16th EMS / 11th ECAC (Trieste, September 12-16, 2016)

Homogenization of daily series: Suitability of additive and multiplicative models and experiences with Climatol 3.0

José A. Guijarro <jguijarrop@aemet.es>

(State Meteorological Agency (AEMET), Balearic Islands Office, Spain)

Introduction

The need to homogenize observational series before its use to assess climate variability emerged long time ago. Efforts were initially focused on annual, seasonal and monthly series, and the successful COST Action ES0601 allowed the exchange of ideas between homogenization specialists and the improvement of the their methodologies. But now the stress is put on the homogenization of daily series, since the study of the variability of indices and extreme values () depends on them.

The Climatol package, now in its version 3.0, works on *normalized* series, where the user can choose three types of *normalization*:

1) Center the series by removing its mean value.
2) Divide all terms of the series by its mean value (unless < 0).
3) Standardize the series by removing its mean value and dividing by its standard deviation. (The default option).

When all series have been normalized by one of these means, any two series y_i and x_i may be related by the simple *Major Reduced Axis* regression model $y_i = x_i + \varepsilon_i$ or, neglecting the error terms, $\hat{y}_i = x_i$, where x_i may be a composite reference from a prescribed number of nearby data.

This approach was taken form the normal-ratio method applied by Paulhus and Kohler (1952) for infilling daily data to avoid too many blanks in the publication of precipitation reports. They used the average of three nearby observations normalized by their annual mean (case 2 above), but this ratio method can be generalized to differences (case 1) and full standardization (case 3).

This procedure allows great flexibility to use nearby data to infill missing data in a problem series, since no common period of observation is needed between the two, and the closest reference data can be chosen in every time step adapting to the different availability of data in the other series. Another important advantage is that the reconstructed series do not have their variance reduced, a drawback of the OLS regression that invalidates its application to daily data where higher order moments need also to be properly estimated (Szentimrey, 2013).

Its main drawback is the estimation of the right means (and standard deviations in the default normalization) of the series when they have missing data, which is normally the case. This problem is solved in the Climatol package by computing initial values with the available data in every series, infilling the missing data, recomputing their means (and standard deviations), and so on, until the maximum difference between the last and the previous means lies below a prescribed threshold.

This normalization method connects also with the methods of differences and ratios for the reduction of climatological averages to a certain period which Conrad and Pollack (1950) apply to temperatures T and precipitations P respectively. If we denote as T or P the averages that we want to compute for a chosen reference period (normals) and as T' or P' the averages of our observations spanning a different period, and we have a reference series R with complete observations in both periods, then we can calculate our normals by:

$$T - T' = R - R' \quad \Rightarrow \quad T = T' + R - R' \qquad \text{(Method of differences)}$$

$$\frac{P}{P'} = \frac{R}{R'} \quad \Rightarrow \quad P = P' \frac{R}{R'} \qquad \text{(Method of ratios)}$$

Conrad and Pollack only favor the use of the ratios method for precipitation, while there is a common tendency to apply it to other strongly biased variables with a natural zero limit (as wind) for which a multiplicative model would be more suitable than the additive model normally applied to temperature and other variables with a near normal distribution (Szentimrey, 2014).

But we can question ourselves about the applicability of these models to the relationship between a problem series Y_i and a reference X_i. Which fits better, $\hat{Y} = a + X$ or $\hat{Y} = b \cdot X$? In fact these are simplifications of the general expression for a linear regression $\hat{Y} = a + b \cdot X$, where $b = 1$ in the additive relationship and $a = 0$ in the multiplicative, and therefore we can adjust

this model to several climatological variables and inspect the values of the regression coefficients to assess the suitability of the additive and multiplicative approach.

Exploratory analysis of linear model adjustments

An exploratory analysis was done on monthly and daily temperature and precipitation datasets, and with daily peak wind gusts as well. The characteristics of these datasets are:

Daily temperatures were obtained from the clean version of Rachel Killick's Wyoming world 1 dataset, containing complete data from 75 stations for the period 1970-2011 (15340 in every station).

Monthly temperatures were computed by aggregating the former data, yielding 504 monthly values for each of the 75 stations.

Monthly precipitations corresponded to a selection of 244 mainland Spanish series that had a maximum of 30 missing data out of the 780 months of the period 1951-2015.

Daily precipitations were compiled from a selection of 110 Balearic stations with a maximum of 120 missing data in the period 2001-2010 (3652 data per station when complete).

Daily peak wind gusts are from 103 Australian series with a maximum of 30 missing data in 2005-2014.

Methodology: Linear regression models were adjusted for every pair of stations in each dataset, gathering the regression coefficients a and b (from $Y = a + bX$), their p-values, and the coefficients of determination r^2. Afterwards, coefficients from adjustments with $r^2 < 0.5$ and not significant at $\alpha = 0.05$ were rejected. Cubic root transformation of the data were also tested with the daily precipitations, and the $log(1+x)$ transformation was also applied to both the daily precipitations and wind peak gusts.

