INTERNATIONAL JOURNAL OF CLIMATOLOGY
Int. J. Climatol. **28**: 465–477 (2008)
Published online 22 June 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/joc.1545

RMetS
ROYAL METEOROLOGICAL SOCIETY

Non-linearity in statistical downscaling: does it bring an improvement for daily temperature in Europe?

R. Huth,[a]* S. Kliegrová[b] and L. Metelka[b]
[a] *Institute of Atmospheric Physics, Prague, Czech Republic*
[b] *Czech Hydrometeorological Institute, Regional Office, Hradec Králové, Czech Republic*

ABSTRACT: Several linear and non-linear statistical downscaling methods are compared for winter daily temperature at eight European stations. The linear methods include linear regression of gridpoint values (pointwise regression) and of predictors' principal components (PC regression). The non-linear methods are represented by artificial neural networks. The non-linearity is also achieved by a stratification of data by classification of circulation patterns and a linear regression conducted separately within each class. As predictors, gridded 500 hPa heights and 850 hPa temperature are used. The verification is conducted in the cross-validation framework. The downscaling methods are evaluated according to four criteria: (1) fit to observations (quantified by the correlation coefficient), (2) shape of the statistical distribution, namely its skewness and kurtosis, (3) temporal autocorrelations with 1 day lag, and (4) interstation correlations. Considering all the criteria together, the pointwise linear regression appears to be the best method. It achieves the best fit with the observations and possesses the best temporal structure. The deviations of statistical distributions from normality are only captured by the neural networks, while the classification methods yield the best spatial correlations. Copyright © 2007 Royal Meteorological Society

KEY WORDS statistical downscaling; daily temperature; non-linearity; neural networks; classifications; Europe

Received 9 December 2005; Revised 2 March 2007; Accepted 17 March 2007

1. Introduction

Statistical downscaling is one of the widely used tools to bridge the gap between what global climate models (GCMs) are able to simulate and what is needed in climate change impact research. It consists in seeking statistical relationships between variables well simulated by GCMs (usually upper-air and/or large-scale fields) and regional or local surface climate variables.

So far, the majority of statistical downscaling studies have employed linear methods, such as multivariate regression and canonical correlation analysis. Since there is no *a priori* reason to suppose the predictor-predictand relationships are linear, non-linear methods have recently begun to emerge. Among them, neural networks (NNs) have been applied most frequently. In order to justify the use of a non-linear, that is more complex, method, at the expense of a simpler linear one, it is important to know whether the introduction of non-linearity leads to an improvement in performance of the downscaling procedure. However, relatively numerous studies are limited to the description and evaluation of a non-linear method, without any comparison with a linear approach (e.g. Cavazos, 1997, 1999, 2000; Hewitson and Crane, 1992; Crane and Hewitson, 1998; McGinnis, 2000; Olsson *et al*., 2001; to name just a few). Those analyses where a comparison is performed give an ambiguous message. The superiority of non-linear methods to multiple linear regression is reported by Mikšovský and Raidl (2005) who tested methods for the downscaling of NCEP/NCAR reanalysis data, using series of daily mean, minimum and maximum temperatures from 25 European stations as predictands. All non-linear techniques (two types of NNs and the method of local linear models) proved to be better than linear regression in the majority of the cases. The next interesting conclusion of their article is that the non-linear character of relations between climate predictors and predictands is not clear at all stations and that the non-linearity exhibits a seasonal variance (a strong detectable non-linear component is present in the relations between predictors and predictand in winter, whereas these relations have a rather linear character in summer). Other examples of results of non-linear methods superior to the linear ones (usually multiple regression) are demonstrated by Trigo and Palutikof (1999) for daily temperature, Weichert and Bürger (1998) for daily temperature, precipitation, and vapour pressure, and Schoof and Pryor (2001) for daily temperature and monthly precipitation. On the other hand, several studies indicate superiority of linear methods: Trigo and Palutikof (2001) for monthly precipitation, Mpelasoka *et al*. (2001) for monthly temperature and precipitation, and Schoof and Pryor (2001) and Wilby *et al*. (1998) for daily precipitation. Zorita and von Storch (1999) demonstrate that a simple analog method outperforms NNs for

* Correspondence to: R. Huth, Institute of Atmospheric Physics, Prague, Czech Republic. E-mail: huth@ufa.cas.cz

daily precipitation. Any generalization of these results is impossible because of differences among studies in the predictand, predictors, both linear and non-linear methods to compare, and the geographical location of the target station or area. Moreover, the comparisons are carried out mostly in terms of correlations and root-mean-square differences between the downscaled and observed values, while other characteristics such as extreme values, various distributional characteristics, and temporal and spatial structure are treated rather scarcely (Weichert and Bürger 1998; Wilby *et al*. 1998).

There is another potential way of introducing non-linearity into a downscaling procedure, viz. a stratification of the dataset by a classification, the downscaling model being built within each class separately (Enke and Spekat, 1997; Cavazos 1999, 2000; Li and Sailor 2000). The idea behind this approach is that the relationship between the predictor and predictand may vary depending on the type of circulation (synoptic) pattern. Similar to non-linear methods, it would also be important to know whether the classification improves the results or not; unfortunately, none of the above mentioned studies presents a comparison of results achieved on stratified data with the unstratified ones. Therefore, nothing is known at present about whether a classification can improve the downscaling and whether there is a sensitivity of the downscaling skill to the classification method and the number of classes. Note that a classification has been utilized in the context of downscaling in two other ways, neither of them being a subject of this article: the mean value of a class is attributed to each member of that class (Saunders and Byrne, 1996, 1999) or the monthly/seasonal frequencies of daily circulation types serve as predictors of monthly/seasonal means or totals (Goodess and Jones, 2002).

