STARDEX

STAtistical and Regional dynamical Downscaling of EXtremes for European regions

EVK2-CT-2001-00115

Deliverable D11

Recommendations on variables and extremes for which downscaling is required

The need for downscaling for extremes: An evaluation of interannual variations in the NCEP reanalysis over European regions

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 3 on 'Analysis of GCM/RCM output and their ability to simulate extremes and predictor variables' which was responsible for the production of this deliverable (D11). Workpackage 3 is co-ordinated by Christoph Frei from the Swiss Federal Institute of Technology, ETH, Zürich, Switzerland.

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

D11 AUTHORS AND VERSION HISTORY

Lead author:
Christoph Frei, ETH

Contributing authors:
Stefano Alberghi, ADGB
Christina Anagnostopoulou, AUTH
András Bárdossy, USTUTT-IWS
Lucía Benito, FIC
Manuel Blanco, FIC
Raphael Borén, FIC
Carlo Cacciamani, ARPA-SMR
Clare Goodess, UEA
Diego Gozalo, FIC
Malcolm Haylock, UEA
Yeshewatesfa Hundecha, USTUTT-IWS
Panagiotis Maheras, AUTH
Valentina Pavan, ARPA-SMR
Guy Plaut, CNRS
Jaime Ribalaygua, FIC
Jürg Schmidli, ETH
Konstantia Tolika, AUTH
Rodica Tomozeiu, ARPA-SMR
Enio Tosi, ADGB

Version 1.0: 8. September 2003
Version 3.0: 18 November 2003
Version 4.0: 28 November 2003

CONTENTS

See also individual partner contributions from UEA, FIC, CNRS, ARPA-SMR, ADGB, ETH, USTUTT-IWS and AUTH – available from http://www.cru.uea.ac.uk/cru/projects/stardex/.

1. Introduction

Changes in the distribution and frequency of anomalous meteorological episodes, such as heat waves, drought periods and heavy precipitation events, are among the most feared impacts to society from future climate change. The planning of mitigation and adaptation measures requires quantitative predictions of changes in extremes. Such predictions are ultimately based on General Circulation Models (GCMs) simulating the response of the climate system to future changes in greenhouse gas and aerosol concentrations. Yet the computational grid of these models has a resolution of around 200 km at best (for coupled atmosphere-ocean models) and this limits the representation of several key processes on finer scales, such as convection and orographic surface forcing, which are relevant to extreme events.

Methods are therefore required to further process GCM data and derive climate change scenarios which take into account more regional-scale physiographic conditions and mesoscale atmospheric phenomena (see e.g. Giorgi et al. 2001). Such downscaling methods are based either on Regional Climate Models (RCMs, dynamical downscaling, see e.g. Jones et al. 1995), on empirical scale-relationships (statistical downscaling, see e.g. von Storch et al. 1993, Wilby et al. 1998) or a combination of both (statistical-dynamical downscaling, see e.g. Fuentes and Heimann 2000). Their development is a major task. It requires region-specific ajustments, laborious performance testing and, in the case of statistical downscaling, different approaches for different target variables (e.g. temperature or precipitation, means or extremes) and different seasons.

The purpose of the STARDEX deliverable 11 is to identify those variables and indices of extremes for which downscaling is particularly needed, i.e., where raw GCM output reveals an insufficient representation of regional climate characteristics. For this purpose, a systematic evaluation of precipitation and temperature extremes from raw GCM data was undertaken by comparison to observations for several European regions. The evaluation will help to focus efforts on the development of downscaling techniques where they are most needed. Moreover, the skill measures derived for the GCM will serve as a benchmark in later comparisons of downscaling techniques and to quantify the added value of downscaling.

The approach taken in this study differs in several respects from conventional GCM evaluations in order to provide specific insights into the need for downscaling:

- The evaluation is undertaken for a palette of indices representative of temperature and precipitation extremes with a potential for climate impacts.

- Observed data is available for five representative European regions at high spatial density and daily resolution to provide a suitably upscaled reference at the scale of the GCM grid-point. Moreover, a European-wide dataset is also used to provide a continental-scale overview.

- The need for downscaling is examined for a reanalysis system (the NCEP reanalysis) rather than a free GCM. Here, the reanalysis is considered as a quasi-ideal GCM, which, by virtue of its assimilation system, is constrained to the observed large-scale flow. Unlike for a free GCM, the skill of the reanalysis is much less affected by large-scale circulation biases, which would require model improvements rather than downscaling.

