Supplemental Material

SUPPLEMENT

Journal of Climate https://doi.org/10.1175/JCLI-D-20-0611.1

S1 Synthetic datasets

In synthetic datasets the climate component is spatially constant (Eq. 4), and six of such datasets are used in our study (Y1…Y6, Table 1). The method of simulation is as follows:

i) One hundred geographical locations are randomly selected as station positions within a 4°x 3° longitude - latitude area.

ii) The time series of common climate signal (C_R) is assigned to one of these stations. This first series is taken from true observed data or from an earlier benchmark dataset, supplying that with an additional 2°C/100yr warming trend. For datasets Y1…Y4, C_R is identical with the quality controlled and homogenized temperature series of Valladolid (Spain, 41.64°N, 4.75°W, 735 m asl) over the period 1951-2010, while for Y5 and Y6 C_R is identical with one series of the HOME benchmark dataset (hotmm04309900d.txt of syn1/000015 network, 1900-1999).

iii) When J stations have data ($0 < J < 100$) and $100–J$ stations are still without data, the distances of all station pairs where one has data (A) and the other does not (B) are examined. The station pair with the shortest distance is retained and the data for station B is generated from the data of station A by adding a series of white noise to the data of A (Eq. S1).

$$\mathbf{X_B} = \mathbf{X_A} + \boldsymbol{\varepsilon} \qquad \text{(S1)}$$

The noise is Gaussian white noise with zero mean and σ standard deviation where σ is arbitrarily chosen for creating datasets of varied spatial correlations. The mean spatial correlation for individual test dadasets ranges between 0.56 and 0.91 (Table 1).

iv) The values of series are shifted to simulate different elevations, and the amplitudes of their seasonal cycles are randomly varied by ±20 %.

With the previous steps, datasets of 100 station series have been created. For the efficiency tests smaller networks are used, which are randomly sampled from these “master” datasets.

Four of the six synthetic datasets (Y1…Y4) are complete, while in Y5 and Y6 missing data occur with similar structures to those in the HOME benchmark, i.e., blocks of missing data of varied lengths occur first of all at the beginning of time series and secondly near to the middle of time series, with shorter maximal duration.

S2 Surrogate homogeneous datasets

For the surrogate datasets, the daily temperature benchmark dataset developed by Willet et al. (2014) and Killick (2016) is adapted. A block of 158 time series (1970-2011) for

Wyoming state (USA) has been selected. It includes all kinds of natural spatial-temporal climate signals (Willett et al., 2014). In our adaptation, monthly means were derived, and the time series were lengthened by repeating the values in the reverse temporal order after the end of the source time series, but keeping the order of calendar months unchanged. Then randomly generated slight linear trends were added to avoid full repetitions of data. For some of the test datasets, spatial correlations were lowered with adding first order autocorrelation of 0.15 to the data. The complete description of this dataset generation was presented by Domonkos and Coll (2017a). Six of such test datasets are used here (U1, U2,…U6). Four of them (U1…U4) are complete, in U5 the missing data structure is the same as in the HOME benchmark, while in U6 the initial 40% of the data is missing in 53% of the time series (or with other words, the time series are 40%shorter in 53% of the time series).

Each of the surrogate datasets consists of 100-500 networks of 5-15 time series. The length of time series varies between 40 and 100 years, while the mean spatial correlation (*R)* is between 0.68 and 0.91 (Table 1).

S3. Differences between residual errors for surrogate and synthetic datasets

We compare the mean results for the 6 synthetic datasets with the results for the 6 surrogate datasets. For obtaining one average error over all error types, they are normalized with the raw data errors in the same way as in Sec. 3.1 before averaging. In other words, the normalized errors of five efficiency measures (i.e., RMSEm, RMSEy, Trb, NetEy, NetTr) are averaged, and the results are shown in Fig. S1.

