# On the verification of climate reconstructions

G. Bürger, U. Cubasch

Clim. Past Discuss., 2, 357–370, 2006
www.clim-past-discuss.net/2/357/2006/


Climate of the Past Discussions

*Climate of the Past Discussions* is the access reviewed discussion forum of *Climate of the Past*

# On the verification of climate reconstructions

**G. Bürger and U. Cubasch**

Institut für Meteorologie, FU, Berlin, Germany

Received: 16 May 2006 – Accepted: 7 June 2006 – Published: 3 July 2006

Correspondence to: G. Bürger (gerd.buerger@met.fu-berlin.de)

## Abstract

The skill of proxy-based reconstructions of Northern hemisphere temperature is reassessed. Using a rigorous verification method, we show that previous estimates of skill exceeding 50% mainly reflect a sampling bias, and that more realistic values vary about 25%. The bias results from the strong trends in the instrumental period, together with the special partitioning into calibration and validation parts. This setting is characterized by very few degrees of freedom and leaves the regression susceptible to nonsense predictors. Basing the new estimates on 100 random resamplings of the instrumental period we avoid the problem of a priori different calibration and validation statistics and obtain robust estimates plus uncertainty. The low verification scores apply to an entire suite of multiproxy regression-based models, including the most recent variants. It is doubtful whether the estimated levels of verifiable predictive power are strong enough to resolve the current debate on the millennial climate.

## 1 Introduction

The validity of proxy based reconstructions of Northern hemisphere temperature (NHT) has attracted a lot of attention in recent years (McIntyre and McKitrick, 2003; von Storch et al., 2004; McIntyre and McKitrick, 2005a–c; Huybers, 2005; Rutherford et al., 2005; Mann et al., 2005; Bürger and Cubasch, 2005; Bürger et al., 2006; Wahl and Ammann, 2006). Aspects of methodology, proxy quality, and verification assessment have been analysed to cover a number of unresolved issues of the Mann et al. (1998) (henceforth MBH98) publication. That study and a follow-up paper (Mann et al., 1999) used a limited number of proxies (dendro, ice-core, corals) as regressors for the main patterns of temperature variability and derived a global temperature history of the past millennium. The method was verified against instrumental data, and for the 22 proxies available back to AD1400 a reduction of error, *RE* (Lorenz, 1956, see below), reported of 42% for the calibration (1902–1980) and 51% of the validation (1854–1901) period.



CPD

2, 357–370, 2006

These validation scores, however, disagree with other scores, such as the coefficient of efficiency, *CE* (−22%; Nash and Sutcliffe, 1970, see below), correlation, $R^2$ (2%), or detrended measures (0%, cf. Mann et al., 1998 SI; Wahl and Ammann, 2006). In the following, we will concentrate on the two scores *RE* and *CE* ($R^2$ is scale independent and thus not really appropriate.)

The scores are estimated from strongly autocorrelated (trended) time series in the instrumental period, along with a special partitioning into calibration and validation samples. This setting is characterized by a rather limited number of degrees of freedom. It is easy to see that calibrating a model in one end of a trended series and validating it in the other yields fairly high *RE* values, no matter how well other variability (such as annual) is reproduced. Note that these few degrees of freedom also initiated the debate on using trended or detrended calibration: The latter had been intuitively applied by von Storch et al. (2004) (noted by Bürger et al., 2006) so as to ensure enough degrees of freedom for the regression (thereby departing from the MBH98 setting).

The problem is that few degrees of freedom are easily adjusted in a regression. Therefore, the described feature will occur with any trended series, be it synthetic or natural (trends are ubiquituous): regressing it on NHT using that special calibration/validation partition returns a high *RE*. McIntyre and McKitrick, 2005b demonstrate this with suitably filtered red noise series. We picked as a nonsense NHT regressor the annual number of available grid points and, in fact, were rewarded with an *RE* of almost 50% (see below)!

If even such nonsense models score that high the reported 51% of validation *RE* of MBH98 are not very meaningful. The low *CE* and $R^2$ values moreover point to a weakness in predicting the shorter time scales. Therefore, the reconstruction skill needs a re-assessment, also on the background of the McIntyre and McKitrick, 2005b claim that it is not even significantly nonzero.

