Electronic supplement - Statistical Framework for evaluation of climate model simulations by use of climate proxy data from the last millennium

Alistair Hind, Anders Moberg and Rolf Sundberg

December 13, 2011

Grid-box autocorrelation

Accounting for the changing area of the grid-boxes with latitude, the proportion of significant lag-1 autocorrelations were calculated over the global land surface that exceed the proportion expected for white noise at the 0.05 (two-sided) significance level (and 0.025 for one-sided). These proportions of grid-boxes are shown in Figure 1, with the black points and fitted line representing the two-sided test and the red and blue points/lines representing the positive and negative one-sided tests respectively. If the CTRL simulation consists of white noise, we should expect to see the black curve lying at 0.05 and the blue/red curves at 0.025. At short time-resolutions, the proportion of grid-boxes with significant lag-1 autocorrelations clearly exceeds that which could be expected from white noise. We see a saturation after approximately 25 years in the lag-1 autocorrelation of the CTRL simulation. This suggests that δ_i (where i is at the grid-box level) can be thought of as white noise for averages of 25 years or larger, in keeping with the statistical model 2 (Sect. 2 of the article). It is known that on larger spatial scales, particularly in the case of the oceans, long-term persistence exists in the climate system. Naturally when the radiative forcings are included in the simulations we can expect the long-term scaling to be far larger if the forcing time-series exhibit that kind of long-term behaviour (Rybski et al., 2008).

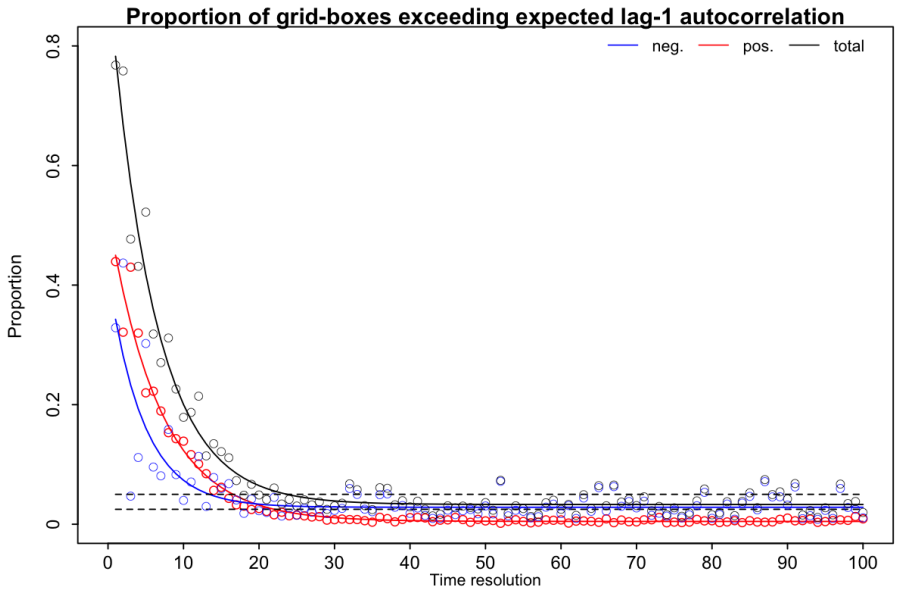

Figure 1: Area-weighted proportion of land-only grid-boxes in the entire 3000-yr CTRL simulation where the lag-1 autocorrelation exceeds the proportion expected from white noise at the 0.05 (two-sided) and 0.025 (one-sided) significance levels. The colours represent positive (red) and negative (blue) autocorrelation and black for combined.

Gobal average autocorrelation

To compliment the grid-box autocorrelation investigation above, the global average land-only lag-1 autocorrelation was calculated for the same time-resolutions (Figure 2). Here the lag-1 autocorrelation becomes insignificant (at the 0.05 level) after a time resolution of 12 or so years. This confirms

that δ_i can certainly be thought of as white noise for the global portion of our analysis on a 30-year time resolution.

