Supplementary Information Report

Number 2

Assessment of Empirical Statistical Downscaling for Climate Change Studies over Singapore

Author: Erasmo Buonomo[1]

Met Office internal reviewers: James Murphy[1], Simon Brown[1] and Doug McNeall[1].

1 - Met Office, Exeter, UK

Contents

S2.1. Introduction

Global General Circulation Models (GCMs) are the main tools to generate physically consistent projections of climate change during the 21st century. However, the horizontal resolution typically used in their atmospheric components (~100km) does not allow them to resolve certain processes important for regional climate change, such as the detailed influences of mountains, coastlines and other physiographic features, or to present information at the finer spatial and temporal scales needed for climate impact studies. Regional climate models, even at their current resolution (~10-50km) only solve this problem partially, since quite frequently climate impact studies depend on the availability of local or site-specific climate information. These problems are particularly relevant for coastal regions and islands, which are affected by small scale processes not explicitly resolved even by regional climate models, at their current resolutions.

Empirical Statistical Downscaling (ESD) methods are widely used to generate climate scenarios at local scales (Wilby et al, 2004). These methods are based on statistical relationships which link predictands, i.e. the local scale meteorological variables, to predictors, the large scale variables describing the state of the atmosphere, by capturing the effects of local processes and physiography. These relationships are built from observed local and large scale atmospheric datasets which cover periods sufficiently long to allow a robust estimation of the required statistical parameters. The relationships are then applied to generate predictands for the future climate by using large scale predictors from GCMs.

At least two main assumptions are necessary for these methods: 1) GCMs are able to describe the large scale state of the atmosphere for the region under investigation; 2) the same statistical relationship found for the calibration period in the present climate is also applicable to the future, changed, climate. The first assumption is common to all downscaling methods, since GCMs are the only tools available to generate greenhouse gas forced climate scenarios producing the physically, spatially and temporally consistent realisations of multiple atmospheric variables needed by these methods. The stationarity of the statistical relationship, however, is a relevant issue for empirical statistical methods, since they are trained with datasets which sample the natural variability in the present climate, and yet are expected to describe climate change trends and possible changes in natural variability in the future climate. An assessment of the stationarity of the relationship is one objective of this study.

Different statistical techniques have been used for ESD, broadly classifiable (Wilby et al, 2004) into weather typing approaches, such as the analogue method, regression methods and weather generators, which include a stochastic component. Wilby et al (op. cit) summarise their main strengths and weaknesses. In this study, linear regression methods are applied to estimate statistical relationships between a set of predictands (observed daily data from Meteorological Service Singapore stations) and large scale predictors suitable to describe the atmospheric state for SE Asia. Methods are chosen mainly on the basis of simplicity of use and ability to test the main assumptions required for their applications, in particular from analysis of the residuals from the trained statistical relationships (Wilks, 1995). Simplicity of use, however, does not imply a lack of accuracy: intercomparisons of different statistical approaches used for ESD (e.g. Zorita and Von Storch, 1999, Goodess et al, 2005) have shown that these methods are comparable in accuracy to more complex statistical approaches.

ESD methods have been more widely used to generate downscaled climate scenarios for mid latitude sites than for tropical regions; one reason may be better availability of long records of station variables, however, another likely factor is that local weather at mid latitudes is typically more strongly influenced by the large scale state of the atmosphere, while smaller scale processes, poorly resolved by climate models, are often more important the tropical and equatorial regions. Two studies are particularly relevant in this context: the application of linear regression methods to daily predictands from West African sites (Manzanas et al, 2012), and daily rainfall downscaled by neural networks for a series of tropical locations (Cavazos and Hewitson, 2005). These studies have shown that, in tropical regions, the large scale predictors can explain less of the natural variability of predictands than is the case for mid latitude results, as expected for sites in climatic regions less constrained by large scale atmospheric flow.

The aim of this study is the application of linear regression methods to estimate statistical relationships for daily mean temperature, minimum and maximum temperature, rainfall and wind for sites in Singapore, with the purpose of using them to downscale GCM projections from the CMIP5 project (Taylor et al, 2011). A potential advantage of ESD is that the methods are computationally cheap, and can potentially be used to downscale results from a wider range of CMIP5 scenarios than is feasible using regional climate model simulations, such as those carried out in the dynamical downscaling work within the V2 project. However, application is contingent on a careful assessment of the quality of the statistical relationships used. Here, these are assessed via analysis of statistical model diagnostics from the training period and from the downscaled climate scenarios, in order to understand strengths and limitations of this approach. RCM data is also used, mainly as a proxy of station data to test the assumption of a stationary relationship in the future climate (Murphy, 2000).

The final outcome is an assessment of possibility of using these linear regression relationships to downscale CMIP5 GCMs.

S2.2. Methodology

In this study, linear regression methods are applied to estimate statistical relationships between a set of predictands (observed daily data from MSS stations, Section S2.2.1 below) and large scale predictors, derived from variables suitable to describe the atmospheric state for SE Asia (Sections S2.2.1 and S2.2.2). In its simplest form, a predictand is expressed as the sum of a linear combination of predictors and a noise term. By minimising the noise term, estimators for the unknown coefficients of the linear combination are obtained. These estimators are applied to derive the linear coefficients using observed predictor and predictand data over a historical period, known as the calibration or training period (Section S2.2.3). Downscaled projections are then obtained by applying the calibrated equations to time series of predictor values derived from historical and future time slices of CMIP5 simulations (Section S2.2.4 below).

S2.2.1 Predictor and predictand variables

Station observations are used as predictands for the calibration period, while predictors can be obtained from observed or quasi-observed datasets, such as those produced by reanalysis projects. The statistical relationship obtained from the calibration step is then applied to generate historical and future downscaled time series of predictands, by using predictors from GCM simulations. When fed by GCM predictor variables, the linear

regression is expected to represent the influence of the time-varying state of simulated large-scale conditions on the local variables of interest in these time series, while also giving results consistent with the mean climatic state of the GCM.

A key step in ESD is the choice of predictors. Criteria that must be fulfilled (Benestad et al, 2008) include: 1) climate model representation, i.e. predictors have to be well simulated by GCMs, 2) description of change, i.e. predictors are expected to describe as fully as possible changes in the predictand under given forcing perturbations, such as changes in greenhouse gases and aerosols, and 3) strong relationships with the predictand. For example, mean sea level pressure, 2m temperature and upper air dynamical variables are usually assumed to be simulated well enough by GCMs to qualify as suitable predictors (Benestad et al, 2008). Some of these variables are also likely to be suitable to describe the effects of climate change on the predictands, although there is no guarantee that predictor-predictand relationships, derived from periods dominated by natural climate variability, will necessarily capture all the main drivers of forced climate change.

Large scale variables with potential to provide strong relationships with a given predictand are usually identified in advance from the scientific literature, in particular climatological studies of processes relevant to the local scale phenomena which determine the observations at the relevant station. However, the region around Singapore has not been studied extensively, at least not to the point of identifying key large scale variables with strong effects on the chosen predictands at the daily time scale, and for all seasons of the year. Therefore, the choice of large scale variables for this study has been based on guidance for forecasters written by Wee Kiong Cheong (unpublished). This document identifies winds, relative humidity and temperature at levels up to 500hPa as the main large scale variables influencing local weather conditions in Singapore.

However, while all these variables can be easily obtained from reanalysis datasets at an appropriate time scale, the choice of predictors is restricted by the availability of GCM data from the CMIP5 archive (Taylor et al, 2012), which is quite limited for daily and sub-daily upper air variables. Nevertheless, a consistent set of upper air variables is available as daily averages. This set includes horizontal winds, specific and relative humidity, temperature and geopotential height at 1000hPa, 850hPa, 700hPa and 500hPa. In addition, we include mean sea level pressure, 10m wind speed and 2m temperature as further potential predictors. Possible effects on longer time scales not efficiently represented by daily predictors have also been investigated, by including the monthly NINO3.4 index as a predictor (with a time lag of -1 month, which gives the largest correlation with all the predictands considered in this study).

S2.2.2 Identifying a few key predictors

In ESD, the dataset describing a potential large-scale predictor variable typically consists of a time series of spatial fields over the chosen domain. A commonly used strategy is to first apply an *Empirical Orthogonal Function* (EOF) analysis (Von Storch and Zwiers,1999) to the dataset, in order to reduce its dimensionality and identify a few important modes of variability. Each EOF expresses a spatial pattern of the relevant mode, and the corresponding *principal component* (PC) represents its time-dependent contribution to the field of values of the relevant variable. The PCs, which depend on time only, are then used as predictors for the linear regressions forming the downscaling equations. EOFs are estimated by diagonalising the spatial covariance matrix of the

large-scale variable: The first few EOFs and their associated PCs are usually sufficient to describe most of the space and time dependence of the variable, thereby allowing a reduction in the complexity of the computational problem of estimating statistical relationships with the predictand. In addition, since most of the noise (i.e. spatiotemporal variability unlikely to be resolved accurately even at the resolution of reanalyses) is captured by higher order EOFs, only the spatially coherent components explaining the dominant components of variability will be retained.

There are different methods to estimate EOFs. The standard approach is to derive a separate set of EOFs to represent spatial variability in each individual variable: For example, this might lead to identification of several EOFs for mean sea level pressure, several more for 2m temperature, several more for 700hPa relative humidity, etc. However, if two or more large scale variables are expected to contribute a combined effect on a predictand, they can be merged into one data structure in space for all the available time levels in the predictor dataset, and this multivariate structure can then be used to generate Mixed EOFs (Benestad et al, 2008). Each mixed EOFs then represents a joint mode of spatial variation for the set of variables considered. The mixed EOF approach ensures full orthogonality between different PCs (thus reducing the risk of overfitting when building predictor-predictand relationships). However, testing revealed that attempting to represent multiple large-scale variables in this way led to reductions in explained variance in calibration, compared with the standard approach. In this study, we elected to use the standard approach to derive predictor sets, with the exception that the mixed EOF approach was used to combine the zonal and meridional wind components, in order to capture the effects of wind direction.

A second issue concerns whether to use only observed data (from historical time series of reanalyses), or to use both GCM and observed data. Using only reanalyses confers the practical advantage that the EOFs only have to be calculated once, and can then be used in downscaling applications with any chosen CMIP5 model. However, a potential disadvantage is that the EOFs describe only historical observed variability, and do not account for the possibility that GCM modes of variability may differ in pattern or frequency due to model error. In this case, estimates of GCM PCs, calculated by projecting GCM variables onto reanalysis EOFs, could be affected by the noise associated with GCM modes with higher order of variability. For this reason, we recommend using the *Common EOF* approach (Barnett, 1999, Benestad, 2001), in which EOFs are obtained from a concatenated dataset of large scale variables consisting of reanalyses of observations for the historical training period plus corresponding GCM data from the baseline and future portions of the relevant CMIP5 simulation. In Section S2.5 we check the results against parallel results obtained using EOFs derived purely from observations. Both methods are available from the software package supplied (see Section S2.3).

The patterns generated by the common EOF procedure will reflect variability and trends from both the calibration and GCM datasets. The extent to which specific EOFs project onto both observations and model data depends on the extent to which they share common patterns of variability (Barnett, 1999). We use reanalysis data from 1980-2010, combined with GCM data from 1980-2005 and 2070-99. While this approach gives the GCM data more overall weight in determining the EOFs (cf Benestad, 2001), this is necessary in order to ensure that sufficient years are available from each of the three

time slices to sample their variability and trend characteristics adequately[1]. Use of common EOFs implies that the calibration of the predictor-predictand linear regression equations (Section S2.2.3) will depend on the specific GCM being used for downscaling.

S2.2.3 Calibrating the downscaling equations

In this study, a procedure used to fit linear multivariate regressions has been constructed using the R software environment (R project), including R objects and functions from the R package clim.pact (Benestad et al). In particular, EOFs and PCs are calculated by using Singular Value Decomposition (Von Storch-Zwiers, 1999). This procedure is applied to the large scale variables after first normalising them by their standard deviation, estimated from the time series of domain averaged variables, in order to obtain EOFs featuring homogeneous standard deviations in non-dimensional units. Since EOF estimates are potentially affected by sampling problems due to the limited length of the time series from which they are calculated (see Section S2.2.2), the numerical accuracy of the EOFs is then tested by using the North's Rule of Thumb (North et al, 1982, Von Storch and Zwiers, 1999).This tests EOFs by comparing sampling errors on their eigenvalues with differences from the eigenvalues of neighbouring EOFs (noting that EOFs are ordered in terms of explained variance). According to this rule, if an eigenvalue has a sampling error larger than the differences with the neighbouring eigenvalues, then the corresponding EOF is assumed to be mixed with the neighbouring EOFs and should be excluded. A backward-forward stepwise algorithm (Venables and Ripley, 1994) is then used to identify those PCs which act as statistically significant predictors in the linear regression, using data from the calibration period. This algorithm adds or removes a predictor to the statistical model, estimates a new linear regression model and applies the Akaike information criterion (Wilks, 1995) to evaluate if the statistical model is improved as a result of the change.

Since we consider a large number of initial predictors, a procedure based on two selection steps is used. The initial screening of PCs is done by estimating correlations between PCs and the time series of observed values of a given predictand, discarding any PC with an absolute value of the correlation smaller than 0.1. The stepwise algorithm is then applied, usually resulting in a number of selected predictors with a relatively high p value, (i.e. with a large probability that their inclusion in the regression has been obtained by chance). In order to avoid overfitting the statistical model, the PCs with p-values lower than a small threshold value (usually 0.01) are then used as predictors in another iteration of the stepwise algorithm, and this step is repeated until a final set of predictors is found, with all p-values lower than the chosen threshold and not modified by additional iterations of the stepwise algorithm. At the end of this two-step selection process, usually there are 5-10 PCs left as predictors, while the variance explained by the linear model is slightly smaller than the variance explained after the first step.

[1] The impact of trends on the common EOFs was assessed using an offline sensitivity test. This consisted of adding a smooth time-independent pattern, typical of a climate change signal, to the time series of a reanalysis variable, and concatenating this perturbed time series onto the original time series. EOF analysis on this combined dataset shows that the first EOF has a significant projection onto the idealised climate change signal, its standard deviation is increased by this climate change contribution and the corresponding PC captures the change in the perturbed period. The ability of the set of PCs to represent historical variability is not affected, since the mean and variance of the leading PC adjusts to account for the absence of a trend in the original unperturbed data, while all the other PCs and their corresponding EOFs remain almost unaltered.

In general, a linear model will be able to capture only a part of the total variance of the predictand. The fraction of explained variance provides an important first order test of the effectiveness of the calibrated equation. A further assessment of the quality of the linear model fit can be done on residuals, i.e. the differences between fitted values and the observed values. In the ideal case, residuals are expected to have a Gaussian distribution with zero mean, no autocorrelation and no trend (Benestad et al, 2008). Any dependence of residuals on the predictors should also be absent if the linear model has successfully captured the dependence of the observed variable on the predictors, in particular residuals should have no trend and constant variance over the range of predictor values (Wilks, 1995). These statistics are used to assess the quality of the linear model fit in Section S2.5.

