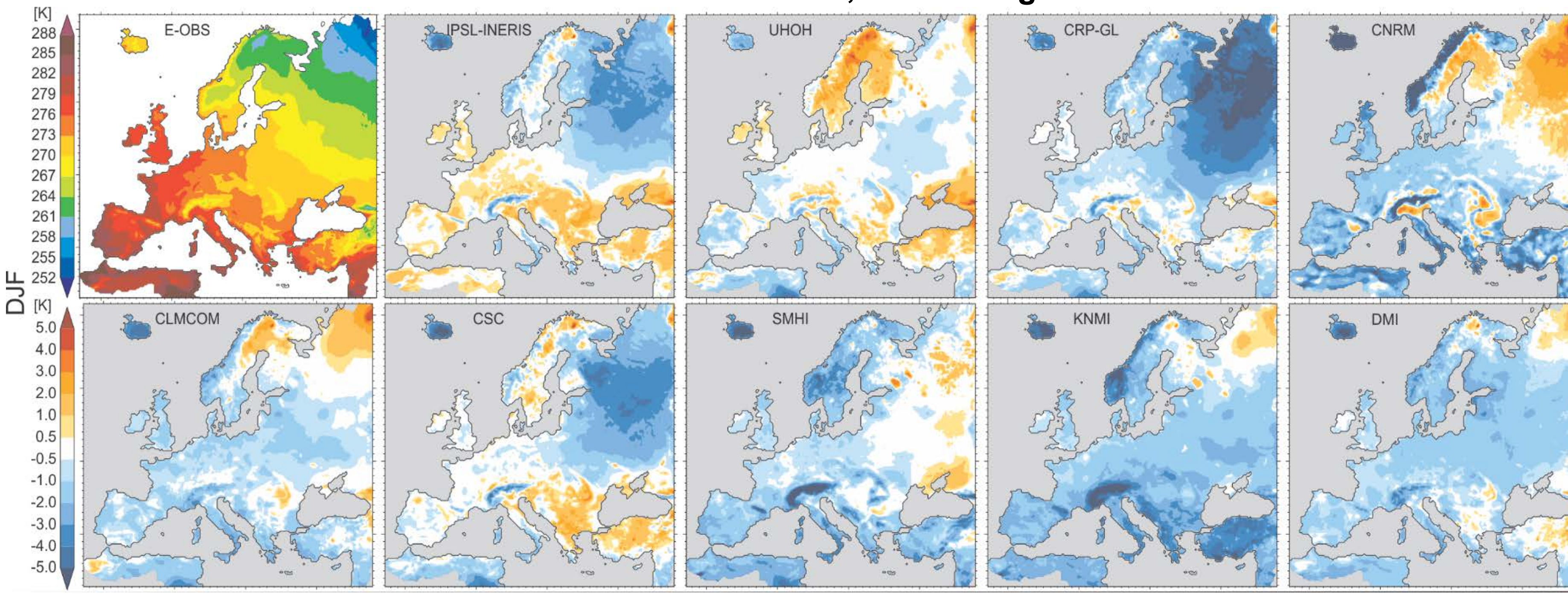
Typically more relevant for obtaining trust in a model.

Probably more relevant for climate change signals.

Often limited availability of reference data.

Example 1: Grid-cell-based mean precipitation bias

Bias of 20-year mean winter temperature (1989-2008)
Models: ERA-Interim-driven EURO-CORDEX RCMs,reference: gridded EOBS dataset



Kotlarski et al., GMD, 2014

Example 2: Spatial Taylor Diagram (Temperature)

