STARDEX

STAtistical and Regional dynamical Downscaling of EXtremes for European regions

EVK2-CT-2001-00115

Deliverable D16 – Summary Report

Recommendations on the more robust statistical and dynamical downscaling methods for the construction of scenarios of extremes

FOREWORD

The STARDEX project on STAtistical and Regional dynamical Downscaling of EXtremes for European regions is a research project supported by the European Commission under the Fifth Framework Programme and contributing to the implementation of the Key Action "global change, climate and biodiversity" within the Environment, Energy and Sustainable Development.

STARDEX will provide a rigorous and systematic inter-comparison and evaluation of statistical and dynamical downscaling methods for the construction of scenarios of extremes. The more robust techniques will be identified and used to produce future scenarios of extremes for European case-study regions for the end of the 21st century. These will help to address the vital question as to whether extremes will occur more frequently in the future.

For more information about STARDEX, contact the project co-ordinator Clare Goodess (c.goodess@uea.ac.uk) or visit the STARDEX web site:
http://www.cru.uea.ac.uk/projects/stardex/

STARDEX is part of a co-operative cluster of projects exploring future changes in extreme events in response to global warming. The other members of the cluster are MICE and PRUDENCE. This research is highly relevant to current climate related problems in Europe. More information about this cluster of projects is available through the MPS Portal:
http://www.cru.uea.ac.uk/projects/mps/

STARDEX is organised into five workpackages including Workpackage 4 on 'Inter-comparison of improved downscaling methods with emphasis on extremes' which was responsible for the production of this deliverable (D16). Workpackage 3 is co-ordinated by Torben Schmith from the Danish Meteorological Institute.

STARDEX PROJECT MEMBERS

UEA	University of East Anglia, UK
KCL	King's College London, UK
FIC	Fundación para la Investigación del Clima, Spain
UNIBE	University of Berne, Switzerland
CNRS	Centre National de la Recherche Scientifique, France
ARPA-SMR	Servizio Meteorologico Regionale, ARPA-SMR Emilia-Romagna, Italy
ADGB	University of Bologna, Italy
DMI	Danish Meteorological Institute, Denmark
ETH	Swiss Federal Institute of Technology, Switzerland
FTS	Fachhochschule Stuttgart – Hochschule für Technik, Germany
USTUTT-IWS	Institut für Wasserbau, Germany
AUTH	University of Thessaloniki, Greece

D16 AUTHORS AND VERSION HISTORY

Lead author:
Clare Goodess, UEA

Contributing authors:
Stefano Alberghi, ADGB
Christina Anagnostopoulou, AUTH
András Bárdossy, USTUTT-IWS
Carlo Cacciamani, ARPA-SIM
Bo Christiansen, DMI
Dimitrios Gyalistras, UNIBE
Yeshewatesfa Hundecha, USTUTT-IWS
Panagiotis Maheras, AUTH
Antonella Morgillo, ARPA-SIM
Valentina Pavan, ARPA-SIM
Jaime Ribalaygua, FIC
Jürg Schmidli, ETH
Torben Schmith, DMI
Konstantia Tolika, AUTH
Rodica Tomozeiu, ARPA-SIM
Luis Torres, FIC
Ennio Tosi, ADGB

Version 1.0: 27 July 2005
Version 2.0: 30 September 2005

CONTENTS

D16 objectives and target audience

STARDEX measurable objective 4 is:

- Inter-comparison of improved statistical and dynamical downscaling methods using data from the second half of the 20th century and identification of the more robust methods.

The first part of this objective is addressed by deliverable D12, while this deliverable (D16) addresses the second part and makes recommendations on the more robust statistical and dynamical downscaling methods for the construction of scenarios of extremes.