The statistical model is also assessed by a jackknife cross-validation procedure. In this procedure, a block of data is removed from the calibration dataset and the two-step calibration procedure is applied to estimate a linear model which is then used to predict the target local variable for the period which has been removed from the calibration dataset. Differences between predictions from the model using the full calibration period and predictions from the jackknife procedure are then analysed, to test the ability of the statistical model in reproducing the observed values and their daily variability.

Since not all the predictands considered in this study are Gaussian distributed, alternative statistical approaches based on Generalised Linear Models (GLM, Venables and Ripley, 1994) have also been used. These statistical models can be used with predictand distributions from the exponential family, in particular for binomial and Gamma distributed predictands. In addition, a functional transformation (link function) is applied to the predictand before expressing it as a linear combination of predictors. Mean daily wind and daily rainfall have been modelled using these methods. A Gamma distribution has been used for mean daily wind, using a log link function (Yan et al., 2006). Rainfall has been modelled in two steps, representing the occurrence of wet days with a Bernoulli distribution and a logit link function, and as a Gamma distribution with a logarithmic link function to model rainfall intensity for wet days. As for other predictands, PCs from common EOF analysis have been used as predictors, applying the same selection criteria used for linear regression models. The calibrated statistical model is then validated using the jackknife approach described above.

S2.2.4 Generating downscaled projections from GCM data

Projections are generated by first calculating daily time series of PC values for historical and future portions of the relevant GCM simulation, and then applying the downscaling equations to derive historical and future time series of station predictands, from which diagnostics of interest can be calculated in the same way as from (say) regional climate model output.

Since both the mean state and the variability in the GCM will in general contain errors, bias corrections are often applied to mean and variance of each GCM predictor, derived from the historical baseline period of the simulation (Wilby, 2004). These empirical corrections are then applied to predictor values derived from the future scenario portion of the GCM simulation, when generating downscaled future time series. Care is needed in applying this approach, however, since the application of separate bias corrections to multiple predictor variables can potentially generate physical inconsistencies in the description of the regional three-dimensional state of the atmosphere, particularly when

applied to daily time series. Some attempts to bias correct GCM boundary conditions involve adjustment steps to avoid such inconsistencies (Xu and Yang, 2012). In the present case, use of predictors derived from pre-defined EOFs reduces the risk of obtaining physically inconsistent results through the application of bias corrections, since these are applied to the PCs, and hence do not affect the spatial coherence built into the EOFs.

In the common EOF framework, the bias correction strategy outlined above is implemented by adjusting the mean and standard deviation of PC time series from the GCM baseline period to match those of the observed PC distribution derived from the reanalysis period (Imbert and Benestad, 2005). For the future section of the GCM simulation, the most straightforward strategy is simply to apply the same corrections directly to the future PC time series. However, this approach would modify any climate change signals reflected in the future time series of the PCs. In particular, any future trend in one or more PCs would be scaled up or down, dependent on the correction applied to the PC standard deviation. Since it is not clear that GCM errors in simulated future trends can be expected to scale with historical errors in simulated natural variability, such a scaling could be problematic, especially for variables such as temperature in which the forced trend is potentially large during the latter stages of the 21st century. For this reason, the future PC time series are detrended prior to application of the bias corrections, by removing significant trends (p-value < 0.1) estimated by linear regression. These trends are then added back into the time series of the bias-corrected future PCs, before generating downscaled predictands.

The common EOF approach is well suited to the bias correction strategy described above (Imbert and Benestad, 2005, Benestad et al, 2008), since the use of historical and future GCM data (alongside observations) to determine the EOFs should reduce the risk of requiring large corrections reflecting differences between observed and simulated patterns of variability. As noted in Section S2.2.2, however, an alternative is to use EOFs derived purely from reanalysis data. Although not our recommended approach, this method is a useful as a sensitivity test of results based on the common EOF method, and can also provide a convenient method of downscaling a large set of GCMs, because the GCM-independent EOFs only have to be estimated once. Methods using reanalysis-based EOFs (e.g. Murphy, 2000) involve the calculation of projection coefficients (analogous to PCs) representing the time-varying projection of the EOFs onto daily spatial fields of the relevant variables simulated by the GCM. The same bias correction strategy outlined above can then be applied to the projection coefficients. In this case, scaling coefficients are expected to be larger than those estimated for the combined EOF approach, since reanalysis modes of variability are not constrained to match the GCM modes as in the combined EOF analysis. Potential problems due to this mismatch of variability modes will lead to scaling factors substantially different from unity.

S2.3. Software

R programs implementing the three regression models described in Section S2.2.3 have been written using object definitions and functions from the R package clim.pact. The basic algorithm includes pre-processing predictands and predictors from ERAInterim and the relevant CMIP5 GCM, estimating EOFs for a given season, calibrating the linear regression relationship, estimating its quality by the jackknife procedure, generating downscaled predictand data for baseline and future time series, and producing a series

of diagnostic plots assessing the quality of the linear relationship and the realism of baseline and future daily time series. Data from the two GCMs used in this report (see Section S2.4) have been saved for future use.

Although the clim.pact package provided a suitable basis for this study, a number of changes were needed to extend its applicability. In particular

- Use of PCs from EOF analyses of individual variables has been included also for downscaling of daily predictands (the original routine only allows the use of one set of mixed EOFs)
- The facility to estimate projection coefficients of GCM variables onto EOFs derived purely from reanalyses of observations has been added, providing an option to generate baseline and future downscaled predictands without recourse to common EOFs
- Bias correction of both GCM PCs (when derived from common EOFs) and GCM projection coefficients has been implemented
- Capability to define any desired non-standard seasons has been added
- The function reading GCM data has been modified to support CMIP5 data, including both the capability to read and pre-process GCM data on the 365-day and 360-day calendars.

Functionality to make plots based on diagnostic checks described in Section S2.2 has been included. Examples of these are shown in Section S2.5.

S2.4. Data

Station observations available for this study include precipitation (27 stations), daily mean temperature, daily minimum and maximum temperature (2 stations) and daily mean 10m wind speed and maximum daily wind speed. In Section S2.5, we investigate precipitation downscaling for Changi airport, s24. Mean daily wind is only available from one station (s24). All these data are available for 1980-2010.

Large scale variables for the calibration procedure have been extracted from the ERAInterim dataset (Dee et al, 2011) as 6-hourly data and averaged to daily values. Days are defined according to the Greenwich meridian to be consistent with the averaging process of CMIP5 data. Station daily averages, however, are constructed according to Singapore local time.

Two GCM models from the CMIP5 set, HadGEM2-ES and GFDL-CM3, have been used to generate examples of downscaled climate scenarios, using predictors derived from the set of large scale variables described in Section S2.2.1, with the exception of variables at 1000hPa; for this level, only the HadGEM2-ES horizontal winds could be used since the other variables contain missing data, since this level is usually lower than the elevation of many grid boxes. Two periods have been used: 1980-2005 from the historical simulation and 2070-2099 from the RCP8.5 integration.

S2.5. Results

S2.5.1 EOF analysis

For the purposes of estimating EOFs, two different areas have been defined, (1N <lat <5N, 99E < lon < 106E for variables likely to influence daily Singapore weather at relatively large scales, -1N< lat <3N, 102E<lon<106E for variables likely to exert an influence over somewhat smaller scales. All variables at 700hPa and 500hPa and mean sea level pressure are defined on the larger area, all others are defined on the smaller area. Optimal sizes and positions for these domains were selected following sensitivity studies on linear regressions for daily temperature.

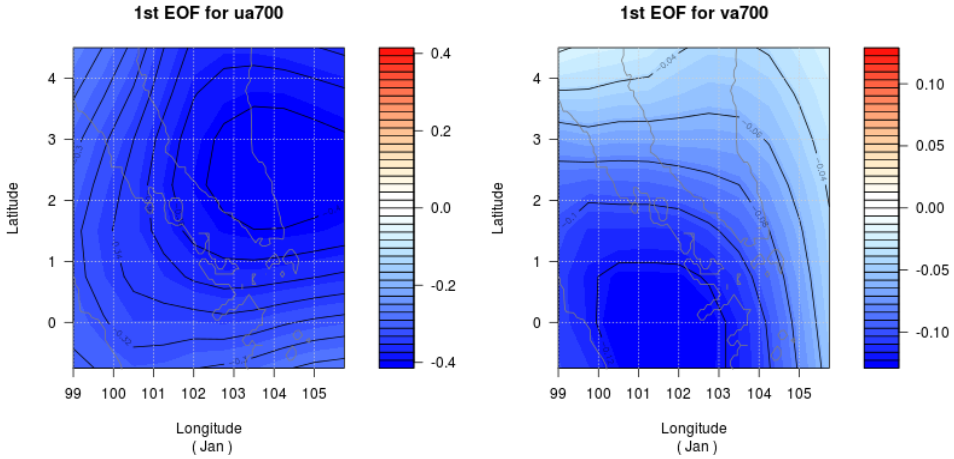

Figure S2.1: Mixed EOFs for horizontal winds at 700hPa from ERAInterim, January (1980-2010)

As explained in Section S2.2.2, EOFs are calculated separately for each large scale variable supplying predictors, with the exception of winds. This is because the direction of the wind plays an important role in driving the weather in Singapore. For this reason, mixed EOFs describing spatial variability in horizontal winds at each given pressure level have been created, from the zonal and meridional components. Figure S2.1 shows the two components of the first mixed EOF for 700hPa, calculated from ERAInterim data for January, 1980-2010. PCs from these mixed EOFs have been combined with EOFs from all the other variables to form the set of candidate predictors for the regression models used in this study.

Results obtained from a test calibration using EOFs based only on ERAInterim data showed that the NINO3.4 index and the geopotential height at all levels could be removed from the set of predictors, since they were never selected by the backward-forward calibration procedure as significant predictors.

Common EOFs are derived for HadGEM2-ES and GFDL-CM3 by combining variables from the historical (1980-2005) and future (2070-2099) periods of the relevant RCP8.5 scenario simulation with ERAInterim variables..

S2.5.2 Daily mean temperature

In order to demonstrate the methodology, linear regressions for daily mean temperature from the Changi airport station (s24) have been fitted for three two-month sub-seasons:

Dec-Jan, Apr-May and Jul-Aug, corresponding to the wet period of the NE monsoon, intra-monsoonal period and the SW monsoon period, using calibration data from 1981-2010. The PCs were derived from common EOFs (see Sections S2.2.2 and S2.5.1). The two selection criteria used to reduce the number of potential predictors (see Section S2.2.3) have been set to an (absolute) predictor-predictand correlation of 0.1 and a p-value (used to avoid overfitting) of 0.01. The efficacy of these criteria should be assessed on a case-by-case basis. However, we find the same values to be suitable for all the calibrations of temperature. Wind speed and rainfall required different settings, specified in Section S2.5.5 and S2.5.6. Despite these rather strict criteria, the calibration procedure was still able to find approximately ten significant predictors for each linear regression model. The variance explained is quite high for the three periods, between 75% and 80%. Tests using the alternative EOF definition based only on observations (see Section S2.2.2) gave similar results. Not surprisingly, the smallest explained variance was obtained for the intra-monsoonal period; however, the result was still comparable with the other two monsoon periods, despite the lack of a consistent large scale flow in this period.

Figure S2.2 shows results from the cross-validation for the months of April and May (1981-2010 calibration period), using PCs from the common EOF analysis based on HadGEM2-ES combined with ERAInterim. This diagnostic plot is useful to compare trends and multi-annual variability of the downscaled predictand with observation, and hence check that the downscaling relationships are producing plausible results. The cross-validation has been performed using the jackknife procedure of Section S2.2.3, in which results for each day in each two month period in each year are obtained by first removing the relevant two months of data from the training dataset, and then predicting daily values in those two months by a linear regression fitted over all the other years. The cross-validation result shows a very good agreement with the results from the best fit obtained over the entire training period. This applies even for the two months in 1998, despite not including an explicit predictor of El Nino phase such as the NINO3.4 index. It is therefore clear that PCs from the chosen large scale variables are able to explain multiannual variability in the station data. However, neither downscaled estimate is able to capture the most extreme daily temperatures, in particular for the colder days. The other two periods (Dec-Jan and Jul-Aug) show similar features, which could be caused by the small negative skewness of daily mean temperature for these periods.

Figure S2.3 shows q-q plots for the normalised residuals for the intra-monsoonal period described above, for the linear regression model used in Figure S2.2. In this plot and corresponding q-q plots for Dec-Jan and Jul-Aug, there is a very good agreement of the distribution of the residuals with the theoretical normal distribution up to +/-2 standard deviations, and quite a small error with respect to observed data (standard deviation ~0.5K). The lack of agreement at the tails of the distribution is a well known problem of the application of linear regression methods to daily meteorological variables (Zorita and Von Storch, 1999), since these methods are not designed to minimise contributions of effects not describable by large scale variables. However, since these errors apply to predictions of only a very small number of daily observations, the distribution of residuals passes all standard tests for normal distributions. The cross-validation results in the right panel are based on the differences between the predictions for the missing periods in the jackknife procedure and the fit using the full calibration period: these results show very small differences between the two estimates (mean difference very close to zero and standard deviation of 0.06). Similar results from the jackknife test have been obtained for all the other predictands.

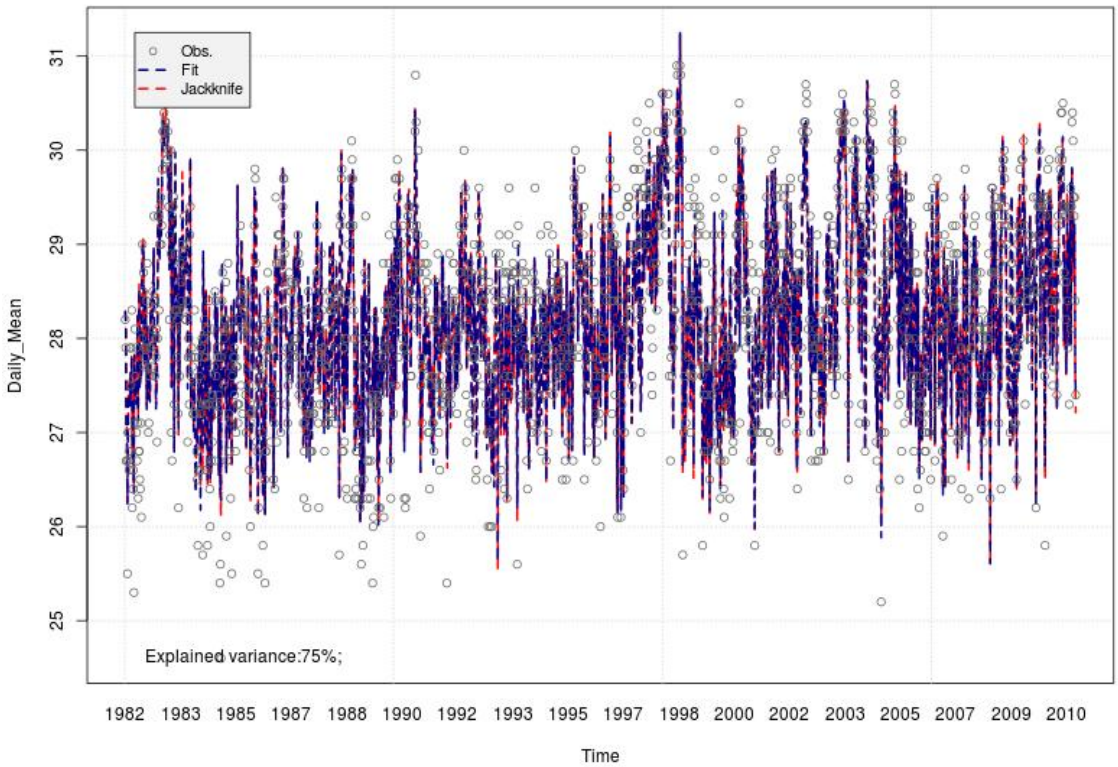

Figure S2.2: Jackknife cross-calibration for downscaling of daily Tmean from the Changi airport station (s24), for the months of April and May of the period 1980-2010, using predictors from the common EOF analysis derived from ERAInterim reanalyses combined with HadGEM2-ES data. The two lines are the fit for the full period (blue), and for the jackknife cross-calibration (red). The points are the observed Tmean values from s24.

