Bias correction (BC) has become a standard procedure in climate change impact studies due to biases in regional climate model (RCM) output that prevent a direct use. There are numerous assumptions and consequences that are connected with applying a BC. Several of those have been discussed and analysed in studies. The effect of the sample size used for BC calibration on BC performance has however so far not been looked into.

Especially in case of precipitation we expect a strong dependence of BC performance on the sample size. Therefore we apply state-of-the-art BC methods based on different sample sizes for calibration and compare their performances to that of a BC based on 30 years.

### Data sets and domain
- precipitation data of 10 RCM reanalysis runs (EU-ENSEMBLES 25 km, 1961–2000, ERA-40)
- Germany as area of interest

### Methods
**BC: 4 quantile matching (QM) approaches**
- eQM: empirical QM
- PTF: Piani Transfer Function
- gQM: transfer function based on a gamma distribution
- GQM: transfer function is a combination of gamma and generalized Pareto distribution (GPD)

**BC performance assessment: 3 skill scores**
- MAE (mean absolute error)
- Perkins skill score
- Ext$_{0.1}$ (measures deviation in the ten highest values)

For every constellation (Fig. 2) and every skill score:
- the skill score is applied to the data of the validation period of all bias corrected data sets as well as of the uncorrected RCM data
- for every bias corrected data set the difference in skill score value to those for the uncorrected RCM data is calculated
- these difference fields are tested versus the ‘best fit’ (one-sided Wilcoxon-test, α = 5%)
- the critical sample size $n_{crit}$ is the largest sample size that shows a significant decrease in skill score values
- $n_{crit}$ quantifies the effect of sample size on BC performance

**Bias correction (30 years)**

- in the calibration period the more complex QM approaches (especially eQM, but also GQM) clearly outperform the less complex QM approaches
- in the validation period the performances of all QM approaches are on a comparable level

### Conclusions
- a small decrease in sample size can result in a significant worse BC performance
- a general critical sample size can not be assessed, since both $n_{crit}$ values vary strongly and are therefore influenced by the choice of the QM approach and also by the choice of the calibration period, but always $>10$ years
- complex QM approaches (eQM, GQM) show larger $n_{crit}$ values
- if unknown data are bias corrected, less complex QM approaches (gQM, PTF) are found to be more robust, show a comparable performance to the more complex ones and are hence favourable

### References

**Assessment of BC performance**
- $n_{crit}$ is the largest sample size that shows a significant decrease in skill score values
- $n_{crit}$ quantifies the effect of sample size on BC performance

**Results for one exemplary constellation**
- the results show a decrease BC performance with decreasing sample size, especially for sample sizes smaller than 10 years
- for this exemplary constellation at a sample size of $n_{crit}$ = 17 years the BC performance is, for the first time, worse than that of the ‘best fit’

**Results for all constellations**
- we found a large spread of $n_{crit}$ values that is comparable for all skill scores ($\approx$ 28–10 years)
- the $n_{crit}$ value that shows the effect of sample size, is found to be strongly influenced by the choice of the QM approach but also by the choice of the calibration period
- overall, QM approach eQM shows the largest $n_{crit}$ values, followed by GQM, PTF and finally eQM.