Models: ERA-Interim-driven EURO-CORDEX RCMs, reference: gridded EOBS dataset

EUR-11 DJF	EUR-44 DJF	ENS-22 DJF	CLMCOM	CSC	SMHI	KNMI	DMI
EUR-11 JJA	EUR-44 JJA	ENS-22 JJA	CNRM	IPSL-INERIS	CRP-GL	UHOH	

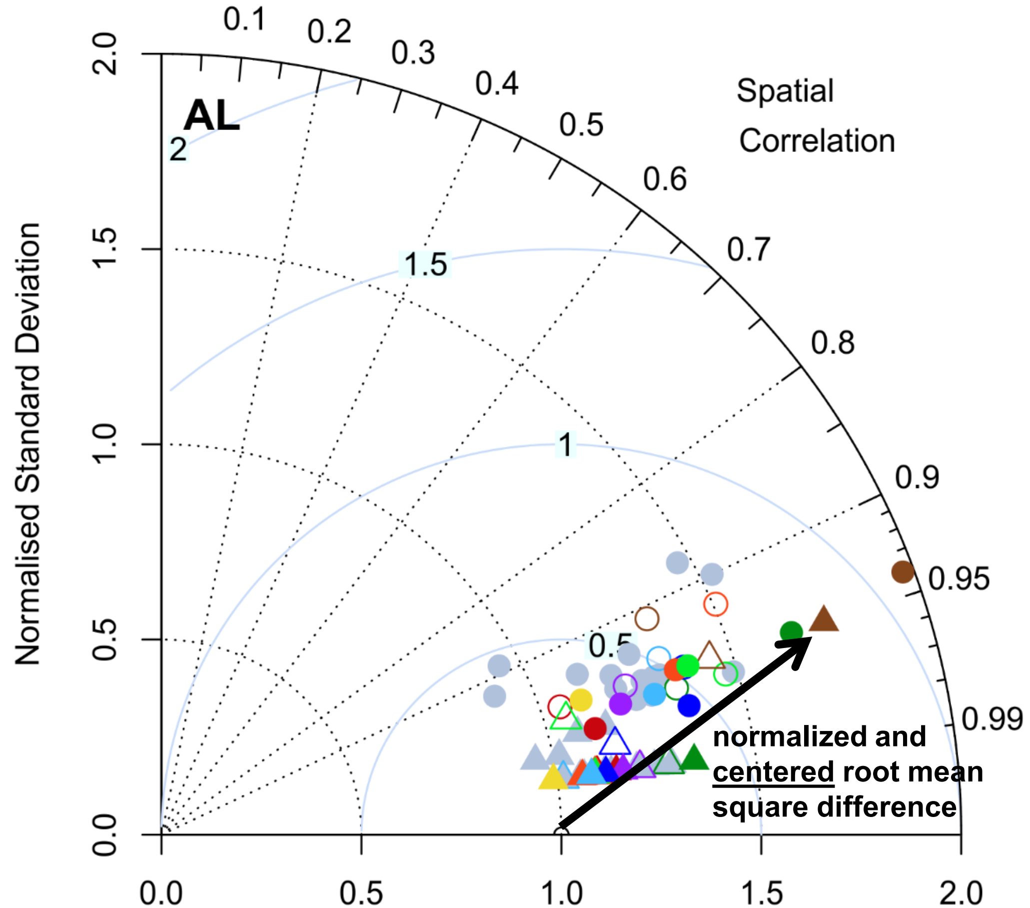


Figure B5. Spatial Taylor diagrams exploring the model performance with respect to the spatial variability of mean winter (circles) and mean summer (triangles) temperature within subdomains AL, BI, FR and MD (see Fig. 9 for subdomains EA, IP, ME and SC). Filled markers: EUR-11 ensemble, nonfilled markers: EUR-44 ensemble, gray markers: ENS-22 ensemble. The diagrams combine the spatial pattern correlation (PACO, cos(azimuth angle)) and the ratio of spatial variability (RSV, radius). The distance from the 1–1 location corresponds to the normalized and centered root-mean-square difference (which does not take into account the mean model bias), expressed as multiples of the observed standard deviation. Note the different number of underlying grid cells per subdomain in the individual ensembles.

Kotlarski et al., GMD, 2014

Example 3: Complex metric

Bellprat et al., 2012

$$\mathrm{PI} = \frac{1}{VRTY}\sum_{v}^{V}\sum_{r}^{R}\sum_{t}^{T}\sum_{y}^{Y}\frac{\sqrt{(m_{v,r,t,y} - o_{v,r,t,y})^2}}{(\sigma_{o_{v,r,t}} + \sigma_{iv_{v,r,t,y}} + \sigma_{\epsilon_{v,r,t,y}})}. \quad (1)$$

Here, $V = 3$ is the number of model variables (T2M, PR, CLCT), $R = 8$ is the number of analysis regions (PRUDENCE regions), $T = 12$ is the number of temporal means (months), and Y is the number of years evaluated, which depends upon the ensemble considered. The variables m and o denote simulated and observed monthly means for the respective variable and region, σ_o is the standard deviation of the interannual variations derived from the observations, σ_{iv} is the standard deviation of the internal variability of the regional model derived from ensemble IV, and σ_ϵ is the standard deviation of the observational error derived from different reference datasets. For each variable (T2M, PR, and CLCT) we use three independent datasets, listed in Table 3, to estimate the observational error.