• It is the model skill in representing climatic variations which is of primary interest in downscaling applications. This evaluation builds on earlier ideas and focuses on the representation of the observed interannual variability (see e.g. Lüthi et al. 1997, Murphy 1999, Widmann and Bretherton 2000, Vidale et al. 2003). This allows quantification of model performance independent from biases in model climatology.

The outline of this synthesis report is as follows: Section 2 describes the common procedure taken in each study region in preparing the reference dataset from station observations and the subsequent comparison to GCM results. A summary of the main findings and synthesis of partner reports is compiled in Section 3. The final Section 4 concludes and discusses some methodological limitations found during the analysis. Contributions from individual partners are available from the STARDEX web site.

2. Evaluation Procedure

The evaluation procedure used in this study consists essentially in a comparison of time-series of seasonal summary statistics, representative of temperature and precipitation extremes, between the NCEP reanalysis (Kalnay et al. 1996) and observational reference datasets over several European study regions and Europe as a whole. The reproduction of the observed interannual variations by the reanalysis (statistics based on raw temperature and precipitation data at reanalysis grid points) is expressed by several skill measures and compared between seasons, summary statistics and regions. This comparison is meant to identify those areas and parameters for which raw GCM output is likely to be deficient and downscaling, if successful, would be needed for deriving reliable climate change scenarios.

The summary statistics under consideration are based on daily temperature and precipitation data. They encompass 10 measures of anomalous warm and cold periods, drought and heavy precipitation. Mean seasonal temperature and precipitation are also considered for comparison. A list of all 13 statistics, which in the following will be referred to as *indices,* is given in Table 1 together with their definitions and acronyms used later in the text. Time series of these indices were determined by applying the STARDEX diagnostic software tool (Haylock 2003, available from http://www.cru.uea.ac.uk/projects/stardex/) to daily (minimum and maximum) temperature and precipitation from the NCEP reanalysis and upscaled observations (see later). Series of indices were calculated for each of the four seasons of the year (DJF: winter, MAM: spring, JJA: summer, SON: autumn).

The observational reference datasets for the evaluation were prepared from daily observations of minimum and maximum temperature and from precipitation at climatological stations. These datasets were compiled by the STARDEX groups for five European subregions (the Alps, England, Emilia Romagna in north-eastern Italy, Greece, and the Rhine basin of Germany) and for Europe as a whole (see Fig. 1). The reference period for the evaluation was 1958-2000. For some of the subregions (e.g. the Alps) a shorter period (1971-1992) was considered for reasons of data quality and availability.

In the case of the subregions, the station data were aggregated onto regional 2.5x2.5 degree latitude-longitude pixels. The upscaling was performed at the daily timescale and resulted in regional temperature and precipitation series, which are compatible with the nominal scale of the global model, i.e. the NCEP grid-pixels. Reference series of extremes indices were then

Table 1: Indices of extremes and acronyms used in the evaluation.

Acronym	*Description*
TNAV	Mean minimum temperature
TXAV	Mean maximum temperature
TXQ90	90% quantile of daily maximum temperature
TNQ10	10% quantile of daily minimum temperature
TNFD	Number of days with minimum temperature below 0°C
THWDI	Heat wave duration: Days with 5K above normal T_{max} (more than 6 consecutive days)
PAV	Mean precipitation
PINT	Precipitation intensity, mean amount on a wet day (>1 mm d^{-1}).
PQ90	90% quantile of daily precipitation on wet days
PF90	Percentage of precipitation at days with more than long-term 90% quantile
PN90	Number of days with precipitation exceeding the long-term 90% quantile
PX5D	Seasonal maximum of 5-day total precipitation
PXCDD	Seasonal maximum number of consecutive dry days (≤ 1 mm d^{-1})

calculated from the upscaled observations using the same diagnostic tool as for the NCEP data. The following upscaling procedures have been used for the different study regions: ordinary block kriging in the Rhine basin and in Greece (e.g. Isaaks and Srivastava 1989), the SYMAP analysis (Shepard 1984, see also Frei and Schär 1998) in the Alps and Emilia-Romagna, and a variance correction method (Osborn and Hulme 1997) for England. The number of stations that could be used in the upscaling for a single 2.5x2.5 degree gridpoint depends on study region and varies between 5 and 50 for temperature and between 5 and 500 for precipitation.