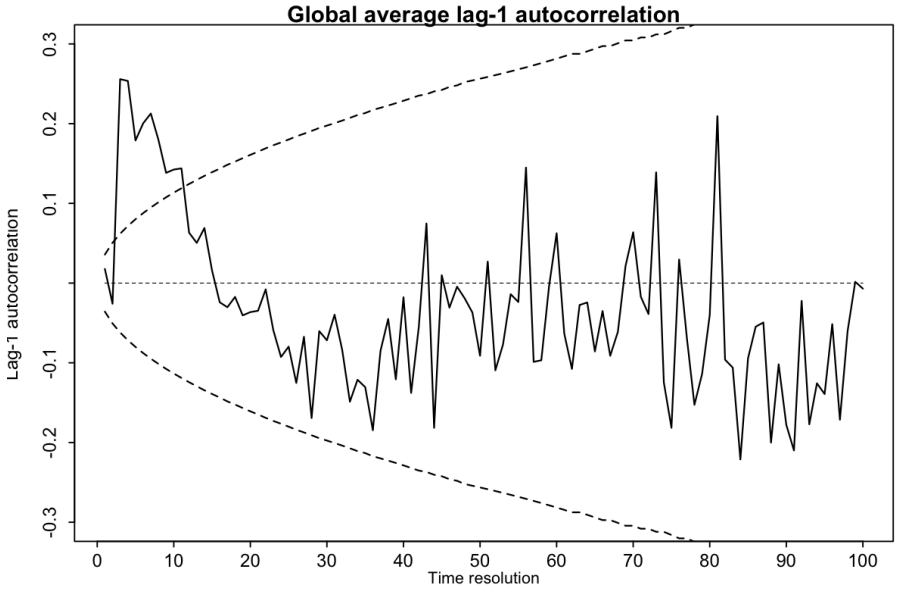

Figure 2: Lag-1 autocorrelation of the land-only global average 3000-yr CTRL simulation against changing time resolution. The curved dashed lines represent the 0.05 (two-sided) significance levels.

Higher and lower noise levels for global analysis

As described in Sect. 9.5 and 10 of the article, the experiments were also conducted for lower and higher noise levels. Figure 3 shows the correlation U_R (*top*) and distance U_T (*bottom*) measures for the E1 (blue) and E2 (red) full-forcing simulations for different % coverage levels. The details of the Figure 3 analysis match those of Sect. 9.5 except that the ≈ 100 noise realizations are for different noise levels in the pseudo-proxies. It should be noted that the pseudo-instrumental data has the same properties as in the main analysis. Figure 3 is with a lower SNR = 0.25 compared with a SNR = 0.71 in the main analysis. The E1 and E2 U_R values show significant correlations, as with the results of Fig. 7, when both E1 (Figure 3a) and E2 (Figure 3b) serve as targets. However, in contrast to the Fig. 7 results, U_T is generally no better than the CTRL simulation when E1 serves as target (Figure 3c). This reflects how the increased noisiness of these pseudo-proxies renders the low solar forcing less detectable than when using the high quality pseudo-proxies of the main analysis. When E2 serves as target, the high solar E2 simulations can still be ranked appropriately using U_T (Figure 3d). Figure 4 is as Figure 3, but for negligible noise levels in the pseudo-proxies. As should be expected given the U_R results for the higher noise cases (Figure 3a and 3b), all full-forcing ensemble U_R values are significantly correlated with the E1 and E2 targets (Figure 4a and 4b). When E1 serves as target, it is clear that the E2 simulations are detectably worse than the CTRL simulation using U_T (Figure 4c). This cannot be concluded so easily from the comparatively higher noise pseudo-proxies used in Fig. 7 in the article. When E2 serves as target, the E2 U_T values are clearly of a higher rank than those of the E1 simulations (Figure 4d). Note how little the U_T values change for different coverage levels in Figure 4.

