Clim. Past Discuss., 7, C1817–C1827, 2011
www.clim-past-discuss.net/7/C1817/2011/

Climate
of the Past
Discussions

Interactive comment on “Benchmarking monthly homogenization algorithms” *by* V. K. C. Venema et al.

V. K. C. Venema et al.

victor.venema@uni-bonn.de

Received and published: 11 November 2011

All suggestions that were simply implemented are not listed below.

Referee #2: “I apologise for the large number of comments but they are mostly very minor in nature.”

No need to apologise. Thank you very much for the effort you put into the manuscript and your valuable suggestions. They have improved the manuscript substantially.

Referee #2: “Specific Comments: Abstract line 26: “Training” training in what? Use of the benchmarks? Building algorithms?”

We have clarified: “Training the users of homogenization software was found to be very important.”

Referee #2: “P 2664 line19+: More information is needed here. Why were those networks so poorly homogenised? What errors were added? We can learn from this if you provide sufficient information and justification for not including them. If they are realistic inhomogeneities then they could be useful for development of improved algorithms to some extent. Also, what is meant by “Selecting stronger did not changes the results anymore.””

See also our longer reply to reviewer #1 on this paragraph. We have made this section longer and now explain: “During the analysis it was found that some of the input data was not homogenized well enough. Consequently, the (long-term) variability of some difference time series in these networks is artificially too strong. The algorithm used to produce the surrogate networks is able and has reproduced this (long-term) variability, which the homogenization algorithms may interpret as inhomogeneities. Consequently, these networks had to be removed and only the best 15 surrogate networks were used in the analysis. Selecting stronger did not change the validation metrics anymore. For the comparison of surrogate and synthetic data, a new dataset was generated using only well homogenized input networks; see Sect. 6.3.1.”

Thus the problem was that some surrogate networks were inhomogeneous before the errors were added.

Referee #2: “P2665 line 1+: Can you say more about the real data. Why were these stations chosen? Are they well studied and well documented? Are there references you can provide relating to them? Can the reader have access to them? Are you 100% all inhomogeneities within these series are ’known’? Are they monthly or have you averaged them? Have they undergone quality control by you or by the National Meteorological Service?”

These networks were chosen as examples of normal data from Europe, from a range of different climates. Some are well documented and include meta-data, which, as always, may be incomplete; some networks only contain monthly data without additional information, just as is often the case. The data is available for download as all other data, was send to us in monthly resolution and were quality controlled. Complete networks for which one is 100% sure that all inhomogeneities are known do not exist. Thus the reason for this dataset was not to provide a dataset with known inhomogeneities; the surrogate and synthetic data sections were included for that purpose. They were included to allow an intercomparison on the most realistic data (which has not been performed yet) and to be able to compare the statistical properties of the detected breaks, which is performed in Section 6.3.2. For future benchmarking exercises it would be advantageous if the inhomogeneous networks of the real data section would have a surrogate counterpart with exactly the same network pattern (cross correlations) and missing data. This would allow for a even more accurate comparison of the realism of the benchmark data.

We have added: "The real data section contains inhomogeneous datasets from various European climates and aims to contain examples of normal Europe datasets."

Referee #2: "Sections 2.2 and 2.3: Are there any limitations of the IAAFT method? Perhaps not. It would be really nice to see an example of both a surrogate and a synthetic time series. Are there features of real data that were not well simulated i.e., ENSO, large volcanic eruptions such as Pinatubo etc."

The limitations of the IAAFT method are described in some more detail in the report on the benchmark dataset (Venema et al., 2011) and in Venema et al. (2006a). The method does not reproduce jumps in the data, it does reproduce the variability on all scales due to this jump, but the localised small-scale variability of a jump would be spread as small-scale variability over the entire surrogate time series. For the benchmark this may even be an advantage, for other applications it is a vice. More generally the method does not reproduce perfectly the intermittence of a measurement in the sense of variability of the variability. In case of ENSO it would reproduce its decadal variability, but not its nonlinear dynamical behaviour. Thus applying a nonlinear pre-

diction algorithm may work better on a real measurement dominated by ENSO as its surrogate. The temperature drop due to a volcanic eruption would be reproduced (via the temperature distribution). The date of the eruption would be random in the standard algorithm, however.

Referee #2: "P2666 line 25: Can you reword the "frequency is drawn from a uniform distribution between 2 and 8%" as I'm not clear what is meant by this."

We had written: "To vary the quality of the data on a station by station basis, this frequency is drawn from a uniform distribution between 2 and 8\,{\%. The break events are independent of each other (Poisson process)."

We have detailed this statement, it now reads: "To vary the quality of the data on a station by station basis, the average break frequency for a station is first drawn from a uniform distribution between 2 and 8\,{\%. The actual break events themselves are drawn with this frequency and are independent of each other (Poisson process)."

Referee #2: "P2667 line 10: Can you provide more information on the seasonal cycle of the breaks added? I wasn't clear on how this was done. Does the seasonal cycle vary from break to break in that sometimes a winter break will be larger than in summer and sometimes it will be smaller? Could breaks be of a different sign in winter compared to summer? This also applies to P2685 lines 6+."