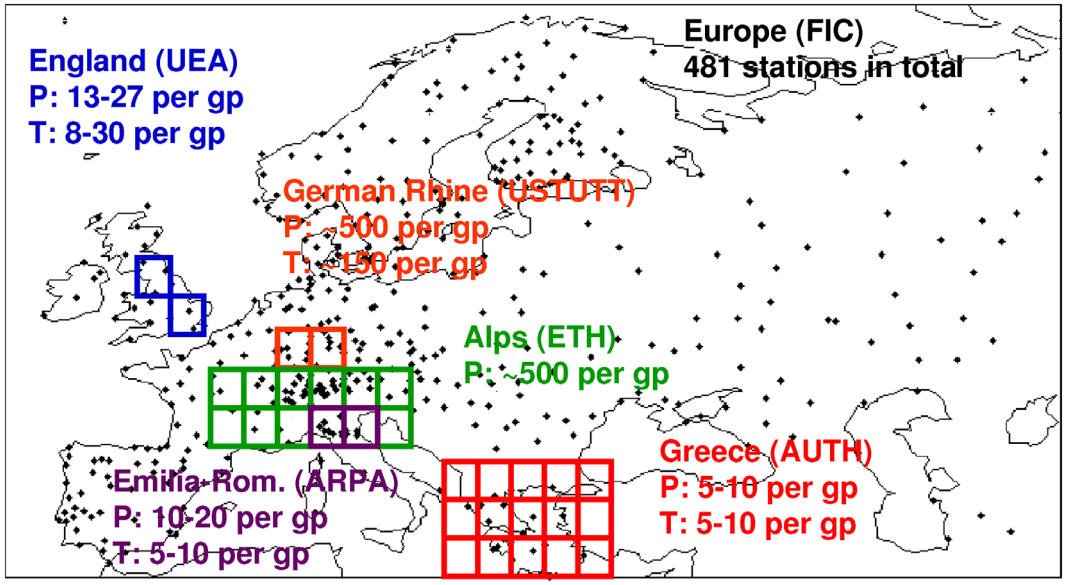

Figure 1: 2.5x2.5 degree grid boxes for which the NCEP reanalyses was evaluated in European subregions. The number of stations available for the upscaling of observations is also listed (P: precipitation, T: temperature, numbers per grid point). Stations used in the evaluation for Europe as a whole are represented as black dots.

The upscaling of the station observations is an essential step in the evaluation procedure. It ensures scale-compatibility between model and observational data, and avoids spurious systematic and random errors between the two time-series (see e.g. Mearns et al. 1995, Osborn and Hulme 1997). For the five selected European subregions, denser station data were available than over the rest of Europe and this lends some confidence to the accuracy of the upscaled index series. Yet for demonstrating the need for downscaling beyond the scale of the model grid-points (e.g. the site scale), skill measures were also calculated between the NCEP series and the series from single stations located in the corresponding grid pixel. These skill measures have served as a benchmark for downscaling to the site scale.

In the case of the European-wide analysis, data were integrated from 481 single stations (an extension of Klein Tank et al. 2001). In this case, the station density was too coarse to allow for an accurate upscaling. The European-wide analysis therefore compares NCEP grid-points directly to index series at single stations. Although affected by systematic and random errors, the comparison provides an overview of the continental-scale variations in the GCM skill.

Several skill measures were used to quantify the correspondence of the resulting time-series between the NCEP and the upscaled observations. These are the standard correlation coefficient, the ratio of standard deviations and the debiased root mean square error. The Taylor diagram is used to display these mutually related skill measures (Taylor 2001). In the present application, the Taylor diagram allows easy comparison of NCEP skills between different indices of extremes and between different seasons. (An example is displayed in Fig. 2.)

It should be noted that the selected skill measures focus on the random error component. Systematic biases of the NCEP model will not be revealed in the correlation skill but only in the ratio of standard deviations for indices where a bias is related to errors in scale (e.g. PINT and PQ90). As the focus of this evaluation is the representation of interannual variations (i.e. the random error component), a distinction from NCEP errors in the long-term climatology (i.e. the systematic errors) was desirable. The random error component is probably more relevant than biases for the identification of downscaling needs. While the former may indicate the lack in resolution of mesoscale processes, systematic model errors may also be affected by errors in, or ill-tuning of, model parameterizations. A systematic analysis of NCEP biases for the indices considered was not specifically included in this study. (Two partners have, however, additionally considered biases (AUTH, USTUTT) and one partner has included an additional skill score (UEA), which combines biases and random errors: LEPS, Potts et al. 1996.) Seasonal precipitation and temperature biases in the NCEP reanalysis over Europe have however been previously examined by Reid et al. (2001) and Hagemann and Dümenil (2001).

