NOTES AND CORRESPONDENCE

Stationarity of Regression Relationships: Application to Empirical Downscaling

TORBEN SCHMITH

Danish Meteorological Institute, Copenhagen, Denmark

(Manuscript received 21 February 2007, in final form 7 January 2008)

ABSTRACT

The performance of a statistical downscaling model is usually evaluated for its ability to explain a large fraction of predictand variance. In this note, it is shown that although this fraction may be high, the longest time scales, including trends, may not be explained by the model. This implies that the model is nonstationary over the training period of the model, and it questions the basic stationarity assumption of statistical downscaling. This is exemplified by using a simple regression model for downscaling European precipitation and surface temperature where appropriate Monte Carlo–based field significance tests are developed, taking into account the intercorrelation between predictand series. Based on this test, it is concluded that care is needed in selecting predictors to avoid this form of nonstationarity. Even though this is illustrated for a simple regression-type statistical downscaling model, the main conclusions may also be valid for more complicated models.

1. Introduction

The best possible projections of future climate are needed for developing optimal mitigation and adaptation strategies in response to anthropogenic climate changes. The widely used tool for achieving these projections is general circulation models (GCMs), which are forced with anticipated increases (scenarios) in the concentrations of anthropogenic greenhouse gases, aerosols, and other forcings.

However, GCMs of the present generation have a horizontal resolution in the order of several hundred kilometers; therefore, the large-scale flow is usually well described, whereas the detailed flow around smaller-scale topographic features, like mountains or land–sea contrasts, is generally inadequately described. In addition, parameters requested by the climate–impact community, like precipitation, surface temperature, and wind, are regarded as less reliable because they are parameterized in the model rather than physically modeled (see Christensen et al. 2007 for a more extensive discussion).

Therefore, to make a projection of these latter variables at smaller spatial scales, downscaling techniques are applied. One possibility is the dynamical downscaling technique in which a model with fine spatial resolution (presently down to 5 km) is applied to the region of interest and “nested” into the coarse-scale GCM. One shortcoming of this method is the expensive computing costs.

Empirical or statistical downscaling is a cost-effective alternative technique. The underlying idea is that, as previously mentioned, variables related to the free atmospheric circulation (geopotential heights and tropospheric temperatures) are usually well simulated by the GCM. These variables are regarded as independent variables (predictors), and they are statistically related to the dependent variables (predictands), for example, surface temperature or precipitation at a given site, which are not well simulated by the GCM. The rationale behind this is that, for example, western European wintertime precipitation is strongly linked to the atmospheric circulation over the Atlantic–European area at 500-hPa height (Wibig 1999) or sea level (Zorita et al. 1992).

The statistical relationship between predictor and predictand variables could be of any complexity ranging from multilinear regression to nonlinear models,
like neural networks or analog methods (e.g., Karl et al. 1990; von Storch et al. 1993; Murphy 1999). The different techniques are all applied in similar ways; we will use a simple multilinear regression as an illustration.

First, the parameters in the statistical relationship (the regression coefficients in a linear regression model) are determined from an observed time series of the predictors and predictand in the training period. We can also get an estimated time series of the predictand from the model in the training period. The model skill is commonly evaluated by calculating measures of the interannual covariability between observed and estimated predictand time series like the multiple correlation coefficient or the fraction of variance explained (e.g., Busuioc et al. 1999). This step also includes some kind of screening procedure to identify significant predictors from a predefined list of potential predictors. The criterion for including a predictor will often be its ability to significantly increase the model skill. At this point, it is important to stress that for real applications, an evaluation of a downscaling model should be done on data outside the training period. However, for the illustrative purpose of this note, this is not important, and both training and evaluation are done on the entire data interval.

Second, this predictor–predictand relationship is applied to time series of predictors from selected periods of GCM runs, which are usually several decades long and represent present and future climate, from which we then obtain the corresponding predictand time series. The difference between the averages of the “future” and “present” time series is the climate change signal for that predictand at the given site.