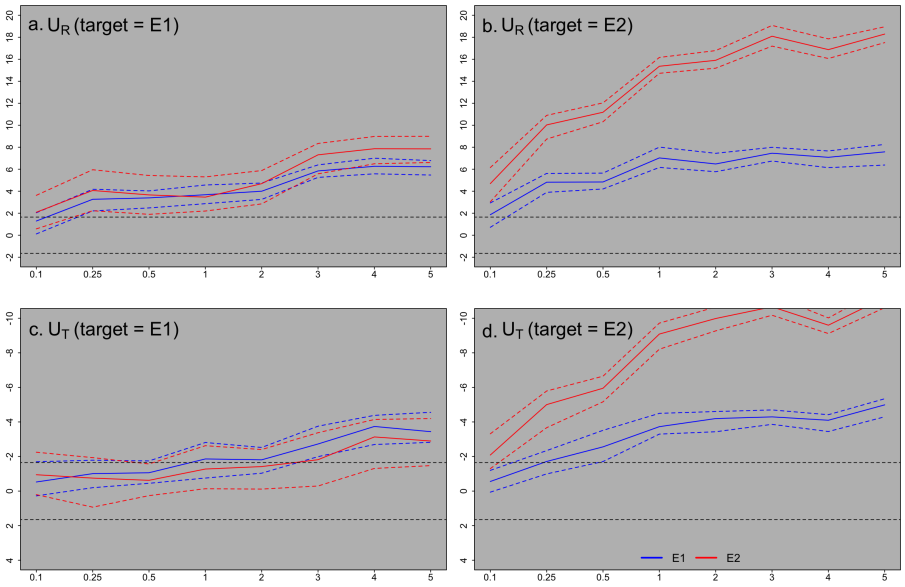

Figure 3: U_R correlation (*top*) and U_T distance (*bottom*) measures for the E1 (blue) and E2 (red) simulations with an SNR = 0.25. The *left* panels are for E1 as target, *right* are E2 as target. The 5% significance level is shown with dashed lines. The filled coloured lines denote the median value, with the dashed coloured lines representing the upper and lower quartiles.

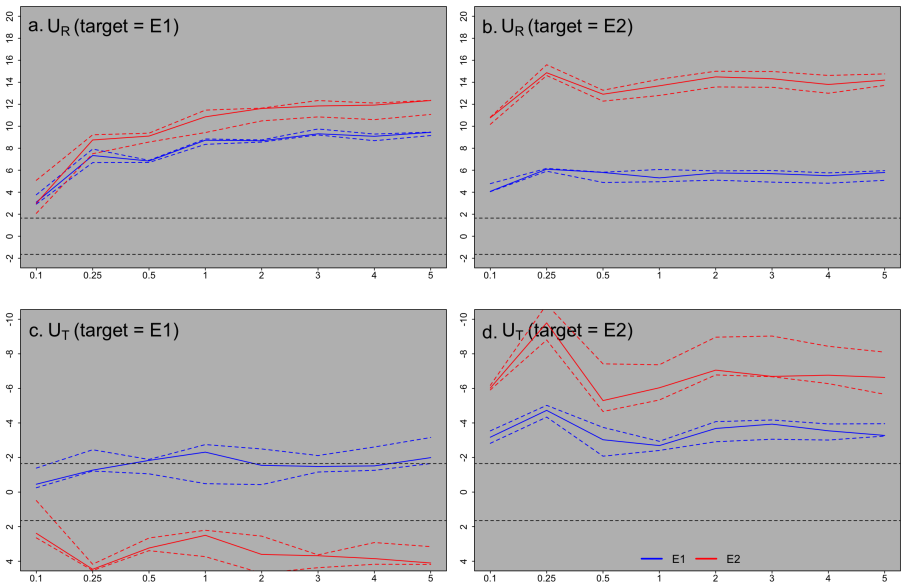

Figure 4: As above, but for negligible noise.

Bibliography

Rybski, D., Bunde, A., and von Storch, H.: Long-term memory in 1000-year simulated temperature records, Journal of Geophysical Research, 113, D02 106 1–9, doi:10.1029/2007JD008568, 2008.