3. Main Findings

The five selected European subregions cover a range of different European climate regions. Nevertheless, several characteristics in the performance of the NCEP reanalysis were found to be common, at least in qualitative terms, across all subregions. (Refer to Table 1 for acronyms of indices.):

- In all subregions, the NCEP reanalysis skillfully reproduces interannual variations for at least some of the temperature and the precipitation indices, in the sense that the correlation

threshold for statistical significance is exceeded (r=0.3 for a significance level of 5% and sample size of 43 years).

• The correlation skills for temperature indices tend to be higher than for precipitation indices and they are mostly statistically significant.

• For precipitation indices, NCEP performance shows a clear seasonality with highest correlations in winter and lowest correlations (in many cases statistically non-significant) in summer. This result is also evident from the European-wide analysis using single station records as a reference. The results for spring tend to be more similar to those for summer and for autumn more similar to winter. (Possibly because May – a month with predominantly convective precipitation in many regions – is dominating the variability in spring). A similar, but less marked, seasonality of results is evident for temperature indices in some regions.

• Among all the precipitation indices considered, PAV and PXCDD are best represented by the NCEP reanalysis in most of the subregions (see e.g. example Fig. 2). Correlation values for these two indices are mostly greater than 0.6 in winter and still clearly significant in summer. For the remaining precipitation indices, performance tends to be around the significance limit in winter and mostly below in summer, but there are exceptions.

• Among the temperature indices, TNFD and THWDI show lowest correlations in many subregions. (This might partly be due to lack of statistical robustness (see also Section 4.) But the performance of NCEP is not overtly different between mean temperatures (TXAV, TNAV) and the temperature quantiles (TXQ90 and TNQ10). The temperature quantiles reach

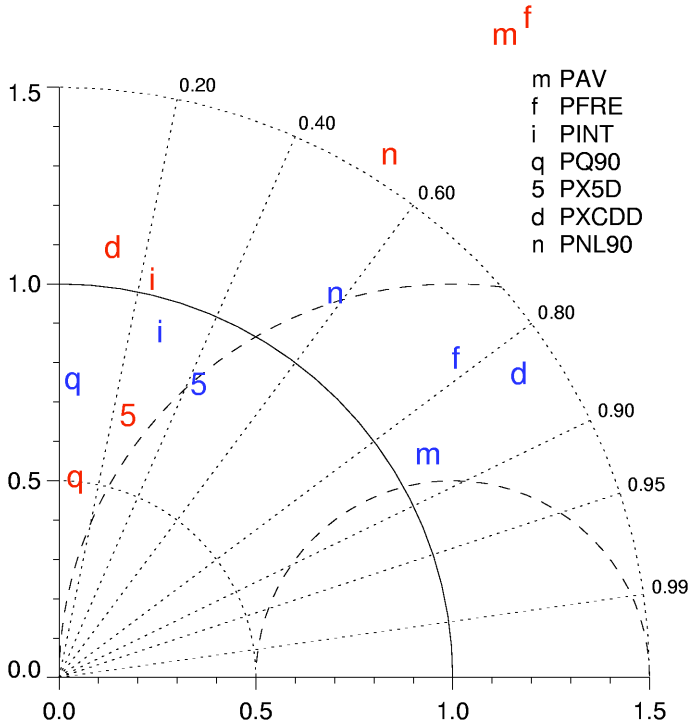

Figure 2: Taylor diagram displaying evaluation results on precipitation statistics for a grid point in the French part of Alpine region (see grid point no 01 for the Alpine region, see ETH partner contribution available from the STARDEX web site). Blue symbols for winter, red symbols for summer. (Courtesy Jürg Schmidli, ETH.)

correlation skills in the order of 0.8 in winter and still above 0.6 in summer.

• The interannual variance of extremes indices is often seriously under- or overestimated by the NCEP reanalysis. This is even the case in regions where upscaling biases are expected to be small (e.g. regions such as Germany and the Alps, where hundreds of stations could be used for the upscaling, see e.g. Fig. 2). The ratio of standard deviations varies typically between 0.5 and 1.5 but is different between regions, indices and seasons. Some part of these errors may be due to systematic long-term biases in the NCEP reanalysis. Smaller errors in variance are found for percentile-based indices, such as PF90 and PN90, where model biases are implicitly adjusted.

