CrossMark

Synchrony between reanalysis-driven RCM simulations and observations: variation with time scale

Ramón de Elía[1,2] · René Laprise[2] · Sébastien Biner[1] · James Merleau[3]

Abstract Unlike coupled global climate models (CGCMs) that run in a stand-alone mode, nested regional climate models (RCMs) are driven by either a CGCM or a reanalysis dataset. This feature makes high correlations between the RCM simulation and its driver possible. When the driving dataset is a reanalysis, time correlations between RCM output and observations are also common and to be expected. In certain situations time correlation between driver and driven RCM is of particular interest and techniques have been developed to increase it (e.g. large-scale spectral nudging). For such cases, a question that remains open is whether aggregating in time increases the correlation between RCM output and observations. That is, although the RCM may be unable to reproduce a given daily event, whether it will still be able to satisfactorily simulate an anomaly on a monthly or annual basis. This is a preconception that the authors of this work and others in the community have held, perhaps as a natural extension of the properties of upscaling or aggregating other statistics such as the mean squared error. Here we explore analytically four particular cases that help us partially answer this question. In addition, we use observations datasets and RCM-simulated data to illustrate our findings. Results indicate that time upscaling does not necessarily increase time correlations, and that those interested in achieving high monthly or annual time correlations between RCM output and observations may have to do so by increasing correlation as much as possible at the shortest time scale. This may indicate that even when only concerned with time correlations at large temporal scale, large-scale spectral nudging acting at the time-step level may have to be used.

Keywords Regional climate model · Spectral nudging · RCM-observations synchronicity

1 Introduction

General statistical properties of nested regional climate model (RCM)-generated simulations may seem at first sight similar to those produced by coupled global climate models (CGCMs). Similarities in fact exist when RCM-simulated fields are studied without consideration of the chronology of events, such as climatological means or frequency distributions. But due to the control exerted by the lateral boundary conditions upon nested RCM simulations, the RCM-simulated fields do exhibit some synchronicity with their driving data, whether GCM-simulated fields or reanalyses. When driven by reanalyses, RCM-simulated fields are also correlated with observations to some extent.

For several applications such as projected future climate changes, synchronicity is not an issue as only the overall trend matters. On the other hand, synchronicity is paramount in applications such as dynamical downscaling of reanalyses aiming at obtaining high-resolution pseudo observations. This last application was first suggested by Anthes (1983) and has become an active area of research over the last few years (e.g. Li et al. 2012; Stefanova et al. 2011; Weisse et al. 2009; Kanamaru and Kanamitsu 2007).

✉ Sébastien Biner
   biner.sebastien@ouranos.ca

1   Consortium Ouranos, 550, Sherbrooke Street West, 19th floor, West Tower, Montréal, QC H3A 1B9, Canada

2   Département des Sciences de la Terre et de l'Atmosphère, Centre ESCER, Université du Québec à Montréal (UQAM), Montréal, QC, Canada

3   Institut de recherche d'Hydro-Québec (IREQ), Varennes, QC, Canada

A simple proof-of-concept exercise of any of these applications is to verify whether and under which conditions RCMs' anomalies correlate well with those of the driving data. von Storch et al. (2000) have shown that the application of large-scale spectral nudging greatly improves the time correlation between a reanalysis-driven simulation and observations, although the match is not perfect (e.g. Alexandru et al. 2009).

In applications of RCM downscaling for which synchronicity matters, an important issue relates to the way chronology-related statistical properties vary with time scale. For example, how does the time correlation between the RCM-simulated fields and the corresponding driving fields or observations change when considering the time series on daily, weekly, monthly, seasonal or annual basis?

A naïve view might lead one to think that time upscaling (aggregation) should necessarily improve time correlation and apparent synchronicity, analogous to temporal or spatial upscaling that generally improves scores such as Root-Mean-Square error or bias by reducing the unpredictable noise. A typical example is "hedging" in weather forecast, operation that uses low-pass filters or horizontal diffusion to remove poorly predicted finer scales (e.g. Jakimow et al. 1992) in order to avoid the "double penalty" problem typical of high-resolution simulations (e.g. Mass et al. 2002; Bougeault 2002). The objective of this work is to investigate in a simple framework whether this conjecture is also correct for time correlation of upscaled RCM-generated data.

This question can be generalized in the following way: How are weather anomalies in the driving data or observations reproduced by the RCM as a function of time scale? If the RCM and driving data are found to be somewhat asynchronous at some short time scale, can they be more synchronous on longer time scales? As simple as it may sound, the problem just stated is not trivial, and as far as the authors are aware, there exists no general solution without some strong assumptions. In what follows we explore analytically four particular cases that help us partially answer this question. In addition, we use observation datasets and RCM-produced data to illustrate our findings.

2 Analytical approaches

2.1 A very simple case

Here we discuss the simplest possible case when both the RCM-simulated and the observations time series have zero time autocorrelation and identical time variance. Despite its simplicity, this assumption can be somewhat realistic in some cases. Here m_t will refer to a reanalysis-driven

RCM-simulation seasonal anomaly time series and o_t the corresponding observations anomaly time series, and their variance $var(m_t) = var(o_t) = \sigma^2$. In the following we will refer to 'observations' only, but all developments would equally apply to driving data as well.

Consider the case when the temporal correlation between these two time series is $corr(m, o) = r$ (correlation hereafter refers to Pearson correlation). It is shown in Appendix 1 that upscaling these time series by averaging two successive time steps or archival times as

$$A_m = \frac{m_t + m_{t+1}}{2}, \, A_o = \frac{o_t + o_{t+1}}{2} \tag{1}$$

yields a correlation

$$corr(A_m, A_o) = r \tag{2}$$

which is identical to that of the original time series.

This may seem a counterintuitive result at first as one might have expected that time averaging would filter-out noise, resulting in an increased correlation (Sect. 2.4 will further illuminate this result). The covariance is indeed decreased, but the variances of the two variables are decreased in the same proportion, keeping the correlation unchanged. This is in fact a well-known result of multivariate statistics (see for example Johnson and Wichern 2007). Although theoretically the correlation is the same for the average, it should be noted that when estimating the correlation with data, the error of the estimation will increase as the sample size decreases with averaging (see Appendix 5). We will explore this with more detail later.

One may wonder if this rule could be generalized to all scales, that is, whether the following rule is valid: lack of correlation at short time scales implies lack of correlation at longer scales. But of course this generalization would violate one of the initial assumptions since daily values are strongly correlated in time. We will consider in Sect. 2.3 a case when autocorrelation is taken into account.

2.2 A case with different variances and correlations

In this example we discuss a somewhat more general case than the previous one. Let us consider a year composed of two seasons for simplicity, summer (S) and winter (W), each one with different variances, both for the simulated (m) and observed (o) variables, $\sigma^2_{S_m}, \sigma^2_{S_o}, \sigma^2_{W_m}, \sigma^2_{W_o}$, respectively, and with different correlations between simulations and observations anomalies, r_S and r_W, for summer and winter, respectively.

We may now ask about the correlation of the annual-mean anomaly r_A, where $A = (S + W)/2$. Assuming that successive seasons are uncorrelated, Appendix 2 shows that the correlation of the annual anomaly r_A can be written as a function of seasonal values as

$$r_A = \frac{r_S \sigma_{S_m} \sigma_{S_o} + r_W \sigma_{W_m} \sigma_{W_o}}{\sqrt{\left(\sigma_{S_m}\sigma_{S_o} + \sigma_{W_m}\sigma_{W_o}\right)^2 + \left(\frac{\sigma_{S_m}}{\sigma_{S_o}} - \frac{\sigma_{W_m}}{\sigma_{W_o}}\right)^2 \sigma_{S_o}^2 \sigma_{W_o}^2}}.$$

(3)

Expression (2) is recovered in the special case of identical variances and identical correlations. An interesting case occurs when the second term in the denominator vanishes, that is, when the modeled variance differs from the observed one but this relative error is similar in both seasons. In this case annual anomaly correlation r_A is then a weighted average of the seasonal anomaly correlations

$$r_A = \frac{a r_s + b r_w}{a + b}.$$

(4)

where $a = \sigma_{S_m}\sigma_{S_o}$ and $b = \sigma_{W_m}\sigma_{W_o}$. This last expression can be easily generalized to more than two seasons or subseasons. The weights are a function of the simulated and observed time variability, and the annual anomaly correlation is dominated by the season with the largest variance. Contrary to what could have been intuitively expected, decreasing the time resolution of the time series from seasonal to annual does not increase the correlation over that of the individual values. Furthermore, if we now note that the neglected term on the denominator is positive, we see that r_A is in fact *smaller* than the seasonal weighted average. Hence, upscaling the time series not only does not increase the correlation but in fact may reduce it; this condition would occur for instance when the model fails to simulate well seasonal variances.