There is a wide range of possible criteria to evaluate statistical downscaling methods. The majority of relevant studies rely, however, on a simple measure of accuracy such as correlation coefficient and root-mean-square error. So far considered sporadically in downscaling studies, although being important in various applications, are the temporal structure of downscaled time series (Huth *et al*., 2001; Huth, 2002), the spatial structure of downscaled fields (Easterling, 1999; Solman and Nuñez, 1999; Huth, 2002), and characteristics of the distribution of a downscaled variable, e.g. its skewness and kurtosis (Huth *et al*., 2003).

The aim of this study is to compare the performance of statistical downscaling of daily temperature by linear methods with corresponding neural network models and to assess the effect on the performance of a stratification of data by classification of circulation patterns. The comparisons are carried out not only in terms of correlation coefficients, but also for temporal and spatial correlations, and higher-order statistical moments. To achieve a more general validity of our results, the analysis is focused on eight stations located in widely differing climatological and geographical settings across Europe.

2. Data

The predictands consist of daily minimum and maximum temperatures at eight European stations: Sodankylä, Finland; Zugspitze, Bamberg, and Hohenpeissenberg, Germany; Praha-Klementinum, Czech Republic; Valentia, Ireland; Salamanca, Spain; and Smolensk, Russia. Their locations and elevations are displayed on a map in Figure 1. Note that two of them, Zugspitze and Hohenpeissenberg, are mountain summit stations. The data are taken from the database of the European climate assessment (ECA) project (Klein Tank *et al*., 2002). There were no missing values at these stations except for a single temperature maximum value at Salamanca, which was linearly interpolated from the values at the preceding and successive days. The selection of the stations was governed by high computational demands of NNs during their training, which allowed only a limited number of stations to be analysed. We selected them with the intention that they represent as wide a range as possible of European geographical and climatological settings.

Three large-scale fields are taken from the National Centers for Environmental Prediction/National Center for Atmospheric Research (NCEP/NCAR) reanalyses (Kalnay *et al*., 1996): 500 hPa and 1000 hPa heights and 850 hPa temperature. The 500 hPa heights and 850 hPa temperature are used as predictors in downscaling models. The 500 hPa and 1000 hPa heights serve a base for circulation classifications. The data are defined on a $5 \times 5^{\circ}$ grid over the area bounded by 25°N and 80°N parallels, and 50 °W and 55 °E meridians (i.e. 264 gridpoints altogether). We use the grid with half a resolution of the original dataset (every second gridpoint was retained both in latitude and longitude) because the original grid with a 2.5° resolution would be unnecessarily dense thanks to a high spatial autocorrelation, and too large (consisting of

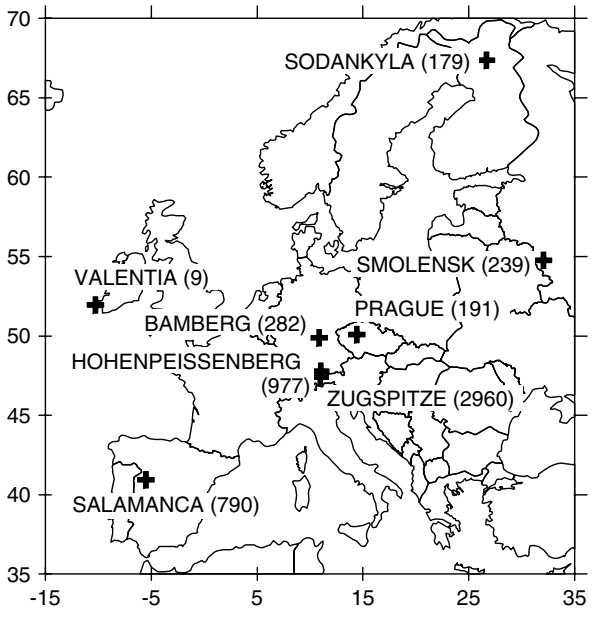

Figure 1. Location of stations; their elevation (in m above sea level) is indicated in parentheses.

DOI: 10.1002/joc

1075 points) to be handled properly by the linear regression models on stratified data. The data from 12 UTC are used. The selection of predictors is guided by results of previous work (Huth, 1999, 2002) where the combination of 500 hPa heights and 850 hPa temperature was shown to be superior at the largest portion of stations in central and western Europe to other combinations of circulation and temperature-based fields (including sea level pressure and 1000/500 hPa thickness).

We analyse 35 winter seasons (December–February), including leap days, in the period 1958/1959–1992/1993.

3. Methods

3.1. Linear

Two linear downscaling models are considered: multiple linear regression (MLR) of principal components (PCs) of the large-scale fields with predictor selection by stepwise screening ('PC regression' hereafter), and MLR of gridpoint values of the large-scale fields with predictor selection by stepwise screening (pointwise regression). The predictors enter the downscaling models in the form of standardized anomalies, which implies that the mean value is reproduced correctly. The inflation procedure (Karl *et al*., 1990) is employed to also reproduce the variance. The methods are described in more detail in Huth (2002). Principal component analysis (PCA) for this purpose is applied in S-mode, which means that columns in the data matrix correspond to gridpoint values, while rows correspond to individual time realizations (days); for the nomenclature of the modes in PCA, refer, e.g. to Richman (1986). Since our goal is a data reduction, not the interpretation of variability modes, unrotated PCs are used. Regression models are built for different numbers of PCs that comply with O'Lenic and Livezey (1988) selection rule, which states that the number of PCs should be cut just behind the 'shelf' (i.e. a section of less steep slope followed by a pronounced drop) on the PC-number versus the logarithm of eigenvalue (LEV) plot. Figure 2 displays the LEV plot for PCA of whole 35 years [the 34-year analyses conducted within the cross-validation (for explanation see below) yield very similar results]. We use 4, 6, 12, 16, and 20 PCs in building the models; these numbers are among the appropriate solutions as indicated by the LEV plot. In the stepwise screening procedure, each potential predictor is evaluated for its individual significance before including it into the regression equation and, after the addition, each variable in the equation is evaluated for its signficance as a part of the model. A variable is included and retained, respectively, in the equation if the corresponding significance level exceeds 90 and 95%.