PI=0 -> perfect match

OUTLINE

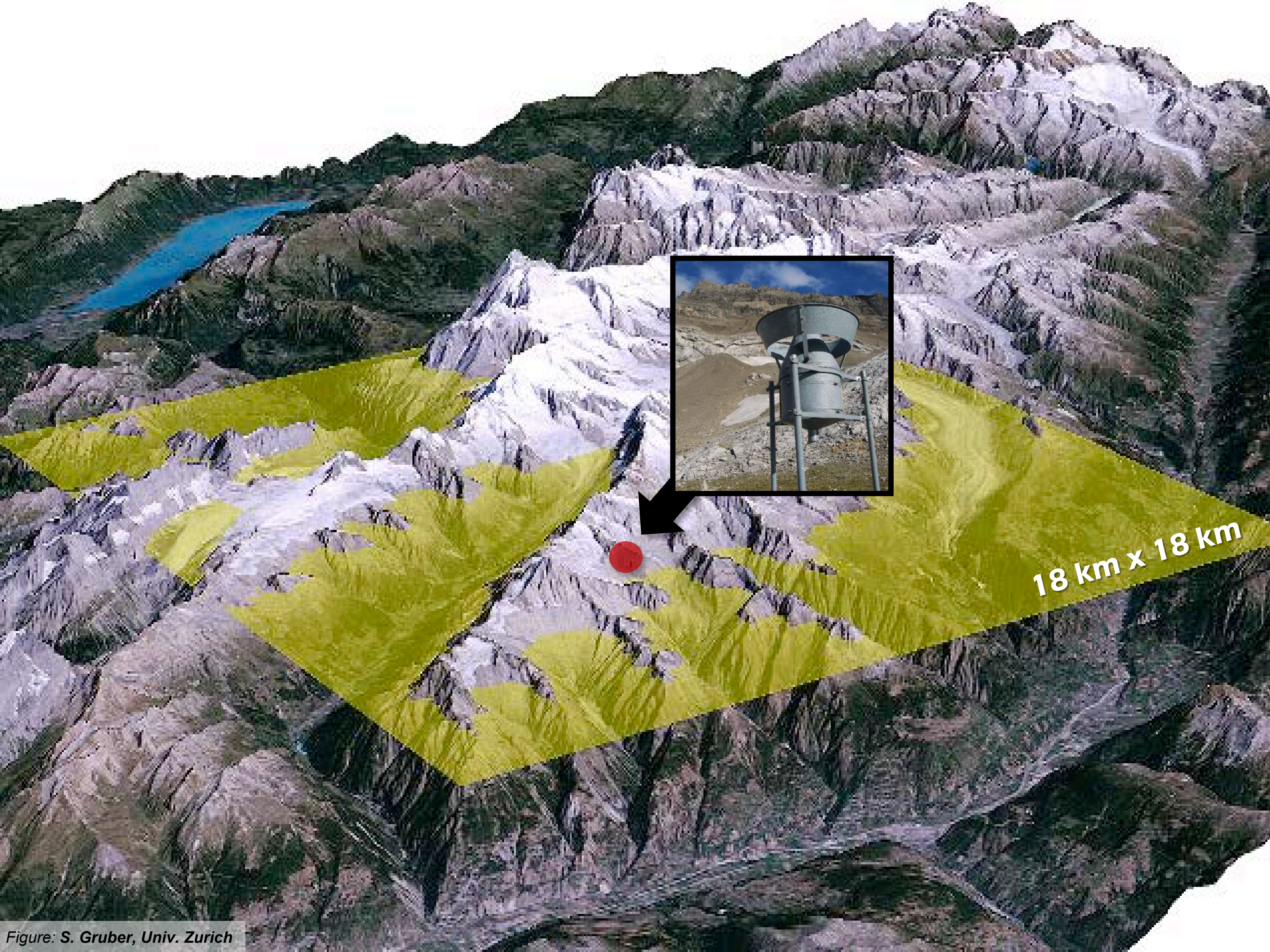
1. The rationale of RCM evaluation

2. Techniques and measures

3. **Potential pitfalls**

4. Summary & conclusions

SCALE ISSUES / SPATIAL REPRESENTATIVITY

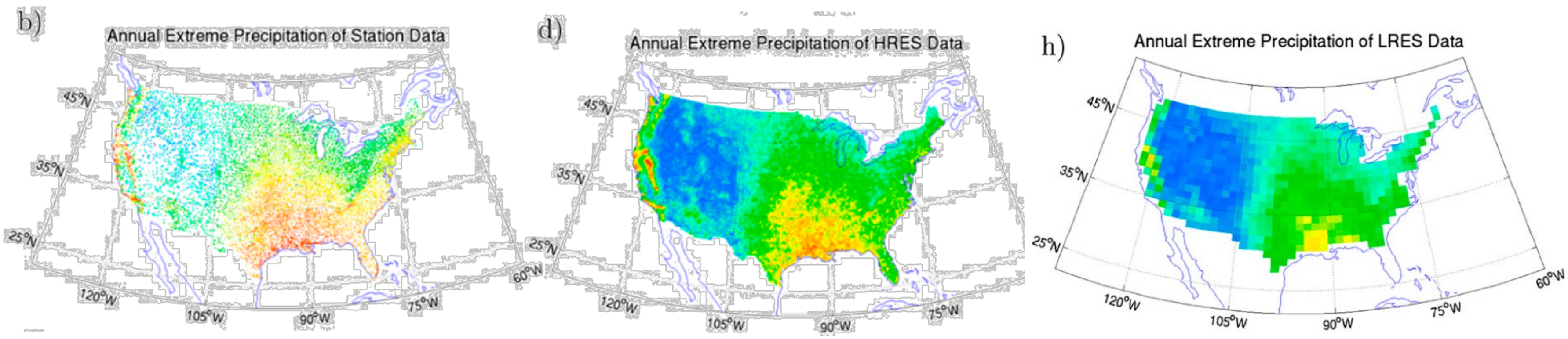

18 km x 18 km
Figure: S. Gruber, Univ. Zurich

The Scale Mismatch

- RCMs operate on **grid cell scale**
- Output typically needs to be interpreted as «**mean over grid cell area**»
- Compared to the site scale, this is associated with
  - **Smoothing of spatial variability**
  - **Smoothing of (localized) extremes**, especially precipitation and winds
  - **Elevation and slope effects** in topographic terrain
  - Neglect of **subgrid variability** (as, for instance, introduced by land surface characteristics): Often not even seen by RCMs

Gridding effects

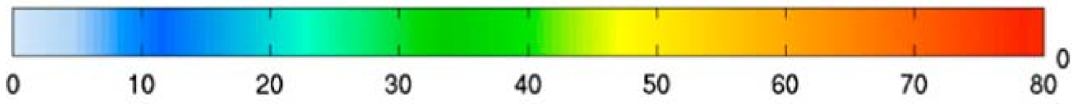
97th percentile of wet-day precipitation (1979-2003):
Stations vs. grids



GHCN stations

Gridded to 0.25°
(Cressman interp.)