What is a good statistical downscaling model? As previously described, the usually applied measure is the multiple correlation coefficient. However, this is not sufficient. As described in basic textbooks on regression analysis (e.g., Sen and Srivastava 1990) the predictand residuals (i.e., observed minus estimated values of the predictand) must be without any structure. In particular, if the residuals have a significant trend over the training period, it is a sign of a missing linearly increasing predictor, or put differently, the relationship between predictors and predictand is nonstationary. But if the relationship between predictors and predictand changes over time in the training period, we have no reason to believe in an unchanged relationship in a future climate. Therefore, nonstationarity over the training period conflicts with the basic stationarity assumption (Wilby 1997) of statistical downscaling, which states that the relationship between predictor and predictand is unchanged between the present and future climates. Thus, it seems reasonable to regard stationarity of the predictor–predictand relationship over the training period as a necessary (but not sufficient) condition for a satisfactory statistical downscaling model.

In this note, we will demonstrate that the stationarity of the predictor–predictand relationship over time is violated for certain choices of predictor and predictand variables by using a simple multilinear regression model for downscaling European winter temperature and precipitation. In addition, nonparametric Monte Carlo–based field significance tests will be developed, taking into account intercorrelations between station series.

2. Model and analysis

Let us investigate the above points using a simple linear regression model relating average surface temperature and total precipitation anomalies (predictands) for the winter season (October–March) to circulation anomalies at different levels (predictors) over the period 1948–99.

Predictor time series are taken from stations in the European Climate Assessment (ECA) dataset (Klein Tank et al. 2002). This dataset consists of daily series of temperature and precipitation that have been quality controlled and tested for homogeneity (Wijngaard et al. 2003). Only series with more than 95% high-quality data coverage and that have been classified as “useful” based on the applied homogeneity tests are used. In total, temperature data from 41 stations and precipitation data from 169 stations are used in the study (see map of stations on Fig. 1). Daily series are aggregated into monthly values, which are then converted to monthly anomalies, and then finally converted into winter anomalies.

Predictors are based on geopotential height fields taken from the National Centers for Environmental Prediction–National Center for Atmospheric Research (NCEP–NCAR) reanalysis (Kalnay et al. 1996; Kistler et al. 2001) in 2.5° × 2.5° resolution over the area (20°–80°N, 60°E–80°W). As for the predictands, monthly anomalies and subsequently winter anomalies are calculated. To overcome collinearity between neighboring points in the predictor field, these are resolved into principal components (PCs), and each predictor field is truncated to a number of significant principal components determined by Preisendorfer’s Rule N. The time series, , correspond to these principal components, which are the potential predictors in the downscaling model. Mathematically, this “full” model for each predictand series can be written as

\[ y_i = b_0 + \sum_{t=1}^{N} b_t a_{ti} + n_i, \]
FIG. 1. Map showing locations of (top) precipitation and (bottom) surface temperature stations.
where \( y_i \) is the observed predictand (winter temperature or precipitation anomaly) for a given station, \( b_0 \) and \( b_i, i = 1, \ldots, N \) are parameters to be determined, and \( \epsilon_i \) is the noise term.

The potential predictors are screened using “forward selection,” and a subset \( II \) consisting of \( M \) predictor PCs is kept. Once the coefficients in the model are determined, the estimated predictand series during the training period can be calculated as

\[
\hat{y}_i = b_0 + \sum_{i=1}^{N} b_i \alpha_i^{(i)}.
\]

The performance of the model is evaluated by the correlation coefficient between the observed and estimated predictand series.

The residual time series is calculated as

\[
r_i = y_i - \hat{y}_i
\]

For each station, the trend of the residual series over the training period is calculated as the linear slope using an ordinary least squares fit. Even for serially correlated series, this estimator has an efficiency of above 0.8 (Prais and Winsten 1954; Chipman 1979). Subsequently, the lag 1 sample autocorrelation function of the detrended residuals is calculated and has absolute values below 0.15, which shows that serial correlation of the residuals in any case is small. In addition, the trend is alternatively calculated using the nonparametric Theil–Sen trend estimator (Theil 1950).