This summary report provides guidelines for those wishing to undertake their own downscaling, building on the IPCC guidelines (Wilby *et al*., 2004; Mearns *et al*., 2003). The checklist of good practice provided here is also likely to be valuable for impacts scientists and other users who want to assess the robustness and reliability of downscaling results from other sources (e.g., to what extent is the recommended good practice followed in these other studies?). The recommendations and guidelines are based on the analyses performed during the project, but some proposals are also made for additional analyses, together with suggestions for how we might do things differently next time. These good practice guidelines form part of the STARDEX downscaling 'toolbox' – the statistical downscaling models available in our toolbox are described in D15 and D12. Application of these tools to construct future scenarios is described in D18.

The D16 partner contributions provide summary documentation concerning the robustness and reliability of the specific statistical downscaling method(s) developed and evaluated by each STARDEX group. This documentation will help users to identify the most appropriate of the STARDEX downscaling methods for their purpose, but is also directed at people wanting to use the scenario results presented in D18 - which are available from the STARDEX central data archive.

Note that D16 (and the criteria in Table 1) focuses on the robustness and reliability of the underlying statistical downscaling models themselves, together with the issue of model stationarity. Additional issues concerning the robustness and reliability of the resulting scenarios (such as the coherency of the predictor and scenario changes) are considered in D18 and D19.

STARDEX assessment criteria and definitions

STARDEX has developed three sets of criteria for assessing the robustness of statistical downscaling methods and the reliability and appropriateness of statistical/dynamical downscaling methods. These are referred to as robustness, performance and application criteria respectively (with the latter most focused on user needs).

The robustness criteria reflect the key assumptions of statistical downscaling (see Section 2.4 of the IPCC guidelines, Wilby *et al*., 2004) and comprise of four elements (Table 1). These are : strength and stability; stationarity; uniformity of performance; and, reliability of the simulation of predictors.

Uniformity of performance is self-explanatory and clearly important, as is reliability of predictor simulation. If predictors are reliably simulated, this implies that one of the assumptions of statistical downscaling – that large-scale circulation variables are better simulated than local-scale surface climate variables, is met.

Stability and stationarity are distinct issues – as indicated by the definitions below and the different assessment methods used for each (Table 1).

Stability: This concerns the stability and sensitivity of the statistical model over the calibration/validation period, i.e., over the observational period. Stability relates to both the selection of predictor variables and the relative weight or influence that particular predictors are given within the model.

Stationarity: This concerns the applicability of the statistical model to future time periods with climate change, i.e., to what extent is it legitimate to extrapolate a statistical model which has been calibrated/validated on the observational period to a future warmer period? Stationarity is considered as a separate issue to stability because a stable statistical model which performs well for the present-day is not necessarily the one that will perform best for the future.

STARDEX deliverables and key conclusions for D16

The recommendations and guidelines presented in this deliverable are based on a synthesis of extensive analyses undertaken during earlier stages of the STARDEX project, which are reported in a number of other deliverables, all available from the public web site. Here, the key messages for D16 from these deliverables are summarized:

D9: spatially coherent changes in European extremes have been observed over the last 40 years - one challenge for downscaling methods is to reproduce these changes.

D10: it is easier to make recommendations about appropriate methodologies for selecting predictors than the specific predictors which should be used in a particular study, although lists of potentially useful predictors can be recommended. Automated methods for predictor selection are generally less suitable and there tends to be a need for user intervention and local expertise.

D11: clearly demonstrates the need for downscaling and provides a baseline for assessing the added value of downscaling. It also demonstrates that some issues can only be properly addressed by upscaling station data to allow fair comparisons to be made.

D12: demonstrates that handling many combinations of different methods, regions, indices, seasons is difficult. These problems can be overcome, but nonetheless non-systematic variations in skill occur and no single best method can be identified. Care is also needed because methods/indices with the highest correlations are often not those with the lowest biases or RMSE - so several different statistical tests should be used to assess performance.
The two key recommendations from D12 are to use a range of statistical downscaling methods and that in some cases skill may be unacceptably low to warrant the construction of scenarios, e.g., in the case of summer rainfall extremes in a number of regions.

D13 – the large-scale predictors employed in the STARDEX statistical downscaling models are generally well simulated by the climate models considered, although there are some exceptions, typically moisture-related variables in summer, particularly over Southern Europe.