A second controversy deals with two criticisms of the regression method itself. von Storch et al. (2004) attribute the low reconstructed amplitudes to an inherent disability of any regression-based method to simulate sufficient variability. In Bürger and

Cubasch (2005) and Bürger et al. (2006) we demonstrate that the method creates an entire spread of millennial histories depending on data processing details. The error grows proportional to both the model uncertainty and the proxy scale, the latter leading to an extrapolation. The criticisms rest on properties of the full proxy-temperature covariance matrix: the first on the scale and the second on the number and corresponding uncertainty of its entries (which are tens of thousands mutual covariances).

In newer studies (Mann and Rutherford, 2002; Rutherford and Mann, 2003; Rutherford et al., 2005; Mann et al., 2005) the estimation of that covariance matrix utilizes a technique called regularized expectation maximization (RegEM; Schneider, 2001). RegEM extends the classical expectation maximization (EM) algorithm (Dempster et al., 1977) to situations with more unknowns than cases. But it is evident that the two critiques above pertain to this newer scheme as well. (A few additional issues regarding the specific use of RegEM are discussed in a supplement.)

In the above literature no millennial verification skill is reported for RegEM (see below); the millennial reconstructions themselves are nevertheless similar to MBH98 (cf. Rutherford et al., 2005). For our study we have decided to choose the RegEM variant instead of the original MBH98 approach, following a suggestion of Rutherford and Mann (2003). But note that their application includes the utilization of the full proxy-temperature covariance, in contrast to MBH98 who explicitly work with a reduced space version of the temperature fields (S. Rutherford, personal communication).

The study is an attempt to thoroughly estimate the skill of current NHT reconstruction methods. The skill evades – as much as possible – any properties that solely reflect the sampling of the calibration period, but instead utilize the maximum possible degrees of freedom available in the instrumental period.

## 2 Stationarity digression

Before we explain our testing procedure, we recall one of the most basic estimation principles: that a regression/verification exercise is generally nonsense if calibration

and validation samples are not drawn from one and the same population. Consequently, differences between sample properties, such as calibration and validation mean, must be considered completely random. Accordingly, verification methodology was developed from and for stationary records, such as weather or riverflow (Lorenz, 1956; Nash and Sutcliffe, 1970; Murphy and Winkler, 1987; Wilks, 1995). In that method, mean and variance are usually seen as population parameters relative to which errors are to be measured. In fact, from the original articles wherein *RE* and *CE* were introduced, (Lorenz, 1956) and (Nash and Sutcliffe, 1970), respectively, they were only two different names for one and the same thing: the reduction of the squared error relative to the variance of the predicted quantity. Only later they appear as distinguished entities (Briffa et al., 1988; Cook and Kairiukstis, 1990; Cook et al., 1994), in that reference is explicitly made to calibration (*RE*) or validation (*CE)* variance. That latter score, *CE*, has been attributed to the hydrologic study (Nash and Sutcliffe, 1970). However, we have not been able to find therein any reference to a validation mean, nor in any of the articles we checked from the hydrologic literature (e.g. Legates and McCabe, 1999; Wolock and McCabe, 1999). It thus appears that *RE* and *CE* were originally envisaged as identical measures but have been mistaken for distinguished entities in dendrochronology. We emphasize that their difference solely reflects sample properties and must be considered random.

In accordance with classical verification we treat calibration and validation as unspecified samples, using a set of resamplings of the full period. Differences in calibration and validation statistics (such as mean and variance) are purely random and contribute to the uncertainty in model skill (instead of serving as a model selection tool).

## 3 Four steps to reconstruction

The predictand, $\boldsymbol{T}$, is defined from the 219 temperature grid points that are most abundant in the full 1854–1980 period (the verification grid points of MBH98; here we

also use them for calibration (cf. Jones and Briffa, 1992). As a predictor, ***P***, we take the 22 proxies of the AD1400 step of MBH98. We thus have 127 years of common proxy and temperature data. Once a calibration subset of these data is defined (next section) everything is set up for the statistical model. These four steps constitute a temperature reconstruction:

**GLB**: the definition of a target quantity

**COV**: the estimation of cross covariances

**MDL**: the calculation of the regression model

**RSC**: the postprocessing (rescaling)

Note that the informational flow goes strictly from **GLB** through **RSC**.

ad **GLB**) – The target can be either the full temperature field on all grid points, a filtered version thereof (EOF truncation), or the average NHT series.