The requested information is in the report on the benchmark dataset (Venema et al., 2011). Because the manuscript is already so long and almost any method is likely okay, we had chosen not to describe this detail in the manuscript. We have now added the sentence: "The seasonal perturbations are computed by smoothing white noise and, if needed, shifting one of its extremes to the summer period."

We start with 12 white noise values, these are smoothed (periodically) and normalised to zero mean and unity standard deviation. If the maximum or minimum is not in the summer, it is moved to the beginning or ending of the summer. Then the random multiplication factor for the standard deviation (seasonal cycle) is applied and the random constant for the mean break is added.

Thus the seasonal cycle can have a different sign in summer and winter and the sign in summer also varies randomly.

In the analysis of the seasonal cycle of the homogenization perturbations (P2685) it is thus possible that a homogeneous subperiod (HSP) has a seasonal cycle of 1°C with a maximum in summer and the next HSP is of the same size, but has its maximum in winter. In that case, this would be counted as no change in the size of the seasonal cycle. Another definition would have been possible, that is also sensitive to such a change in the phase, but for a comparison of the two data sections, this would probably not make any difference.

Referee #2: “P2668 line 0: What about random missing data? Many time series have missing months dotted throughout the series.”

Thank you for reminding us. This was something I had wanted to write about, but forgot. We have now added to Section 7.2: “A few remaining outliers were found to have little influence on homogenization; a future dataset could do without outliers. The periods with much missing data clearly made homogenization more difficult; in future also inserting random missing data may thus be interesting and enhance the realism of the benchmark.”

Referee #2: “Section 4: You define “relative homogenisation” here. Please can you define “absolute homogenisation” here too as it is referred to in later sections.”

After the sentence on relative homogenisation we have added: “One absolute homogenization algorithm is employed, in this case only the station time series itself is used for homogenization.”

Referee #2: “P2671 line 19: Does “reconstitution of missing data” mean infilling of missing data?”

The sentence: “Since some methods do not perform reconstitution of missing data” has been changed to: “Since some methods do not fill data gaps”.

Referee #2: “P2672 line 10: How do you obtain the values for “true negatives”? Do you treat each month/year without a break as a potential one or do you assume that breaks could occur with a certain frequency e.g., every 6 months. Also, how do you deal with location of a break 1 month too late or early? Is there any margin of error in the assignment?”

The data is analysed at the yearly scale. If there is a break in any month in a year, this year is considered to contain a break. True negative thus means that the homogenization algorithm did not find a break in a year and that there was no break inserted in that year.

It is customary in homogenisation to allow for a margin of error in the date of the break. One can easily do so for the true positives (hits) and thus increase the hit rate for algorithms that did find the break, but not exactly at the right date.

To increase this margin was one of the reasons to analyse the data at a yearly scale. Still, even at a yearly scale, it may happen that predicted break was one month off and thus considered to be a miss. This may be considered to be unfair. On the other hand, in an intercomparison study all methods will have this additional difficulty and it does not seem to be too unfair to punish methods, which do not predict the date accurately.

Furthermore, for the other terms (false positives, false negatives and true negatives) using definitions with a margin of error is problematic and especially so for the skill scores, which are based on them. We were unable to come up with definitions for the skill scores in which the number of data points is still equal to the number of hits, false alarms, misses and true negatives.

Referee #2: “Section 6.1.4: Do you think the decrease in CRMSE from older to more recent data in the time series is anything to do with the way in which corrections are applied? Are older chunks of inhomogeneous data adjusted to match all of the more recent data or just the most recent homogeneous chunk? If there is a background trend in the data this may also have an effect. This perhaps links in with the last sentence “This fits to Climatol stating the correction of the breaks at the beginning of the series”. I’m not quite sure what is meant by this.”

Yes, the decrease is related to the way the corrections are applied for many algorithms. In algorithms in which one HSP after another is corrected errors can accumulate. An advantage of the ANOVA-type correction method used by PRODIGE is that this accumulation is avoided. We have removed the sentence on Climatol. The main factor is, however, most likely the missing data in the middle of the dataset and in the first quarter of the dataset. Looking at the temporal behaviour of the individual contributions, 1945 and 1925 are often points where the errors jump or start to grow.

In Section 7.1 we wrote previously: “Most, but not all contributions, showed much larger errors in the beginning quarter or half of the century. This may point to possibilities for developers of homogenization algorithms to improve the handling of missing data and of networks with few stations.”

The temporal behaviour of the errors in now discussed in more detail: “Most, but not all contributions, showed much larger errors in the beginning quarter or half of the century. Partially this is unavoidable due to the sparser density of the networks for the earlier periods. Consequently, detection of the changes is less precise, consequently also the corrections. These errors may also point to possibilities for developers of homogenization algorithms to improve the handling of missing data and of networks with few stations. Another reason may be that most algorithms perform no corrections for the more recent period and compute break sizes from one homogeneous subperiod to the next, which may lead to an accumulation of errors.”

Referee #2: “P2679 line 18: What is a “predicted break”? Are these the same as detected breaks?”