2.3 Autoregressive case

Cases treated thus far have assumed no autocorrelation within each time series, making them unrealistic in many instances. To allow for autocorrelation, we will now use a representation of simulated and observed time series by first-order autoregressive processes. Assuming that the simulated time series m and that of the observations o follow first order autoregressive processes, and knowing $corr(o_t, m_t)$, we want to know the upscaled value for $corr(A_o, A_m)$, where $A_0 = (o_{t+1} + o_t)/2$ and $A_m = (m_{t+1} + m_t)/2$.

In Appendix 3 it is shown that the solution is

$$corr(A_o, A_m) = corr(o_t, m_t) \frac{1}{2} \left(\frac{2 + \rho_m + \rho_o}{\sqrt{1+\rho_o}\sqrt{1+\rho_m}} \right)$$

(5)

where ρ_o and ρ_m are the autoregressive parameters of each time series for a first order autoregressive process (these parameters need to be constrained between -1 and 1 for the process to be second-order stationary, and are equal to the autocorrelations at lag 1). The simplest case result

presented in (2) is recovered when the autocorrelations vanish. The more interesting cases occur when the multiplying factor on the right-hand side differs from unity. As shown in Appendix 3, its value is always greater than or equal to 1, but only marginally so. For the reasonable case when the autoregressive parameters of simulations and of observations time series do not differ much, this factor exceeds unity by about 1 %, and it can reach 6 % when autocorrelations are the most different from each other but remaining positive (which means model and observations have quite different behaviours). These results suggest that

$$corr(A_o, A_m) \approx corr(o_t, m_t)$$

(6)

is a very good approximation. So this result also indicates that time-upscaling an RCM time series does not improve noticeably its time correlation with observations under this approximation.

2.4 Noise plus sinusoidal signal

For the next case we assume that both time series—that of the observations and that of simulated data—consist in a random noise superimposed upon a slowly evolving sinusoidal signal. To simplify we will postulate that parameters of both time series are equal, differing only in the instantaneous values of the random noise, which are uncorrelated here. This corresponds to the case in which the model is able to reproduce the long time-scale signal in both amplitude and phase, but unable to reproduce the phase of the random noise features.

Under these assumptions, it is shown in Appendix 4 that the correlation of the upscaled time series is

$$corr(A_o, A_m) = corr(o_t, m_t) \left[1 + \frac{\sigma_R^2/\sigma_S^2 \cos(\omega \Delta t)}{\sigma_R^2/\sigma_S^2 + 1 + \cos(\omega \Delta t)} \right]$$

(7)

where σ_R^2 and σ_S^2 represent the variances of the random terms and of the sinusoidal signal, respectively. The factor ω in the cosine function represents the angular frequency, which is inversely proportional to the period. For the sake of shortening the expressions we write it as a function of the angular frequency $\omega = 2\pi/T$, where T is the period of the signal.

The factor on the right-hand side is greater than 1 for a short time interval and/or long time-scale signal, $2\pi \Delta t/T < \pi/2$ or when $\Delta t < T/4$. If in keeping with the notation used in the previous sections we assume $\Delta t = 1$, we obtain that the period should be larger than the time covered by four successive time intervals. In this case the two successive data points to be averaged in the upscaling will have in most cases the same anomaly sign. For longer

Table 1 The first column indicates the ratio of variances between noise R and sinusoidal signal S

σ_R^2/σ_S^2	$corr(o, m)$	$corr(A_o, A_m)$	Factor
1	0.5	0.66	1.33
¼	0.8	0.89	1.11
4	0.2	0.33	1.66

The second and third column show the correlation before and after the time upscaling, respectively. The forth column indicates the increase factor by upscaling

time periods this property becomes dominant. Table 1 gives a sense of the importance of this factor for a very long period signal fully encompassed in the time series.

When the noise term is small compared with the sinusoidal signal amplitude, $\sigma_R^2 < \sigma_S^2$, the correlation of the original time series is high and averaging increases it further. When noise is dominant, $\sigma_R^2 > \sigma_S^2$, the original correlation is low and averaging increases this correlation substantially, but correlation remains modest.

This result sheds some light upon what our intuition may suggest: that when there is a distinct signal, reduction of noise by averaging does indeed increase the temporal correlation. The conflicting results obtained thus far indicate that analytical developments cannot reach the bottom of all our questions. One may wonder whether real cases have properties that resemble more one case or another.

3 Some results from RCM simulations and observations

In order to illustrate the previous developments with actual data, in this section some results for precipitation and temperature obtained from RCM simulations, weather stations, reanalysis, and gridded observation datasets are presented. The RCM used is the fifth-generation Canadian Regional Climate Model (CRCM5), described in Martynov et al. (2013) driven by Era Interim (Dee et al. 2011). For reasons of availability, data coming from different simulations will be discussed: The two simulations used in the first subsection were performed on a subcontinental domain over eastern Canada on a 11-km grid mesh, while the three used in the second subsection were performed over a domain encompassing the entire North America a 22-km grid mesh. One of these three simulations uses a moderate large-scale nudging on the wind field, with a 6-h relaxation time from the top of the troposphere until 500 hPa, then diminishing linearly with height to become unforced at the surface (Separovic et al. 2011). For seasonal values observations from CRU3.20 dataset are used (Harris et al. 2013). For daily observations data from Era-interim and


Fig. 1 Location of selected Environment Canada weather stations

from Environment Canada's homogenized weather stations are used (see Mékis and Vincent 2011; Vincent et al. 2012). Stations were chosen using two criteria: geographical coverage within the integration domain and data quality and completeness. Five stations were selected: Bagotville, Gaspé, Kuujjuaq, Kuujjuarapik, and Roberval (see Fig. 1).

3.1 Time upscaling from daily to monthly values

In this section we will analyse the time-correlation upscaling from RCM-downscaled reanalysis and the corresponding observed values. Downscaled daily data is taken from two CRCM5 simulations differing only in initial conditions (known internally at Ouranos as bba and bbb, hereafter mentioned as "twin" simulations), integrated for 32 years from 1980 to 2011. The gridpoints used are those that are closest to the selected weather stations. Daily values are divided in two time series, one for January and the other for July, in order to remove seasonality that may increase artificially the correlation. January and July are considered to have 32 days to facilitate computations. Daily values (a total of 1024 for each series) are upscaled using expression (1) to a 2-day average time series (a total of 512 for each series), to a 4-day average (a total of 256 for each series), to a 8-day average (a total of 128 for each series), to a 16-day (a total of 64 for each series), and to a 32-day average (a total of 32 for each series).