• Correlation values are consistently higher with the upscaled observations than with the index series from single reference sites. (There are a few random exceptions.) Also, the variance at NCEP grid-points is generally lower than at the reference sites. The differences in skill measures seem to be more pronounced for the precipitation indices than for temperature.

Besides these general results, the various evaluations reveal interesting regional variations:

• NCEP reproduces interannual variations of extremes less skillfully in mountain regions compared to flatlands. This is evident from the station-based European-wide analysis, which reveals particularly low correlation skills in the Alpine region, near the Pyrenees and Scandinavia. However, a considerable portion of these errors seems to be associated with the lack of scale compatibility, i.e., with the strong site dependence of interannual variations in complex topography. Direct comparison between grid-pixels in the Alps and the Rhine basin of Germany (i.e., the immediate flatland to the north of the Alps), reveals that skill measures between mountains and flatland do not differ as much for upscaled data as for single stations.

• The performance of NCEP in mountainous regions also depends markedly on the delineation of topographically determined climate regimes by the grid structure. For example, in the Alpine region, the correlation skills for precipitation indices are particularly low for a grid-pixel extending across northern and southern slopes (gridpoint number 10), but reasonably high for a grid-pixel restricted to the southern slopes and adjacent flatland (gridpoint number 03). Also, the variations of extremes at a gridpoint may be more representative for a certain subdomain within the pixel as is illustrated in the analysis for Emilia-Romagna.

• Temperature quantiles are found to be particularly well represented over England. Correlation values here reach 0.8 for TNQ10 in winter and 0.9 for TXQ90 in summer. Skills for these indices in Greece are lower (typically between 0.6 and 0.8) particularly for TXQ90 in summer. The role of physical parameterizations over Southern Europe in summer (radiation and clouds) as opposed to air-mass advection over Northern Europe may explain this difference.

• Over both the Rhine basin and the northern parts of the Alpine region, NCEP appears to overestimate the interannual standard deviation of mean summer precipitation by up to a factor of two (see e.g. example in Fig. 2). This may partly reflect the overestimate in mean summer precipitation over the European continent (see also Hagemann and Gates 2001). However, surprisingly this bias does not simply scale in the standard deviation for PINT, PQ90, PX5D. Overestimates in standard deviation for these indices are much lower in the Rhine basin. In the Alpine region, there is even an underestimate of standard deviation by a

factor of two. Together, these results pinpoint the difficulties experienced by the reanalysis in reproducing the observed precipitation recurrence process, which, in summer, is governed by the triggering of convection. Clearly the performance for all precipitation indices is affected, including that for PXCDD. A related sign of model deficiency may be found in the overestimate of TXAV and TXQ90 standard deviation in summer, evident in the Taylor diagrams for the German Rhine basin and for Greece.

4. Conclusions and Discussion

In this analysis, observed interannual climate variations over Europe are used to test the representation of regional extremes by the NCEP reanalysis. Considering this reanalysis as a GCM in assimilation mode, the evaluation is intended to reveal the potential and limitations to be expected in GCMs used for climate-change studies. The approach is complementary to conventional GCM evaluations with a focus on systematic errors, in that it tests for the links between large-scale circulation and grid-point scale extremes. The results provide some guidance on the need for downscaling and its dependence on season, type of extremes and region.

Indices of temperature extremes are found to be significantly better represented than most precipitation extremes. For temperature quantiles, high correlations (comparable to those for mean temperature) are found in areas and seasons when a strong circulation control is expected on surface temperatures (i.e., northern and central Europe, winter). Lower but still statistically significant skills are found over southern Europe in summer when clouds and surface radiation balance contribute to temperature extremes. These results indicate that GCMs could provide valuable information on variations of temperature extremes even at the scale of the nominal GCM resolution. This information should not a priori be rejected in an analysis of regional climate change. General mistrust in GCM temperature extremes, simply because of large systematic biases, is inappropriate. Clearly, this does not exclude the need for downscaling. However such a need derives primarily for regions and seasons with weak advective control and for scales not resolved by the GCM.