Remapped to 0.9° x 1.25°
(Conservative remapping)



Gervais et al., 2014

Gridded Reference Data

Use of gridded reference data

A) Station measurements interpolated onto a regular grid

- Measurements and interpolation subject to considerable uncertainties! (see later)

B) Re-analysis products

- Observations only indirectly represented (data assimilation)
- Uncertainties due to assimilation scheme, re-analysis model and changing mix of underlying observational data
- For instance: introduction of satellite data in 1970s

C) Remote sensing products

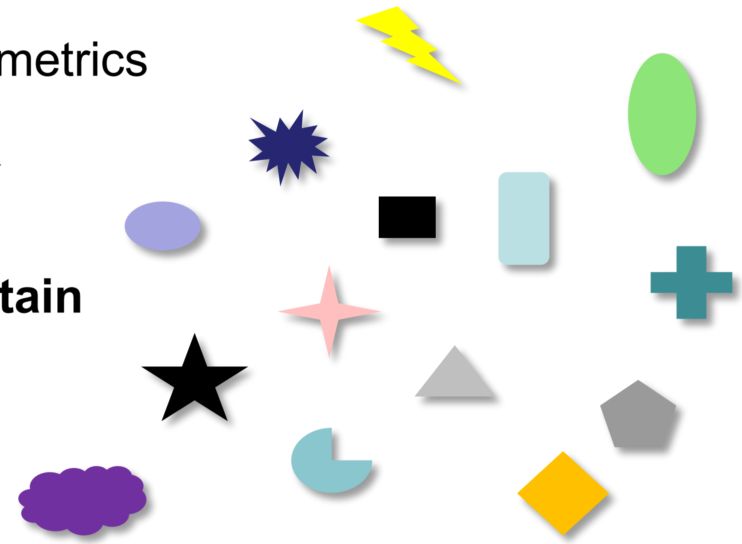
- Also involve models and assumption
- Good spatial, but typically limited tem

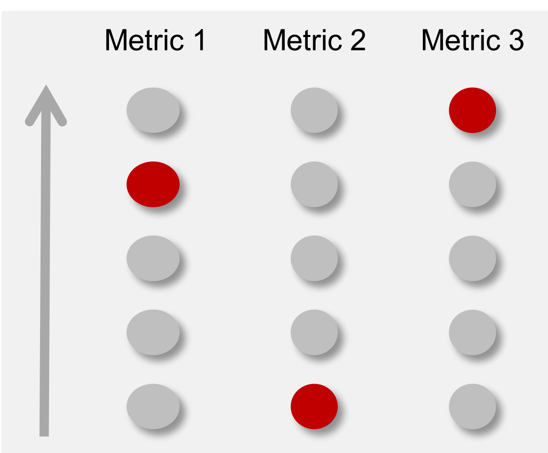
Exception:
Validation of RCMs in idealized «**single column mode**» (RCM development)

METRIC SELECTION

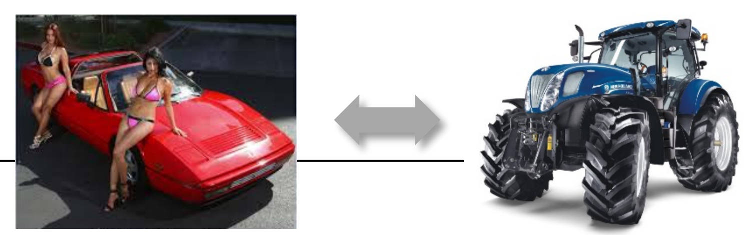
Choice of Performance Metric (1)

- «**Metric Zoo**»: Infinite number of potential metrics
- No well-defined common set of benchmark metrics; but several «standard» metrics
- **One single metric ALWAYS neglects certain aspects of model performance**
- RCM: Metrics typically consider climatology or trend!





- **Subjective choice**
- Outcome of evaluation exercise typically strongly depends on metric
- **Concept of one best model is ill-defined! (but there may be a best model for a given purpose)**



MODEL CALIBRATION

The Role of Model Calibration (1)

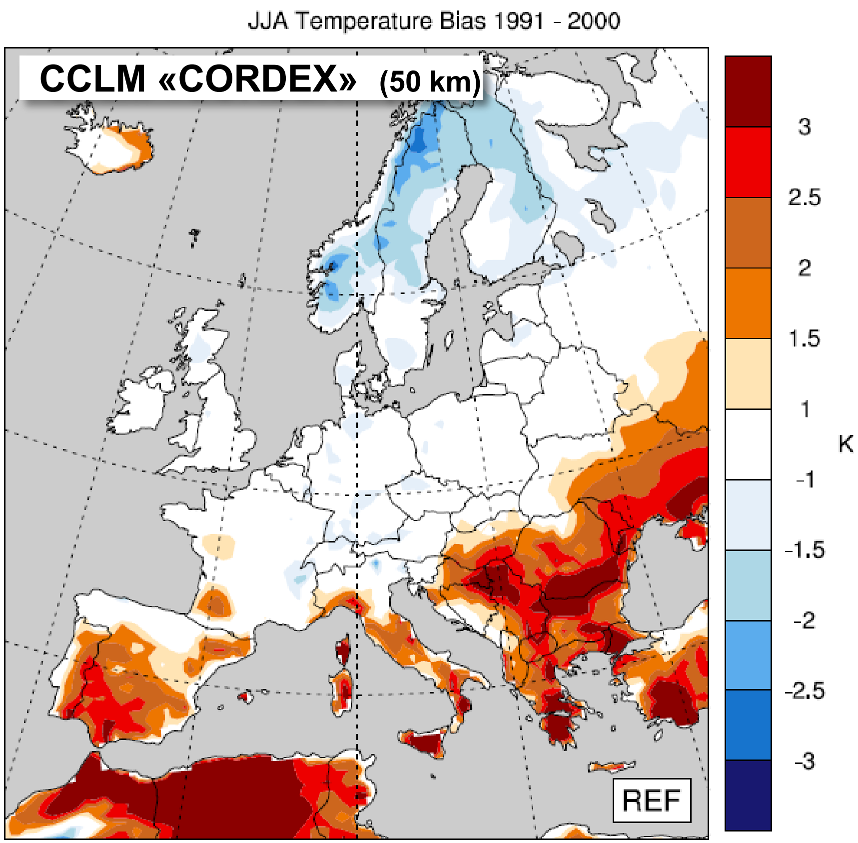
- RCMs physically based, but especially model **physics** typically include a large number of poorly constrained parameters that need to be calibrated («**tuning**»)
- **Calibration will affect model performance**!
- The same is true for further choices concerning model setup (domain size, time step, relaxation procedure, horizontal and vertical resolution, etc.)
- Calibration is typically **intransparent** (calibration procedure and target not known)