Figure 2 depicts time correlation for surface temperature between the CRCM5 downscaled reanalysis and the Environment Canada observations for the different time scales studied. Blue lines indicate January values and red ones July values. Different symbols are used for the twin simulations (a star and a cross). During winter, correlation seems to generally increase for longer timescales. During summer,
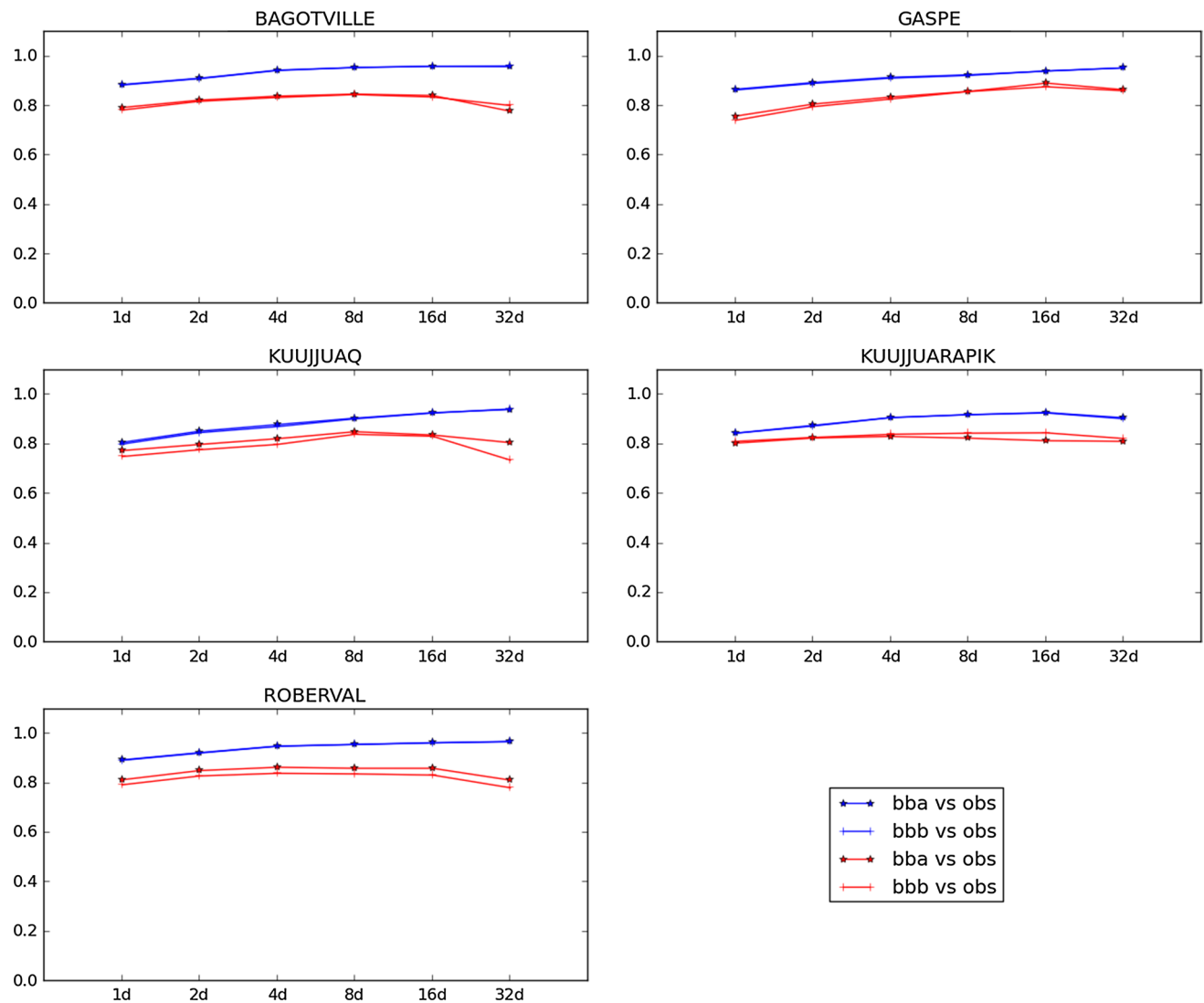
Fig. 2 Time correlation of surface temperature between different CRCM5 simulations and station observations from Environment Canada, for different timescales ranging between 1 and 32 days. Blue and red lines identify January and July values, respectively. Difference in symbols indicates results obtained from twin simulations. Each graphic depicts one of the selected weather stations (see Fig. 1)

correlation is lower than during winter, and the existence of an increase with timescale is not obvious. Twin simulations produce very similar although not identical results for both seasons. Figure 3 is similar to Fig. 2 but for precipitation. Without surprise correlations are lower than for temperature. Winter precipitation seems to have an overall increase with time scale, with twin simulations producing very similar results. Summer precipitation shows no clear signs of increase and twin simulations may differ considerably in some locations.

It is important to recall that the estimations of correlation presented at longer time scales are associated with larger random errors as the sample number is reduced by upscaling. Appendix 5 discusses and illustrates the magnitudes of such errors. Single-day time series contain near 1000 values and hence error bars can be found to the right of the diagram in Fig. 8. Series of 32-day averages contain around 30 values and can be found near the left side of the same diagram. As we can see considering the color scale, high correlation values are associated with small but asymmetric intervals. For winter temperatures, correlation values between 0.8 and 0.9 are common, which gives intervals of confidence (total length) of around 0.03 for 1-day series and near 0.3 for 32-day series. This puts into question the statistical robustness of the apparent increase of correlation by time upscaling during winter noted in Figs. 2, 3. For summer temperature and precipitation in both seasons the situation is even less clear since lower overall correlation values are associated with larger intervals of confidence.
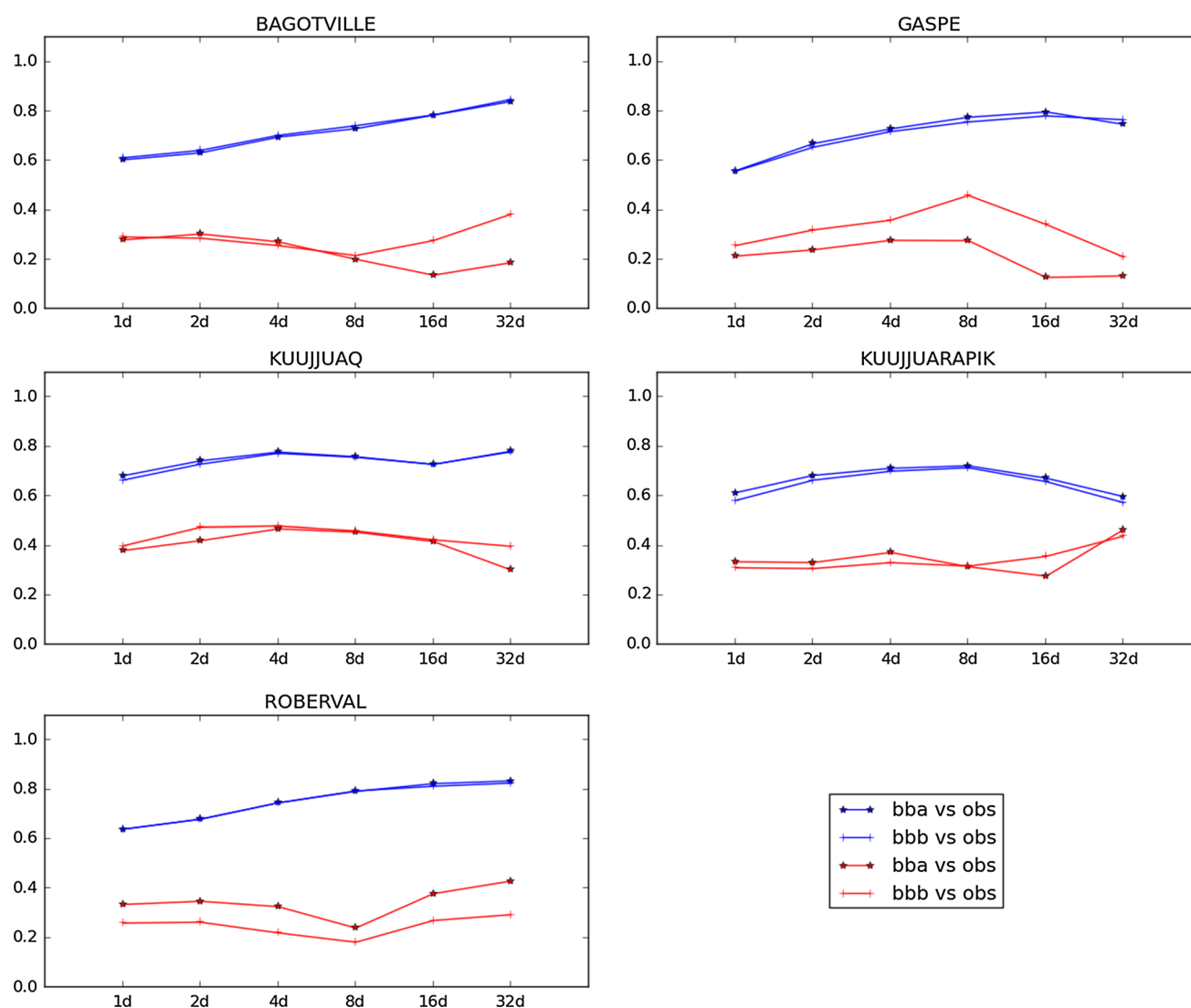
Fig. 3 As in Fig. 2 but for precipitation

This suggests that even if time upscaling might appear to produce an increase in correlation, establishing its statistical robustness may need more data. But naturally this prompts the following question: Is a barely detectable increase in correlation by upscaling worth the statistical battle?