As regards precipitation extremes, the NCEP interannual variations are found to be significantly correlated with observations in winter only. In most regions the performance is clearly below that for mean precipitation, except for the maximum consecutive dry days, for which similar performance is reached, probably due to the link of this index to the duration of persistent and well-resolved high-pressure situations. In summer, the performance is near or below the significance limit. The implication of the analysis is that downscaling is desirable for precipitation extremes in both seasons (but particularly in summer) and even on spatial scales resolved by the GCM. It is, however, yet to be determined to what extent downscaling can improve the performance and hence improve on the relationship between large and regional scales compared to the GCM itself.

There are several limitations in the approach taken and its application, which complicate the unequivocal identification of downscaling needs:

- The focus of the present analysis is on the representation of *interannual* variations. However, in a climate-change application the relevant time scale is multi-decadal. High model skill on interannual scales does not necessarily assure reliability on time-scales of climate change. In this analysis the comparatively short series of the NCEP reanalysis and the

possibility of assimilation inhomogeneities (see below) has prevented consideration of longer time-scales. Therefore, although the present approach addresses a critical issue, it is not necessarily a sufficient model test. Needs for downscaling may also result for reasons not evident in the testbed of this study. For example the representation of the annual cycle (a relevant criterion for some impact models) or the representation of long-term trends.

• The utility of GCM output in regional impact assessments (e.g. as input for impact models) is often complicated by the existence of systematic model errors. In practice, an interface between the GCM and an impact model is required, which adjusts for the biases in the GCM output for the particular impact application based on some suitable stationarity assumption. In the conclusions of the present analysis, the biases of NCEP are not explicitly discussed, although some partners have listed biases in their contribution. Firstly, this is because the biases of a GCM may be different from those of the NCEP reanalysis (generality of results) and secondly, because biases are more indicative of the *need for bias adjustment* rather than the *need for downscaling*. The focus on the correlation skill measure (Taylor diagram) is on downscaling (i.e. the prediction of regional anomalies using large-scale predictor anomalies). However, it is recognized that the existence of GCM biases can further limit the utility of direct GCM output in impact studies, and that this additional limitation is not included in the skill measure discussed here.

• Two partners have visually compared the evolution of mean seasonal precipitation from the NCEP reanalysis with the corresponding upscaled observations (for the French part of the Alps and the Rhine basin of Germany). They found trends and long-term variations in the NCEP reanalysis (long-term decreases in summer, spring and autumn), which are not evident in the observations. A likely reason for these drifts is inhomogeneities in the reanalysis process. While the underlying method of data assimilation is the same for the entire reanalysis period, the amount of available data varies and can give rise to inhomogeneities (see e.g. Kalnay et al. 1996). It would, however, be inappropriate to interpret these discrepancies directly as a *need for downscaling*. A free GCM is usually not constrained to observations and therefore is not affected by trends of that origin.

• Results for some of the indices turn out to be difficult to interpret with the methodology of this analysis. This applies particularly to indices of count data defined by fixed thresholds such as TNFD and THWDI. These two indices have very low counts in some regions (e.g. THWDI in England and TNFD in Southern Europe). As a result, linear regression is an inappropriate technique to measure the correspondence to observations for these low number count data. Moreover, there are high limits of predictability due to high rarity (see also below). An alternative index definition using a variable threshold and allowing for higher number counts would be desirable.

• The present evaluation has exclusively focussed on the NCEP reanalysis. Therefore, at this stage, the representativity of the results for a wide spectrum of GCMs is not clear. Further insights can be gained when data from other reanalyses (such as those from the ECMCF: ERA15 Gibson et al. 1999 and ERA40 Simmons and Gibson 2000) are included in a similar analysis.

• Our interpretations with respect to the need for downscaling are based essentially on comparisons of correlation values. The implied notion is that poorer values indicate a higher need for downscaling. However, low correlation values could not only result from model errors in representing regional extremes but also from inherent limitations in predictability.

Regional extremes are in part non-deterministic phenomena and a climate model (in this case the reanalysis) cannot be blamed for all discrepancies from observations. Likewise, a downscaling model cannot be expected to compensate for limited predictability. At this stage, it is not entirely clear to what extent the lower correlation values for (most) precipitation extremes and for the summer season, as found in the present analysis, are truly attributable to model limitations which could be overcome by downscaling. Quantification of predictability limits would require several model realizations using reanalysis ensembles. In view of these complications, the present analysis indicates, at best, where downscaling is desirable but not necessarily where it will be successful. Moreover, its primary value is to serve as a benchmark against which the added value of downscaling can be assessed. This is the subject of further joint activities in the STARDEX project.