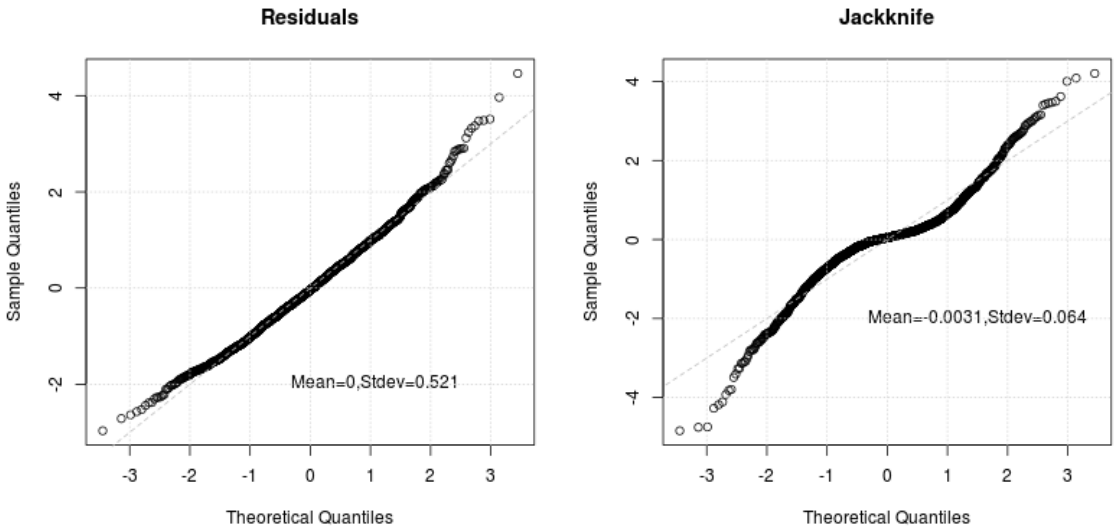

Figure S2.3: q-q plot for the residuals from the linear regression model for Tmean at Changi airport estimated for the months of April and May, 1981-2010, normalised by standard deviation, left panel. The right hand panel shows residuals between the regressed estimates derived from the full training period (blue curve in Figure S2.2), and the cross-validated estimates (red curve in Figure S2.2).

Estimates of autocorrelation in the time series of residuals during the calibration period are shown in Figure S2.4, in comparison with the correlogram obtained from the observed daily temperature for April-May and for December-January. Encouragingly, the

residuals show considerably smaller autocorrelations than the station data for the first few days in both periods. The residual autocorrelation approaches zero quite rapidly for the Dec-Jan period, while for the April-May period the residuals still show a small but significant autocorrelation out to 20 days. For the period not shown, Jul-Aug, residuals have a correlogram quite similar to the April-May period. Furthermore, correlograms from the cross-validation predictions (not shown) are quite similar to the correlograms estimated from observed data in Figure S2.4. These cross-validation results indicate that the linear regression model is able to reproduce quite well the correlation structure of the observed daily mean temperature distribution.

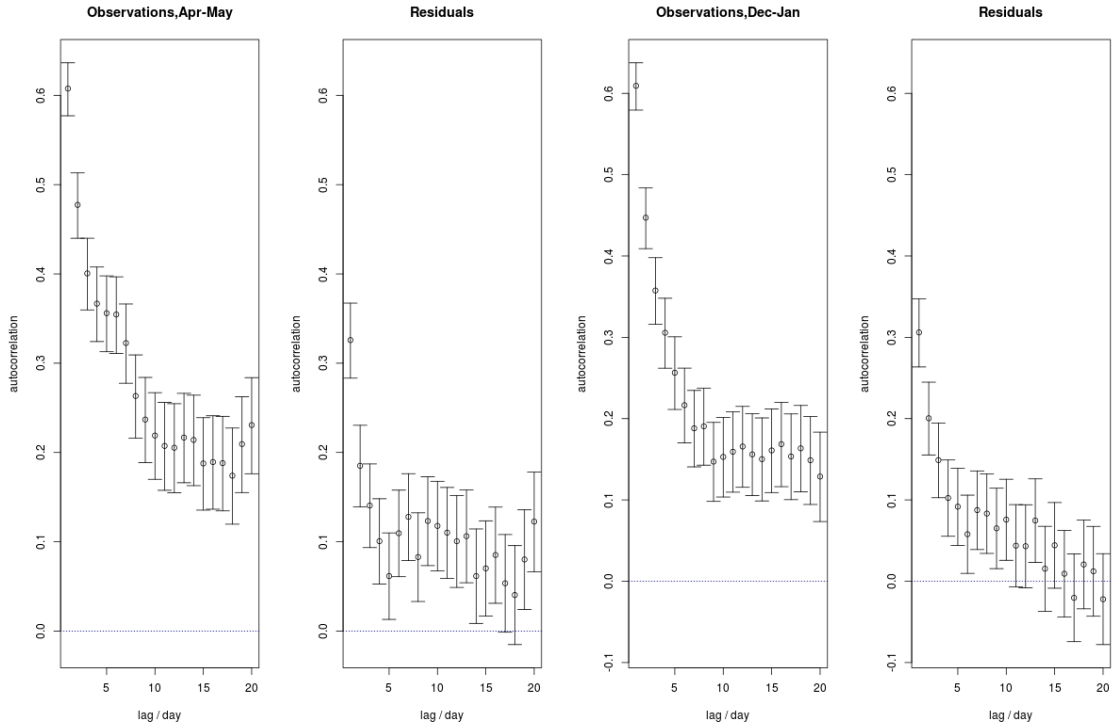

Figure S2.4: Autocorrelation plots for the observed daily mean temperatures from Changi airport (left panels) and for the residuals from the linear regression (right panels), for Apr-May and Dec-Jan. Confidence intervals at 95% have been estimated by applying the method based on the Fischer's z transformation (von Storch, book)

Nevertheless, these results are not fully consistent with the prior methodological hypothesis of zero autocorrelation, and Gaussian distributed residuals. It is important to test these assumptions, since they could be used to justify the introduction of a white noise, additive term to the linear regression (Von Storch, 1999), aimed at describing the small scale noise stochastically in order to improve the statistical downscaling model. Tests on the effects of the inclusion of a linear noise term are shown in Figure S2.5, for the period Apr-May and Jul-Aug. The noise has been generated using random sampling from a Gaussian distribution with zero mean, the standard deviation of the residuals from the linear model from the full calibration period, and assuming no autocorrelation. The result shows a small improvement at the tails for the Apr-May distribution and a mixed outcome for the Jul-Aug period, since small improvements at the lower end are counterbalanced by spurious values introduced at the upper tail; therefore, for this predictand, the addition of a linear noise term to the linear regression model does not necessarily improve the overall credibility of the downscaled estimates.

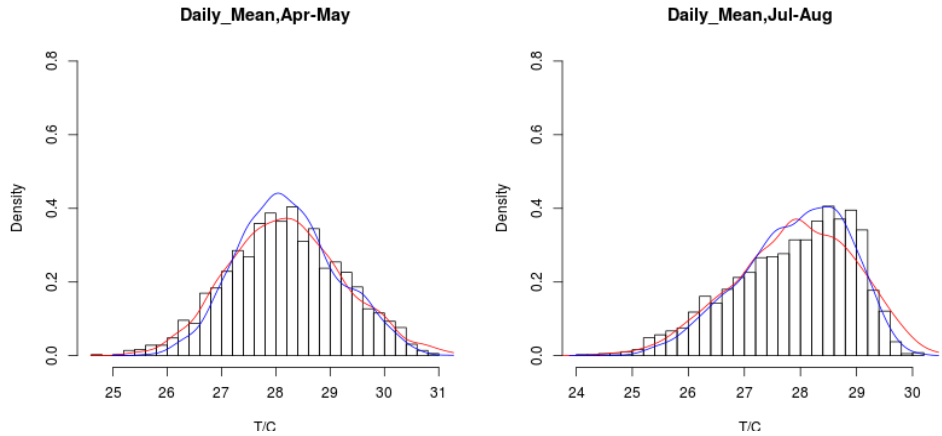

Figure S2.5: Histogram of mean daily temperature for Changi airport compared with the distribution from the linear regression fit (blue curve), and from the linear model with the addition of white noise distributed as the residuals (red curve). The statistical models have been estimated for the period 1980-2010, for April-May (left panel) and June-July (right panel), using PCs from the HadGEM2-ES/ERAInterim common EOF analysis.

Since Figure S2.4 indicates the presence of small but significant autocorrelation out to 20 days in the residuals, it is useful to assess whether the large scale signal associated with the PCs used in the linear regression has been completely accounted for in the regression equations. A comparison of correlations of the selected PCs with both the predictand and the residuals (Figure S2.6, filled circles in top and bottom panels respectively) shows that these PCs are almost completely uncorrelated with the residuals. The same applies to additional PCs not selected by the backward-forward stepwise algorithm, despite showing relatively high predictor-predictand correlations. This outcome indicates that the long range persistence of residual autocorrelations seen in Figure S2.4 cannot be ascribed to a failure to capture the influence of large-scale variables considered in the EOF analysis.

In summary, the above analyses on the residuals for Changi airport mean temperature suggested that the calibration of these linear models is acceptable. However, this assessment is not sufficient, in isolation, to guarantee that these statistical models can be applied to downscale GCM projections.

Downscaled future projections are generated by applying the fitted linear regression to the PCs describing the future HadGEM2-ES climate scenario obtained from the common EOF approach, after applying the bias correction procedure described in Section S2.2.4. Figure S2.7 shows the results for Changi airport, obtained from the April-May period discussed above. Downscaled temperatures are in the same range of the GCM average daily temperatures for the Singapore grid box (also shown in Figure S2.7), with higher variability at the lower end of the distribution. This result is consistent with the time series of downscaled HadGEM2-ES baseline daily temperatures (not shown), which also features an increased variability at lower temperatures and temperature in the same range. The leading bias-corrected PCs selected for this linear model are shown in Figure S2.8. Scaling factors representing the fractional change to the standard deviation of the PCs introduced via bias correction are reported in the Figure. These are reasonably small, in particular for the leading PCs; similar results are obtained from bias corrected projection coefficients obtained from the linear regression model based on PCs derived only from ERAInterim data. The leading PC (top left panel) corresponds to the leading EOF of the 2m temperature variable: This PC determines the main climate change

between the 2070-99 and baseline periods, including also the trend within the 2070-2099 period.

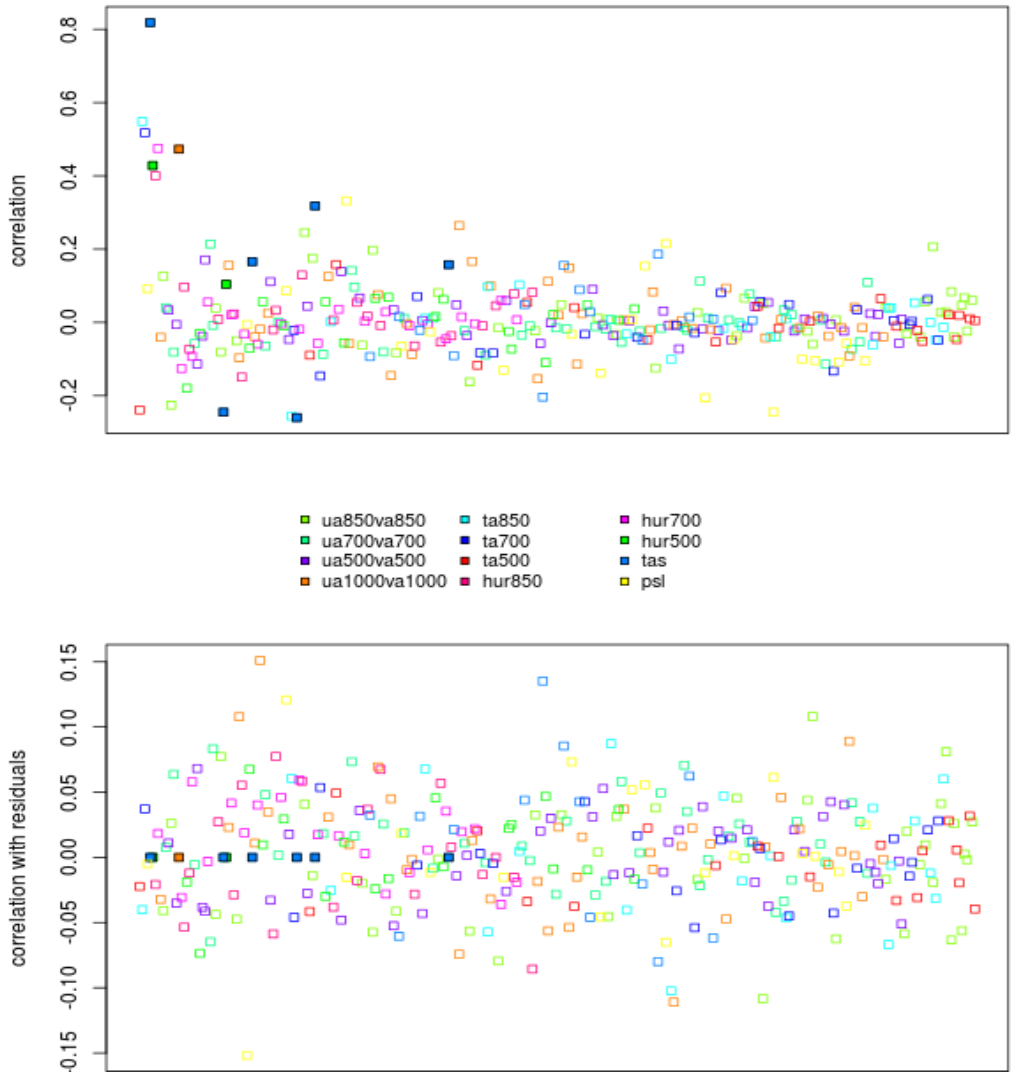

Figure S2.6: Correlation of PCs forming potential predictors of Changi airport daily mean temperature, top panel, and with residuals from values predicted by the calibrated linear regression, bottom panel. PCs from different variables are shown with different colours, the full circles indicating the PCs which were selected for use in the linear regression model.