Results and discussion: Boxplots of these selected coefficients a and b can be seen in Figure 1. Both monthly and daily temperatures show b values near to 1 and a values that, although unbiased with respect to 0, show a clear dispersion, thereby validating the applicability of additive models to this variable at both time scales. Monthly precipitations have both coefficients far from their target values, and therefore the simplification of the linear model is not adequate for them. Daily precipitations, on the contrary, display a coefficients very near to 0, hence indicating the suitability of the multiplicative model. Daily peak wind gusts behave very much like monthly precipitations, not allowing the simplification of any of the coefficients of the OLS regression. Cubic root and logarithmic transformations do not alter these appreciations.

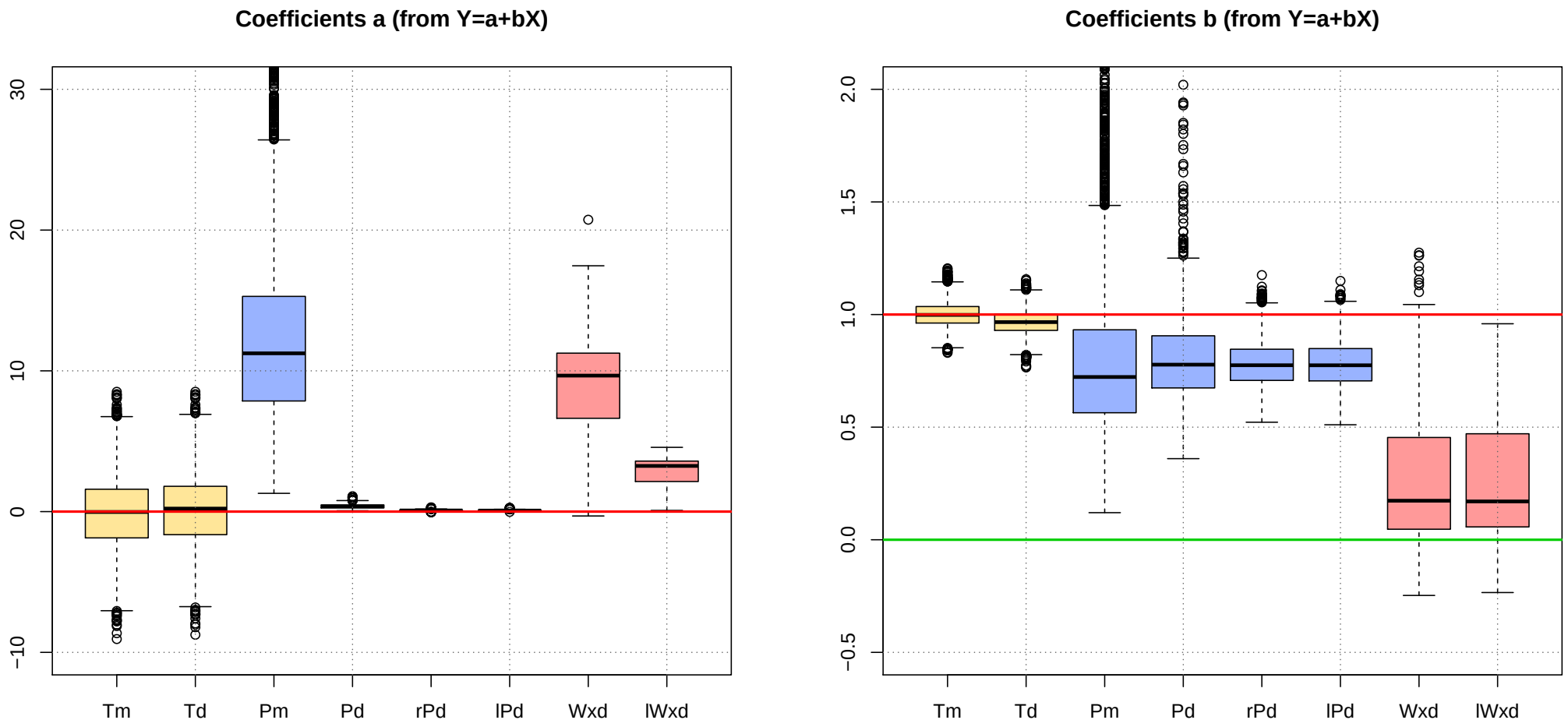

Fig. 1: Boxplots of the coefficients *a* (left) and *b* (right) from significant pairwise adjustments of linear models to the following data: Tm : Monthly temperatures; Td : Daily temperatures; Pm : Monthly precipitations; Pd : Daily precipitations; Wxd : Daily peak wind gusts.