Yes, in principle. The problem with the word “detected break” is that it suggest that the detection was correct (a hit), the word “predicted break” does not carry this connotation. Therefore, we have chosen to use the word “predicted break” in the parts about contingency scores, as is common in detection theory.

Referee #2: “P2680 lines 11-13: I’m not quite sure what is meant by this sentence.”

It was hard to explain the pre-homogenization of ACMANT with just one sentence: “A pre-homogenization is applied in which to avoid biases a candidate that used certain reference stations is not reversely used as reference for those stations.”

If station B is used to compute the reference to homogenize station A in one step and later station A is used to compute the reference for station B, this may lead to biases and this was one reason for the modest performance of the old blind ACMANT version. Therefore, ACMANT late performs a pre-homogenization for the largest breaks in which such pairs are always used in only one direction.

We hope this formulation, first stating the problem, is clearer: “Using station B to compute the reference for station A and later station A to compute the reference for station B can lead to biased results. Therefore, a pre-homogenization for large breaks is applied in which this is forbidden.”

Referee #2: “P2686 line 3-4: Are cross-correlations the percentage of variance in one time series that is explained by another time series? I wasn’t quite sure what metric should be used here.”

These two measures are related, but not the same, which can already be seen by the negative cross correlations. The cross-correlation is the cross covariance (between x and y) divided by the standard deviations of x and y. Thus the cross-correlation can not only tell how much variance of one time series is explained by the other, but also whether when one goes up, the other also tends to go up (positive cross-correlation) or tends to go down (negative cross-correlation). The latter is also an advantage of this measure.

Referee #2: “Section 7.1: Do you have any ideas of how to improve/enable true objective intercomparison of algorithms in future research? This is a really valuable assessment but some guidance on how to progress given the lessons you have learned from this research would be really helpful.”

This is not discussed in Section 7.1, but in Sections 7.2 (including the above mentioned addition) and 8. Is there anything missing in these sections you would like to hear more about?

Referee #2: “P2687 line 19: How do you define the “all-over best”?”

Subjectively. Averaging the rankings over all computed measures (best without the contingency scores) would do the trick. Normalising all measures and averaging would also work. The difference between the mentioned best algorithms and the rest is quite consistent, thus almost any averaging method would work.

Referee #2: “P2693 line 15: I disagree with this statement and it undervalues the very considerable and useful work that you have done. I think you have shown great value in the benchmarks both for validation of algorithms and also for development although as you say, these should be separate components as you cannot validate with a benchmark that has been used to tune an algorithm. A cyclical program would be useful and you do discuss creation of more benchmarks every few years. Something like: design benchmarks, invite blind testing, release results and then allow development using old benchmarks, design new benchmarks, invite blind testing etc. etc.”

We fully agree with you. The statement “benchmarking does not allow for systematic studies” was not intended to devaluate our study. It only pleads for a diversity of approaches. In future we should not only be performing “benchmarking studies”, but rather perform a mix of benchmarking studies with other validation studies, as all methodologies have their advantages and disadvantages.

We now explicitly state so: “Given that every methodology has its own advantages and disadvantages, we expect that progress is best served by a diversity of methodologies. Benchmarking is important for its ability to obtain reliable accuracy metrics, due to the blind testing of the contributions and the realism of the data.”

Referee #2: “P2693 line 21+: This is a really useful summary of recommendations. Can you structure this more explicitly as recommendations to the community, perhaps using bullet points. I think this would make this stand out much more to the reader.”

A subsection “Recommendations” has been added. Because the recommendations are mixed with explanations and disclaimers, we feel that bullets would not improve legibility.

Referee #2: “P2694 line 29+: Why should networks without added inhomogeneities be studied separately?”

This is an important network to analyse as it would show what homogenization algorithm do with data which does not have any inhomogeneities. While in the other networks we can also study the false alarm rate, this false alarm rate is also partially due to algorithms finding other solutions. If there is one break in 1949 and one in 1951, both in the same direction and the algorithm finds it in 1950, this is a false alarm. If a small break is detected, but the date of the break is a bit off, it is a false alarm. In this network with out inhomogeneities you can study the false alarm rate of the algorithms very well. Additionally, it is not only about detecting, but also about how large the perturbations are in case no perturbations should be added. Such perturbations should be as small as possible. This can probably only be studied on such a network.

We have changed the sentence to: “For instance, in networks without breaks homogenization algorithms should change as little as possible, this can be studied in the network without inserted inhomogeneities.”

Referee #2: “Gradual inhomogeneities are discussed in the input errors to simulate urbanisation or a growing environmental feature. However, the ability of the algorithms to find such inhomogeneities is not discussed.”

That is a true omission. Thank you. We have added the sentence: ”How well the gradual local trends are removed by homogenization would warrant a dedicated study, as well.”

Referee #2: ‘Table 1: Please can you define “DP”, “HBS” and “MLR” in the table header?’

DP = Dynamic programming (optimization method)

HBS = (semi-)hierarchic binairy splitting

MLR = Maximum Likelihood Ratio test

Interactive comment on Clim. Past Discuss., 7, 2655, 2011.