ad **COV**) – the main statistical quantity to be determined is the full $\Sigma$ cross covariance matrix $\Sigma$, consisting of the 4 submatrices $\Sigma_P$, $\Sigma_{PT}$, $\Sigma_{TP}$, and $\Sigma_T$, between the proxy and temperature fields. Classically, $\Sigma_{PT}$ would simply be the covariance between ***P*** and ***T*** estimated from the calibration sample. In RegEM, an iterative procedure is applied to estimate the full $\Sigma$, wherein only the validation ***T*** is withheld, and $\Sigma$ and ***T*** are mutually approximated using the expectation of ***T*** given ***P*** under the current iterate of $\Sigma$. The expectation is determined via some form of (regularized) regression.

ad **MDL**) – any regression model can be derived from the full , as follows:

1. $R = \Sigma_P^{-1}\Sigma_{PT}$ – least squares (LS) regression.

2. $R^I = \Sigma_{TP}^{+}\Sigma_T$ – inverse regression ("$^+$" denoting pseudo inverse).

3. truncated total least squares (TT; Fierro et al., 1997).

4. ridge regression (RR; Hoerl, 1962).

Note that the calculations 1.–4. are based on standardized variables with subsequent rescaling. Here we follow most regularizations schemes, such as 3. and 4., as well as the RegEM implementation of (Schneider, 2001).

ad **RSC**) – The result is rescaled to match the calibration variance (cf. Bürger et al., 2006).

The details are found in a supplement (http://www.clim-past-discuss.net/2/357/2006/cpd-2-357-2006-supplement.zip). Table 1 illustrates the various settings.

By varying the criteria a set of 48 model variants or flavors is defined, identifiable by a quadruple from [0,2]×[0,1]×[0,3]×[0,1] as in Bürger and Cubasch, 2005. For example, the MBH98 method corresponds to variant 1011 and Rutherford et al., 2005 to 0130.

## 4 Resampling proxies and temperature

Suppose we have fixed a calibration set **C** consisting of **P** and **T** values, and we want to build one of the models **M** above. Since **M** does not explicitly contain the time variable, it only depends on the set **C** of selected **P** and **T** values, and not on their ordering. In an ideal world, any other sampling $\mathbf{C}'$ would result in the same model $\mathbf{M}(\mathbf{C}')= \mathbf{M}(\mathbf{C})$. In the real world of statistics there are sampling errors, and the estimates $\mathbf{M}(\mathbf{C})$ vary more or less about the "true" model **M** – if such model exists at all. We assume **M** exists, and for each of the above flavors we are now estimating it along with a corresponding uncertainty range. We are confident that 100 random sets **C** are sufficient to obtain reasonable estimates.

Let the full, 127-year long **P–T** record be given as $(P_i, T_i, i \in I)$. A "random" calibration set is defined by picking a random permutation $\pi$: $I \rightarrow I$ and letting

$$\mathbf{C}(\pi) = \{ i \in I \quad | \quad \pi(i) \leq n/2 \} \tag{1}$$

($n$ being the length of $I$). This divides the original record into two sets, calibration $\mathbf{C}(\pi)$ and validation $\mathbf{V}(\pi)=I \setminus \mathbf{C}(\pi)$, of roughly equal size. Applying now the 4 steps from above yields 48 flavors per $\pi$, and doing this 100 times generates for each flavor a distribution of 100 regression experiments.

Their predictive power is evaluated in terms of NHT, calculated from observed, $x$, and predicted values (from the validation part), $\hat{x}$, of the 219 grid points. In accordance with (Lorenz, 1956) and (Briffa et al., 1988) we use the scores

$$RE = 1 - \frac{\left\langle (\hat{x} - x)^2 \right\rangle}{\left\langle (x - \bar{x}_c)^2 \right\rangle} \quad ; \quad CE = 1 - \frac{\left\langle (\hat{x} - x)^2 \right\rangle}{\left\langle (x - \bar{x}_v)^2 \right\rangle}, \tag{2}$$

with brackets indicating expectation. They are only distinguished by the different reference value of calibration and validation mean, $\bar{x}_c$ and $\bar{x}_v$, respectively.