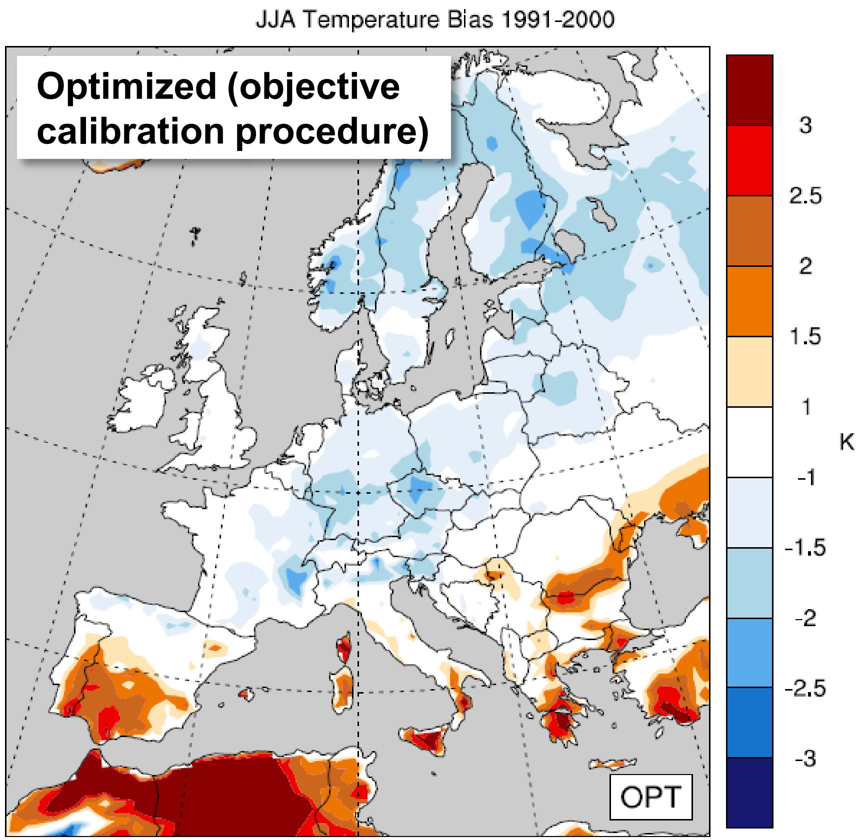
➡ **Evaluation might not be independent (if the same evaluation period, reference data and performance measures were used during calibration)**

➡ **Weak test of performance!**

(However, calibration not as explicit as in statistical downscaling)

The Role of Model Calibration (2)





Bellprat et al., submitted

INTERNAL VARIABILITY

Internal Variability (IV) in RCMs

Unforced random variability in climate due to internal non-linear processes in the climate system

Introduces sample uncertainties in climate model output

- Even **with identical boundary forcing**, slightly differently initialized or perturbed RCM experiments with exactly the same setup will differ from each other to some extent
- This effect is random!!
- Furthermore: Observational reference just reflects one realization of possible climates

Internal Variability (IV) in RCMs

IV influence is

larger for short analysis periods (partly averages out on longer time scales)

larger for small analysis domains (partly averages out by spatial averaging)

larger for (rare) extremes

typically **larger for precipitation** than for temperature

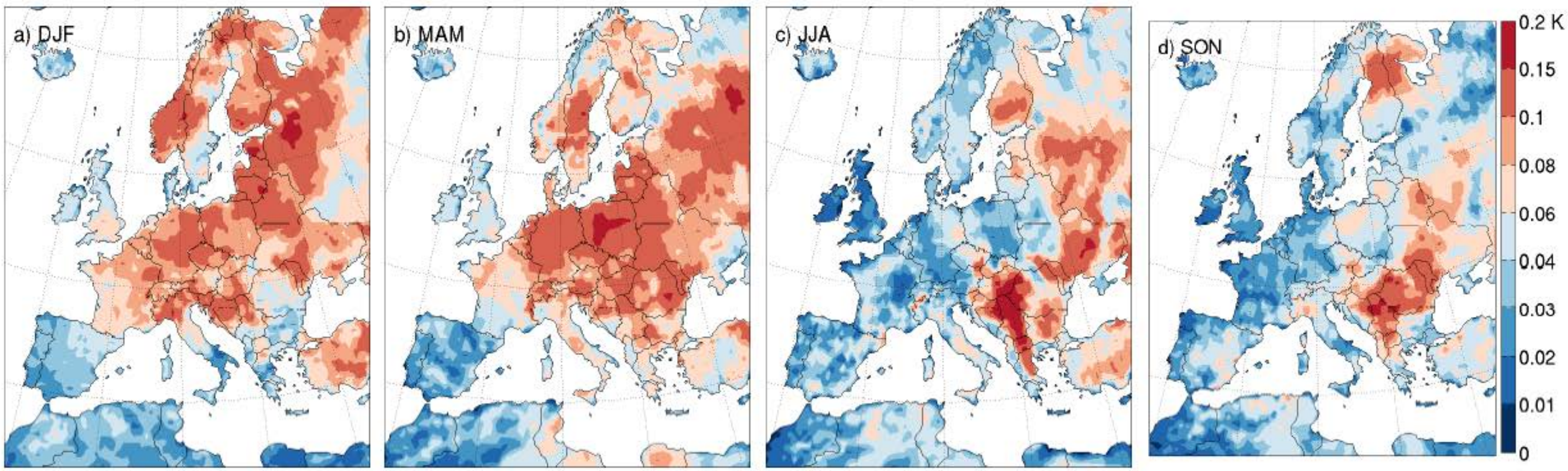
typically **larger in summer** (RCM solution less constrained by boundary forcing)