3.2 Time upscaling from seasonal to yearly values

The previous section omitted the impact of seasons when studying correlation at different timescales. In this section, as in Sect. 2.2 we will concentrate in the upscaling from seasonal to yearly sampling.

Figure 4 depicts correlations between a number of datasets and the CRU3.20 data for temperature (Fig. 4a) and precipitation (Fig. 4b). On the vertical axis the correlations

obtained for annual values are plotted, while in the horizontal axis values that estimate the annual quantity by weighted average of seasonal values are depicted [here we use an extension to four seasons of expression (4)]. Each color corresponds to a different dataset, three for different simulations of the CRCM5 (bao, ban and bar2), and one (era_int) for the Era Interim reanalysis. Runs bao and ban have identical setups differing only in initial conditions (twin simulations). Run bar2 is also identical but large-scale nudging is used. In order to give some statistical robustness correlations are computed for all grid points over the province of Québec, Canada (812 points in total; but note that spatial autocorrelation makes these points far from independent, particularly for the case of temperature). Colored stars indicate the median value for each dataset.
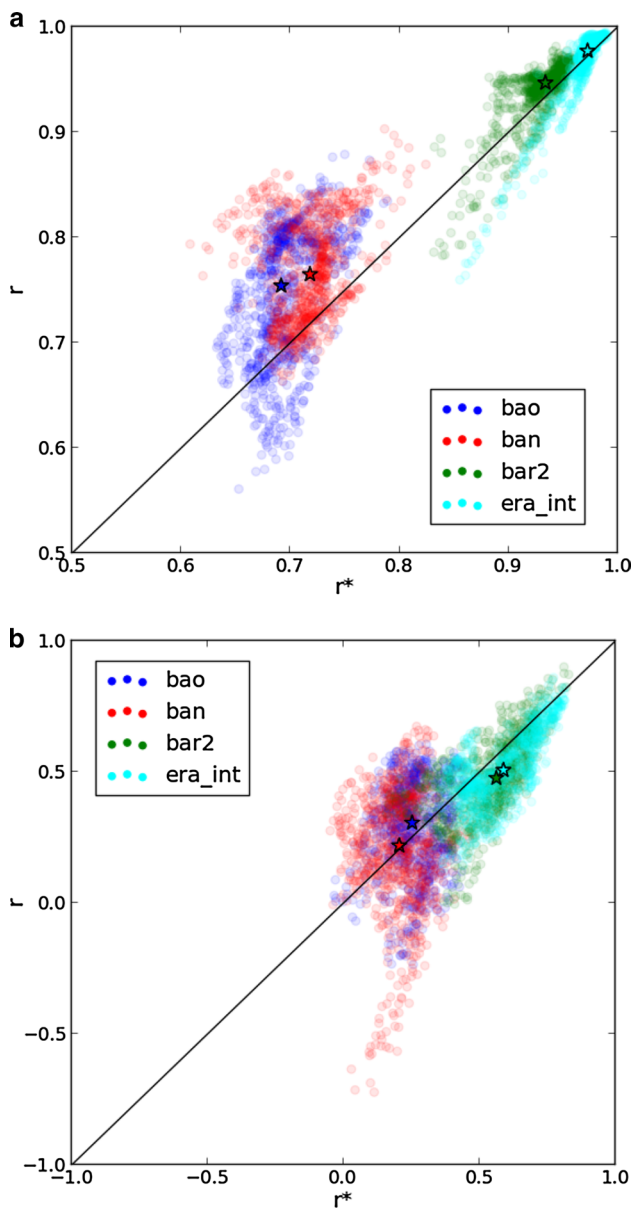
Fig. 4 Correlations between a number of datasets and the CRU3.20 data for temperature (**a**) and precipitation (**b**). On the vertical axis the correlations obtained for annual values are plotted, while in the horizontal axis values estimates of the annual quantity by weighted average of seasonal values are depicted [see expression (4)]. Each color corresponds to a different dataset, three (bao, ban and bar2) to different versions of CRCM5, and one (era_int) to the Era Interim reanalysis. Runs bao and ban are identical simulations differing only in initial conditions. Run bar2 is also identical but uses a moderate form of large-scale spectral nudging. Values are taken for a period from 1981 to 2010 over the province of Québec, Canada (main delimited area in Fig. 1). Correlations are computed at all grid points (which corresponds to 812 points). Colored stars indicate the mean value for each dataset

The graphs should be interpreted as follows: when points lay above the diagonal it indicates that upscaling from seasonal to annual does in fact improve over

simple combinations of seasonal correlation. As discussed in Appendix 5, errors in the estimations of correlation with a 30-year long time series are quite large, hence values in the scattergram have a considerable spread.

Figure 4a shows that for temperature most of the data cloud and their mean values lay above the diagonal, suggesting that upscaling does improve correlation. This is particularly clear for simulations bao and ban performed without large-scale nudging. For ERA-Int and the spectrally nudged simulation bar2, the improved correlation is more modest; but it should be kept in mind that it is difficult to improve very high correlations since it is a bounded quantity.

The case for precipitation is somewhat different (Fig. 4b, notice change of range in both axes). Correlations between observed and simulated precipitation are overall significantly lower than that for temperature, and upscaling seems not to have a distinctive positive effect. The highest correlations are slightly degraded by upscaling and the lower ones are slightly improved.

4 Conclusions

The aim of this study was to analyse the effect of time upscaling (aggregating) on the correlation between a time series produced by a reanalysis-driven RCM and observations. Lacking a general solution, we have approached the issue from four different simple analytical perspectives: 1- Two stochastic correlated time series, without autocorrelation and with identical variances, 2- As in 1, but allowing different variances and correlations between series, as typical in the case of seasonal statistics, 3- Time series with an autoregressive behaviour, typical of short time interval series, and 4- Stochastic time series that are modulated by a sinusoidal signal, such as a interannual oscillations.

The four approaches delivered different results. The first two suggested that in the absence of autocorrelation, upscaling does not improve correlation and may in fact deteriorate it if the model fails to produce the appropriate seasonal interannual variances. Substantial autocorrelation is however a typical property of daily time series—for fields such as surface temperature for example–, and this case was discussed in the third approach. Results indicated that no degradation occurs, but no substantial improvement is to be expected by upscaling either. The last case represented a stochastic time series with no autocorrelation— and no cross correlation—with a superimposed sinusoidal signal (which in fact introduces autocorrelation and cross correlation in the time series). This case can be thought of, for example, as yearly values modulated by interannual variability with a single dominant mode. In this case, upscaling produces a substantial increase in correlation for long time scale oscillations.

The answer to the question asked in the introduction is hence not straightforward from a theoretical perspective. Results for real time series will depend on which property dominates and which assumptions are more realistic for the case chosen. Examples shown here with RCM-simulated data and observations corroborate in part the more cautious analytical results, producing some hope of correlation increase with scale for winter but almost none for summer. More work with real data is needed to provide more robust answers for a variety of cases, although the hope of substantial (and practical) gains by upscaling seems to be limited.

Before more is known about the topic, the safest approach may be to assume that lack of correlation at short time scales will express itself as lack of correlation at longer scales. If one is interested in achieving high correlation with observations at time scales such as monthly, seasonal or annual, one may be constrained to use some form of large-scale nudging within the RCM in order to strengthen the control exerted by the driving data on short time scales. It is still to be seen how intense this nudging should be to attain satisfactory results.