Their distribution is depicted in Fig. 1. For each flavor, the scores from the random calibrations show a considerable spread, but that spread is remarkably similar for *RE* and *CE*. This demonstrates that, in fact, both measure the same thing, and possible differences merely reflect sampling properties. Most flavors have difficulty predicting the entire **T** grid (0xxx), except maybe the variant 0130 favored by Mann et al. (2005). Overall, prefiltering the **T** grid using EOF truncation noticeably improves the performance. Here also the use of RegEM gives a few percent of additional score. Interestingly, using NHT itself as a predictand (2xxx) does not seem to be favorable as many calibrations show very poor performance, thus increasing the uncertainty. Most of the higher scores lie in a range somewhere between 10% and 30%. Given this uncertainty, it is hard to pick one flavor as optimal. From the Figure, the flavor 1120 shows the best results with a moderate uncertainty, scoring between 15% and 40%.

The additional dots in the Figure represent the "classical" calibration $\mathbf{C}_0$ of the period 1902–1980 (with validation 1854–1901) used in previous studies. $\mathbf{C}_0$ obviously assumes the role of an outlier, in a positive sense for *RE* and in a negative one for *CE*. While *RE* values approach 60% (for 1120) the *CE* values are negative throughout. It

appears that in the same sense that that particular calibration rewards trended predictors with high *RE* values it penalizes them with small *CE* scores. In other words: the trend (which is "invisible" to *CE*) dominates calibration and validation.

Note that the RegEM variant 0130 only scores 30% for $\mathbf{C}_0$. This method, whose millennial performance is assessed here for the first time, was advertised by Mann et al. (2005) and earlier to replace the original MBH98 method 1011, which scores almost 50% in our emulation. Hence, even if $\mathbf{C}_0$ did not reflect a sampling bias, our results do not suggest a transition to that variant.

The nonsense predictor mentioned above (number of available grid points) scores *RE*=46% (and *CE*=−23%), which is more than any of the flavors ever approaches in the 100 random samples. And it is not unlikely that other nonsense predictors score even higher. On this background, the originally reported 51% of verification *RE* are hardly significant. This has already been claimed by McIntyre and McKitrick (2005b) in a slightly different context. In addition to that study, we have derived more stable estimates of verifiable scores for a whole series of model variants, the optimum of which (1120) scoring with *RE*=25%±7% (90% confidence).

Note that such a random calibration set **C** very likely destroys the original temporal ordering of the **P** and **T** series (albeit synchronously for both), along with the observed 20th century warming trend. To someone more used to dynamical models (which contain the time variable explicitly) this "shuffling" may appear irritating as it would destroy the main "physical process" that one attempts to reflect. We therefore emphasize that empirical models of this kind do in no way contain or reflect dynamical processes other than can be sampled in instantaneous covariations between variables. The trend may be an integral part of such a model, but only as long as it represents these covariations.

One might nevertheless attempt to "help the sampling" by picking only those **C** that preserve contiguous time spans of a length typical for **P–T** interactions, say 5 years. We have tested for 1, 5, and 10 years, but did not observe significant changes to Fig. 1 apart from a slight decrease of skill and increase of spread (see supplement). For longer time scales the diminishing degrees of freedom is a limiting factor.



CPD
2, 357–370, 2006



EGU

## 5 Conclusions

Previous estimates of climate reconstruction skill, especially *RE*, are founded on the particular partitioning into calibrating and validating portions of the trended instrumental period, and thus mainly reflect sampling properties. This leaves very few degrees of freedom, and they can easily be matched by nonsense regressors. To accommodate for this sampling bias we have proposed a strategy that is based on repeated resampling of the instrumental period, similar to other techniques not unusual in statistical estimation theory (cf. Efron and Gong, 1983).

The results pose a number of questions. (1): Are the results representative, i.e. are 100 experiments per flavor enough to estimate the uncertainty? Given the huge amount of possible permutations of 127 years the number of 100 experiments is quite small. On the other hand, if there is sense at all behind the idea to distill an empirical model out of the 127 proxy and temperature records, sampling 100 is probably sufficient. The 1902–1980 calibration, as an outlier, is very hard to "sample" randomly. – (2): Are we in a position to advertise a "best" flavor? The flavor 1120 – EOF truncated predictand, RegEM, TT regression, and no rescaling – with an *RE* of 25%±7% shows the highest scores; but other scores (1011, 1101) are well within the uncertainty bound. – (3): Are 25% *RE* enough to decide the millennial NHT controversy? This is the crucial question. 25% *RE* translates to an amplitude error of $\sqrt{}$(100–*RE*) ~85%. If one were to focus the controversy into the single question: Was there a Medieval Warm Period (MWP) and was it possibly warmer than recent decades? – we doubt that question can be decided based on current reconstructions alone.