The STARDEX robustness criteria

Table 1 lists the STARDEX robustness criteria and then summarises the key questions and assessment methods used, indicating the STARDEX deliverables and papers where the latter are presented. Thus this table provides an overview of much of the work undertaken in the STARDEX project. In some cases, novel approaches are proposed for exploring the issues further in future studies.

Table 1: The STARDEX robustness criteria and summary of assessment methods. The final column indicates the STARDEX deliverables and papers where the latter are presented. For deliverables, the most relevant partner reports are indicated in brackets.

Robustness criteria	**Key questions and recommended assessment methods**	**STARDEX examples**
'Strength and stability'	Can strong predictor/predictand relationships be identified, supported by, for example, high correlation values?	D10; Haylock and Goodess, 2004
	Are these relationships physically meaningful, i.e., supported by literature review, theoretical considerations and/or local meteorological/synoptic climatology evidence and expertise?	D10; Haylock and Goodess, 2004; Maheras *et al*., 2004; Schmidli *et al*., 2005a
	If different methods (e.g., correlation, stepwise multiple regression, compositing) or time periods are used for predictor selection, are similar sets of predictors obtained?	D10 (KCL); D12
	Is the strength of the predictor/predictand relationships and/or the performance of the statistical downscaling model sensitive to changes in calibration/validation period (e.g., relatively longer/shorter periods and/or swapping the calibration/validation periods)?	D12 (ARPA-SMR), Tolika *et al*., 2005
	Is the statistical model performance sensitive to other user choices, such as predictor domain, number of predictors and model parameters (e.g., choice of misfit term in neural network models)?	D10; D12 (UNIBE); Gyalistras and Schuepbach, in preparation

'Stationarity'	Minimise potential problems by the incorporation of predictors which are expected to change due to global warming (e.g., moisture flux and specific/relative humidity), based on literature review and theoretical considerations. Assess whether the direction and magnitude of observed trends in the predictand, together with low-frequency variability, are reproduced by the statistical model (ideally using cross-validation in order to maximize the analysis period). Assess whether the projected changes in predictor variables lie outside the range of variability observed over the calibration/validation period. Assess whether predictor/predictand relationships calculated from GCM/RCM output change between the control and perturbed periods. Calibrate the statistical model on a 'cold' period and validate it on a 'warm' period and *vice versa*. Calibrate the statistical model in one region and apply it (without re-calibration) in a warmer region with equally simple topography.	D10 (FIC) D12 (ARPA-SIM, AUTH) D18 USTUTT-IWS, paper in preparation Proposed method not tested.
Uniformity of performance	Uniformity of statistical model performance across: - stations - regions - seasons - variables (i.e., temperature *vs* precipitation, means *vs* extremes) - indices of extremes (e.g., occurrence *vs* magnitude) Evaluated for present-day conditions using: - BIAS (mean difference between simulated and observed values) - CORR (Spearman rank-correlation coefficient) - RMSE (Root Mean Square Error) - ratio of observed : simulated standard deviations Plotted using: - Maps - Histograms - Box-whisker plots - Q-Q diagrams - Taylor diagrams	D12; D16 – see performance criteria tables; Goodess *et al*., 2005; Schmidli *et al.*, 2005b; Haylock *et al*., 2005
Reliability of simulation of predictors	Compare predictors calculated from climate model output with those calculated from Reanalysis data, taking into consideration: - raw values (e.g., sea level pressure, 500 hPa geopotential height, relative/specific humidity) - derived indices (e.g., principal components, blocking indices, synoptic circulation types) - spatial patterns - temporal trends - frequency and persistence, and day-to-day transitions, of circulation/weather types.	D13

	Is performance of the statistical model for the control period degraded when predictors are taken from climate model output rather than Reanalysis data?	D18, e.g. Tolika *et al*., 2005

STARDEX application criteria

Good practice in scenario development should include demonstration of the need for downscaling (D11 and Wilby *et al.* 2004) for each user's specific application. For some applications, this need will be more pressing, e.g., those requiring information about extremes at a high temporal and spatial resolution. The applications user needs will also determine which downscaling approach (dynamical or statistical) is most appropriate and, if statistical downscaling is adopted, which specific methods are most appropriate. The STARDEX application criteria have been developed for this purpose. They encompass spatial and temporal scale and consistency, together with resource (computing and data) requirements.