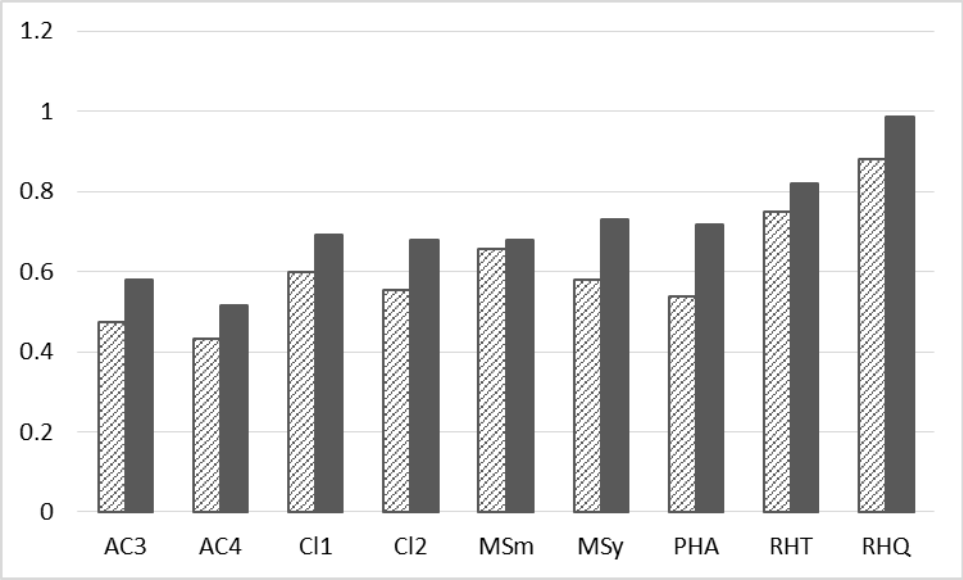

Figure S1. Mean residual errors for 5 kinds of efficiency measures (RMSEm, RMSEy, Trb, NetEy, NetTr) normalised by the raw data errors. Striped bars: means for the 6 synthetic datasets, filled bars: means for the 6 surrogate datasets.

It can be seen that the mean results for synthetic datasets are better for all tested methods than the results for the surrogate datasets, and the increase of the residual errors for the surrogate datasets is about 10% of the raw data error with only moderate differences according to homogenization methods. These differences can be seen more detailed in Table S1, which presents the ratio of residual errors in surrogate datasets in comparison with synthetic datasets, separately for three groups of error types, i.e. for the errors in individual time series, network mean errors, and all error types.

The mean residual error is 18% higher for surrogate datasets than for synthetic datasets, and the ratios do not differ significantly between the groups of error types. Note, however, that the dataset properties in synthetic and surrogate datasets differ in many ways, hence Fig. S1 and Table S1 show the combined effects of surrogate – synthetic difference and other differences. The differences of ratios of Table S1 according to homogenization methods are smaller than ±10% except for PHA and

Table S1. Mean ratio of residual errors in surrogate datasets in comparison with synthetic datasets.

Method	Individual series	Network means	All error types
AC3	1.204	1.234	1.219
AC4	1.194	1.218	1.206
Cl1	1.113	1.194	1.155
Cl2	1.228	1.230	1.229
MSm	1.037	1.034	1.036
MSy	1.253	1.271	1.262
PHA	1.307	1.363	1.332
RHT	1.135	1.066	1.094
RHQ	1.174	1.060	1.117
Mean	1.183	1.186	1.184

MSm. The large ratio for PHA is likely linked to the high performance of PHA in removing semi-synchronous breaks from datasets Y5 and Y6, as this elevates the mean performance of PHA for the tested synthetic datasets.

The generally low variation of the ratios presented in Table S1 indicates that the kind of homogeneous set (i.e. synthetic or surrogate) does not have a strong effect on the efficiency rank order in error reduction. Note that during HOME a similar comparison between the residual errors of surrogate and synthetic datasets showed only 5% error increase for the surrogate HOME temperature benchmark. The comparison was made with an early version of ACMANT, and its results have not been published. However, that comparison was more reliable than the comparison of this study, as the synthetic and surrogate sections of HOME temperature benchmark have the same properties except for the surrogate – synthetic difference.