5. References

Frei C. and C. Schär, 1998: A precipitation climatology of the Alps from high-resolution rain-gauge observations. *Int. J. Climatol.,* **19,** 873-900.

Fuentes, U. and D. Heimann, 2000: An improved statistical-dynamical downscaling scheme and its application to the Alpine precipitation climatology. *Theor. Appl. Climatol.*, **65,** 119-135.

Gibson, J.K., P. Kallberg, S. Uppala, A. Hernandey, A. Nomura and E. Serano, 1999: ERA-15 description (version 2). ECMWF Reanalysis Project Report Series (Reading UK), 1, 74 pp.

Giorgi, F. and al. et, 2001: Regional climate information - Evaluation and projections. In: *Climate Change 2001: The scientific basis.* The Third Assessment Report of the Intergovernmental Panel on Climate Change (IPCC), Houghton J.T. et al. (eds.), 583-638.

Hagemann, S. and L. Dümenil, 2001: Validation of the hydrological cycle of ECMWF and NCEP reanalyses using the MPI hydrological discharge model. *J. Geophys. Res.*, **106,** 1503-1510.

Haylock M. 2003: STARDEX diagnostic extremes indices software: User information. Version 3.2.6, May 2003. Available from the STARDEX web-site: http://www.cru.uea.ac.uk/projects/stardex/

Isaaks, E.H. and R.M. Srivastava, 1989: *An introduction to applied geostatistics.* Oxford University Press, 561 pp.

Jones, R.G., J.M. Murphy and M. Noguer, 1995: Simulation of climate change over Europe using a nested regional-climate model I: Assessment of control climate, including sensitivity to location of lateral boundaries. *Q. J. R. Meteorol. Soc.*, **121,** 1413-1449.

Kalnay, E., M. Kanamitsu and al. Kistler, 1996: The NCEP/NCAR 40-year reanalysis project. *Bull. Amer. Meteorol. Soc.*, **77,** 437-471.

Klein Tank, A.M.G. and 38 others, 2002: Daily surface air temperature and precipitation dataset 1901-1999 for European Climate Assessment (ECA). *Int. J. Climatol.,* **22,** 1441-1453.

Lüthi, D., A. Cress, C. Frei and C. Schär, 1996: Interannual variability and regional climate simulations. *Theor. Appl. Clim.,* **53,** 185-209.

Murphy, J.M., 1999: An evaluation of statistical and dynamical techniques for downscaling local climate. *J. Climate*, **12,** 2256-2284.

Mearns, L.O., F. Giorgi, L. McDaniel and C. Shields, 1995: Analysis of daily variability of precipitation in a nested regional climate model: comparison with observations and doubled CO^2 results. *Global Planet. Change,* **10,** 55-78.

Osborn, T.J. and M. Hulme, 1997: Development of a relationship between station and grid-box rainday frequencies for climate model evaluation. *J. Climate,* **10,** 1885-1908.

Potts, J.M., C.K. Folland, I.T. Jolliffe and D. Sexton, 1996: Revised LEPS scores for assesing climate model simulations and long-range forecasts. *J. Climate,* **9,** 34-53.

Reid, P.A., P.D. Jones, O. Brown, C.M. Goodess and T.D. Davies, 2001: Assessments of the reliability of NCEP circulation data and relationships with surface climate by direct comparisons with station based data. *Climate Res.,* **17,** 247-261.

Shepard, D.S., 1984: Computer mapping: the SYMAP interpolation algorithm. In Gaile, G.L. and Willmott, C.J. (eds), *Spatial Statistics and Models*, Dordrecht, pp. 133–145.

Simmons, A.J. and J.K. Gibson, 2000: *The ERA-40 project plan*. Available from ECMWF, Reading UK., 60 pp. (http://www.ecmwf.int/research/era/Project/Plan/)

Taylor, K.E., 2001: Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.,* **106,** 7138-7129.

Vidale, P.L., D. Lüthi, C. Frei, S. Seneviratne and C. Schär, 2003: Predictability and uncertainty in a Regional Climate Model. *J. Geophys. Res.,* **108(D18)**, art. no. 4586.

Widmann, M. and C.S. Bretherton, 2000: Validation of mesoscale precipitation in the NCEP reanalysis using a new gridpoint dataset for the northwestern US. *J. Climate*, **11**, 1936-1950.