Experiences with Climatol 3.0

The first operational homogenization of daily data with Climatol was done by colleagues at the AEMET Murcia Office for a project studying maximum temperatures in their region. They used the former version 2.2 and realized that SNHT detection thresholds had to be set very high (several hundred units) to adapt the method to their dataset.

More recent is the experience of homogenizing daily peak wind gusts series from Portugal and Spain (Azorín-Molina *et al.*, 2016), still using the same version. In that work, direct homogenization of the daily series (adjusting the SNHT thresholds as in the Murcia case) yielded better results than homogenizing at the monthly scale and then interpolating the corrections at the daily resolution. This latter is the recommended procedure for homogenizing daily temperatures (Vincent *et al.*, 2002; Brunet *et al.*, 2006), but it cannot give good results with strongly biased variables.

This was confirmed in ongoing homogenization works on Australian daily peak wind gusts and daily precipitation databases for the Horn of Africa. Moreover, root or logarithmic transformations were not of much help. In these cases, the ratio normalization of the series was the best option when applying Climatol to these datasets, a fact put in evidence also in the first benchmarking results of the MULTITEST Spanish project, as shown in Figure 2.

Homogenization of daily data with Climatol 3.0

It is well known that the detection of shifts in the mean is much more difficult on daily series than on monthly, seasonal or annual time resolution, and indeed Climatol detected more breaks in the monthly series than in the daily data. But as we have discussed, the interpolation of monthly corrections is only recommended for temperatures (and likely for other variables with an unbiased, normal like distribution), and only to adjust the mean, but not the variance. However, to study the variability of the mean we do not really need to work with daily data!

The final procedure recommended to homogenize daily series with the Climatol 3.0 package consists in the following steps:

- Build monthly series from the daily data.
- Homogenize the monthly series.
- Optionally: Edit the file of detected break dates to adjust them to relevant metadata when available.
- Obtain the homogenized daily series honoring the metadata file. (No break detection is attempted at the daily scale).

This methodology can be tested with the example data accompanying the package that, once installed and R started, can be reproduced with the following steps ('#' and the rest of the line is a commentary):

```
library(Climatol) # load the functions of the package
```

First you must generate the input files from example data:

```
data(Ttest)
write(dat, 'Ttest_1981-2000.dat')
write.table(est.c, 'Ttest_1981-2000.est', row.names=FALSE, col.names=FALSE)
rm(dat,est.c) #remove loaded data from memory space
```

Now you can begin the homogenization example:

```
dd2m('Ttest', 1981, 2000) #aggregation of the daily data into monthly averages
homogen('Ttest-m', 1981, 2000) #Homogenization of the monthly data
```

At this point, you can edit the file Ttest_1981-2000_brk.csv to adjust the dates of the breaks to known relevant metadata.

Finally, you can obtain the homogenized daily series by splitting them by the break dates and reconstructing complete series from the homogeneous sub-periods:

```
homogen('Ttest', 1981, 2000, metad=TRUE)
```

Results include a PDF file with lots of diagnostic graphics, lists of breaks and outliers and an R binary file containing input and homogenized series, from which you can get statistics and grids with accompanying post-processing functions. More details in the documentation of the package, which is available at http://cran.r-project.org/web/packages/climatol/index.html