In the examples presented here we have focused mainly on correlations between reanalysis-driven RCM simulations and observed values. The formal considerations discussed in Sect. 2 are however relevant for other situations as well. Consider the following examples of two different sources of a given variable with significant cross correlation:

1. Two nested simulations driven by the same driving data, either using the same RCM (such as twin simulations as in Alexandru et al. 2009), variants of a given RCM (such as spectrally nudged or not), or different RCMs.
2. An RCM simulation and its driving dataset, whether a reanalysis or some CGCM-simulated data. Note that this applies even if the variable to correlate is not used to drive the RCM, such as precipitation. Studies of correlation between reanalysis-driven RCMs and reanalysis are common features in the production of high-resolution pseudo-observations or poor's man high-resolution analysis by dynamical downscaling (e.g. Kanamaru and Kanamitsu 2007).
3. CGCMs simulations nudged towards reanalysis and observations (e.g. Eden et al. 2012).
4. Two observational datasets. Studies on temporal correlation have been carried out, for example by Brands et al. (2012), who computed time correlations on a daily timescale between two different reanalysis.

Finally, here we have concentrated on time upscaling but this discussion could be extended to spatial-upscaling too. For example, whether for daily precipitation time

correlation between RCM and observations improves when instead of a single station a regional scale is considered. This is also a question that deserves attention.

Appendix 1

Let us consider a reanalysis-driven RCM-simulated anomaly time series m_t and the corresponding observations anomaly time series o_t. Let us assume for simplicity that the elements of each time series are random, decorrelated in time, with the same variance $\text{var}(m_t) = \text{var}(o_t) = \sigma^2$, and with correlation $corr(m_t, o_t) = r$ between the two time series.

Consider now a time series made of the average between adjacent time elements of these time series:

$$\bar{m}_t = \frac{m_t + m_{t+1}}{2}, \quad \bar{o}_t = \frac{o_t + o_{t+1}}{2} \tag{8}$$

The variances of these time series are obtained from

$$\text{var}(\bar{m}_t) = \text{var}\left(\frac{m_t + m_{t+1}}{2}\right)$$
$$= \frac{1}{4}\{\text{var}(m_t) + \text{var}(m_{t+1}) + 2\text{cov}(m_t, m_{t+1})\} \tag{9}$$
$$= \frac{1}{2}\text{var}(m_t) = \frac{1}{2}\sigma^2$$

and similarly $\text{var}(\bar{o}_t) = \frac{1}{2}\sigma^2$, and the covariances between these time series are

$$\text{cov}(\bar{m}_t, \bar{o}_t) = \text{cov}\left(\frac{m_t + m_{t+1}}{2}, \frac{o_t + o_{t+1}}{2}\right)$$
$$= \frac{1}{4}\left\{\begin{array}{c} \text{cov}(m_t, o_t) + \text{cov}(m_{t+1}, o_{t+1}) \\ +\text{cov}(m_t, o_{t+1}) + \text{cov}(m_{t+1}, o_t) \end{array}\right\} \tag{10}$$
$$= \frac{1}{2}\text{cov}(m_t, o_t) = \frac{1}{2}r\sigma^2$$

Then using the definition of correlation we get

$$corr(\bar{m}_t, \bar{o}_t) = \frac{\text{cov}(\bar{m}_t, \bar{o}_t)}{[\text{var}(\bar{m}_t)\text{var}(\bar{o}_t)]^{1/2}} = \frac{\frac{1}{2}r\sigma^2}{\left[\frac{1}{2}\sigma^2 \frac{1}{2}\sigma^2\right]^{1/2}} = r. \tag{11}$$

Appendix 2

Consider a summer time series S and a winter one W for both model simulations (m) and observations (o), with correlations

$$corr(S_m, S_o) = r_S \tag{12}$$

and

$$corr(W_m, W_o) = r_W. \tag{13}$$

We want to know what is the correlation of the annual time series—here considered as the simple average of winter and summer only—

$$corr(A_m, A_o) = r_A \tag{14}$$

where $A = (S + W)/2$. Using basic properties of the variance we get

$$\begin{aligned}\text{cov}(A_m, A_o) &= \text{cov}\left(\frac{S_m + W_m}{2}, \frac{S_o + W_o}{2}\right) \\ &= \tfrac{1}{4}(\text{cov}(S_m, S_o) + \text{cov}(W_m, W_o) \\ &\quad + \text{cov}(S_m, W_o) + \text{cov}(W_m, S_o))\end{aligned} \tag{15}$$

If we assume for the sake of simplicity that the last two terms are negligible—season anomalies are uncorrelated—and using the first two expressions plus the definition of correlation, we get

$$r_A \sigma_{A_m} \sigma_{A_o} = \tfrac{1}{4}\left(r_S \sigma_{S_m} \sigma_{S_o} + r_W \sigma_{W_m} \sigma_{W_o}\right). \tag{16}$$

The variance of the yearly signal can be obtained from the properties of the variance as

$$\text{var}\left(\frac{S_o + W_o}{2}\right) = \tfrac{1}{4}(\text{var}(S_o) + \text{var}(W_o) + 2\text{cov}(S_o, W_o)) \tag{17}$$

As before, assuming no correlation between seasons we get

$$\sigma_{A_o}^2 = \tfrac{1}{4}\left(\sigma_{S_o}^2 + \sigma_{W_o}^2\right). \tag{18}$$

Replacing this in (16) we get

$$r_A = \frac{\left(r_S \sigma_{S_m} \sigma_{S_o} + r_W \sigma_{W_m} \sigma_{W_o}\right)}{\sqrt{\left(\sigma_{S_m}^2 + \sigma_{W_m}^2\right)\left(\sigma_{S_o}^2 + \sigma_{W_o}^2\right)}} \tag{19}$$

and this can be rewritten as

$$r_A = \frac{r_S \sigma_{S_m} \sigma_{S_o} + r_W \sigma_{W_m} \sigma_{W_o}}{\sqrt{\left(\sigma_{S_m} \sigma_{S_o} + \sigma_{W_m} \sigma_{W_o}\right)^2 + \left(\frac{\sigma_{S_m}}{\sigma_{S_o}} - \frac{\sigma_{W_m}}{\sigma_{W_o}}\right)^2 \sigma_{S_o}^2 \sigma_{W_o}^2}} \tag{20}$$

The last expression is obtained after adding and subtracting $2\sigma_{S_m}^2 \sigma_{S_o}^2 \sigma_{W_m}^2 \sigma_{W_o}^2$ under the square root in the denominator of (19) and reorganizing.

Appendix 3

Consider now that the RCM-produced time series (m) and the observed time series (o) can be represented as autoregressive process with a given standard deviation σ and autocorrelation ρ. Then, they can be represented as

$$o_{t+1} = \rho_o o_t + \varepsilon_{t+1}^o \tag{21}$$

and

$$m_{t+1} = \rho_m m_t + \varepsilon_{t+1}^m \tag{22}$$

where t refers to a particular time and ε represents a random noise. To obtain the property that interests us we impose that the terms that represent the random part are cross-correlated between model and observations so that

$$corr(\varepsilon_t^o, \varepsilon_t^m) = r_\varepsilon \tag{23}$$

Knowing the correlation of the time series at successive times t, we now wonder about the correlation when the time series is transformed into averages of successive data. We define these new time series as

$$A_0 = \frac{o_{t+1} + o_t}{2} \tag{24}$$

and

$$A_m = \frac{m_{t+1} + m_t}{2} \tag{25}$$

Correlation between the original time series

Applying the covariance operator to the series defined in (21) and (22) we get

$$\text{cov}(o_{t+1}, m_{t+1}) = \text{cov}\left(\rho_o o_t + \varepsilon_{t+1}^o, \rho_m m_t + \varepsilon_{t+1}^m\right) \tag{26}$$

and using the bilinear property we get

$$\begin{aligned}\text{cov}(o_{t+1}, m_{t+1}) &= \rho_o \rho_m \text{cov}(o_t, m_t) + \text{cov}(\varepsilon_{t+1}^o, \varepsilon_{t+1}^m) \\ &\quad + \rho_o \text{cov}(o_t, \varepsilon_{t+1}^m) + \rho_m \text{cov}(\varepsilon_{t+1}^o, m_t)\end{aligned} \tag{27}$$