<table>
<thead>
<tr>
<th>Expt No.</th>
<th>Predictors</th>
<th>Predictand</th>
<th>Unadjusted</th>
<th>Adjusted</th>
<th>0.25/0.75 fractile</th>
<th>Two-sided residual trend</th>
<th>Two-sided residual trend</th>
<th>Two-sided residual trend</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1000-hPa height (S PCs) + year</td>
<td>Precipitation</td>
<td>0.440071</td>
<td>0.440071</td>
<td>0.75</td>
<td>0.08</td>
<td>0.76</td>
<td>0.35</td>
</tr>
<tr>
<td>2</td>
<td>1000-hPa height (S PCs) + 850-hPa height (S PCs)</td>
<td>Precipitation</td>
<td>0.440071</td>
<td>0.440071</td>
<td>0.75</td>
<td>0.08</td>
<td>0.76</td>
<td>0.35</td>
</tr>
<tr>
<td>3</td>
<td>1000-hPa height (S PCs) + 850-hPa height (S PCs)</td>
<td>Precipitation</td>
<td>0.440071</td>
<td>0.440071</td>
<td>0.75</td>
<td>0.08</td>
<td>0.76</td>
<td>0.35</td>
</tr>
<tr>
<td>4</td>
<td>1000-hPa height (S PCs) + 850-hPa height (S PCs)</td>
<td>Precipitation</td>
<td>0.440071</td>
<td>0.440071</td>
<td>0.75</td>
<td>0.08</td>
<td>0.76</td>
<td>0.35</td>
</tr>
<tr>
<td>5</td>
<td>1000-hPa height (S PCs) + 850-hPa height (S PCs)</td>
<td>Precipitation</td>
<td>0.440071</td>
<td>0.440071</td>
<td>0.75</td>
<td>0.08</td>
<td>0.76</td>
<td>0.35</td>
</tr>
<tr>
<td>6</td>
<td>1000-hPa height (S PCs) + 850-hPa height (S PCs)</td>
<td>Precipitation</td>
<td>0.440071</td>
<td>0.440071</td>
<td>0.75</td>
<td>0.08</td>
<td>0.76</td>
<td>0.35</td>
</tr>
<tr>
<td>7</td>
<td>1000-hPa height (S PCs) + 850-hPa height (S PCs)</td>
<td>Precipitation</td>
<td>0.440071</td>
<td>0.440071</td>
<td>0.75</td>
<td>0.08</td>
<td>0.76</td>
<td>0.35</td>
</tr>
<tr>
<td>8</td>
<td>1000-hPa height (S PCs) + 850-hPa height (S PCs)</td>
<td>Precipitation</td>
<td>0.440071</td>
<td>0.440071</td>
<td>0.75</td>
<td>0.08</td>
<td>0.76</td>
<td>0.35</td>
</tr>
<tr>
<td>9</td>
<td>1000-hPa height (S PCs) + 850-hPa height (S PCs)</td>
<td>Precipitation</td>
<td>0.440071</td>
<td>0.440071</td>
<td>0.75</td>
<td>0.08</td>
<td>0.76</td>
<td>0.35</td>
</tr>
</tbody>
</table>

The table shows the results from the linear downscaling model. Column 1 shows the expt No., in columns 2 and 3, predictors and predictand are given. Columns 4 and 5 give the range (0.25 and 0.75 quantiles) of correlation coefficient between observed and hindcasted station precipitation (unadjusted and adjusted, respectively). Columns 6 and 7 display the fraction of stations with a positive residual trend and the associated p value. This p value is calculated using a Monte Carlo simulation as explained in the text. Columns 8 and 9 correspond to columns 6 and 7, but where the residual trend is calculated with the nonparametric Theil–Sen estimator. Columns 10 and 11 show corresponding results for the differential intercept.
with positive trends is a stochastic variable whose statistical distribution we can approximately determine by carrying out this procedure 10,000 times. Experimentation gives robust results from which the two-sided \( p \) value corresponding to the fraction calculated from the directly observed series can be calculated. Field significance is calculated along these lines based both on the least squares and the Theil–Sen trend estimator.