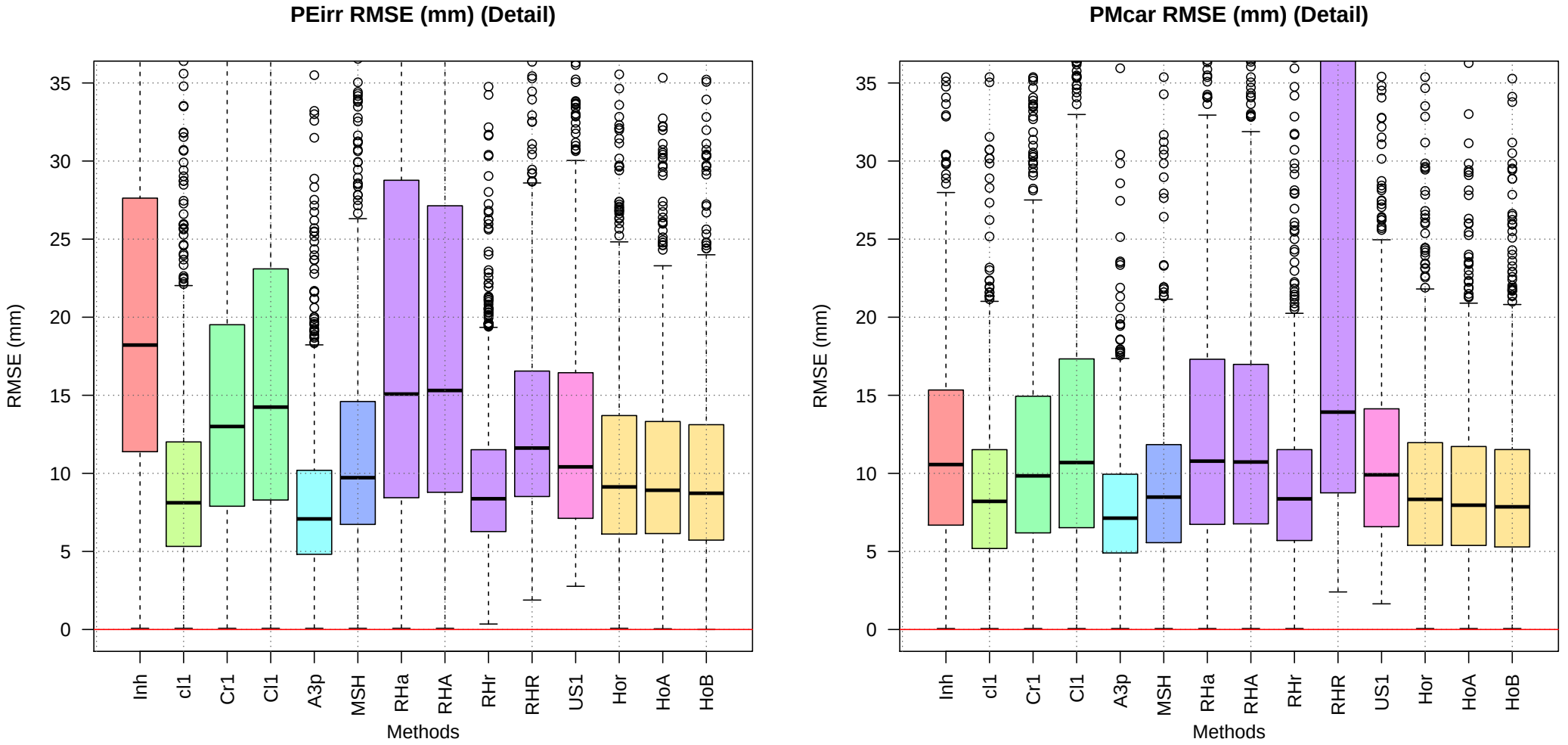

Fig. 2: RMSE of different homogenization methods applied to synthetic monthly precipitations series simulating an Atlantic Temperate (left) and a Mediterranean (right) climate. The first box (Inh) corresponds to the problem series, and the following three to the results obtained by Climatol, with different settings: ratio normalization (cl1), cubic root transformation (Cr1) and $log(x+1)$ transformation (Cl1). (The latter two series were normalized by the $y_i = \frac{Y_i - \bar{Y}}{\sigma_Y}$ standardization after their transformation). Ratio normalization gives much better results (left). Mediterranean precipitations are more difficult to homogenize (right).

Conclusions

- The exploratory analysis confirm the suitability of the additive and multiplicative models to temperature and precipitation respectively, although the best model for monthly precipitation remains debatable.
- It is neither clear the applicability of the multiplicative model to wind daily series.
- The Climatol package seems a good tool to homogenize monthly and daily series while preserving their variance, which is an important feature to assess the variability of extreme values in the homogenized series.

Acknowledgements

Many thanks to Rachel Killick for providing the Wyoming temperature series and to Tim McVicar and the Australian Bureau of Meteorology for the daily peak wind series. Spanish data were extracted from the AEMET climatic database.

References

Azorín-Molina C, Guijarro JA, McVicar TR, Vicente-Serrano SM, Chen D, Jerez S, Espirito-Santo F (2016): Trends of daily peak wind gusts in Spain and Portugal, 1961-2014. *Journal of Geophysical Research Atmospheres*, DOI: 10.1002/2015JD024485

Brunet M, Saladié O, Jones P, Sigró J, Aguilar E, Moberg A, Lister D, Walther A, Lopez D, Almarza C (2006): The development of a new dataset of Spanish daily adjusted temperature series (SDATS) (1850-2003). *Int. J. Climatol.*, 26:1777-1802.

Conrad V, Pollack LW (1950): *Methods in climatology*. Harvard Univ. Press, Cambridge-Massachusetts, xi+459 pp.

Paulhus JLH, Kohler MA (1952): Interpolation of missing precipitation records. *Month. Weath. Rev.*, 80:129-133.

Szentimrey T (2013): Theoretical questions of daily data homogenization. *Időjárás*, 117:113-122.

Szentimrey T (2014): Multiple Analysis of Series for Homogenization (MASH v3.03). http://owww.met.hu/pages/seminars/seeera/downloads.htm

Vincent LA, Zhang X, Bonsal BR, Hogg WD (2002): Homogenization of daily temperatures over Canada. *Journal of Climate*, 15:1322-1334.