*Acknowledgements.* This work was funded by the EU project SOAP.

## References

Briffa, K. R., Jones, P. D., Pilcher, J. R., and Hughes, M. K.: Reconstructing summer temperatures in northern Fennoscandinavia back to A.D.1700 using tree ring data from Scots Pine,

Arctic Alpine Res., 20, 385–394, 1988.
Bürger, G. and Cubasch, U.: Are multiproxy climate reconstructions robust?, Geophys. Res. Lett., 32, L23711, doi:10.1029/2005GL0241550, 2005.
Bürger, G., Fast, I., and Cubasch, U.: Climate reconstruction by regression – 32 variations on a theme, Tellus A, 58(2), 227–235, doi:10.1111/j.1600-0870.2006.00164.x, 2005.
Cook, E. R. and Kairiukstis, L. A.: Methods of dendrochronology: applications in the environmental sciences, Kluwer Acad. Publ., 394 pp., 1990.
Cook, E. R., Briffa, K. R., and Jones, P. D.: Spatial regression methods in dendroclimatology: a review and comparison of two techniques, Int. J. Climate, 14, 379–402, 1994.
Dempster, A., Laird, N., and Rubin, D.: Maximum likelihood estimationfrom incomplete data via the EM algorithmm, J. Royal Stat. Soc. B, 39, 1–38, 1977.
Efron, B. and Gong, G.: A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation, Am. Stat., 37(1), 36–48, 1983.
Fierro, R. D., Golub, G. H., Hansen, P. C., and O'Leary, D. P.: Regularization by Truncated Total Least Squares, SIAM J. Sci. Computing, 18(4), 1223–1241, 1997.
Hoerl, A. E.: Application of ridge analysis to regression problems, Chem. Eng. Prog., 58, 54–59, 1962.
Huybers, P.: Comment on "Hockey sticks, principal components, and spurious significance", Geophys. Res. Lett., 32, L20705, doi:10.1029/2005GL023395, 2005.
Jones, P. D. and Briffa, K. R.: Global surface air temperature variations during the twentieth century: Part 1, spatial, temporal and seasonal details, The Holocene, 2, 165–179, 1992.
Legates, D. R. and McCabe, G. J.: Evaluating the use of "goodness of fit" measures in hydrologic and hydroclimatic model validation, Water Resour. Res., 35, 233–241, 1999.
Lorenz, E. N.: Empirical orthogonal functions and statistical weather prediction, Sci. Rept. No. 1, Dept. of Meteorol., M. I. T., 49 pp., 1956.
Mann, M. E. and Rutherford, S.: Climate reconstruction using "Pseudoproxies", Geophys. Res. Lett., 29(10), 139, 2002.
Mann, M. E., Bradley, R. S., and Hughes, M. K.: Global-scale temperature patterns and climate forcing over the past six centuries, Nature, 392, 779–787, 1998.
Mann, M. E., Bradley, R. S., and Hughes, M. K.: Northern Hemisphere temperatures during the past millennium: Inferences, uncertainties, and limitations, Geophys. Res. Lett., 26(6), 759–762, 1999.
Mann, M. E., Rutherford, S., Wahl, E., and Ammann, C.: Testing the Fidelity of Methods Used