Since the random terms are decorrelated from the crossed time series we obtain

$$\text{cov}(o_{t+1}, m_{t+1}) = \rho_o \rho_m \text{cov}(o_t, m_t) + \text{cov}(\varepsilon_{t+1}^o, \varepsilon_{t+1}^m) \tag{28}$$

and finally

$$\text{cov}(o_t, m_t)(1 - \rho_o \rho_m) = \text{cov}(\varepsilon^o_{t+1}, \varepsilon^m_{t+1}) \tag{29}$$

Similarly, using the properties of the variance in (21) and (22) (assuming no covariance between crossed terms) we get

$$\text{var}(o_{t+1}) = \rho^2_o \text{var}(o_t) + \text{var}(\varepsilon^o_{t+1}) \tag{30}$$

and finally

$$\text{var}(o_t)(1 - \rho^2) = \text{var}(\varepsilon^o_{t+1}) \tag{31}$$

Returning to (29) we can now rewrite it using (31) and the definition of covariance on both sides as

$$corr(o_t, m_t)\sigma_o\sigma_m(1 - \rho_o \rho_m) = r_\varepsilon \sigma_o \sigma_m \sqrt{1 - \rho^2_o}\sqrt{1 - \rho^2_m} \tag{32}$$

and finally as

$$corr(o_t, m_t) = r_\varepsilon \frac{\sqrt{1 - \rho^2_o}\sqrt{1 - \rho^2_m}}{(1 - \rho_o \rho_m)} \tag{33}$$

It can be shown that this can also be written as

$$corr(o_t, m_t) = r_\varepsilon \sqrt{\frac{(1 - \rho_o \rho_m)^2 - (\rho_m - \rho_o)^2}{(1 - \rho_o \rho_m)^2}} \tag{34}$$

The second factor on the right hand side is smaller than 1, and its pattern is shown in Fig. 5.

Correlation between the averaged series and the original

Using the information obtained above we will now try to estimate the correlation of the averaged signal. Knowing that

$$A_0 = \frac{o_{t+1} + o_t}{2} = \tfrac{1}{2}(o_t(1 + \rho_o) + \varepsilon^o_{t+1}) \tag{35}$$
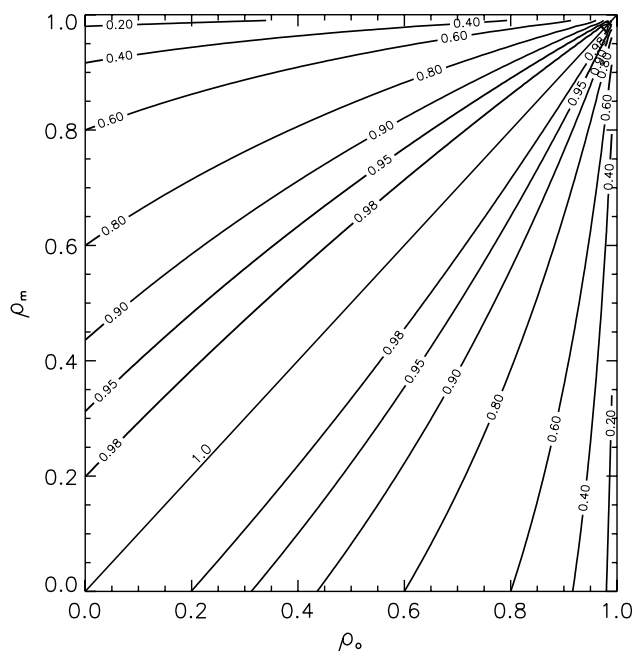
and

$$A_m = \frac{m_{t+1} + m_t}{2} = \tfrac{1}{2}(m_t(1 + \rho_m) + \varepsilon^m_{t+1}) \tag{36}$$

we can write the covariance as

$$\text{cov}(A_o, A_m) = \tfrac{1}{4}(1 + \rho_o)(1 + \rho_m)\text{cov}(o_t, m_t) + \tfrac{1}{4}\text{cov}(\varepsilon^o_{t+1}, \varepsilon^m_{t+1}) \tag{37}$$

replacing for the terms developed in the previous section we get

Fig. 5 Factor multiplying the correlation of the noise present in an autoregressive case for the correlation discussed in (34) from Appendix 3

$$\text{cov}(A_o, A_m) = \tfrac{1}{4}(1 + \rho_o)(1 + \rho_m)\sigma_o\sigma_m r_\varepsilon$$
$$\frac{\sqrt{1 - \rho^2_o}\sqrt{1 - \rho^2_m}}{(1 - \rho_o \rho_m)}$$
$$+ \tfrac{1}{4}r_\varepsilon \sqrt{1 - \rho^2_o}\sqrt{1 - \rho^2_m}\sigma_o\sigma_m \tag{38}$$

and then

$$\text{cov}(A_o, A_m) = \tfrac{1}{4}r_\varepsilon \sigma_o \sigma_m \sqrt{1 - \rho^2_o}\sqrt{1 - \rho^2_m}\left(\frac{2 + \rho_m + \rho_o}{1 - \rho_o \rho_m}\right) \tag{39}$$

To continue from here we need to write the variance of the average as a function of the original time series. Then applying the variance operator on (35) we get
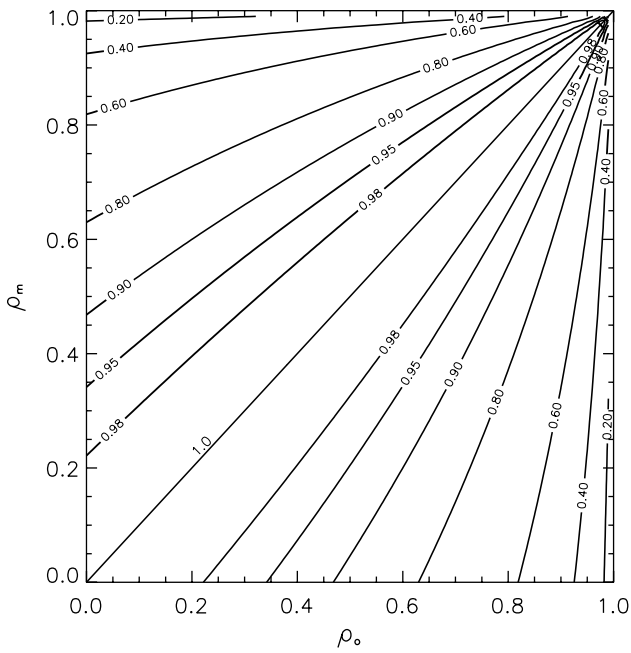
$$\text{var}(A_0) = \tfrac{1}{4}\text{var}\left[(o_t(1 + \rho_o) + \varepsilon^o_{t+1})\right] \tag{40}$$
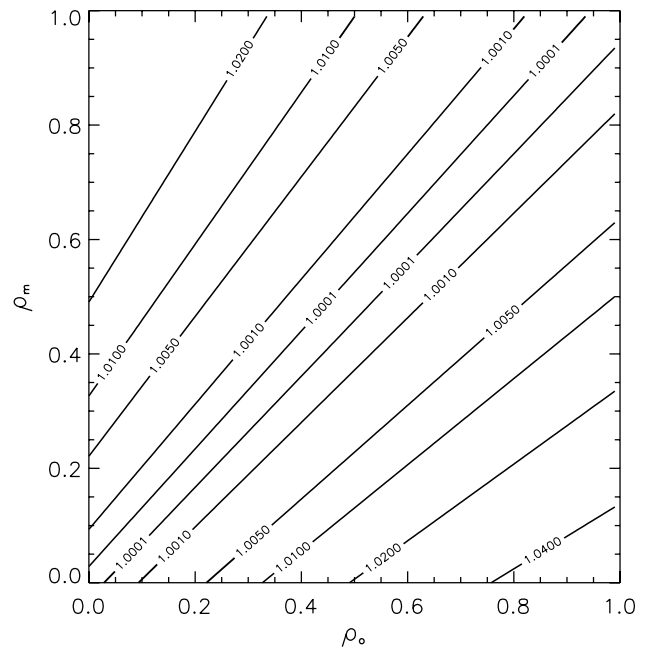
and then

$$\text{var}(A_0) = \tfrac{1}{4}(1 + \rho_o)^2 \text{var}(o_t) + \tfrac{1}{4}\text{var}(\varepsilon^o_{t+1}) \tag{41}$$

Using (31) we can finally write the previous expression as

$$\text{var}(A_0) = \tfrac{1}{2}(1 + \rho_o)\text{var}(o_t) \tag{42}$$

Fig. 6 Factor multiplying the correlation of the noise present in an autoregressive case for the upscaled correlation as defined in expression (44) from Appendix 3

Fig. 7 Factor multiplying the correlation of the time series present in an autoregressive case for the upscaled correlation as defined in expression (45) from Appendix 3

A similar expression is also valid for A_m.