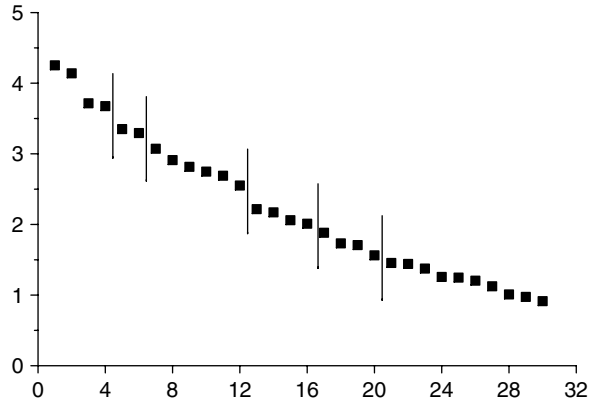

Figure 2. PC number *versus* logarithm of eigenvalue plot for S-mode PCA of 500 hPa heights and 850 hPa temperature. Vertical lines indicate the cutoff points separating the PCs retained for further analysis.

3.2. Neural networks

Neural networks have become very popular in various scientific areas as a convenient tool for time series analysis and data processing. Here we do not repeat a general description of artificial NNs, which can be found in many textbooks and application papers. A brief introduction into the NNs and their applications to atmospheric sciences can be found, e.g. in Gardner and Dorling (1998) and Hsieh and Tang (1998).

Among several possible network architectures, the multilayer perceptron (MLP) proved to be the best in our analysis. Such a type of network is the most widely used one in the context of statistical downscaling (cf. Trigo and Palutikof, 1999; Mpelasoka *et al*., 2001), although the radial basis function NNs seem also to be convenient and can offer better application properties than MLPs in many respects (Mikšovský and Raidl, 2005). The application of MLPs is potentially difficult. There is a risk that the learning will not reach the global minimum of the error function and will end in a local minimum. There is also a possibility of over-learning: if the learning procedure runs for too long a time, the network becomes overoptimized only for learning dataset and is unable to generalize. The MLP consists of one input layer (where the number of neurons is equal to the number of predictors), one output layer (where the number of neurons is equal to the number of predictands), and one or more hidden layers in between. Preliminary tests indicated that the MLPs with two and more hidden layers were not superior to the MLPs with just one hidden layer; therefore we selected the MLP with one hidden layer as the NN model for this study. In the neurons of the hidden layer, the sigmoidal transformation $y = 1/(1 + \exp(-x))$ is employed, whereas the input and output layers use a linear transformation. The learning of the NN is realized by the back error propagation algorithm.

The NNs were built with help of the Intelligent Problem Solver of the STATISTICA Neural Networks package, which tests various combinations of potential predictors and different configurations of the MLPs to find the best network. The dataset was divided into three parts: the training, verification, and test sets. The network is trained on the training set by an iterative process in which its weights are adjusted at each step. The verification set is used to track the error performance

DOI: 10.1002/joc

of the network, to identify the best network, and to stop training to avoid an over-learning. The test set provides an independent assessment of the network's performance when an entire network design procedure is completed. The division into the training, verification, and test sets was in the ratio of 50 : 25 : 25 for most networks. The cases were shuffled randomly between the subsets. This proportion had to be changed when too complex NNs (with the number of neurons in the input layer exceeding 98) were trained because the number of parametres, and hence the risk of over-learning, increase dramatically with the number of internal nodes. The proportion of training, verification and test sets was in this case 66.7 : 16.7 : 16.7.

Two different kinds of NN models were examined. The first one is an analogy to the pointwise regression: potential predictors include standardized gridpoint values of 500 hPa heights and 850 hPa temperature at all gridpoints. In the other model, potential predictors include 20 leading PCs of the 500 hPa heights and 850 hPa temperature, so it is analogous to the PC regression model for 20 PCs. The predictand in both NN models is the standardized temperature at stations. The pointwise NN model was built at all the eight stations. Since the first results of the PC-based NN model were similar to the results of the pointwise NN model, but generally slightly inferior, and because of high computational demands, only a limited number of the PC-based NN models was built and trained (for four selected stations only: Sodankylä, Salamanca, Hohenpeissenberg, and Prague). The numbers of neurons in the input and hidden layers of the optimum networks are displayed in Table I.

3.3. Classification

We apply two classification methods: k-means and PCA in T-mode. They are briefly described below. Separate classifications are performed for the 500 and 1000 hPa heights.

The data matrix of T-mode PCA consists of columns corresponding to time observations and rows representing gridpoint values. To achieve a classification, the PCs must be obliquely rotated (Huth, 1996a); here, we use the Direct Oblimin rotation method. Every daily pattern is classified with the mode on which it has the highest loading (in absolute sense). For a particular number of PCs retained and rotated, one can potentially get twice as many classes because high positive and high negative loadings can define two opposite types for each PC. In practice, for each PC, the loadings of only one sign (usually positive) dominate; therefore, the number of classes is equal to the number of PCs. Of the possible options, we selected 4, 11, and 18 PCs (and hence circulation types) for the 500 hPa heights, and 4, 12, and 18 PCs for the 1000 hPa heights; (Figure 3).

K-means is a method of a non-hierarchical cluster analysis, widely used in a broad range of climatological studies; its description is provided, e.g. in Gong and Richman (1995). As the clustering methods usually require the input variables to be uncorrelated, the clustering is performed in the space of leading 12 PCs (in an S-mode). (The results are not sensitive to a particular choice of the number of PCs.) The dissimilarity is measured by the Euclidean distance, and the seed points are selected by a random choice from the data. Classifications for the same numbers of types as for the T-mode PCA are performed in order to allow a fair comparison.