In addition, a test for a structural break in the predictor–predictand relationship (Chow 1960) is applied for each series using record splitting. A residual trend in a series is equivalent to a change in the intercept (the constant) in the predictor–predictand relationship. Therefore, each series is split into two segments of equal length by separating them at year 1973 and tested to determine whether the difference between the two intercepts obtained by running the regression on the two segments individually, the differential intercept, is positive. This is most easily done using a dummy variable technique (Gujarati 1970), which directly outputs the differential intercept. The fraction of station series with a positive differential intercept can then be field significance tested by calculating the \( p \) value using a similar bootstrapping procedure as described above.

3. Results

The model described above is used with different combinations of predictors and predictands as given in Table 1. In experiment 1 the predictor is 1000-hPa geopotential height, and the predictand is precipitation. The model has a satisfactory overall performance as measured by the multiple correlation coefficients between observed and estimated predictand, which are in the range 0.44 to 0.71 (see Fig. 2). The corresponding “adjusted” values in which unduly small sample infla-
tion is accounted for (e.g., Sen and Srivastava 1990, p. 40) are in the range 0.41 to 0.69.

As an example, observed and estimated precipitation anomalies are shown in Fig. 3, together with the residual time series for the station Hamburg Fuehlsbuttel, which exhibits a trend in the residuals that is discernible by the eye. The interannual variations are well reproduced by the model, giving an overall multiple correlation coefficient of 0.74.

The fraction of positive trends in the corresponding residual time series is 0.75, calculated using least squares; the geographical distribution of these is shown in Fig. 4 and is 0.76, calculated using the Theil–Sen estimator (geographical distribution not shown). The field stationarity tests described above of the residuals gives a \( p \) value of 0.08 for the least squares method and 0.05 for the Theil–Sen method, meaning that the observed spatial pattern of trends in the residuals is not very likely under the null hypothesis of no overall residual trend. From this we conclude that there is a nonstationarity in the predictor–predictand relationship. The result of the test for differential intercept is a \( p \) value of 0.13. Overall, we conclude from this experiment that a high multiple correlation coefficient does not guarantee that the longest time scales, including trends, are well captured by the actual choice of predictors.

To illustrate that the global test on the fraction of positive residual trends works, an additional experiment, experiment 2, is carried out in which “year” is used as an additional (unphysical) predictor, which we a priori expect will remove the nonstationarity. In this experiment, we get multiple correlation coefficients that are very similar to experiment 1, but with \( p \) values of 0.33 and 0.25 for the two residual trend tests and 0.61 when testing the differential intercept. This illustrates that year added as a predictor removes, as expected, the nonstationarity in the residuals and that the applied tests work. We also note that there is no real difference in the multiple correlation coefficients. The nonstationarity does not show up here, and it illustrates that a high correlation coefficient does not ensure the absence of nonstationarity.

Additional subsequent experiments (3–6) in which PCs of 1000-hPa geopotential height were replaced by 850-, 700-, 500-, and 300-hPa geopotential height as predictors were carried out (see Table 1). The range of multiple correlation coefficients in these experiments is very similar to experiments 1 and 2, but the \( p \) values for the residual and differential intercept tests are much higher than experiment 1, which implies no indication of nonstationarity of the predictor–predictand relationship. These results confirm our previous observation that stationarity versus nonstationarity in the residual trends cannot be judged from the multiple correlation coefficients.

Finally, two experiments (7–8) with PCs of 1000- and 500-hPa geopotential height as predictors and surface temperature as predictand were carried out. The predictor–predictand relationship was found stationary in both of these experiments based on all three field tests. The geographical distribution of the least squares trend in the residuals is shown on Fig. 5.

4. Discussion and summary

Wilby and Wigley (2000) did an extensive and systematic study on the procedures for selecting predictors in a statistical downscaling model based on predictor–predictand correlations. The purpose of the present note is to shed some light on a largely ignored aspect, namely, their ability to describe the longest time scales including trends, which is not ensured by considering correlations only.