Returning now to (39), operating on the left-hand side using (42) we get

$$\text{cov}(A_o, A_m) = corr(A_o, A_m)\sigma_{A_o}\sigma_{A_m} = \frac{1}{\sqrt{2}}\frac{1}{\sqrt{2}}\sigma_o\sigma_m\sqrt{1+\rho_o}\sqrt{1+\rho_m}$$
(43)

Then replacing in (39) we get

$$corr(A_o, A_m) = \frac{1}{2}r_\varepsilon\sqrt{1-\rho_o}\sqrt{1-\rho_m}\left(\frac{2+\rho_m+\rho_o}{1-\rho_o\rho_m}\right)$$
(44)

The factor multiplying r_ε is smaller or equal than unity as can be seen in Fig. 6.

This can also be rewritten as a function of the correlation of the time series using (33) so that

$$corr(A_o, A_m) = corr(o_t, m_t)\frac{1}{2}\left(\frac{2+\rho_m+\rho_o}{\sqrt{1+\rho_o}\sqrt{1+\rho_m}}\right)$$
(45)

Here the factor is larger than unity, but marginally so as can be seen in Fig. 7. From here we can see that the $corr(A_o, A_m) \approx corr(o_t, m_t)$ seems to be a good approximation of (45).

The same result can be obtained using a multivariate approach (not shown).
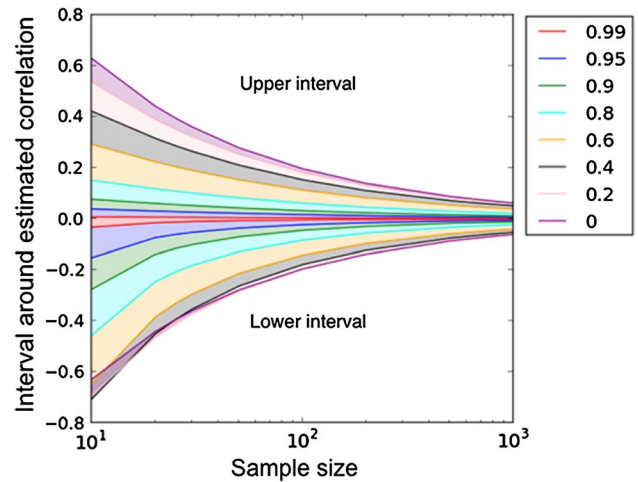
Fig. 8 Confidence intervals for the Pearson correlation. The *horizontal axis* represents the sample size in logarithmic scale. The ordinate represents the upper 95 % confidence interval (when values are positive), and the lower 95 % confidence interval (when values are negative). Since these values are dependent on the actual correlation value, intervals are defined for different correlation values (in *color*)

Appendix 4

Here we assume that both time series—that of the observations and that of the model—consist in a random noise

superimposed on a sinusoidal signal, and may be written as

$$o_t = \sigma_{Ro}\varepsilon_t^o + \sqrt{2}\sigma_{So}\cos(wt) \qquad (46)$$

and

$$m_t = \sigma_{Rm}\varepsilon_t^m + \sqrt{2}\sigma_{Sm}\cos(wt), \qquad (47)$$

where σ_R^2 and σ_S^2 represent the variances of the random term and of the sinusoidal signal, respectively. For the sake of shortening the expressions we write it as a function of the angular frequency $w = 2\pi/T$, where T is the period of the signal (in years, for example). The random variables ε have zero mean and unit variance.

In the following we will assume that parameters in both time series are equal, differing only on the instantaneous values of the random noise. We will also assume this time for the sake of simplicity that there is no correlation between the variables ε.

Correlation of original series

With these definitions we can estimate the variance of each time series and we obtain

$$\mathrm{var}(o_t) = \sigma_R^2 + \sigma_S^2 \qquad (48)$$

and

$$\mathrm{var}(m_t) = \sigma_R^2 + \sigma_S^2. \qquad (49)$$

The covariance between both times series can be written as

$$\mathrm{cov}(o_t, m_t) = \sigma_R^2 \mathrm{cov}(\varepsilon_t^o, \varepsilon_t^m) + 2\sigma_S^2 \mathrm{cov}[\cos(wt), \cos(wt)] \qquad (50)$$

Also knowing that the variance of a sinusoidal function is ½, and assuming lack of correlation between noise instantaneous values we obtain that

$$\mathrm{cov}(o_t, m_t) = \sigma_S^2 \qquad (51)$$

and using the definition of correlation that

$$corr(o_t, m_t) = \frac{\sigma_S^2}{\sigma_R^2 + \sigma_S^2} \qquad (52)$$

Correlation of averaged series

With this information we can now try to estimate the correlation for the filtered or averaged time series, defined as in the previous appendices as

$$A_0 = \frac{o_{t+1} + o_t}{2} \quad \text{and} \quad A_m = \frac{m_{t+1} + m_t}{2}$$

we obtain that

$$A_o = \tfrac{1}{2}\sigma_R(\varepsilon_t^o + \varepsilon_{t+1}^o) + \frac{\sqrt{2}}{2}\sigma_S[\cos(wt) + \cos(w(t + \Delta t)] \qquad (53)$$

and

$$A_m = \tfrac{1}{2}\sigma_R(\varepsilon_t^m + \varepsilon_{t+1}^m) + \frac{\sqrt{2}}{2}\sigma_S[\cos(wt) + \cos(w(t + \Delta t)] \qquad (54)$$

Applying the variance operator to (53) we get

$$\mathrm{var}(A_o) = \tfrac{1}{4}\sigma_R^2\mathrm{var}(\varepsilon_t^o + \varepsilon_{t+1}^o) + \tfrac{2}{4}\sigma_S^2\mathrm{var}[\cos(wt) + \cos(w(t + \Delta t)] \qquad (55)$$

and then

$$\mathrm{var}(A_o) = \tfrac{1}{2}\sigma_R^2\mathrm{var}(\varepsilon_t^o)$$
$$+ \tfrac{2}{4}\sigma_S^2\{\mathrm{var}[\cos(wt)] + \mathrm{var}[\cos(w(t + \Delta t)]]$$
$$+ 2\mathrm{cov}[\cos(wt), \cos(w(t + \Delta t)]\} \qquad (56)$$

Simplifying we obtain

$$\mathrm{var}(A_o) = \tfrac{1}{2}\sigma_R^2 + \tfrac{1}{2}\sigma_S^2\{1 + corr[\cos(wt), \cos(w(t + \Delta t)]\} \qquad (57)$$

and finally

$$\mathrm{var}(A_o) = \tfrac{1}{2}\sigma_R^2 + \tfrac{1}{2}\sigma_S^2[1 + \cos(w\Delta t)] \qquad (58)$$

An identical result is found for A_m.