Completed application criteria for dynamical and statistical downscaling approaches are shown in Tables 2 and 3 respectively. These have been completed for the general approaches. Application criteria for specific statistical downscaling methods are shown in the D16 partner contributions.

STARDEX performance criteria

The third set of STARDEX criteria is the performance criteria. Evaluating model skill using independent data is a crucial element of any downscaling application (Wilby *et al*., 2004). The STARDEX downscaling methods have been extensively evaluated for present-day conditions using Reanalysis data and focusing on extreme events (D12; Goodess *et al*., 2005). Particular emphasis was given to how well interannual variability is reproduced (measured by correlations – see Table 1), as a proxy for climate change. The reasoning is that if year-to-year changes can be modelled, you can have much more faith that you do indeed have all the relevant predictors and so your scenarios are more meaningful (than if only the biases are small). If interannual variability can't be well modelled, this implies that relevant predictors may be missing or that noise far overshadows any model skill.

Whilst such evaluation using rigorous statistical testing is vital, the volume and detail of results are likely to be more than most users require, particularly if they are trying to inter-compare and identify a handful of most appropriate methods for their specific application. Thus the STARDEX performance criteria tables (an example is shown in Table 4) attempt to summarise the relative performance confidence and overall performance of each method, as well as indicating the optimal spatial scale and recommended impact applications.

Key recommendations

Previous studies have identified and summarised the theoretical advantages and disadvantages of different downscaling methods (e.g., Goodess *et al*., 2003). STARDEX has advanced the science by developing assessment methods and criteria for the more objective identification and selection of appropriate downscaling methods.

The STARDEX criteria shown in Tables 1-4 are flexible enough to be used in any study. They can be used to assess existing methods and resulting scenarios, and can also be used in new studies, e.g., studies considering variables other than temperature and precipitation which were the two variables considered by STARDEX.

Recommended good practice in statistical downscaling consists in following the IPCC guidelines (Wilby *et al*., 2004) and undertaking sufficient analyses and assessments to address the STARDEX criteria shown in Tables 1-4.

Further recommendations and points of good practice based on the STARDEX experience are summarized below.

Planning downscaling studies

- Identification of methodologies for ensuring consistent and fair comparisons requires a lot of thinking and planning (e.g., agreeing principles of verification). STARDEX chose a case-study matrix approach, allowing more than 20 statistical downscaling methods to be tested in a number of different regions (see Goodess *et al*., 2005). For other inter-comparison studies, it may be more appropriate to focus on a single common region.

- Don't underestimate the time/importance of 'preliminary' work for statistical downscaling, i.e., assembling, reformatting and quality control of data sets; identification and selection of predictors; and, testing model sensitivity.

Calibration and validation of statistical downscaling models

- Use independent calibration/validation periods and/or cross-validation.

- Undertake sensitivity studies (e.g., using different predictors, domains, calibration/validation periods, parameter choices).

- Don't forget the underlying physical processes – or underestimate the importance of local expertise and knowledge in predictor selection and assessing model performance.

- Validate potential predictor variables, i.e., are they reliably simulated by the climate models used?

- Demonstrate the added value of downscaling (see, for example, Schmidli *et al*., 2005a, 2005b).

Multi-model approaches

- Test/use a range of statistical downscaling methods (this is a key message from Goodess *et al*., 2005 and STARDEX deliverable D12), focusing on those that best fit the user application requirements (see Table 2).

- Two somewhat contradictory messages emerge from the STARDEX work: use multi-model ensembles of methods, but avoid black-box approaches. The lack of fullyautomated methods implies a lot of work for statistical downscaling. But neither should dynamical downscaling be used without appropriate validation work (which can also be time consuming, particularly if appropriate observed data are not readily available).