larger towards the outflow boundary (RCM solution less constrained by boundary forcing)

Influnec of IV on 42-year RCM Climate

4 COSMO-CLM simulations for 1958-2000 driven by ERA40 re-analysis with slightly shifted start dates

Mean seasonal temperature difference (42-year means) between the sensemble members



Roesch et al., 2008

"It can thus be concluded that the model's performance in predicting climate extremes cannot be properly evaluated using only one model simulation"

OBSERVATIONAL UNCERTAINTY

Observational Uncertainty: Origins

- **Measurement errors** (e.g., automatic weather stations)
- **Deficient translation** of measured quantities into validation parameters (e.g. radiances to temperatures, cloud coverage or precipitation rates)
- Inappropriate **gridding procedure** and/or target resolution
- **Spatial and/or temporal inhomogeneities** of underlying station dataset
- **Representativeness errors**, including physiographic effects (Does a grid point of an observational grid really represent **areal averages**? Is the **reference altitude** of observations and models the same?)

Measurement Errors: Precipitation

- **Systematic undercatch** of rain gauges due to deformation of wind field and evaporative losses
- Strongly depends on site characteristics, ambient weather conditions and measurement device
- Most important for snowfall and during strong winds (less than 50% of true precipitation)
- Usually **not corrected for** in gridded products)

A wet model bias of 10-20% can well be explained by deficient observations!