The climatological distributions of baseline and future downscaled temperature derived from HadGEM2-ES predictor data are shown in Figure S2.9. The distribution for the historical period is centred on the corresponding distribution downscaled from ERAInterim predictor data, reflecting satisfactory performance of the bias correction strategy applied to the PC time series from the GCM. This Figure also shows the downscaled future climate distribution, derived from the detrended downscaled variable, in order to compare the historical and future realisations of local natural variability. In this case, the future results show an increase in variability, consistent with an increase in the variability of daily temperature in HadGEM2-ES for the grid box containing Singapore (not shown).

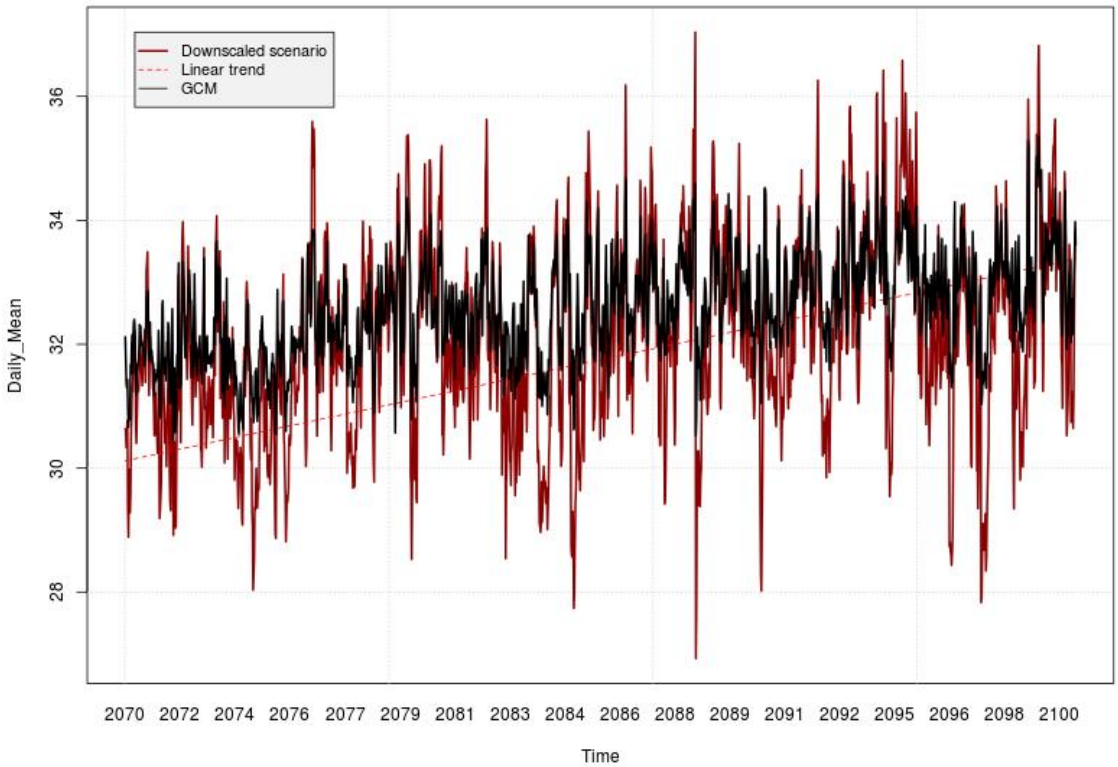

Figure S2.7: Downscaled daily mean temperature for the HadGEM2-ES RCP8.5 projection (red line), for April and May in the period 2070-2099, using PCs from the common EOF analysis. Daily mean temperatures extracted from the HadGEM2-ES Singapore grid box (black line) are also shown.

Downscaled results for Dec-Jan and Jul-Aug historical temperature distributions (not shown) show mean values with a small cool bias, and a small reduction in variability, in particular at the lower end of the distribution, when distributions derived from HadGEM2-ES predictors are compared with those derived from ERAInterim predictors. However, downscaled temperatures derived from HadGEM2-ES have a more realistic spread than values taken directly from the grid box containing Singapore, increasing their potential utility for impact studies. Similar results are obtained from an alternative linear regression model based on the EOFs derived purely from ERAInterim data For the GFDL-CM3 model, the climatological average of daily near-surface temperature for the Singapore grid box is ~2K smaller than the observed and downscaled temperature from the GFDL-CM3 historical period, and also shows higher variability. Therefore, the downscaling procedure leads to a derived distribution showing a better fit to observations than the raw GCM grid box data. This result is obtained both from the linear regression based on PCs from the common EOF analysis, and also from projection coefficients obtained using EOFs derived purely from ERAInterim data. The warm shift found in the baseline period between the downscaled temperatures and the raw GFDL-CM3 values is also found in the future climate results.

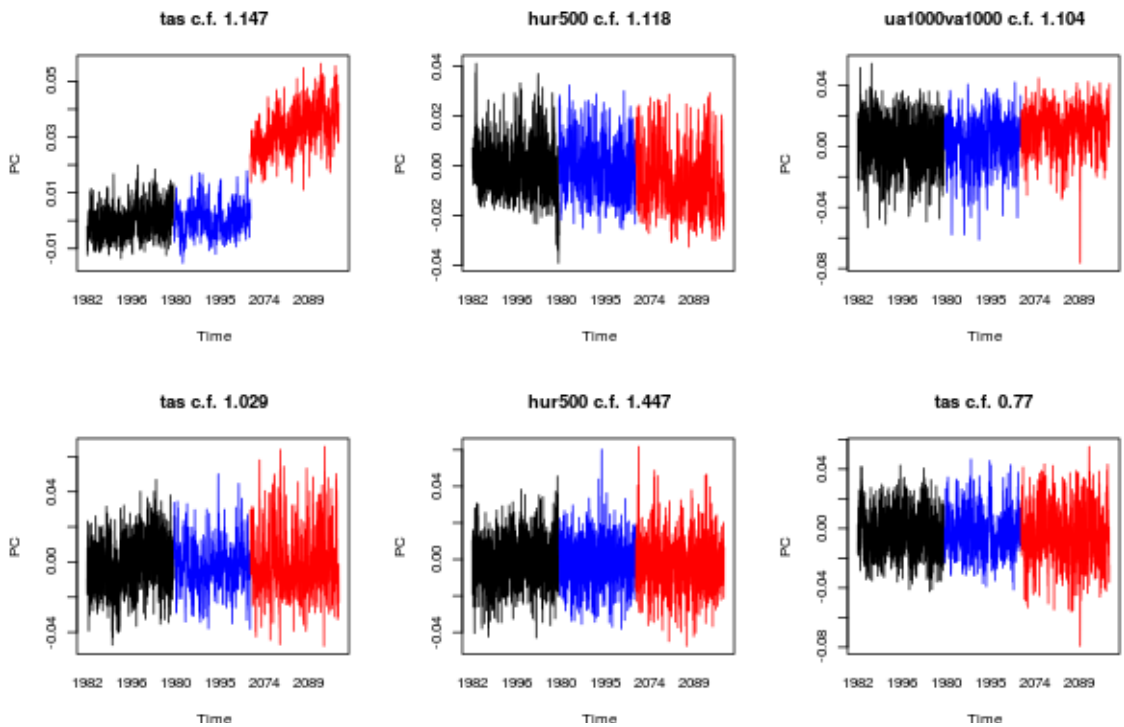

Figure S2.8: Principal components from the common EOF analysis based on ERAInterim and HadGEM2-ES data, selected for the linear regression for Changi airport mean daily temperature, for the period Apr-May (1980-2010). The results show future PC time series processed by detrending, bias correction and reinstatement of trends as described in Section S2.2.4. The black curves are the principal components for the calibration period (ERA-Interim), the blue curve for the HadGEM2-ES historical period and the red curve for the HadGEM2-ES future period 2070-99, under RCP8.5 scenario forcing. tas, 2m temperature; hur500, relative humidity at 500hPa, ua1000va1000 is the PCs for the mixed horizontal winds at 1000hPa. Three PCs from the 2m temperature variable, two PCs from 500hPa relative humidity and one from the mixed EOF of 1000hPa wind components were selected in the linear regression procedure.

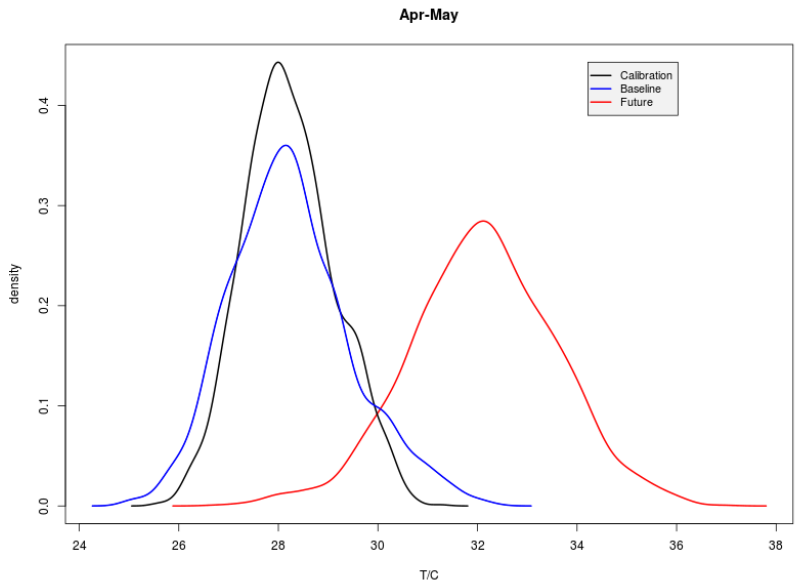

Figure S2.9: Distributions of downscaled daily mean temperatures for Changi airport using calibrated downscaling relationships applied to PC predictor time series from ERAInterim

for the calibration period (black), the HadGEM2-ES historical period (blue) and the HadGEM2-ES future period (2070-99), in response to the RCP8.5 emissions scenario.

S2.5.3 Daily Maximum Temperature

Daily maximum temperatures for the station under investigation (Changi airport, s24) have highly negative skewed distributions (skewness ~ -(0.8-1.3)), with corresponding standard deviations ranging between 1.3C and 1.8C. Nevertheless, an attempt to fit a linear regression model to this predictand was made. Since monthly distributions are quite similar within the bimonthly periods used in Section S2.5.2, the same two month periods are used here.

The calibration plot in Figure S2.10, obtained from a linear regression based on a common EOF analysis for HadGEM2-ES and ERAInterim, shows a fit which does not capture extreme values, in particular for the lowest temperatures. The model therefore fails to capture the skewness of Tmax for this period. However, the linear regression model still explains 58% of the variance. Similar values have been obtained for the other two periods (55% for Apr-May and 65% for December-July).

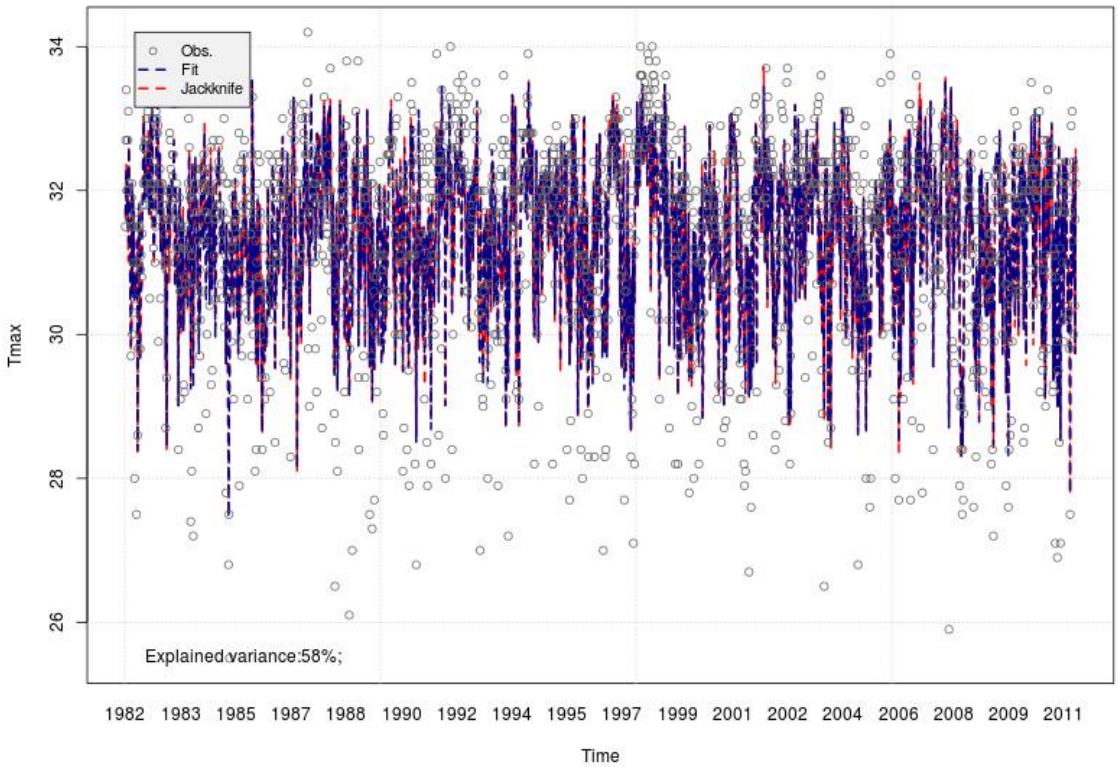

Figure S2.10: Calibration, for July-August 1980-2010, of daily maximum temperatures (Tmax) for Changi airport, using PC predictor variables based on common EOFs obtained from HadGEM2-ES and ERAInterim data. Fitted Tmax (blue curve), fit from the cross-calibration procedure (red curve) and observed values (open circles).

The analysis of residuals in Figure S2.11 shows a rather good agreement with a Gaussian distribution and a very low residual autocorrelation, in particular when compared with the autocorrelation in the observations. This is despite the failure of the downscaling to capture the low Tmax extremes, which occur on only a small number of days. From this stage of the assessment, the linear regression model appears fit for purpose.