in Proxy-Based Reconstructions of Past Climate, J. Climate, 18, 4097–4107, 2005.
McIntyre, S. and McKitrick, R.: Corrections to the Mann et al. (1998) Proxy Data Base and Northern Hemispheric Average Temperature Series, Energy Environ., 14(6), 751–771, 2003.
McIntyre, S. and McKitrick, R.: Hockey Sticks, Principal Components and Spurious Significance, Geophys. Res. Lett., 32(3), L03710, 2005.
McIntyre, S. and McKitrick, R.: The M&M Critique of the MBH98 Northern Hemisphere Climate Index: Update and Implications, Energy Environ., 16(1), 69–100, 2005.
Murphy, A. H. and Winkler, R. L.: A general framework for forecast verification, Mon. Wea. Rev., 115, 1330–1338, 1987.
Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models – Part I – A discussion of principles, J. Hydrol., (10), 3, 282–290, 1970.
Rutherford, S. and Mann, M. E.: Climate Field Reconstruction under Stationary and Nonstationary Forcing, J. Climate, 16, 462–479, 2003.
Rutherford, S., Mann, M. E., Osborn, T. J., Bradley, R. S., Briffa, K. R., Hughes, M. K., Jones, P. D.: Northern Hemisphere Surface Temperature Reconstructions: Sensitivity to Methodology, Predictor Network, Target Season and Target Domain, J. Climate, 18, 2308–2329, 2005.
Schneider, T.: Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values, J. Climate, 14, 853–871, 2001.
von Storch, H., Zorita, E., Jones, J. M., Dmitriev, Y., and Tett, S. F. B.: Reconstructing Past Climate from Noisy Data, Science, 306, 679–682, 2004.
Wahl, E. R. and Ammann, C. M.: Robustness of the Mann, Bradley, Hughes reconstruction of Northern hemisphere surface temperatures: Examination of criticisms based on the nature and processing of proxy climate evidence, Climatic Change, in press, 2006.
Wilks, D. S.: Statistical Methods in the Atmospheric Sciences. An Introduction. Academic Press, San Diego, 467 pp., 1995.
Wolock, D. M. and McCabe, G. J.: Explaining spatial variability in mean annual runoff in the conterminous United States, Climate Res., 11, 149–159, 1999.

**Table 1.** The 3×2×4×2 = 48 regression flavors.

| | *GLB* | *COV* | *MDL* | *RSC* |
|---|---|---|---|---|
| 0 | 219 grid points | conventional | LS | no rescaling |
| 1 | 1 EOF | RegEM | inverse | rescaling |
| 2 | global average | | TT | |
| 3 | | | RR | |

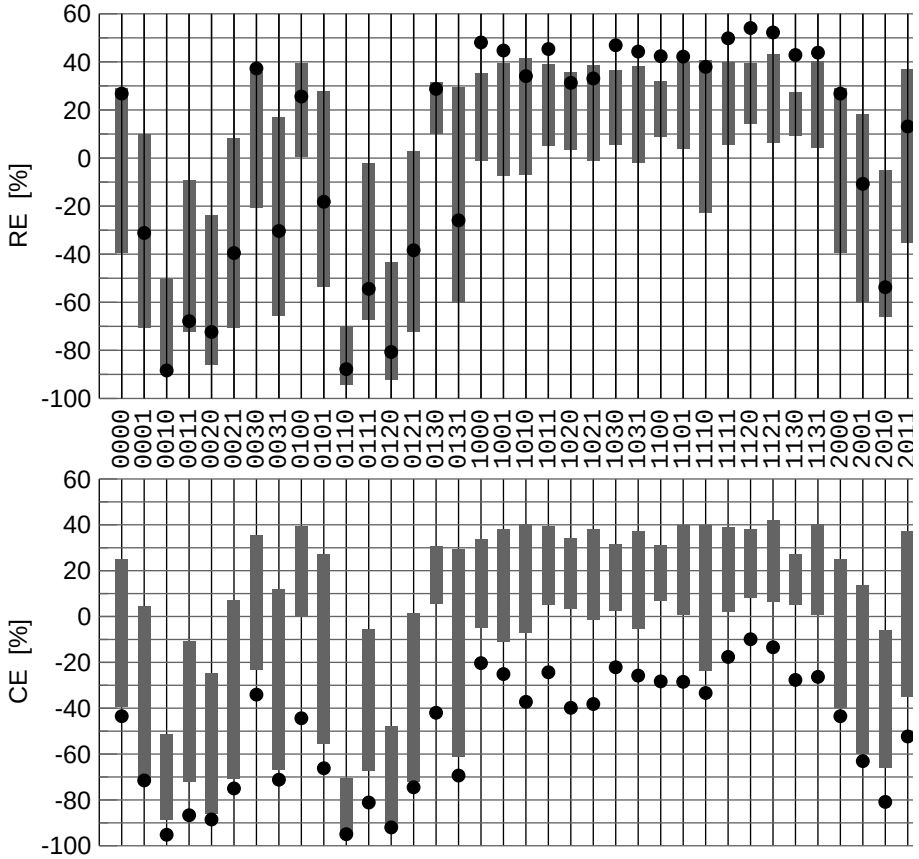


**Fig. 1.** The verification scores *RE* and *CE* of all 48 flavors. A bar indicates the spread of all 100 experiments using randomly resampled calibrations, a black dot the experiment with 1902–1980 calibration.



CPD
2, 357–370, 2006



EGU