Only of minor importance for statistical downscaling

Complicates comparison of SD and RCM performance

Influence on Model Evaluation

RCMs versus **E-OBS**

RCMs versus **national grid** with high underlying network density

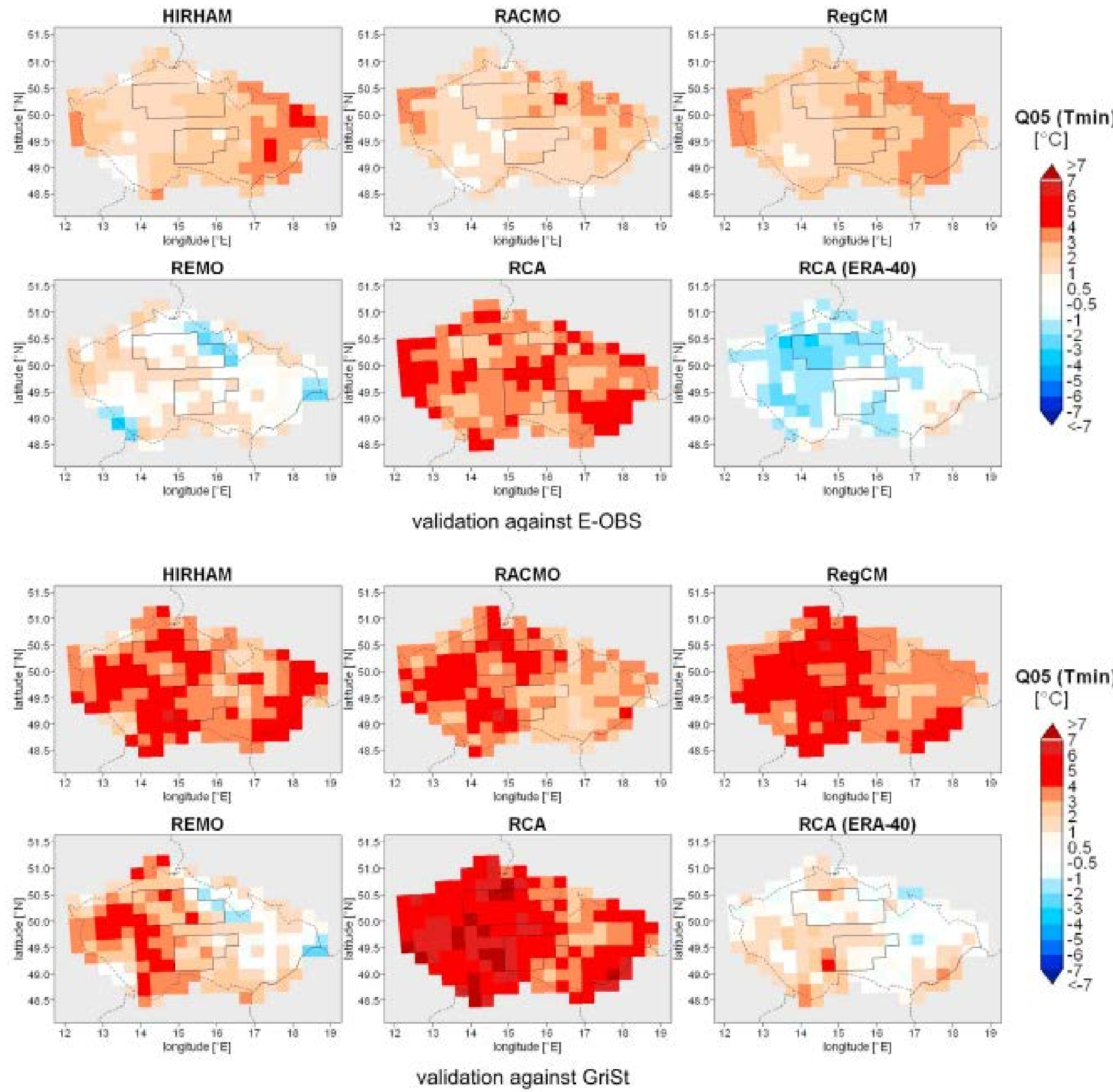


Figure 6. Differences in the 5% quantile of T_{min} in DJF between control simulations of RCMs (1961–1990) and gridded observed data for (top) GriSt and (bottom) E-OBS.

Kysely and Plavcova, 2010

PRESENT-DAY PERFORMANCE VS. CLIMATE CHANGE SIGNAL

&

NON-STANTIONARITY OF MODEL BIASES

Bias Non-Stationarities

- Model bias **cannot necessarily be assumed to be stationary in time**, particularly if two different climatic states are considered
- Limited significance of evaluating performance in historical periods; bias changes will distort simulated climate change signal!
- Observational and historical simulation record typically too short to diferentiate between two climatic states
- No future observations available for assessing future model biases (**pseudo realities** can partly help out)

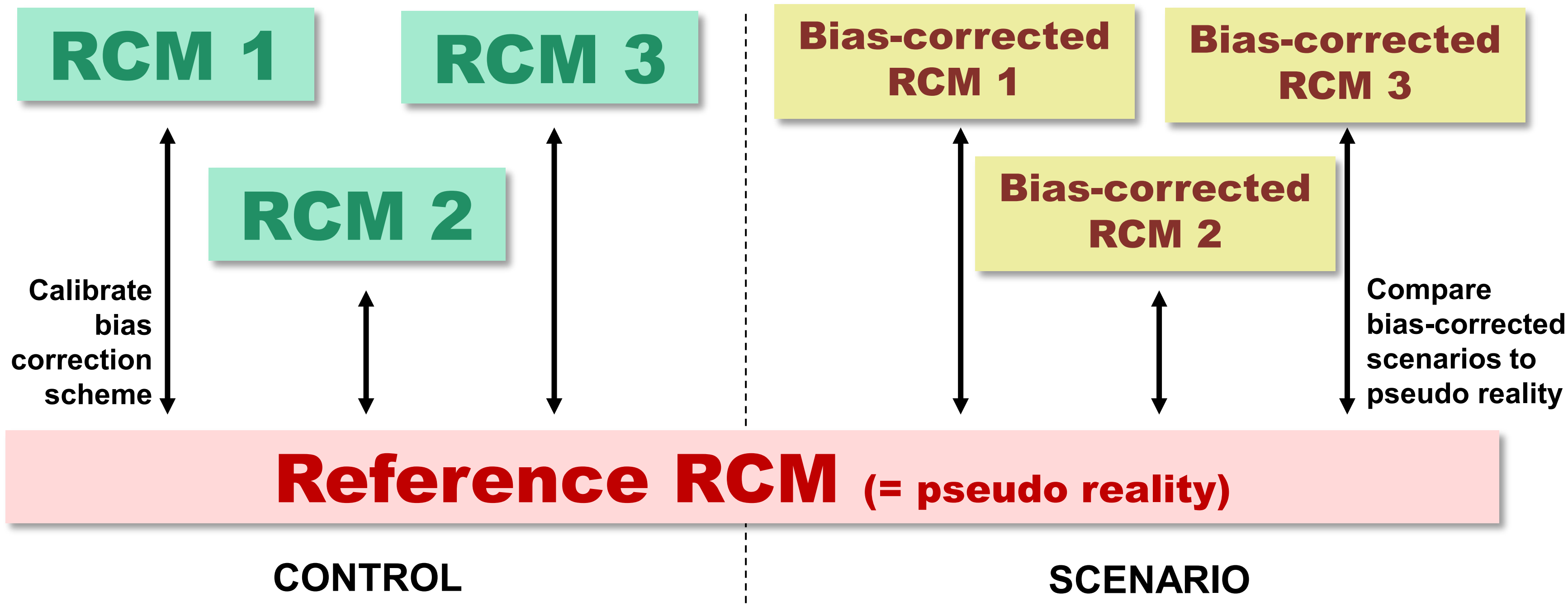
Indeed

clear relation between skill in present-day climate and simulated climate change signal usually not found

strong indications for non-stationary biases (Boberg and Christensen 2012, Bellprat et al. 2013, Maraun 2012)