The classification methods have their own distinct properties, advantages and drawbacks (Huth, 1996b):

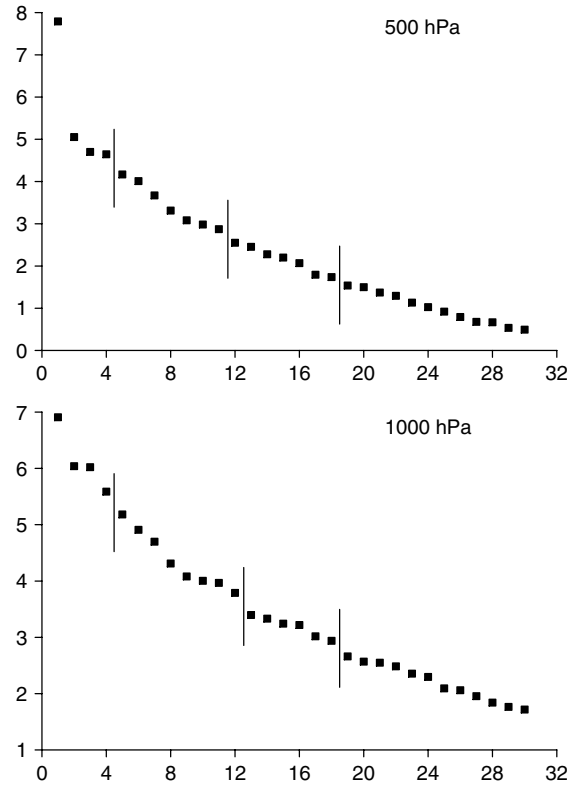

Figure 3. PC number *versus* logarithm of eigenvalue plot for T-mode PCA of 500 hPa heights (top) and 1000 hPa heights (bottom). Vertical lines indicate the cutoff points separating the PCs retained for further analysis.

Table I. Numbers of neurons in the input and hidden layers in the NN downscaling models.

Station	Soda	Zugs	Sala	Bamb	Hohe	Vale	Smol	Prag
pointwise – Tmax	87/10	9/17	101/11	32/17	6/17	90/15	71/17	98/10
pointwise – Tmin	127/17	100/18	76/17	100/14	78/11	48/15	99/17	79/16
PC – Tmax	20/7	–	17/7	–	16/11	–	–	15/12
PC – Tmin	18/11	–	19/11	–	13/7	–	–	15/11

DOI: 10.1002/joc

The k-means method is rather sensitive to the selection of initial seedpoints and is biased towards producing equal-sized classes. Its main advantage is a very good separation among classes. The T-mode PCA method is excellent at reproducing the known structure of data, at the expense of a lower between-group separation. Both the methods have been successfully used in a number of various studies of atmospheric circulation. T-mode PCA is applied e.g. in Bartzokas and Metaxas (1996), Compagnucci and Salles (1997), Huth (1997), Jacobeit *et al*. (2003), Müller *et al*. (2003), and Romero *et al*. (1999); while the k-means method is used, e.g. by Brinkmann (1999), Cassou *et al*. (2004), Esteban *et al*. (2005), Luterbacher *et al*. (2001), Santos *et al*. (2005), Solman and Menéndez (2003), and Terray *et al*. (2004).

To summarize, altogether 12 different classifications were constructed: 2 methods $\times 2$ levels $\times 3$ numbers of clusters.

3.4. Validation of downscaled values

The downscaling methods are evaluated within the cross-validation framework, which allows an unbiased estimate of potential 'predictability' without a risk of overfitting the models. The cross-validation consists in omitting one case in turn, building the whole statistical model on the remaining dataset, and applying the statistical model to the omitted case (Michaelsen, 1987). Since time series of daily temperatures exhibit considerable autocorrelation, the above procedure would lead to an optimistic bias in the skill. We therefore hold out one season at a time, the statistical model being built on the remaining 34 seasons and verified on the omitted period. All the statistical models are thus built 35 times. It is important to note that the classifications are calculated for the 35-year period as a whole and are not subject to the cross-validation. For the NNs, the optimal architecture (shown in Table I) was not involved in the cross-validation, i.e. it was determined from the whole dataset, and the training itself was performed under the cross-validation.

The accuracy of specification of the downscaled values is quantified in terms of correlation coefficient between the downscaled and observed values. The results were qualitatively the same if other measures of correspondence, such as mean absolute error or root-mean-square error, were used. Therefore, we chose the correlation coefficient, which is easy to interpret. To obtain one number for each method to characterize its performance, the correlations are averaged over the stations.

As different properties of the downscaled temperature series may be relevant in climate change impact studies in different sectors, we employ more criteria for the evaluation of downscaled values than a mere correspondence to observations. We evaluate the time structure in terms of lag-1 autocorrelations (persistence) and the spatial structure in terms of spatial autocorrelations. The degree of asymmetry and peakedness of statistical distributions are evaluated in terms of the standardized third and fourth moments, i.e. conventionally defined skewness and kurtosis.

4. Accuracy of specification

4.1. Linear methods

The performance of linear methods in terms of correlation coefficients is displayed in Figure 4. The two most notable facts confirm the findings achieved by Huth (1999, 2002) on a geographically less extensive dataset. First, the pointwise regression (black bars) yields a consistently better fit with observations than the PC regression (grey bars). Second, the accuracy of downscaling increases with an increasing number of PCs: even many of the higher order, i.e. seemingly less significant, PCs add to the temperature variance explained. In general, the increase in correlations is slowing down with the increasing number of PCs, and at several stations, there are hints of saturation after 12 or 16 PC (e.g. at Salamanca, Valentia, and Smolensk). Adding more PCs does not lead to a further general improvement of fit (not shown). This implies that the pointwise regression is truly superior to PC regression for any reasonable number of PCs at all stations, which is in accord with results of Klein and Walsh (1983) obtained in a somewhat different context. The root-mean-square errors range (Table II) from 1.4 °C at Valentia to 5.9 °C at Sodankylä for maximum temperature, and from 2.0 °C at Zugspitze to 8.2 °C at Sodankylä for minimum temperature.