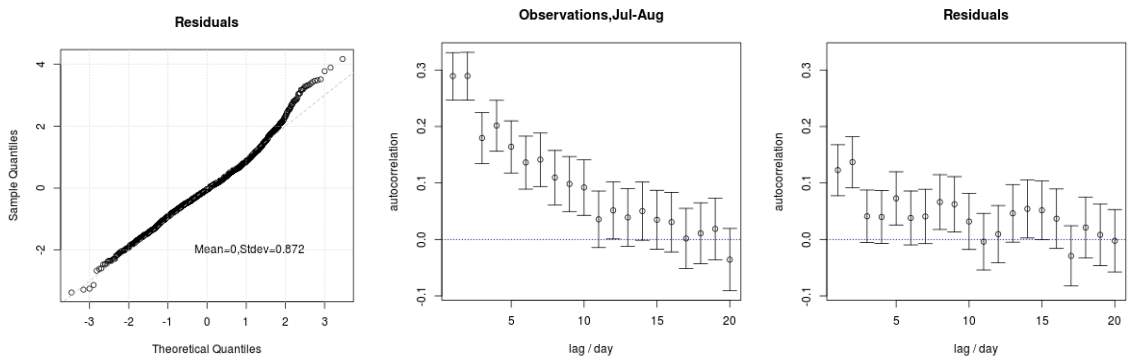

Figure S2.11: q-q plot of Tmax residuals normalised by standard deviation (left panel), autocorrelation of the observed Tmax (middle panel) and autocorrelation of the residuals from the linear regression of Tmax, for July-August, 1980-2010, corresponding to the downscaling for Changi airport shown in Figure S2.10.

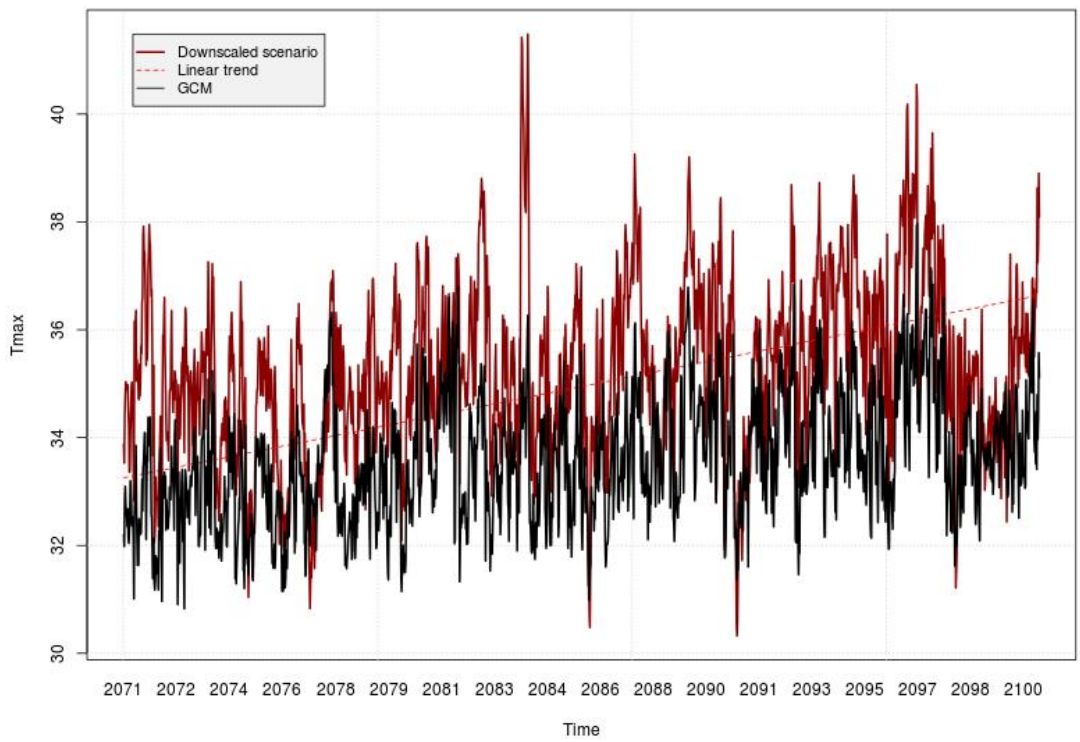

Figure S2.12: Downscaled Tmax from the HadGEM2-ES RCP8.5 projection (red line) and Tmax taken from the same HadGEM2-ES integration for the grid-box including Singapore, Jul-Aug, 2070-2099.

The application of the linear fit to downscale the HadGEM2-ES RCP8.5 simulation is shown in Figure S2.12, using bias corrected PCs from the common EOF analysis. The downscaled Tmax projection shows a shift to higher absolute temperatures with respect to raw HadGEM2-ES values extracted from the grid box including Singapore. The shift of future downscaled temperatures to higher values with respect to its driving model is comparable to a similar shift obtained by downscaling HadGEM2-ES for the present climate period, which produces results comparable with station observations. This

indicates that the downscaling is successfully addressing biases in the HadGEM2-ES Tmax data, while the projected future changes are similar between the downscaled values and those taken directly from the GCM output. The daily variability of the downscaled Tmax is slightly larger than the HadGEM2-ES grid-box Tmax for both future and baseline periods (not shown); however, given the limitations in explained variance and the negative skewness of the distribution of observed values, daily variability in the downscaled results could still be underestimated. Similar results were obtained from an alternative linear regression using EOFs derived only from ERAInterim data.
The possibility of including an additive noise term to the linear model for Tmax should be considered with caution. Tests showed mixed results: spurious extreme high values of Tmax were created, together with some improvement at the lower tail of the distribution, when applied to the HadGEM2-ES historical period. Results for Tmax downscaled from GFDL-CM3 are consistent with these findings.

S2.5.4 Daily Minimum Temperature

Distributions of daily minimum temperature (Tmin) for the observed dataset (s24) show negative skewness, but the extent is smaller than found for daily average temperature (Tmean). For this reason, the use of a linear regression model is even more justifiable than for Tmean. The method was fitted to the same bimonthly periods used for Tmean and Tmax, since distributions of Tmin are also quite similar for the individual months within these periods.

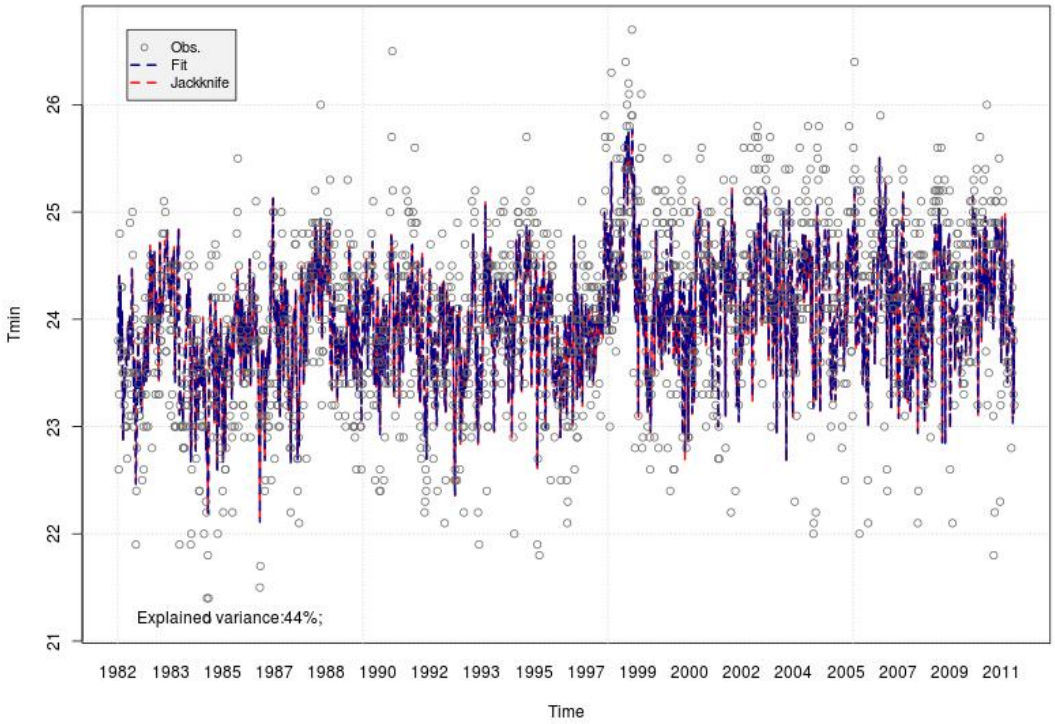

Figure 13: Calibration for Tmin at Changi airport, for Dec-Jan 1980-2010. Fitted Tmin (blue curve), fit from the cross-calibration procedure (red curve) and observed values (open circles).

The calibration for the period July-August (1980-2010), using HadGEM2-ES/ERAInterim common EOF analysis, is reported in Figure S2.13. Fitted values have a spread which does not cover the full range of values from observations, and explained variances are

quite low (between 40% and 50% for the period considered in this study). Cross-validation results are similar to those obtained from fitting the entire training period, while the correlogram of the cross-validation predictions shows slightly larger autocorrelations for the first few days with respect to the correlogram of the station observations.

Figure S2.14 shows results from the analysis of the residuals for the Dec-Jan period for the linear fit: the q-q plot indicates very good agreement with a Gaussian distribution (up to 3 standard deviations) and the correlograms show very small values of autocorrelation left in the residuals.

Downscaled minimum temperatures from the baseline and future HadGEM2-ES integration, derived from bias corrected common EOFs (not shown), are lower than the corresponding HadGEM2-ES values for the Singapore grid box. Downscaled Tmin shows daily variability comparable with the variability of the Tmin values for the Singapore grid-box from HadGEM2-ES, possibly reflecting the small explained variance explained by this linear model. The autocorrelation of the residuals drops to zero after few days, suggesting that the linear regression model has effectively removed all dependence on the larger scales seen in the correlogram of the observations. This is a rather surprising result, given its small explained variance (41% for this period). Therefore, results of the analysis of the residuals in figure S2.14 support the conditions for introducing an additive, white noise term to the linear regression fit.

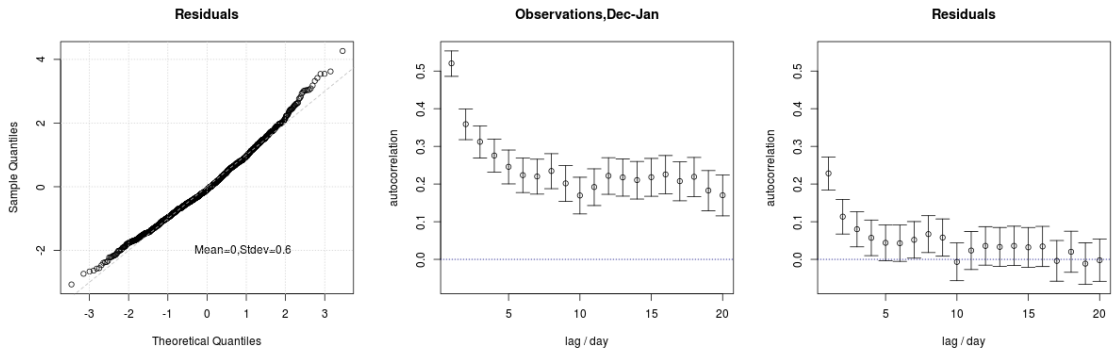

Figure S2.14: q-q plot (left panel), autocorrelation of the observed Tmin (middle panel) and autocorrelation of the residuals from the linear regression of Tmin, Dec-Jan, 1980-2010

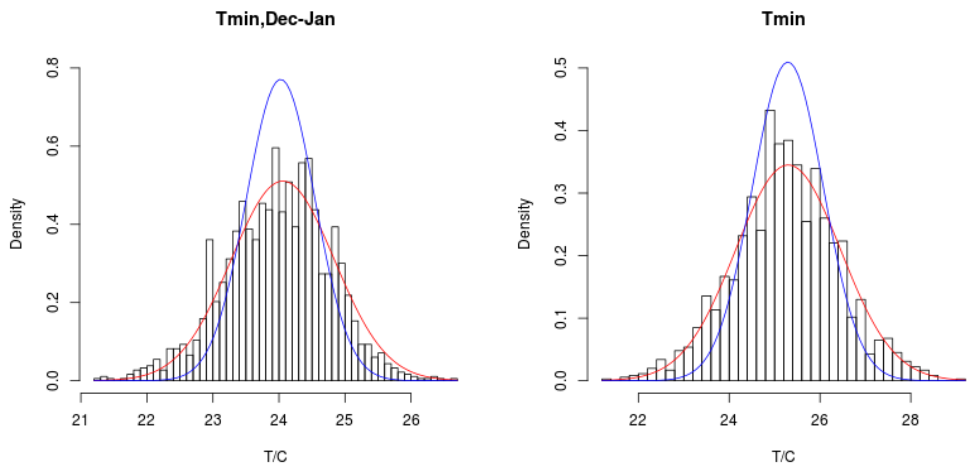

Figure S2.15: Histogram of observed daily Tmin compared with the distribution obtained from the linear fit (blue) and the distribution of the model which includes the linear fit and a white noise term (red). Left panel, Dec-Jan, 1980-2010; right panel, Apr-May, 1980-2010, from the linear fit using the HadGEM2-ES/ERAInterim common EOF analysis.

The results of adding white noise can be seen from Figure S2.15, showing bimonthly distributions of Tmin compared with the linear fit and with the model which includes an additive Gaussian noise term, with mean equal to zero and the variance of the distribution of the residuals from the linear model. The large scale constraint on this predictand, as captured by the linear regression model, is rather small (blue curves). The agreement of the model including additive noise (red curves) with the observed station distribution is much improved. These findings therefore suggest that most of the daily Tmin variability is caused by small scale processes which cannot be described by the PCs used in the linear regression, since the explained variance due to the large scale predictors is 41% and the white noise term, constructed from the almost uncorrelated residuals, is able to reproduce the shape of the distribution of the observed values. These findings are reproduced by the linear model based on the alternative EOF analysis using ERAInterim alone. Similar results are also obtained by downscaling the GFDL-CM3 model.

The linear model with additive noise, therefore, can be used to downscale future scenarios as well. However, in this case, the assumption that the weight of the additive noise will be the same in the future climate is worth studying as part of a recommended assessment of the issue of the stationarity of the statistical relationship, using regional climate model data (see Section S2.6).

S2.5.5. Average Daily Wind Speed

Daily average 10m wind speed, only available for the Changi airport station (s24), was modelled using the GLM approach, since, for most of the months, the distributions of observed values are strongly positively skewed (not shown). An obvious choice for modelling these distributions is the Gamma distribution (Yan et al, 2006); in particular, the mean parameter of the Gamma distribution is written as the exponential of a linear combination of predictors.

In this case, the calibration dataset consists of one month blocks for the period 1981-2010 since monthly distributions of daily mean 10m wind speed are quite different. PCs were taken from a common EOF analysis using HadGEM2-ES and ERAInterim variables. We consider the same set of candidate variables used to downscale station temperatures, except for the addition of the daily average 10m wind speed (justified by previous work, e.g. Manzanas et al, 2012). We also withheld near-surface and upper level temperature from the set of potential predictors, since including these caused non-physical behaviour of the fitted regression models in the downscaling of the future, changed, climate; in particular, the inclusion of temperature PCs with a strong climate change signal produced very large and unrealistic future 10m wind speeds, amplified by the exponential dependence from predictors. This illustrates the importance of checking whether relationships trained on historical climate variability produce credible realisations of future climate, in cases where the predictor variables project significantly onto both natural variability and the future response to changes in greenhouse gases. The correlation threshold for identification of potential PC predictors was set to 0.1 (as for temperature variables), however the p-value threshold for significance in calibration was increased to 0.05. Despite this increase, the smaller set of variables available for calibration and the smaller sample result in a smaller number of significant predictors being selected at the end of the stepwise procedure (usually not more than six PCs).
Figure S2.16 shows results obtained from the calibration for the month of July. The fitted model is able to capture most of the larger values from the observational dataset,

although there is a systematic overestimation of the smallest values. There is also a good match with the ERAInterim series of daily 10m wind speed, which shows smaller daily variability with respect to fitted model and observations. The explained variance is also relatively high (58% for the month of July, in general between 55% and 60%).