Now, the covariance between the two averaged time series can be written from (53) and (54) as

$$\mathrm{cov}(A_o, A_m) = \tfrac{1}{4}\sigma_R^2\mathrm{cov}(\varepsilon_t^o + \varepsilon_{t+1}^o, \varepsilon_t^m + \varepsilon_{t+1}^m)$$
$$+ 2\tfrac{1}{4}\sigma_S^2\mathrm{cov}\{\cos(wt) + \cos[w(t + \Delta t)],$$
$$\cos(wt) + \cos[w(t + \Delta t)]\} \qquad (59)$$

then as

$$\mathrm{cov}(A_o, A_m) = 2\tfrac{1}{4}\sigma_S^2(2\mathrm{cov}\{\cos(wt), \cos(wt)\}$$
$$+ 2\mathrm{cov}\{\cos(wt), \cos[w(t + \Delta t)]\}) \qquad (60)$$

Using that the correlation of two sinusoidal waves can be written as a cosine of the difference in phase we get

$$\mathrm{cov}(A_o, A_m) = \tfrac{1}{2}\sigma_S^2(1 + \cos(w\Delta t)) \qquad (61)$$

and finally using (58) as

$$corr(A_o, A_m) = \frac{\sigma_S^2(1 + \cos(w\Delta t))}{\sigma_R^2 + \sigma_S^2(1 + \cos(w\Delta t))} \qquad (62)$$

Writing as function of the original correlation we get

$$corr(A_o, A_m) = corr(o, m) \frac{\sigma_R^2 + \sigma_S^2}{\sigma_S^2} \frac{\sigma_S^2(1 + \cos(w\Delta t))}{\sigma_R^2 + \sigma_S^2(1 + \cos(w\Delta t))}$$
(63)

and finally as

$$corr(A_o, A_m) = corr(o, m) \frac{(\sigma_R^2 + \sigma_S^2)(1 + \cos(w\Delta t))}{\sigma_R^2 + \sigma_S^2(1 + \cos(w\Delta t))}$$
(64)

Extracting the variance of the signal we obtain

$$corr(A_o, A_m) = corr(o, m) \left[1 + \frac{\sigma_R^2/\sigma_S^2 \cos(w\Delta t)}{\sigma_R^2/\sigma_S^2 + 1 + \cos(w\Delta t)} \right]$$
(65)

The factor on the right-hand side is larger than 1 for short time interval increments with respect to the period. Table 1 gives a sense of the importance of this factor for long periods fully encompassed in the time series.

Appendix 5

Distribution of the estimated correlation coefficient for bivariate data

Fisher (1915) derived the explicit distribution of the estimate ($\hat{r}$) of the correlation coefficient (r). Since the form of this distribution is rather complex, it is sufficient here to recall the expectation and variance of the estimated correlation coefficient which are respectively given by

$$E(\hat{r}) = r \left\{ 1 - \frac{(1 - r^2)}{2n} \right\},$$
(66)

and by

$$\text{var}(\hat{r}) = \frac{(1 - r^2)^2}{n} \left(1 + \frac{11}{2n} r^2 + \cdots \right).$$
(67)

It is thus clear that the variance decreases as a function of n, the number of independent replicates of the bivariate variables.

Confidence interval for the estimation of correlation

The Pearson correlation coefficient r between two populations is usually estimated by means of two random samples. This estimation can be associated to an error bar, the same way that is the case with other statistical estimations.

Since in this work we are interested in changes in correlations, it is crucial to know whether changes are within the expected error.

The Fisher transform can be used to this aim (e.g. von Storch and Zwiers 1999). Figure 8 displays how confidence intervals vary with sample size and correlation value. The horizontal axis represents the sample size in logarithmic scale, the ordinate represents the upper 95 % confidence interval (when values are positive), and the lower 95 % confidence interval (when values are negative). Since these values are dependent on the actual correlation value, intervals are defined for different correlation values (in color). Note how the interval of confidence tightens for a given sample size for increasing values of r. Similarly, decreasing sample size for a given r diminishes confidence in a single estimation and makes the upper and lower intervals more symmetric. This last case is very relevant in the issue of upscaling or aggregation since these operations may reduce sample size. For example, a time series of monthly anomalies is three times longer than one of seasonal anomalies.

References

Alexandru A, de Elia R, Laprise R, Separovic L, Biner S (2009) Sensitivity study of regional climate model simulations to large-scale nudging parameters. Mon Weather Rev 137:1666–1686. doi:10.1175/2008MWR2620.1

Anthes R (1983) Regional models of the atmosphere in middle latitudes. Mon Weather Rev 111:1306–1335

Bougeault P (2002) WGNE survey of verification methods for numerical prediction of weather elements and severe weather events. CAS/JSC WGNE Report No. 18, Appendix C. WMO/TD.No.1173, Toulouse, France

Brands S, Gutiérrez J, Herrera S (2012) On the use of reanalysis data for downscaling. J Clim 25:2517–2526

Dee DP, Uppala SM, Simmons AJ et al (2011) The ERA interim reanalysis: configuration and performance of the data assimilation system. Q J R Meteorol Soc 656:553–597

Eden JM, Widmann M, Grawe D, Rast S (2012) Skill, correction and downscaling of GCM-simulated precipitation. J Clim 25:3970–3984

Fisher RA (1915) Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. Biometrika 10:507–521

Harris I, Jones PD, Osborn TJ, Lister DH (2013) Updated high-resolution grids of monthly climatic observations–the CRU TS3.10 dataset. Int J Climatol. doi:10.1002/joc.3711

Jakimow G, Yakimiw E, Robert A (1992) An implicit formulation for horizontal diffusion in gridpoint models. Mon Weather Rev 120:124–130

Johnson RA, Wichern DW (2007) Applied multivariate statistical analysis, 6th edn. Prentice Hall, New York

Kanamaru H, Kanamitsu M (2007) Fifty-Seven-year California reanalysis downscaling at 10 km (CaRD10). Part II: comparison with North American regional reanalysis. J Clim 20:5572–5592. doi:10.1175/2007JCLI1522.1

Li H, Kanamitsu M, Hong S-Y (2012) California reanalysis downscaling at 10 km using an ocean-atmosphere coupled regional

model system. J Geophys Res 117:1–16. doi:10.1029/2011JD017372

Martynov A, Laprise R, Sushama L, Winger K, Separovic L, Dugas B (2013) Reanalysis-driven climate simulation over CORDEX North America domain using the Canadian Regional Climate Model, version 5: model performance evaluation. Clim Dyn 41:2973–3005. doi:10.1007/s00382-013-1778-9

Mass CF, Ovens D, Westrick K, Colle BA (2002) Does increasing horizontal resolution produce more skillful forecasts? Bull Am Meteorol Soc 83(3):407–430

Mekis E, Vincent LA (2011) An overview of the second generation adjusted daily precipitation dataset for trend analysis in Canada. Atmos Ocean 49:163–177

Separovic L, Elía R, Laprise R (2011) Impact of spectral nudging and domain size in studies of RCM response to parameter modification. Clim Dyn 38:1325–1343. doi:10.1007/s00382-011-1072-7

Stefanova L, Misra V, Chan S, Griffin M, O'Brien JO, Smith TJ (2011) A proxy for high-resolution regional reanalysis for the Southeast United States: assessment of precipitation variability in dynamically downscaled reanalyses. Clim Dyn 38:2449–2466. doi:10.1007/s00382-011-1230-y

Vincent LA, Wang XL, Milewska EJ, Wan H, Yang F, Swail V (2012) A second generation of homogenized Canadian monthly surface air temperature for climate trend analysis. J Geophys Res 117:D18110. doi:10.1029/2012JD017859

von Storch H, Zwiers FW (1999) Statistical analysis in climate research. Cambridge University Press, Cambridge, UK/New York

von Storch H, Langenberg H, Feser F (2000) A spectral nudging technique for dynamical downscaling purposes. Mon Weather Rev 128:3664–3673

Weisse R, von Storch H, Callies U, Chrastansky A, Feser F, Grabemann I, Gunther H, Pluess A, Stoye T, Tellkamp J, Winterfeldt J, Woth K (2009) Regional meteorological–marine reanalyses and climate change projections. Bull Am Meteorol Soc 90:849–860. doi:10.1175/2008BAMS2713.1