Generally speaking, the fit is better where free atmospheric conditions can easily affect the surface weather. The downscaling methods, therefore, perform best at the two mountain stations, Zugspitze and Hohenpeissenberg, which are followed by two stations located in an open terrain, Valentia, situated on a coast and surrounded by sea, and Smolensk on flat plains of western Russia. With a few exceptions, the fit is better for maximum temperature, which is also an expected feature: The minimum temperatures are, especially in winter, more likely to be affected by local conditions and weather peculiarities, such as surface inversions.

4.2. Neural networks

Figure 5 compares the correlations achieved by the neural network models with their linear counterparts. The NN models based on 20 PCs were built at four stations only; an improvement over the linear regression of 20 PCs is achieved at one station for both maximum and minimum temperature (Sodankylä and Prague, respectively); for one station (Salamanca) for maximum temperature is the performance of NNs and linear regression approximately the same. In the remaining five cases, the linear method outperforms the NNs.

Table II. Root-mean-square errors for the pointwise linear regression (in °C).

Station	Soda	Zugs	Sala	Bamb	Hohe	Vale	Smol	Prag
Tmax	5.9	2.1	2.9	3.2	2.5	1.4	3.7	3.1
Tmin	8.2	2.0	3.1	4.8	2.5	2.5	5.0	3.7

DOI: 10.1002/joc

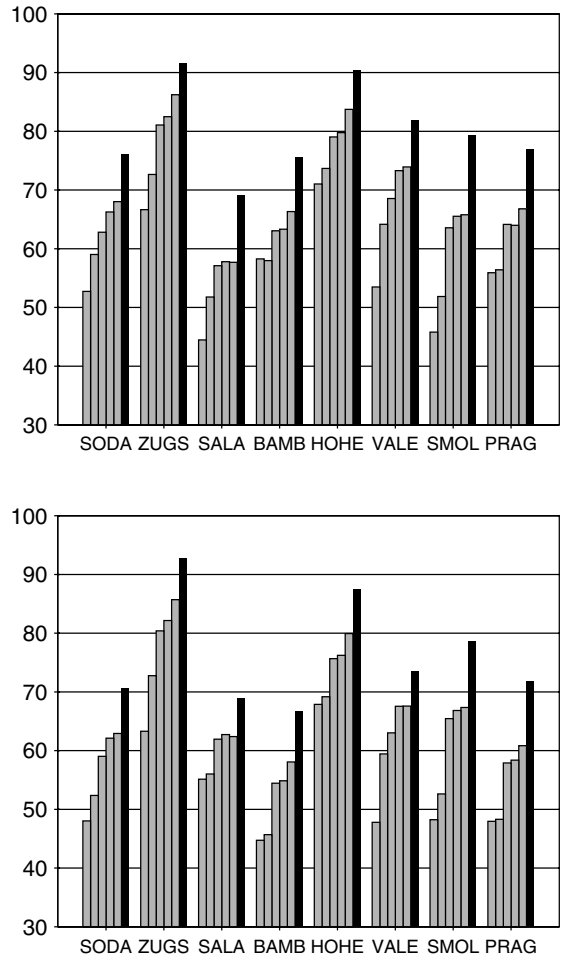

Figure 4. Correlation coefficients (×100) between observed and downscaled values for linear models: maximum temperature (top) and minimum temperature (bottom). Each cluster of bars corresponds to one station; within each cluster, the grey bars refer to PC-regression for (from left to right) 4, 6, 12, 16 and 20 PCs, and the black bar refers to pointwise regression.

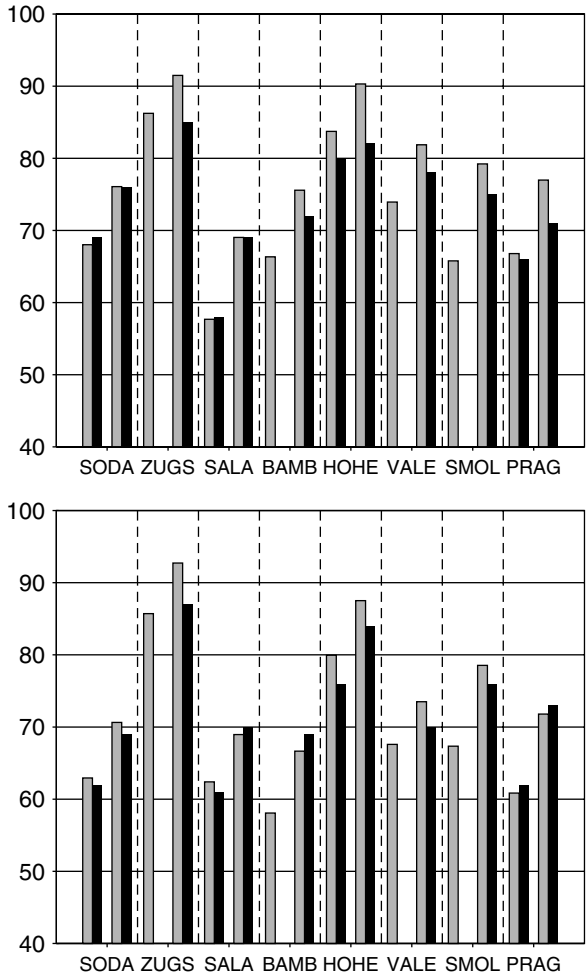

Figure 5. Correlation coefficients (×100) between observed and downscaled values for neural network models and their linear counterparts: maximum temperature (top) and minimum temperature (bottom). Each cluster of bars corresponds to one station; the left (right) pair of bars within each cluster corresponds to the 20 PC (pointwise) models; the grey (black) bars refer to the linear regression (NN) models.