The results of the jackknife cross-validation are similar to those reported for temperatures, i.e. the standard deviation of differences between the downscaled time series based on full-period and cross-validated regressions is smaller than the standard deviation of the residuals with respect to the predictand data, and the correlogram of jackknife time series is very similar to the correlogram of the observed values. These results indicate that overfitting has been successfully avoided in the calibration procedure and the statistical model is able to reproduce the daily variability of the observed 10m wind speed.

Figure S2.17 shows q-q plots for two months, July and April. The match with a Gaussian distribution is quite good up to two standard deviations in both cases, similar to the results obtained for temperatures. However, in April the low and high tails are not well reproduced.In this respect, the statistical model can be improved by including a random noise component. In this case, since the model is not linear, the procedure is more complex than for temperature: the predictable part is used together with a dispersion parameter, estimated from the residuals, to derive the two parameters of the gamma distribution, from which random samples including both predictable and random components can be generated. The resulting distributions (not shown) are in good agreement with observed distributions, in particular lower and upper tails are quite well reproduced.

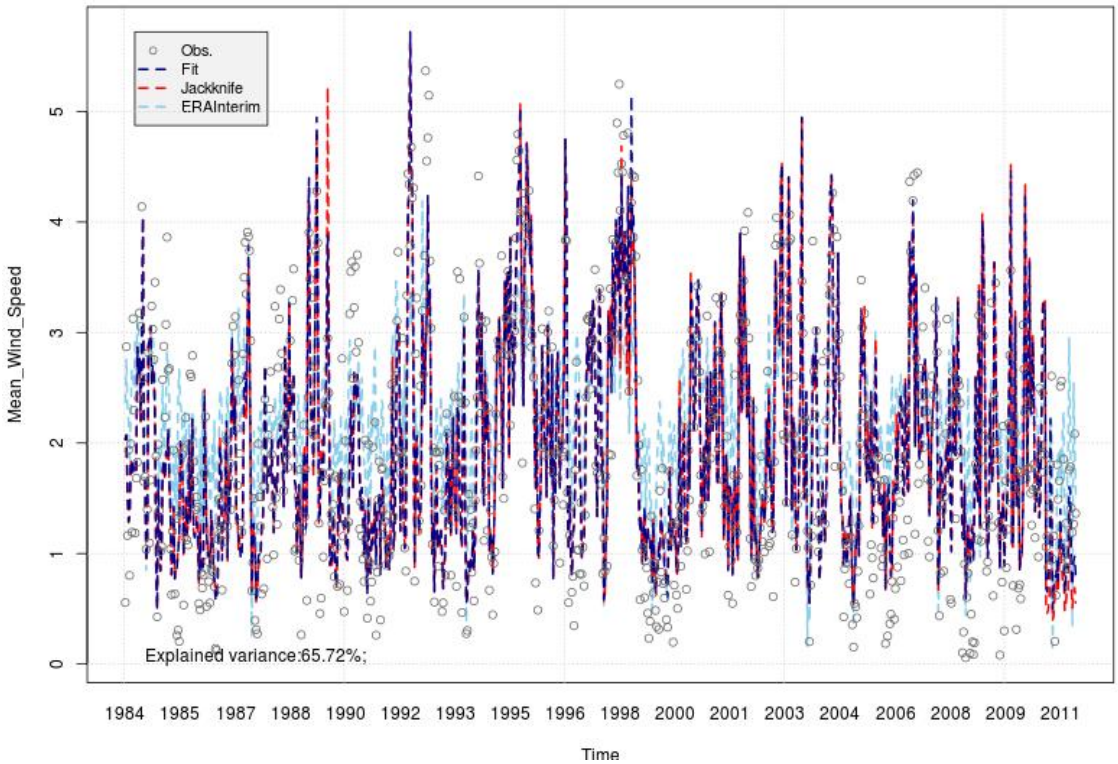

Figure S2.16: Calibration for 10m wind speed at Changi airport during July, 1981-2009, from the statistical model fit using PCs from the HadGEM2-ES/ERAInterim common EOF analysis. Fitted daily 10m wind speed (blue curve), fit from the cross-calibration procedure (red curve) , ERAInterim 10m wind speed for the Singapore grid box (skyblue curves) and observed values for Changi airport (open dots).

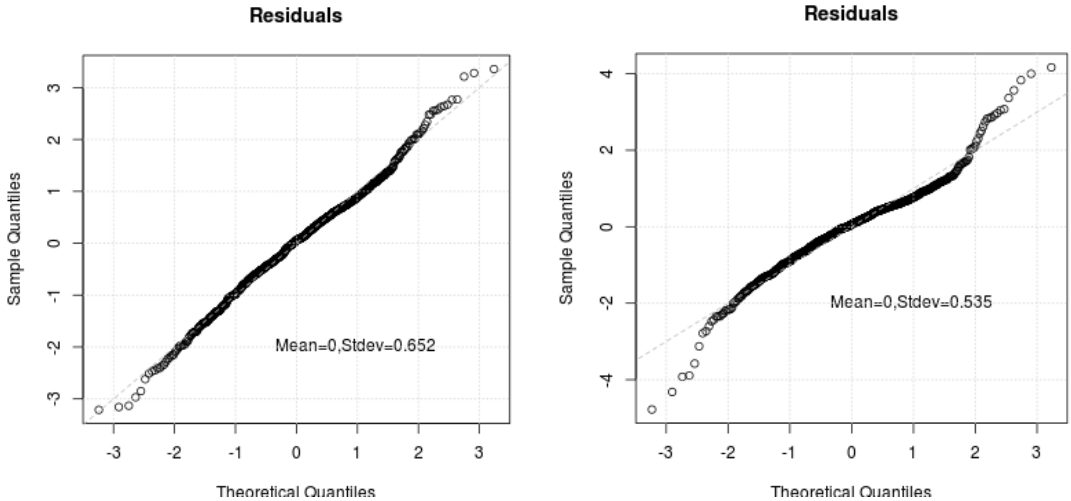

Figure S2.17: q-q plots for the residuals (per unit standard deviation) from the calibration for 10m wind speed at Changi airport for 1981-2009, using HadGEM2-ES/ERAInterim common EOFs: July (left), April (right).

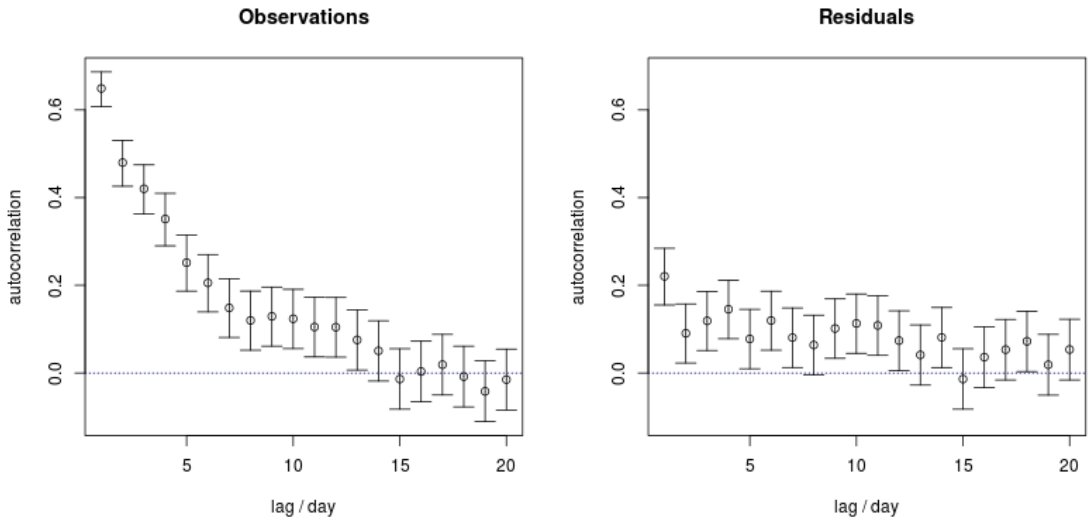

Figure S2.18: Autocorrelation for the observed daily 10m wind at Changi airport (left panel), and from the residuals of the downscaling calibration, from the analysis for April, 1981-2009.

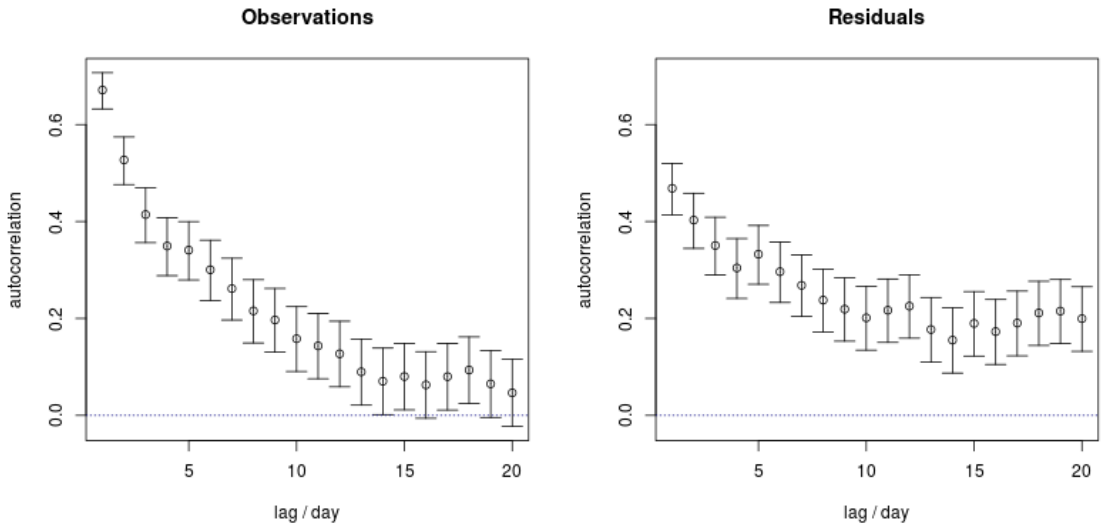

Figure S2.19: As Figure S2.18, for January, 1981-2009.

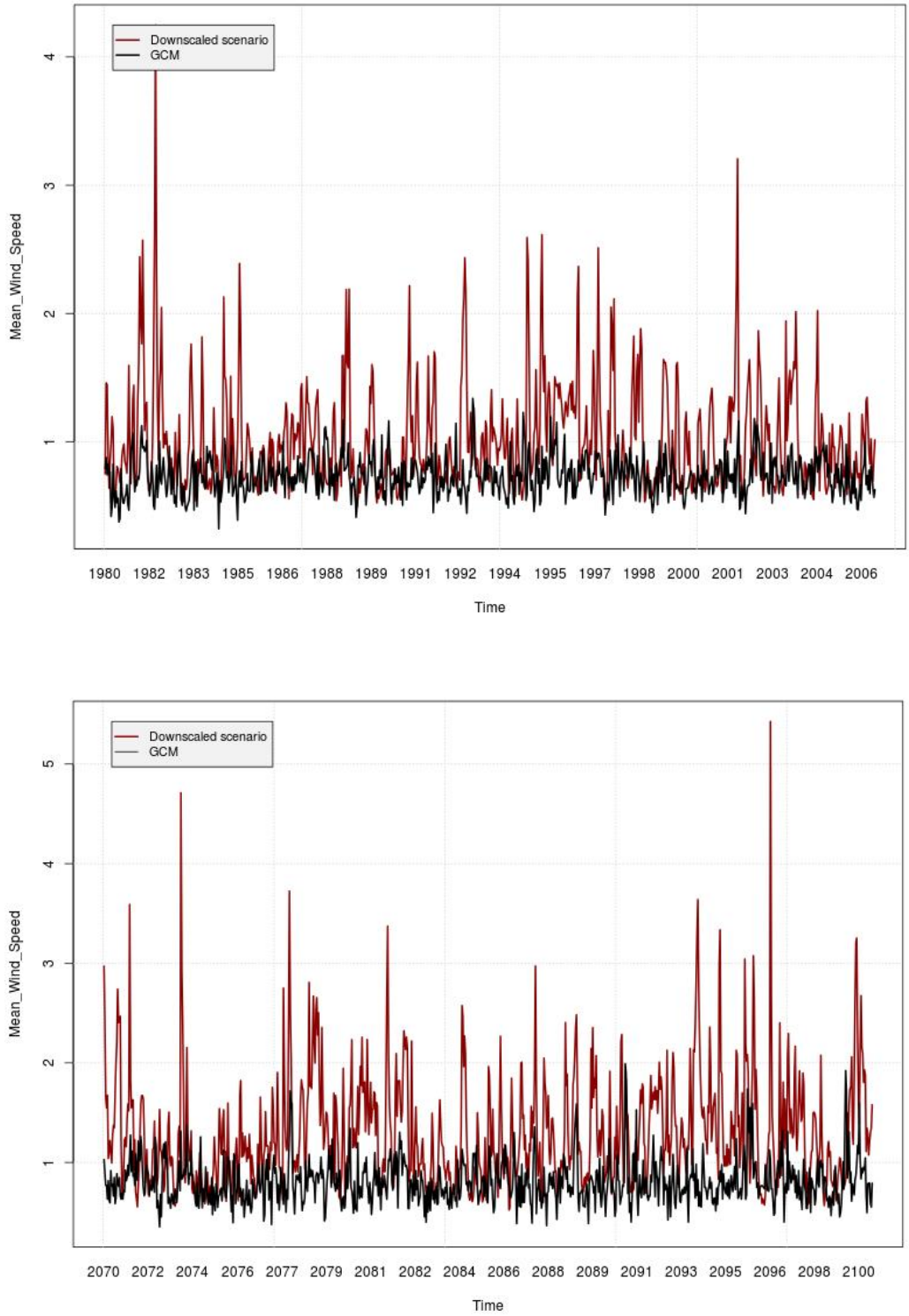

Figure S2.20: Downscaled HadGEM2-ES projections of average daily 10m wind speed at Changi airport for April, present climate (1980-2005), top panel, and future climate (2070-2099, RCP8.5), bottom panel, from the statistical model based on the HadGEM2-ES/ERAInterim common EOF analysis (red curves). The black curves are average daily 10m wind speed values for the Singapore grid box taken directly from HadGEM2-ES.

Figure S2.18 shows the correlograms for observations and residuals for April and supports the hypothesis that the statistical model has successfully fitted (and accounted for) the large scale influence on wind variability at Changi airport. However, the corresponding analysis for January (Figure S2.19) indicates a larger long range persistence of residual autocorrelations than the observed dataset.