Frias et al. (2006) studied the performance of statistical downscaling methods for precipitation in the surrogate climate of a coupled climate model. The advan-
The advantages of this approach are that downscaled values can be directly compared with the "observed" ones and, furthermore, that inhomogeneities in the ingoing predictors–predictand data are nonexistent. The aim of the study is to investigate the use of sea level pressure (SLP) versus 500-hPa geopotential height as predictor and, furthermore, the role of humidity as an additional predictor. The study concludes that SLP performs better as a predictor, which is not in agreement with our results. Interestingly, the study points to humidity predictors in some regions as increasing the skill. We will revert to this later.

We can conclude from our experiments described above that a fair performance of a regression downscaling model, as measured by the multiple correlation coefficients between observed and estimated predictand series over the training period, does not guarantee a good representation of the lowest frequencies including trends. To test whether these longest time scales are well captured by the model, a Monte Carlo–based nonparametric test on the observed fraction of positive residual trends and the observed fraction of positive differential intercepts has been presented. The test yields two-sided p values under the null hypothesis that no overall trend is present in the residuals.

An overall trend in the residuals will indicate a nonstationarity in the predictor–predictand relationship. This nonstationarity could have two causes: nonhomogeneities in either predictors or predictands, or a missing and slowly varying predictor variable. Is it likely that the geopotential height fields (predictors) are inhomogeneous? Despite all efforts to avoid inhomogeneities in the NCEP–NCAR reanalysis project, they do occur (Reid et al. 2001) and because the 1000-hPa surface is more influenced by surface observations than the levels above, it cannot be ruled out that this geo-

![Fig. 4. Trend of residual (observed minus estimated) winter precipitation over the training period for experiment 1 (predictor and predictand same as Fig. 2) using the least squares method. The height of the triangle is proportional to the relative residual trend (i.e., the residual trend divided by the climatological value of the predictand).](image-url)
potential height surface is most inhomogeneous. However, in that case we would have detected nonstationarity with both precipitation (experiment 1) and surface temperature (experiment 7) as predictand and because this is not the case, this possibility is not very likely.

Then there is the issue of predictand series being inhomogenous. An argument against this is that only precipitation and surface temperature series, which have passed all homogeneity tests applied, are included in the study. Furthermore, if inhomogeneities of a substantial fraction of the precipitation series are responsible for the result of the experiment, then we could expect that the outcome of experiments 3–6 with predictors derived from geopotential heights at upper levels would also be nonstationarity, contradicting what we have found.

Therefore, in the likely case that inhomogenities are not the cause for the detected nonstationarity in the predictor–predictand relationship, it is worth discussing what physical causes could be in play. This was investigated by Charles et al. (1999), who found that statistical downscaling models with only circulation variables as predictors did not produce a satisfactory climate change signal (in the surrogate climate of a limited area model) and, furthermore, that relative humidity at 850 hPa increased the performance of the downscaling model. This is also investigated in the present study by experiment 9, which is similar to experiment 1 with PCs of relative humidity at 850 hPa added to the predictor list. The outcome of the field tests from this experiment confirms that the nonstationarity is removed as judged from the residual trend field tests, whereas the result from the differential intercept test is somewhat “in between.”

In conclusion, the present study addresses that something useful may be learned by considering structural changes over the period of observed data when statistical downscaling models are identified and validated.

Fig. 5. Same as Fig. 4 but for experiment 7 (predictor is 1000-hPa geopotential height and predictand is surface temperature using the least squares method).
This may lead to more reliable estimates of climate changes.

Acknowledgments. This work was supported by the European Commission as part of the statistical and regional dynamical downscaling of “extremes for European regions” (STARDEX) Contract EVK2-CT-2001-00115. NCEP reanalysis–derived data were provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado (available online at http://www.cdc.noaa.gov). European Climate Assessment data are available online (http://eca.knmi.nl). The constructive comments received from the three anonymous reviewers are appreciated.

REFERENCES