For the pointwise models, the NNs are slightly better than the linear regression at three stations (Salamanca, Bamberg, Prague) for minimum temperature, while nowhere for maximum temperature. The performance of NNs and linear regression is the same at two stations (Sodankylä, Salamanca) for maximum temperature. In the remaining majority of cases, the linear regression is superior. Its superiority is particularly pronounced at the mountain stations (Zugspitze, Hohenpeissneberg).

The overall inferiority of the NN pointwise model is most likely due to a very large number of its parameters to be fitted. On the other hand, this may not be the case for the NN PC model, which contains a much lower number of parameters. A potential alternative explanation is, however, that the processes governing the relationships between the predictors and daily temperature are actually close to linear.

4.3. Classification

The correlations for the pointwise regression on data classified by T-mode PCA of 500 hPa heights are displayed in the top panel of Figure 6 for maximum temperature. Obviously and without exception, the increasing number of classes degrades the fit between the observed and downscaled values. The decline in the correlations is steepest at Salamanca where the overall skill is worst, while it is weakest at the two mountain stations (Zugspitze and Hohenpeissenberg) where the overall skill is best. The classification with four classes leads to a slight improvement over the unstratified data at two stations only (Sodankylä and Valentia). At the other six stations, the classification does not bring any improvement in downscaling skill. The two facts, derived from the top panel of Figure 6, also apply to the other classification method (k-means) and level (1000 hPa), as well as to minimum temperature.

A comparison of best performing classifications (i.e. those with four classes) for different classification methods is provided in the bottom panel of Figure 6, again for maximum temperature. Sodankylä and Valentia are the only stations where at least one of the classifications brings some gain in downscaling skill. A more

DOI: 10.1002/joc

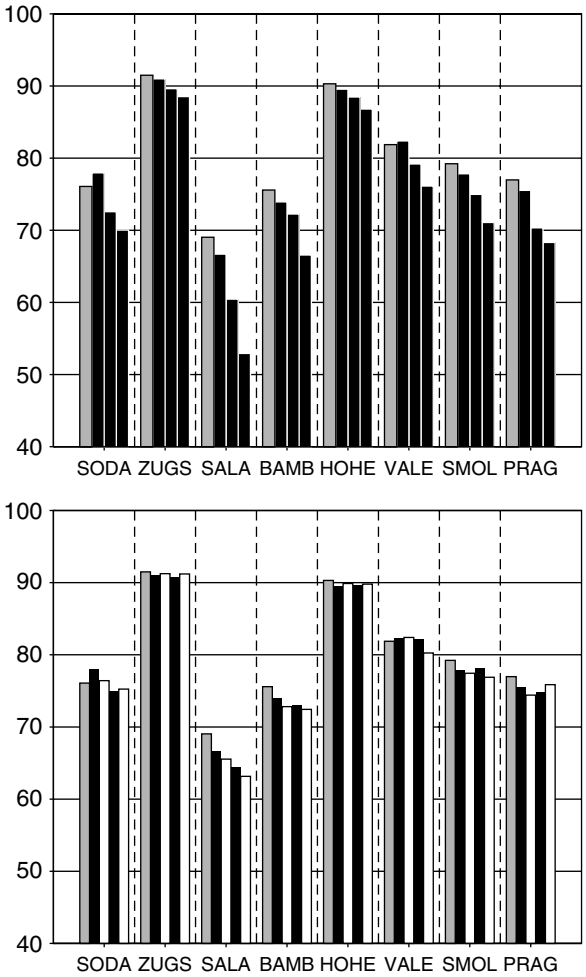

Figure 6. Correlation coefficients (×100) between observed and downscaled values for pointwise regression on classified data, all for maximum temperature. Each cluster of bars corresponds to one station. Top: for T-mode classification of 500 hPa heights; bars correspond, from left to right, to the pointwise regression on unclassified data (grey) and to the pointwise regression on classified data with 4, 11, and 18 classes (black). Bottom: for classifications with four classes; bars correspond, from left to right, to the pointwise regression on unclassified data (grey), and to the pointwise regression on data classified by T-mode PCA of 500 hPa heights (black), T-mode PCA of 1000 hPa heights (white), k-means clustering of 500 hPa heights (black), and k-means clustering of 1000 hPa heights (white).

general picture, summarizing the performance of pointwise regression on stratified data, is provided in terms of area mean correlations in Table III. None of the classifications results in even a marginal improvement in the area averaged downscaling skill. There is a slight tendency towards the T-mode PCA performing better than k-means, and the classifications based on the 500 hPa level being better than those based on the 1000 hPa level.

The results of downscaling by the PC regression on stratified data are presented in Table IV for T-mode PCA of the 500 hPa heights, in terms of mean correlation coefficients, for maximum temperature only. Results for minimum temperature and other classification methods are analogous. An improvement over the unstratified data is gained by classifications for all numbers of classes

Table III. Performance of downscaling by the pointwise regression on classified data: Mean correlation coefficient (×100) for different methods, levels where classification was performed, and numbers of clusters; for maximum (Tmax) and minimum (Tmin) temperature. The mean correlation coefficient (×100) for the reference method (pointwise regression without classification) is 80.1 for maximum temperature and 76.3 for minimum temperature.