Downscaled scenarios from bias corrected PCs from the HadGEM2-ES/ERAInterim common EOF analysis are reported in Figure S2.20, showing downscaled wind speed from the HadGEM2-ES present period and future climate (RCP8.5). The downscaled results show greater variability than the HadGEM2-ES values, and are closer to the distribution of station observations in this respect. Similar results have been obtained for January and July, and from a parallel analysis using bias corrected projection coefficients obtained from EOFs based on ERAInterim alone.

.

Linear statistical models generated for GFDL-CM3 show a different behaviour, with GCM model data usually showing larger mean values than the downscaled scenario, but with a consistent representation of daily variability in the GCM grid-box values and the downscaled wind speed. Since wind speeds are particularly strongly influenced by local effects, comparison against GCM values may not be a particularly appropriate test of the plausibility of the downscaled scenarios. Therefore, additional tests, possibly using RCM wind speed, could be useful. Indeed, this caveat applies, to some extent, to all the future scenario results given in this section. Comparisons against raw GCM output for the Singapore grid box are provided as a basic sanity check of the downscaled results, however it is precisely the purpose of the downscaling to add value to the GCM output. Therefore, the discovery of differences in the characteristics of the downscaled and non-downscaled results is not necessarily a cause for concern, provided the differences can be understood as expected consequences of the benefits conferred by downscaling, for example in achieving a better representation of the natural variability found in the station observations.

S2.5.6. Rainfall

As outlined in Section S2.2.3, a two step procedure has been used to model precipitation. Occurrence of wet days (defined with a threshold of 1mm/day) has been modelled by a Bernoulli distribution (i.e. the binomial distribution with n=1), with the logit of the probability expressed as a linear combination of predictors. Rainfall intensity has been modelled by a Gamma distribution, expressing the logarithm of rainfall intensity as a linear combination of the PC predictors. This statistical procedure has previously been used in other statistical downscaling studies (e.g. Sapiano, 2004, Fealy and Sweeney, 2007). Results presented in this section were obtained from a study focusing explicitly on model calibration, wherein PCs were obtained from EOF analysis on ERAInterim data only, using predictors derived from the same set of variables used for temperature (u, v and relative humidity up to 500hPa, mean sea level pressure and 2mT). Since a very small number of predictors show correlations with rainfall intensity and occurrence larger than 0.1, a threshold-based criterion for identification of potential predictors was not used. However, the total number of potential predictors was limited by considering only the five PCs with the largest explained variance. The threshold p-value used in the step wise algorithm was increased to 0.1, to ensure selection of a number of predictors consistent with the results obtained for the other predictands. Explained variances were lower than 50%, but analysis of correlations of PCs with residuals gave few predictors

with an absolute value of correlation larger than 0.05 for both rainfall intensity and occurrence (not shown). This result indicates that large scale signal has been efficiently captured by the two linear regression models.

,

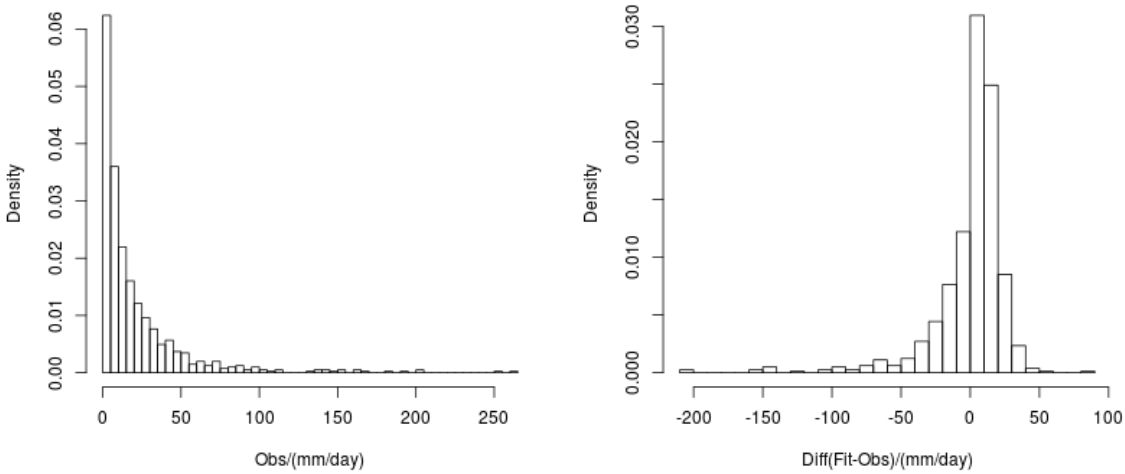

Figure S2.21: Histogram of observed daily rainfall at Changi airport for January (left panel), and of residuals from the downscaling calibration, based on common EOFs from HadGEM2-ES and ERAInterim (right panel)

Figure S2.21 shows the rainfall distribution for the month of January: the distribution is very wide and has a very long tail. This month falls in the wettest phase of the NE monsoon, however, even in this monsoonal month rainfall still occurs predominantly as a result of small scale convective activity, implying that strong links to large-scale predictor variables may be hard to find. During the inter-monsoonal periods the large scale forcing will be even weaker.

The calibration relationship for rainfall intensity produced fitted values in the middle range of the distribution of observed intensities (not shown), basically failing to reproduce any event with intensity smaller than 10mm/day. This result is consistent with the histogram of the residuals in Figure S2.21, which shows, not surprisingly, a negative tail of large underestimates resulting from a failure to reproduce the most extreme observed intensities, but also a rather large number of days with intensities overestimated by up to 50mm/day. This outcome is summarised by the rather small correlation between observed and fitted intensities, with a value of 0.5. This is consistent with the strong influence of local convective events referred to above. In this context, it is worth noting that the lack of a dominant large scale driver for the heavier rainfall events is likely to limit the predictability of any downscaling technique in a historical evaluation based on time series correlations. This is supported by the very small (absolute) values of the correlations between residuals and predictors and by the lack of significant autocorrelations in the residuals at any time lag. While the limitations in explained variance may not, in isolation, be sufficient to assess the methodology as unsuitable, the utility of the statistical model is further compromised by the fact that it cannot reproduce the mode of the observed distribution (Figure S2.21, left panel).

In addition, it turns out that the statistical model for intensity is also rather unstable. Specifically, for some periods the cross-calibration procedure could not be completed because the procedure to estimate the parameters did not converge. These numerical

convergence problems may indicate that the statistical procedure for representing intensity, while based on a state of the art approach from current literature, is not suitable to model the characteristics of daily rainfall in Singapore. It is unlikely that any other statistical downscaling approach available from the current literature would improve the description of the relationship with large scale variables, which needs to be explicitly described for applications in climate change studies. Therefore a new method, specifically designed for Singapore, would likely be required to downscale daily rainfall intensities.

S2.6. Stationarity of statistical relationships

A fundamental assumption in statistical downscaling is that relationships calibrated from the present climate can also be applied to the future, changed, climate. This assumption cannot be tested directly, as it would require the availability of observed data for the future climate. However, an indirect test can be performed by using RCM model output as a proxy of station data and the driving GCM as a source of large scale predictors (Murphy, 2000). This test assumes that RCM surface variables for a grid box can be assumed as representative of corresponding station data. An analysis of results obtained from this approach is discussed in this section.

The stationarity test is performed using HadGEM3-RA integrations at 12km resolution, driven by the HadGEM2-ES integration for the present climate and the RCP8.5 future concentration pathway. In particular, this analysis focuses on comparison of future downscaled variables predicted using relationships calibrated from the simulated present climate, against either RCM variables taken directly from the future climate integrations, or from alternative downscaled results based on relationships calibrated from the future data. This is sufficient to test the robustness to potential non-stationarities *of the final results from the multivariate regression relationships*. However, we do not attempt a more detailed assessment of the effects of climate change on *individual predictor-predictand relationships* contributing to the multivariate regression (Murphy, 2000). Such a study would be challenging for the methodology used in this work, since the iterative procedure used for calibration is liable to return somewhat different predictor sets when calibrated on present and future climate conditions. In particular, the choice is potentially sensitive to the presence of strong correlations between some of the PCs considered as potential predictors, which can lead to exclusion of a PC from the final calibrated statistical model simply because its effects might be captured by the correlated PCs, rather than because the predictor itself is not significant.

For this study, linear regression relationships between HadGEM2-ES predictors and HadGEM3-RA surface variables have been calibrated by using an EOF analysis from the HadGEM2-ES present climate integration (1980-2005). The analysis has been carried out using the RCM output (driven by the same HadGEM2-ES simulation) for the grid box which includes the Changi airport station (s24) as a proxy of station observations. From these relationships, downscaled scenarios for the period 2070-2099 from the HadGEM2-ES RCP8.5 integration have been generated, from predictors obtained by projecting HadGEM2-ES large scale variables from the future climate integration onto the present climate HadGEM2-ES EOFs. These downscaled variables have been compared with: 1) simulated RCM variables for 2070-99; 2) alternative statistically downscaled values derived by recalibrating the linear regression relationships using RCM surface variables, and EOFs and PC predictor values from HadGEM2-ES, all calculated from 2070-99 period of the simulations, rather than from

1980-2005. These statistically downscaled values represent the component of the predictand which is expected to be reproduced by predictions from linear regression models, since predictions are only based on the linear regression term of the statistical model and do take into account the contribution of residuals. However, a comparison of residuals from present and future climate calibrations is also relevant this study, as a simple test for the introduction of additive white noise terms in statistical models. It is worth noting that this test is only based on GCM large scale variables, therefore there is no need to include bias correction of GCM predictors.

An analysis of future daily temperature for the Apr-May period has shown a good agreement between predicted (i.e. statistically downscaled) variables and RCM grid box values. The explained variance is similar to the observed historical results obtained using station data and ERAInterim predictors. Differences between predicted values and future RCM grid box values during 2070-99 do not show any significant trend, indicating that the statistical model trained on present climate data is able to reproduce the climate change trend. However, there is a non negligible mean bias between predictions and RCM future values (0.6K). The comparison of daily variability is also good, as can be seen from the left panel in Figure S2.22. Linear trends have been removed from all datasets used for this Figure. The differences are distributed normally and their standard deviation (0.7K) is comparable to the standard deviation between fitted values and RCM values from the calibration period (0.6K). This result, which is common to all the other predictands in this study, supports the use of additive white noise terms (Von Storch, 1999), since they require a constant variance of residuals from present and future climate. The right panel in Figure S2.22 shows a q-q plot which compares two sets of predicted values, derived from calibrations based on historical and future data respectively: The differences are normally distributed and the standard deviation is rather small compared with the direct validation against RCM data (0.25K cf 0.7K in the left panel).

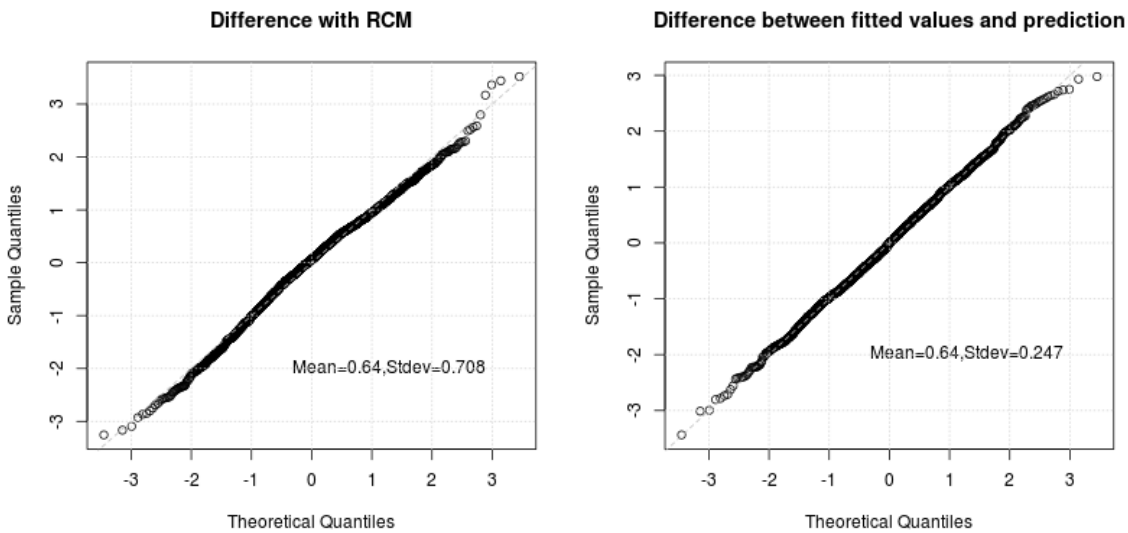

Figure S2.22: q-q plots of detrended daily temperature predicted from the linear regression model based on HadGEM2-ES and HadGEM3-RA present climate data with 1) detrended future RCM daily temperature from the HadGEM3-RA temperature for the Changi airport grid box (left panel) and 2) with detrended fitted values from the future HadGEM2-ES calibration period (right panel).

Similar results have been obtained from the other two periods (Dec-Jan and Jun-Jul), although the mean future bias is smaller in both periods (-0.1 for Dec-Jan and 0.2 for Apr-May).

Corresponding analysis for daily maximum temperature also does not show any trend in differences between predicted values and RCM variables for the future period. However, mean biases can be quite large (1.5K for Jun-Jul and Apr-May, 0.5K for Dec-Jan). The spread of differences between predicted values and future RCM values is also quite large (1.3-1.5K) for all sub-seasons. The RCM Tmax distribution is more variable than the station data distribution and the variance explained by fitting the statistical model to RCM values for the present climate is smaller than that obtained from calibration using observed station data. These results indicate that RCM Tmax may not be a good proxy for observed Tmax; therefore, it is difficult to draw any inference regarding the stationarity of the Tmax relationship from these results.

For daily minimum temperature, results are similar to those obtained from daily average temperature, both for mean bias and daily variability. The explained variances for both present and future climate statistical models, in application to RCM data, are larger than the observed values obtained from the station record. Analysis of the differences between predictions from relationships calibrated from present climate data and RCM grid box values for the future climate shows a significant trend (p-value 0.01), but with very small values (0.1K/decade) compared with the rates of change simulated by the RCM itself.

Statistical predictions of RCM daily mean wind speed for the Changi airport grid box give similar results: Calibration based on the historical GCM and RCM time series give. explained variance similar to observed downscaling relationships using station data; Future predictions using the historically trained relationships compare well with both the future RCM values, and (to an even greater extent) with fitted values from a calibration based on future climate model data. The mean bias in the future projections is also quite small (0.1-0.2 m/s). The comparison between predictions based on historical relationships, and fitted values from the future climate calibration, is shown in Figure S2.23 (Jul, 2070-2099). This demonstrates the good agreement between predicted (blue) and calibrated (red) values referred to above, although some of the largest values in the future-calibrated time series are overestimated by the predictions derived from historical relationships.