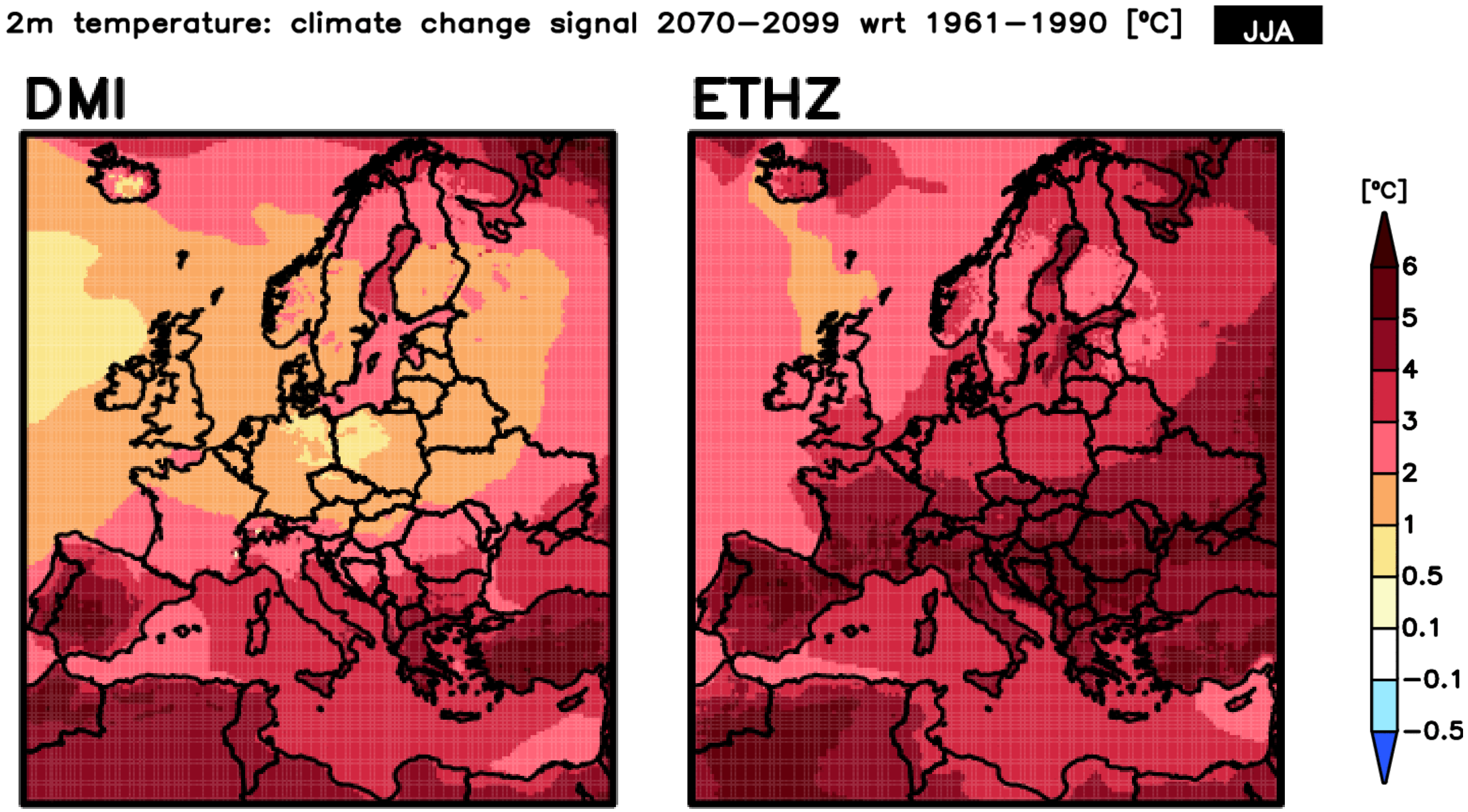
Pseudo Realities

(e.g. Vrac et al. 2007, Maraun 2012, Bellprat et al. 2013)



- **Cannot uncover all kinds of bias non-stationarities (common non-stationaries possible)**
- **But: Provides strong evidence for bias non-stationarities over some regions and for some parameters**

Stationary Model Bias?



Do these models show a stationary temperature bias on the spatial and temporal scales considered?

Further Issues

SKILLFUL SCALE

Can a climate model really be analysed and evaluated at its nominal spatial resolution?
(Several grid cells are required to represent atmospheric phenomena!)

REPRESENTATIVENESS

Should we assume that the simulated location of some phenomenon is identical to the «true» location?
(or are there systematic spatial shifts in the climate model output)

QUALITY OF BOUNDARY FORCING

The skill of an RCM depends on the quality of the supplied boundary forcing!

SPATIAL CORRELATION OF MODEL BIAS

Biases at individual grid cells cannot be assumed to be independent of each other
(important for hypthesis testing)

OUTLINE

1. The rationale of RCM evaluation
2. Techniques and measures
3. Potential pitfalls
4. **Summary & conclusions**

Summary and Conclusions

- **(Regional) Climate model evaluation as an important component of model development and application**
- **Important to provide trust into models and their scenarios**

- **Infinite number of evaluation schemes!**
- **Choice of scheme can strongly determine final outcome**
- **RCM evaluation ALWAYS has a subjective component**
- **Large number of issues to consider during evaluation exercise and interpretation of results**

RCM versus SD Evaluation

RCM evaluation ...

should not be carried out a the point scale but at the RCM grid cell scale or coarser (**scale mismatch**)

can typically **not be carried out event-wise, particularly not if small spatial scales are considered (IV!)**

has to account for the fact that only a «**global calibration**» is possible

is directly influenced by issues of **spatial representativeness** and **measurement errors**

A Final Note

- **Skill in the present does not imply skill in the future**
- **But: A model has to reflect the behaviour of the real system in order to be suitable for scenario development (minimum requirement)**

THANK YOU

sven.kotlarski@meteoswiss.ch