		T-mode PCA		k-means	
	No. of classes	500 hPa	1000 hPa	500 hPa	1000 hPa
Tmax	4	79.4	78.8	78.5	78.1
	11/12	76.0	75.1	75.4	74.9
	18	72.5	73.0	70.7	72.1
Tmin	4	75.5	74.8	75.1	74.9
	11/12	72.0	71.0	72.0	70.9
	18	69.2	68.5	67.5	68.6

in the regression of 4 PCs, and for four classes in the regression of 12 and 20 PCs. The improvement is, nevertheless, marginal only and the performance of these downscaling models remains far below the pointwise regression.

One might argue that there is a redundancy in the above downscaling models because atmospheric circulation (i.e. geopotential heights) enters the models twice: in the classification, and again in the regression. To evaluate the degree of the redundancy, we built regression models with the 850 hPa temperature as the only predictor, both on unstratified data and for various classifications. Results are summarized in Table V. Comparing the two rows, one can see that the inclusion of geopotential heights among the predictors improves the performance of the downscaling models both on unstratified and classified data, that is, it adds extra information on atmospheric circulation that is not contained in the classifications themselves.

In the models based on classification, two effects influencing the agreement with observations counteract: the division into subsamples (classes) allows the regression to better tune to the predictor–predictand relationships within each subsample, leading to a gain in skill; whereas

Table IV. Performance of downscaling by PC regression on classified data: Mean correlation coefficient (×100) for the T-mode PCA classification of 500 hPa heights for different numbers of PCs and clusters. Also shown are results for the reference methods, viz., the PC regression without classification and the pointwise regression.

		No. of PCs			
		4	12	20	pointwise regr.
Without classification		56.0	67.4	71.1	80.1
No. of clusters	4	56.4	67.7	71.3	79.4
	11	56.8	64.9	68.1	76.0
	18	56.4	63.7	66.4	72.5

DOI: 10.1002/joc

Table V. Area mean correlation coefficients (×100) between the observed and downscaled maximum temperatures; pointwise regression, T-mode PCA of 500 hPa heights as a classification method, for the 850 hPa temperature (T8) and for the 500 hPa heights and 850 hPa temperature (Z5 + T8) as a predictor.

		No. of classes		
	Unstratified	4	11	18
T8	76.0	76.2	73.1	70.8
Z5 + T8	80.1	79.4	76.0	72.5

smaller sample sizes results in a greater sampling variability, and hence larger errors, in the estimation of regression parameters. The results indicate that the loss in skill due to a smaller sample size is more important than the gain due to a better tuning.

5. Higher-order statistical moments

First of all, several examples of temperature-empirical distributions are displayed in terms of histograms. The top row in Figure 7 shows three differently shaped observed minimum temperature distributions: a normal-like one for Salamanca, a strongly negatively skewed and highly peaked one for Bamberg, and a broad distribution almost limited from above at 0 °C for Smolensk. Both the pointwise regression and classification (second and last row, respectively, in Figure 7) yield distributions that appear to be close to normal, regardless of the shapes of the observed one. On the other hand, the neural networks seem to reproduce some of the specific features of the observed distributions to at least a certain extent: for Bamberg, the distribution produced by NNs is clearly peaked, although not as strongly as the observed, with hints of a negative skewness. For Smolensk, the NNs produce a relatively broad distribution with a strong decline in frequencies at the right tail, which is, however, far from the sharp upper limit of the observed distribution.

The ability of downscaling models to reproduce the shapes of temperature distributions is quantified in terms of the skewness and kurtosis coefficients in Tables VI and VII, respectively. Supposing the number of independent realizations in the time series is 300, which is a very conservative estimate, the skewness test for normality (Thode, 2002) indicates that the hypothesis of a zero skewness is rejected at the 95% significance level if the skewness coefficient exceeds (in absolute sense) 0.275. For an easier perception, skewness values exceeding 0.4 are in Table VI printed in bold. One can see that a large negative skewness is observed at five (four) stations for minimum (maximum) temperature. This is strongly underestimated by both the linear and classification methods, while neural networks yield values much closer to those observed. In one case, on the other hand, a neural network model strongly overestimates the observed skewness that is close to zero (minimum temperature at Salamanca, NN model based on PCs).

The kurtosis test of normality (Thode, 2002) yields the critical values for 300 independent realizations and the 95% significance level, equal to −0.47 and +0.63. In Table VII, the values exceeding these limits are shown in bold. For minimum temperature, a highly negative kurtosis, indicating that the distribution is flatter than the normal one, is observed at two stations, Sodankylä and Smolensk, whereas at Bamberg and Prague, the kurtosis is highly positive, resulting from the distribution being peaked. For maximum temperature, positive kurtosis prevails with values less extreme than for temperature minima. Similarly to skewness, large kurtosis values are only simulated by neural networks, although we note a fairly strong overestimation in some cases and a failure for minimum temperature at Sodankylä. Neither linear methods nor classifications are able to produce kurtosis values significantly different from zero.

To summarize, only the neural networks are capable of simulating the deviations of temperature distributions from normality, quantified by the third and fourth statistical moments. The linear methods only transfer the statistical properties of predictors to predictands; there is no mechanism in the linear downscaling able to produce deviations from the predictors' distributions. Since the predictors (geopotential heights, upper-air temperature, and their PCs) are more or less normally distributed, the linearly downscaled values cannot deviate from normality; this was already discussed by Huth *et al*. (2003). The classifications, on the other hand, have the potential to introduce non-normality by mixing several normal distributions, separate for each class. This effect appears, however, to be rather weak and insufficient to produce significant deviations from normality, or even a correct skewness and kurtosis.