Figure S2.23 also shows the HadGEM2-ES time series of wind speed for the Singapore grid-box. Here, we show raw GCM rather than raw RCM wind speeds, in order to illustrate the caveats involved with using future GCM values to check the downscaling results (cf Figures S2.7 and s2.12, for example). The two downscaled curves are shifted to higher values with respect to the GCM time series, but not with respect to the RCM time series for Changi airport (not shown). Therefore, the bias between the statistically downscaled and GCM time series does not, in this case. indicate a shortcoming in the downscaling methodology. Rather, this difference occurs because the RCM grid box wind speeds (not shown) give a smaller frequency of low wind speeds than the corresponding GCM data, which effectively represent values spatially averaged over a much larger area.

The wind speed assessment also supports the use of a set of predictors which excludes temperature at all levels (see earlier discussion in Section S2.5.5), since the remaining predictors can still produce realistic estimates of downscaled wind speed for the future climate. Recalling that successful downscaling of wind speed requires a random

component derived from the calibration residuals (Section S2.5.5), the stationarity of the random component was also assessed: dispersion parameters from the present and future climate calibrations were found to be in good agreement. Furthermore, the statistical model including a random component, estimated from RCM present climate output, gives distributions of future 10m daily wind speed which agree quite well with the distribution of future 10m wind speed simulated by the RCM.

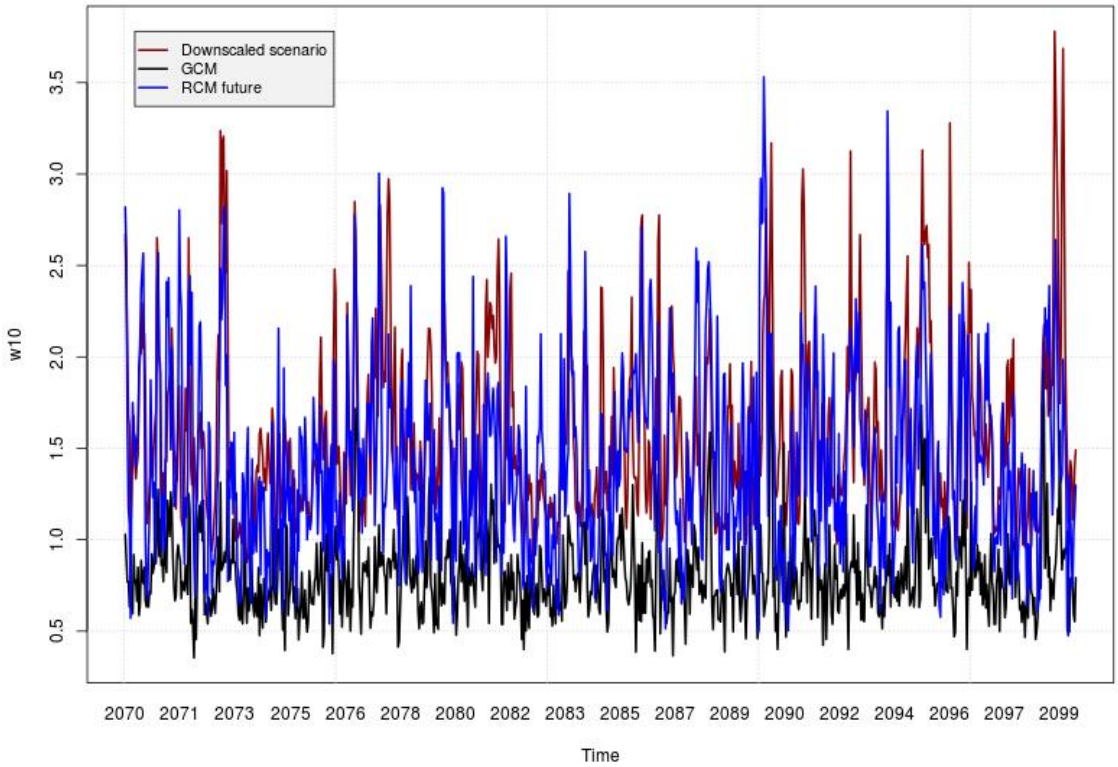

Figure S2.23: Comparison between downscaled time series of daily average 10m wind speed predicted from the statistical model trained on the HadGEM2-ES/HadGEM3-RA present climate simulations (red curve), fitted values from the statistical model calibrated using future HadGEM2-ES/HadGEM3-RA data (blue curve), and average 10m wind speed for the Singapore grid-box taken from the HadGEM2-ES RCP8.5 integration, for April (2070-2099).

S2.7. Summary

An assessment of the applicability of Empirical Statistical Downscaling, based on linear regression models calibrated using historical data from 1980-2010 as a training period, has been presented in this report. Five observed station datasets, for mean daily temperature, daily minimum and maximum temperature, daily mean wind speed and rainfall, have been used as predictands. Downscaling relationships are built using linear regression models, calibrated using large scale predictor variables derived from a set of key variables describing the day-by-day evolution of the state of the atmosphere in a wider region surrounding Singapore.

Before defining predictor variables, it is necessary to reduce the dimensionality of the large scale data. This is done using a common EOF (Empirical Orthogonal function) analysis (Benestad). In this approach, EOFs, which describe leading spatial patterns of variability, are obtained from a concatenated dataset of large scale variables consisting

of a time series of ERAInterim reanalyses of observations during for the historical period, plus corresponding time series from both historical and future portions of a global (CMIP5) climate model simulation selected to provide future downscaled projections. The common EOF approach identifies, by construction, patterns of variability present in both the observational and GCM data, thus facilitating the application to GCM predictor variables of downscaling relationships calibrated using observations.

Predictor variables for the downscaling relationships are then drawn from principal components (PCs), which represent the time-dependent contributions of each common EOF to the regional state of the atmosphere, in either reanalysis or GCM data. Following calibration of the relationships, downscaled projections are then obtained by applying the relationships to PC time series taken from both historical and future time slices of the relevant CMIP5 simulation. For this purpose, bias corrections are applied to the historical and future PC time series, designed to correct empirically for biases in the mean and standard deviation with respect to observations, without altering future climate change trends simulated by the GCM. Example results are provided using two CMIP5 GCMs, HadGEM2-ES and GFDL-CM3.

The results show that:

- For mean daily temperature and maximum daily temperature, good quality linear regression models can be calibrated from the observed climate. These are able to reproduce observed daily variability consistent with station observations and ERAInterim data, although the downscaled results do not capture well the lower end of the station distribution, particularly for daily maximum temperature. With this caveat, the regression model is assessed as suitable for use in downscaling other GCMs.
- For minimum daily temperature, the calibration step shows only moderate predictability of local values from large scale weather variables. However, in this case, the linear regression models can be extended to include noise terms which describe the remaining variance not predictable from the large scale variables. When applied to GCM data, the variance of the predictable term of the statistical model improves when the predictor time series are bias corrected.
- Daily station wind speed can be statistically modelled from observed data quite well, with a Generalised Linear Model (GLM) approach based on a Gamma distribution in which the location parameter is expressed as the exponential of a linear combination of predictors while the dispersion parameter is estimated from the residuals. The downscaled wind speed from bias corrected GCM predictors is reasonable, and is improved by inclusion of a non-linear random noise component, achieved by using a dispersion parameter estimated from fitting residuals in the calibration of the GLM. This produces a statistical model which improves the description of the tails of the distribution.
- An attempt to downscale rainfall was made, using a more complex, two-step procedure, in which occurrence and intensity are both represented. However, it was not possible to produce acceptable calibration relationships for Singapore, so we assess the method to be unsuitable for rainfall downscaling.
- Since the downscaled relationships are all based on linear regressions, daily extremes are in general not well represented. Use of predictors from EOF analysis may also contribute to this problem, since higher order (and less frequent) modes of variability are not used in these statistical models
- The plausibility of the downscaled future changes is checked through comparison with changes in corresponding variables simulated by the relevant CMIP5 model, at the grid box containing Singapore. In general, the downscaling results show

similar long term signals of change, but differences in future variability, reflecting the benefits conferred by accounting better for the characteristics of local station distributions through the predictor-predictand relationships. However, for variables in which such local influences are relatively dominant, such as 10m wind speed, comparison against GCM results is of limited utility in assessing credibility. In such cases, comparison with dynamically downscaled results using regional climate models is an important additional check.

- The availability of regional climate model (RCM) simulations from the dynamical downscaling work provided an opportunity to carry out this further test. In particular, this involved testing a key assumption in these empirical downscaling results, that the observed predictor-predictand relationships can be assumed stationary (i.e. invariant) in the future. The HadGEM3-RA integrations (historical and RCP8.5 scenario) driven by HadGEM2-ES were used, taking HadGEM3-RA surface variables for the Changi airport grid box as predictands and HadGEM2-ES large scale variables as predictors. Predicted values from linear regressions calibrated using data from the historical integration were compared with future values taken directly from the RCM, and also with fitted values from alternative linear regressions using a calibration based on the future climate period. The latter describes the fraction of predictand variance explainable by relationships with large scale variables. Results showed that Tmean, Tmin and 10m daily wind speed were predicted quite well by the linear regression models used in the present study. Tmax results were more difficult to interpret, since the RCM grid box values have a larger spread than observed Tmax; this result indicates that the RCM Tmax for the Changi airport grid box may not be a good proxy for its observed counterpart, hence the stationarity test for this predictand is inconclusive. For daily mean temperature and Tmin, residuals from the calibration using historical climate data are comparable to the residuals obtained from calibration using future climate data, supporting the use of additive white noise terms to model the variance not explained by the linear regression models. Similar results were obtained for 10m daily wind speed, supporting inclusion of a stochastic component in the GLM model used for statistical downscaling of this variable, as described above.
- Since the software is not computationally expensive, the methodology provides an opportunity to obtain downscaled projections from a wider range of CMIP5 models than can be covered using the dynamical approach of Chapter 5 in the V2 project. However, the method needs careful assessment on a case-by-case basis, since a separate common EOF analysis and statistical calibration is needed for each CMIP5 model. Credibility also depends on the quality of the simulation of regional climate by the CMIP5 model (for example, large errors can necessitate excessively large bias corrections), and on whether the future downscaled scenario provides local climate change estimates consistent with the grid-scale changes given by the CMIP5 model (see above) .
- An alternative method of defining EOFs was tested, in which only reanalysis data is used to define the optimal patterns of variability (e.g. Murphy, 2000), which can then be projected onto either observational data (when calibrating predictor-predictand relationships), or GCM data (when obtaining downscaled climate scenarios using these relationships). This method was used to test the robustness of results derived using the common EOF approach, and was found to confirm the main conclusions above. We recommend use of the common EOF method where feasible, since the bias correction procedure described above is more readily justifiable for EOFs representing both observed and GCM modes of variability. However, the alternative method is also available from the

software provided, noting that it may provide a more convenient option in applications involve downscaling many CMIP5 models, since the EOFs only have to be calculated once, whereas common EOFs have to be calculated separately for each climate model.

References

Barnett, T. P. 1999. Comparison of near-surface air temperature variability in 11 coupled global climate models. J. Climate ,**12**, 511–518.

Benestad, R. E. 2001. A comparison between two empirical downscaling strategies. Int. J. Climatol., **21**, 1645–1668.

Benestad R.E., 2012, clim.pact: Climate analysis and empirical-statistical downscaling (ESD) package for monthly and daily data.. R package version 2.3-10, http://CRAN.R-project.org/package=clim.pact

Benestad, R., Hanssen-Bauer I., Chen D., 2008, *Empirical-Statistical Downscaling*. World Scientific Publishing Co., Singapore, 215 pp.

Cavazos T and Hewitson BC, 2005, Performance of NCEP–NCAR reanalysis variables in statistical downscaling of daily precipitation, Clim. Res., **28**, 95-107

Dee, D. P. and 35 co-authors., 2011, The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Q. J. R. Meteorol. Soc., **137**, 553–597.

Fealy, R and Sweeney, J,(2007. Statistical downscaling of precipitation for a selection of sites in Ireland employing a generalised linear modelling approach. Int J of Climatology, **27**, 2083-2094.

Goodess C.M., Anagnostopoulou C., Bárdossy A., Frei C., Harpham C., Haylock M.R., Hundecha Y., Maheras P., Ribalaygua J., Schmidli J., Schmith T., Tolika K., Tomozeiu R. and Wilby R.L., 2005, An intercomparison of statistical downscaling methods for Europe and European regions – assessing their performance with respect to extreme temperature and precipitation events, Climatic Research Unit Research Publication 11 (CRU RP11), http://www.cru.uea.ac.uk/cru/pubs/crurp/CRU_RP11.pdf

Imbert A, Benestad RE. 2005. An improvement of analog model strategy for more reliable local climate change scenarios. Theoretical and Applied Climatology, **82**, 245–255

Manzanas R, Gutierrez J.M., Cofino A.S., San-Martın D., Tuni M., Herrera S.,2012; D3.1.d: Regional seasonal predictions, with uncertainty estimations, in the target regions, and inclusion in the QWeCI Statistical Downscaling Portal, accessible to partners in user-friendly formats, QWeCI, http://www.liv.ac.uk/media/livacuk/qweci/Deliverable,3.1.d_CSIC_SUBMITTED.pdf

Murphy J. 2000. Predictions of climate change over Europe using statistical and dynamical down-scaling techniques. Int. J. of Climatology, **20**, 489–501.

North, G. R., T. L. Bell, R. F. Cahalan, and F. J. Moeng, 1982: Sampling errors in the estimation of empirical orthogonal functions. Mon. Wea. Rev., **110**, 699–706

R Core Team (2012). R: A *language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, *http://www.R-project.org/*

Sapiano M, 2004, *Trends and variablity in observations of winter precipitation*, PhD thesis, Reading University.

Taylor, K.E., R.J. Stouffer, G.A. Meehl, 2012, An Overview of CMIP5 and the experiment design." Bull. Amer. Meteor. Soc., **93**, 485-498.

Venables, W.N. and Ripley, B.D., 1994, *Modern Applied Statistics with S-PLUS*. Springer-Verlag, New York.

von Storch H, 1999, On the Use of "Inflation" in Statistical Downscaling, J. Climate, **12**, 3505-3506.

von Storch, H. and Zwiers, F.W., 1999, *Statistical analysis in climate research*, Cambridge University Press, Cambridge.

Wilby RL, Charles SP, Zorita E, Timbal B, Whetton P, Mearns LO, 2004, Guidelines for Use of Climate Scenario Developed from Statistical Downscaling Methods, TGICA IPCC, 1-27

Wilks D.S., 1995, Statistical methods in the atmospheric sciences, second edition International Geophysics Series, Vol 59, Academic Press.

Xu, Z and Yang, Z, 2012.: An Improved Dynamical Downscaling Method with GCM Bias Corrections and Its Validation with 30 Years of Climate Simulations. *J. Climate*, **25**, 6271–6286.

Yan, Z., Bate, S., Chandler, R.E., Isham, V. and Wheater, H., 2006, Changes in extreme wind speeds in NW Europe simulated by generalized linear models. Theoretical and Applied Climatology, **83**, 121-137.

Zorita, E., and H. von Storch, 1999: The analog method as a simple statistical downscaling technique: Comparison with more complicated methods. J. Climate, **12**, 2474–2489