- Comparison of statistical and dynamical downscaling is desirable (see, for example, Schmidli *et al*., 2005a,b; Haylock *et al*., 2005), ideally this should involve upscaling to allow fair comparisons to be made.

- The European Union-funded ENSEMBLES integrated project (http://www.ensembles-eu.org/) will involve extension of the multi-model ensembles approach (and also consider the sensitivity of impacts to multi-model ensembles).

Application of statistical downscaling models

- Scenarios should not be constructed for cases where skill is unacceptably low.

- Do not assume that evaluation results for robustness and, in particular, performance criteria, are applicable to other regions. Thus some assessment of these issues is always likely to be needed before transferring methods to other regions.

- Document methods, user choices and underlying assumptions and ensure that this information is readily available in appropriate forms for users.

- Robust models do not guarantee reliable and plausible scenarios. The coherency of predictor changes and the sensitivity of changes to methods and choices needs to be considered (see STARDEX deliverables D18 and D19, and associated papers, e.g. Haylock *et al*., 2005; Schmidli *et al*., 2005b).

Concluding remarks

The STARDEX project has demonstrated that while statistical downscaling is less demanding than dynamical downscaling in terms of completing resources, the development and evaluation of robust, improved methods is scientifically demanding and time consuming. This D16 summary report attempts to synthesise the many lessons learnt from the project (which are reported in detail in other deliverable reports and journal papers) in the form of criteria and recommendations which are more readily accessible to developers and users of statistical downscaling. The D16 partner contributions which accompany this summary report document how the STARDEX criteria have been implemented by partners – thus providing detailed evidence of their utility.

References

Goodess, C.M., Hanson, C., Hulme, M. and Osborn, T.J., 2003: Representing climate and extreme weather events in integrated assessment models: A review of existing methods and options for development, *Integrated Assessment*, **4**, 145-171.

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 Change,* submitted.

Gyalistras, D. and Schuepbach, E., 2005: Statistical downscaling of seasonal temperature statistics in the European Alps: Sensitivity of model skill and climate change scenarios to choice of large-scale predictors, *Climate Dynamics*, in preparation.

Haylock, M.R. and Goodess, C.M., 2004: 'Interannual variability of European extreme winter rainfall and links with mean large-scale circulation', *International Journal of Climatology,* **24**, 759-776.

Haylock, M.R., Cawley, G.C., Harpham, C., Wilby, R.L. and Goodess, C.M., 2005: 'Downscaling heavy precipitation over the UK: a comparison of dynamical and statistical methods and their future scenarios', *International Journal of Climatology,* submitted.

Maheras, P., Tolika, K., Anagnostopoulou, Chr., Vafiadis, M., Patrikas, I. and Flocas, H., 2004: 'On the relationships between circulation types and changes in rainfall variability in Greece', *International Journal of Climatology,* **24**, 1695-1712.

Mearns, L.O., Giorgi, F., Whetton, P., Pabon, D., Hulme, M. and Lal, M., 2003: *Guidelines for Use of Climate Scenarios Developed from Regional Climate Model Experiments*, Intergovernmental Panel on Climate Change (IPCC) Task Group on Data and Scenario Support for Impacts and Climate Analysis (TGICA), available from: http://ipcc-ddc.cru.uea.ac.uk/guidelines/dgm_no1_v1_10-2003.pdf.

Tolika, K., Maheras, P., Vafiadis, M., Flocas, H. and Arseni- Papadimitriou, A., 2005b: 'Simulation of seasonal precipitation and raindays over Greece: A statistical downscaling technique based on Artificial Neural Networks (ANNs)', *Climatic Change,* submitted.

Schmidli, J., Frei, C. and Vidale, P.L., 2005a: 'Downscaling from GCM precipitation: A benchmark for dynamical and statistical downscaling', *International Journal of Climatology,* in press.

