Clim. Past Discuss., 3, S151–S157, 2007
www.clim-past-discuss.net/3/S151/2007/

Climate of the Past Discussions

Interactive comment on "On the verification of climate reconstructions" *by* G. Bürger

G. Bürger

Received and published: 7 April 2007

This paper offers a rather critical view on current climate reconstructions of hemispheric or global scale. Its point of view had been very controversially debated in a first, and rejected, version (cpd/2006-2-357). Although I was quite unhappy with the review process then, I gradually began to see where the presentation could actually be improved, leading to this second version. This time I enjoyed the review process, which still raised a number of critical points, but in a purely scientific manner.

Verification of climate reconstructions bears a sizable intricacy and is loaded with non-scientific reverberation at the same time. It is thus prone to heavy debate. In that situation, a critical inspection of the methods is mandatory, to which this paper tries to contribute.

Following are some final comments on the reviews. What is not mentioned here will be handled appropriately in a revision.

Referee 1 (J. Guiot)

All comments prove a thorough reading and understanding of the study, and are very welcome! They will be considered accordingly. The following items deserve special attention.

p.252. l14: I was unaware of the peculiarities of sampling with/-out replacement, etc. I had used the most straightforward approach of selecting random samples of fixed size (here n/2) and verifying with the complement. This is closer to single and double crossvalidation (CV), and also to the original multi CV as introduced by Geisser 1975 or Krus and Faller 1982. Repeating the analysis using proper bootstrapping is an interesting idea, but infeasible for this study in a realistic timeframe, since already the single regression estimates take too long, esp. those using RegEM.

p.255, l.15: The equation $CE \leq R_c^2$ is in fact incorrect. But the main conclusion, that skill-less predictions with $R_c = 0$ must have $CE \leq 0$, follows directly from Eq. 2.

p.258, l.13: Yes, Fig. 3 depicts, in x and y, the average of 500 experiments, each one resulting from 100 sequential swappings of the original order.

p.263, l.23-24: Using a single NHT PC is probably not the optimum, but required for the MBH98 emulation.

p.265, l.11: The “allowance for nonsense regressors” certainly needs clarification. The significance of RE for simulating mean NHT is derived, not from the null model of a 1-dim red noise process but from the null model of 22 red noise processes *regressed on NHT*. This allows for memory effects to inflate the verification skill.

p.266: If I‘m not mistaken, that argument is applicable only if the RE distribution is obtained by shifting a null distribution that is centered about 0 to the mean RE value. But looking at Fig. 6 that is not the case here. First, the null distribution is not centered at 0, that is, nonsense regressors achieve positive RE skill, and second, that distribution is more dispersive. It is unlikely that the nonsense regressors share *all* characteristics

CPD
3, S151–S157, 2007

with its real counterparts (how could they?). For any method to be skillful (and for any skill score to be useful) it is essential of not being outperformed by random-based, a priori skill-less methods.

p.269: Whether or not the MWP has a global or only a regional signature cannot be decided - in a verifiable sense - on the basis of these proxies and methodologies. For this, a *substantial* increase of skill is required.

Referee 2

As a small addition to my interactive comment, I think much of the confusion was caused by me trying to avoid an overly formalistic style. I just don't like it. In this case it had the unfortunate consequence of words like "skill" becoming ambiguous, which is surely bad style.

For further discussions on "skill" see below.

Referee 3 (S. McIntyre; review to be added to CPD)

Steve McIntyre is known to be a rather critical mind, for which I have much sympathy. Accordingly, I am quite critical with his review. A very good setting for a lively debate at CPD, if not his review came in so late (it is available at *www.climateaudit.org/?p=1302*). Steve formulates the following criticisms:

- The use of multi CV (which is Steve's "multicalibration")

- The under/mis-representation of spurious relationships.

- The use of "model skill"

- The inclusion of the North American (NA) PC1.

"multi CV" - I do not know how one can contest multi CV as a valid and well established verification method (compare, e.g., the comments of J. Guiot, as well as Steve's own comment "252 10"). Being now more than 3 decades old (Geisser 1975: *The predictive*

CPD

3, S151–S157, 2007

sample reuse method with applications) multi CV is obviously not a novel application. But what is more important, it is based on **fewer** assumptions than simple or double CV (with their dependence on specific calibration sets). Hence, if those can be applied to climate reconstructions, even more so can multi CV.

Noone, including myself, has ever claimed to have found the "magic bullet" of climate reconstruction assessment. Steve presents Yule's 1926 classic example of spurious relationships - it is, as much as I know, between marriages and *mortality*, not alcoholism - as a counterexample to my Fig. 3, suggesting that the approach taken in the study fails in this case. But that is not the purpose of Fig. 3. That Figure should, instead, demonstrate the effect of *temporal separation* in a verification context, for scores like RE and CE. That the Yule data have $RE \geq 80\%$, no matter what calibration set, is not a sign of very good quality, say on a scale between 0%=miserable to 100%=perfect. It just means that there simply is no further artificial skill from temporal separation. For my trivial model (the original Fig. 3), on the other hand, that effect accounted for half of the attained RE score, the other half being ... "spurious"?

"spurious" - I fully agree with Steve that spurious relationships should generally be a concern, and econometrics provides a number of alarming examples. Blind application of standard tests (such as Students t-test or Fishers F) often produce false results, for example, in the presence of nonstationarity and/or autocorrelated residuals. But for climate records, at least those considered here, that is not really an issue. Among the 22 proxies are merely two that reveal signs of nonstationarity (St. Anne river and the NA PC1, by using a unit root test); and regressing temperature on the proxies produces no significant autocorrelated residuals (Durbin-Watson statistic of about 1.9). The exact numbers depend of course on the chosen calibration set.

But can't we put it much simpler? Isn't limited sample size, aggravated by trends, the main issue here, in full accordance with Yule 1926? That easily creates artificial RE skill, one half by way of temporal separation, the other by degrading the independence of the validation set.

CPD
3, S151–S157, 2007

EGU

But be there nonstationary proxies and autocorrelated residuals or not, none of the problematic significance tests has been applied in this study. They were explicitely avoided in favor of a more rigorous Monte Carlo approach, covered in section 6. That section, forming the basis of any of the significance statements, somehow went unnoticed by Steve. A complete verification assessment (in the sense of this study) of his "counterexample" would require another set of Monte Carlo simulations, with memory parameters estimated from the Yule data. And I am confident that the resulting skill would be insignificant. Since Steve is otherwise a careful reader it is probably on my side to clarify the purpose of that section.

""model skill"" - I finally began to understand why Steve persistently uses quotes in referring to "model skill". It looks as if he's guided by the - fully comprehensible - quest for the ideal score, one that shows -1 (or 0, or whatever) for something miserable and 1 for perfect. But that score does not exist, not for climate reconstructions, not for weather forecasting (see, e.g. the excellent article by A. Murphy, 1996: *The Finley Affair: A Signal Event in the History of Forecast Verification*). One is left with "some" score whose mapping to "bad" and "good" has to be explored ex post, for example, by utilizing models of a priori known quality such as random-based forecasts.

"NA PC1" - According to Steve, one "needs to establish that the model is even valid", alluding to my inclusion of the NA PC1 proxy. First of all, the purpose of this study was to assess verification scores of *existing* models and compare the results with previous estimates. It was not the purpose to modify or improve any of the models. Having said that, it is of course important to establish a model's validity and, e.g., discard "invalid" predictors. But what exactly is an invalid predictor? - Based solely on statistical criteria, such as Durbin-Watson, the NA PC1 cannot be discarded so easily. And if that proxy serves to improve "model skill", I doubt it will ever be dicarded only because its definition relied on an unusual application of PCA.

In this situation, a full-fledged, best-subset multiple regression analysis using all available proxies (in the sense of, e.g., Chapter 12 of Seber and Lee 2003) is perhaps

CPD
3, S151–S157, 2007

EGU

desirable. But given the large number of proxies vs. the small number of cases (observations) this is a very difficult, maybe unrealistic undertaking. Moreover, a deeper physico-biological understanding is always a better guide than any data mining.

Minor points:

Of Steve's minor points, the following need further commenting; the others are evident and will be considered accordingly.

254 14: The SS_{clim} of Wilks 1995 is arithmetically identical to CE, not RE. "Climatology" is then the climatology of the validation period.

254 15-17: Fritts: "...differences of the actual data from the mean of the dependent data set used for calibration." - Briffa: "... with values assumed to be equal to the calibration period mean."

254ff: Eqs. (2) and (3) are rather handy in explaining the different behavior of the scores R_c, RE and CE. And that $CE \leq R_c$ is definitely needed for the later argument that RE and CE are positively biased in finite samples with memory (section 6, and Fig. 6).

256 7: Supposedly there was a gradual enlightenment with regard to the shrinkage of predictive skill, starting in fact as early as Larson 1931. But at that time no clear distinction existed between population *validity* (the adjusted R^2) and *cross-validity* (the R_c^2). Maybe this should be mentioned in the text.

256 15-20: I could not find any script.

257 1-16: As mentioned above, this is all about artificial verification RE through temporal separation. That idea is crucial, as it explains about half of the arificial skill of the reconstructions.

257 18 (not 13): discrepancy between RE and CE, see also comment above.

265 20-25: As mentioned above, misspecification of a proxy model is not easily estab-

CPD
3, S151–S157, 2007

EGU

lished empirically.

266 13: Yes, Huybers reports a critical value of $RE = 0\%$. As mentioned in the text, for $RE = 36\%$ I used his script “Monte_Carlo_RE.m” from the supplement.

266 14: MM05b say: “We did new simulations in which we took 1000 simulated PC1s saved from the simulations described by MM05; for each PC1 in turn, we made a ’proxy network’ of 22 series with the other 21 being white noise (replicating the 22 series of the MBH98 AD1400 network).” - It is important to know here whether the 21 white noise series are mutually correlated or not, in other words, how the structure of the full 22-dim predictor covariance matrix is determined.

266 21: For the benchmarking, the main question is: Given a model $T = f(P)$, is it necessary to reflect for the noise model as many properties of f and P as possible (which is what I meant by “convergence”), and replace P with appropriate noise N to apply $T = f(N)$? - This approach represents the question: Can a given reconstruction such as MBH98 be distinguished, in terms of skill, from fitting noise? - Or is it, as in my (and supposedly also MM05’s) view, only necessary to provide *lower bounds* on the significance level of skill. This approach represents the question: Is there any chance that the reconstruction is outperformed by a random-based (nonsense) reconstruction?

267 24: Significance testing is done, see Fig. 7.

269 10-15: This is discussed earlier, see 252 25.

Interactive comment on Clim. Past Discuss., 3, 249, 2007.