Schmidli, J., Goodess, C.M., Frei, C., Haylock, M.R., Hundecha, Y., Ribalaygua, J. and Schmith, T., 2005b: Statistical and dynamical downscaling of precipitation: Evaluation, intercomparison, and scenarios for the European Alps, *Journal of Geophysical Research*, in preparation.

Wilby, R.L., Charles, S.P., Zorita, E., Timbal, B., Whetton, P. and Mearns, L.O., 2004: *Guidelines for Use of Climate Scenarios Developed from Statistical Downscaling Methods*, Intergovernmental Panel on Climate Change (IPCC) Task Group on Data and Scenario Support for Impacts and Climate Analysis (TGICA), available from: http://ipcc-ddc.cru.uea.ac.uk/guidelines/StatDown_Guide.pdf.

Table 2: STARDEX application criteria for dynamical downscaling.

Method provides:	**Yes/No**	**Comments/Notes**
Station-scale information	No	
Grid-box information	Yes	Currently, 50 km resolution (25 km resolution in the ENSEMBLES project runs which will start in Autumn 2006)
European-wide information	Yes	
Daily time series	Yes	Not always available for all variables
Seasonal indices of extremes	No	Can be calculated from daily output
Temporally consistent temperature and precipitation[1]	Yes, in theory	Requires evaluation – dependant on model parameterisations
Spatially consistent multi-site information[2]	Yes, in theory	Requires evaluation – dependant on model parameterisations
Temporally consistent multi-site information[3]	Yes, in theory	Requires evaluation – dependant on model parametrisations
Information at sites with no observations	Yes	
Method requirements :	**Relatively high/medium/low**	**Comments/Notes**
Computing resources	High	
Volume of data inputs	High	GCM forcing fields and observed gridded data for validation
Availability of input data	Currently restricted to a few GCMs	

[1] i.e., the temperature/precipitation co-variance is similar for the downscaled validation series and observed series
[2] i.e., the downscaled validation series has a similar spatial pattern to the observed series
[3] i.e., the downscaled validation series has similar daily inter-site correlations to the observed series

Table 3: STARDEX application criteria for statistical downscaling.

Method provides:	**Y/N**	**Comments/Notes**
Station-scale information	Yes	
Grid-box information	Potentially	Requires gridded observations
European-wide information	Some methods	
Daily time series	Yes – for indirect methods No – for direct methods	
Seasonal indices of extremes	Yes – for direct methods	Can be calculated from daily series for indirect methods
Temporally consistent temperature and precipitation[1]	Some methods available	
Spatially consistent multi-site information[2]	A few methods available	
Temporally consistent multi-site information[3]	A few methods available	
Information at sites with no observations	No	
Method requirements :	**Relatively high/medium/low**	**Comments/Notes**
Computing resources	Medium/low	
Volume of data inputs	Medium/low	
Availability of input data	Medium/low	

[1] i.e., the temperature/precipitation co-variance is similar for the downscaled validation series and observed series
[2] i.e., the downscaled validation series has a similar spatial pattern to the observed series
[3] i.e., the downscaled validation series has similar daily inter-site correlations to the observed series

Table 4: STARDEX performance criteria: example of a hypothetical entry for a statistical downscaling method.

	Relative	**Performance**	**Confidence**
	High	**Medium**	**Low**
Temperature			
Indices	Txav, tnav	Txq90, tnq10, tnfd, txhw90	
Seasons	Winter	Spring, Summer, Autumn	
Regions	SE England, NW England	W Iberia	SE Iberia
Precipitation			
Indices	Pav	Pxcdd	Pq90, px5d, pint, pxcdd, pfl90, pnl90
Seasons	Winter	Spring, Autumn	Summer
Regions	NW England	SE England, W Iberia	SE Iberia
Overall performance:			
Mean temperature	Good		
Temperature extremes	Average		
Mean precipitation	Average		
Precipitation extremes	Poor		
Optimal spatial scale:	Regional averages (5-10 stations)		
Recommended impact applications:	Any which require high-spatial resolution information about seasonal extremes, but do not require spatially-correlated time series for multiple sites		