6. Temporal and spatial structure

The temporal structure of the downscaled series is characterized by 1-day lag autocorrelations, displayed in Table VIII for maximum temperature. Results for minimum temperature are analogous. Generally, all the PC regressions lead to the overestimation of persistence, the highest autocorrelations appearing for the lowest number of PCs, 4. On the other hand, the classifications with 11 and 18 classes lead to a consistent underestimation of persistence, the bias increasing with the number of classes. The pointwise NN model tends to moderately overestimate autocorrelations. The pointwise regression and the classification models with four classes perform similarly, showing both an over- and underestimation, although the persistence yielded by the classifications methods is consistently lower. Quantifying the accuracy of simulated persistence by its mean absolute error over stations, one can see that the linear regression performs best in this respect, followed by the classifications with four classes, which in turn outperform the neural networks.

DOI: 10.1002/joc

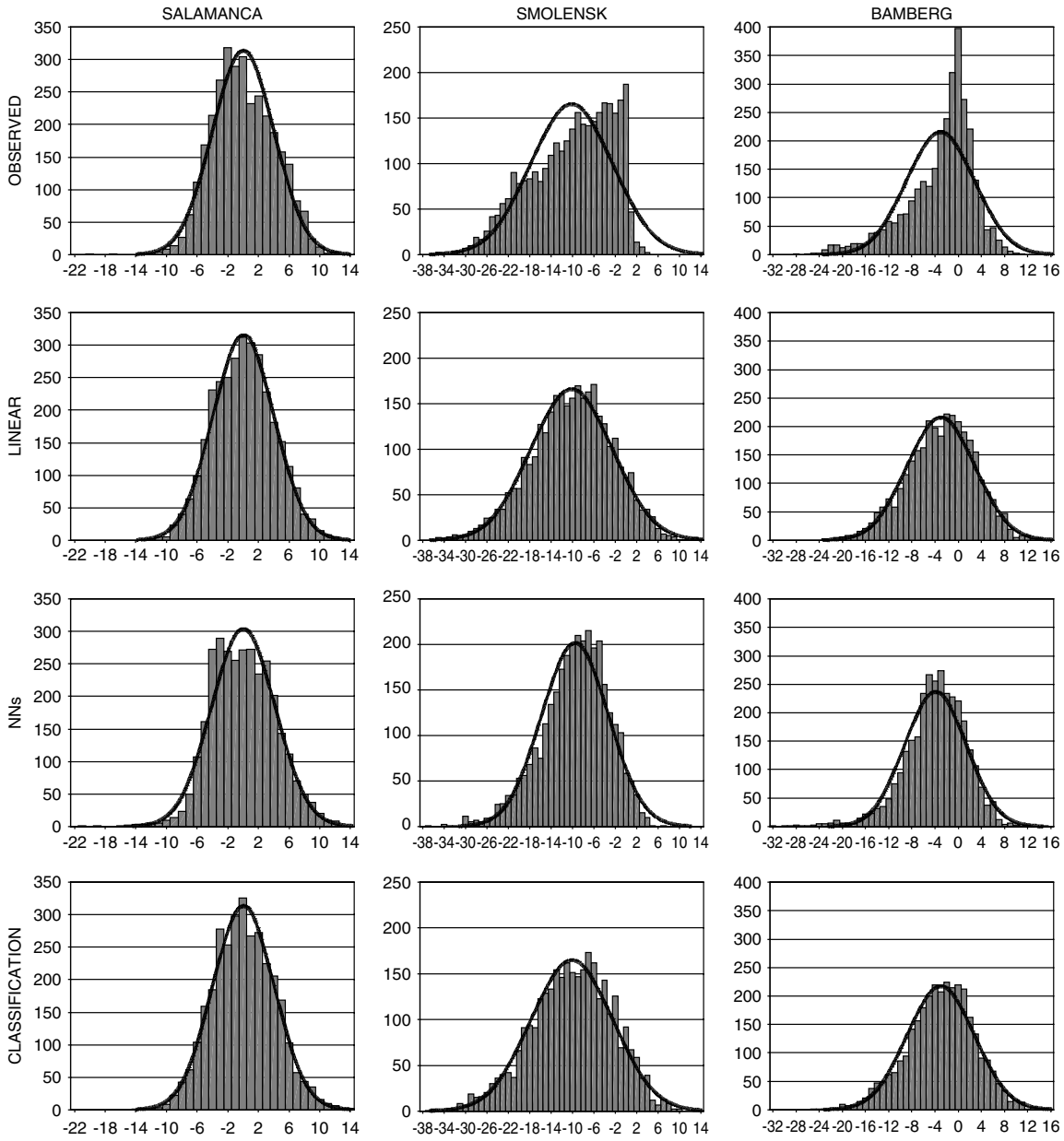

Figure 7. Histograms of minimum temperature at (from left to right) Salamanca, Smolensk, and Bamberg, for (from top to bottom) observed data, pointwise linear regression, pointwise neural networks, and pointwise regression on the T-mode PCA classification of 500 hPa heights with four classes. Over each histogram, an estimate of a pdf of normal distribution is superimposed.

As an example of spatial structure, the interstation correlations with one of the central European stations, Prague, are displayed in Table IX. Virtually all of the downscaling methods overestimate the spatial correlations. The bias is almost an order higher than for the temporal autocorrelations. The overestimation is strongest for the PC regression, while smallest (occasionally with a slight underestimation) for the classifications with a high number of classes. With a few exceptions, the classifications with four classes yield spatial correlations closer to the observed than the pointwise regression and the neural networks.

The reason for a weak performance of the PC regression models in terms of both temporal and spatial autocorrelations consists in the fact that several leading PCs, even though determined from daily data, describe large-scale patterns of low-frequency variability; their time series are therefore highly persistent. The higher-order PCs become less spatially extensive and more temporally variable, but since their variance is much smaller, the temporal and spatial properties of the downscaled series are dominated by the slowly temporally a spatially varying leading PCs. The decrease in persistence with an increasing number of classes in the classification methods is natural: the larger number of classes leads to a shorter lifetime within a class and a larger number of transitions between classes, i.e. a larger number of shifts from one regression model to another. This finally results in larger day-to-day variations than if a single model is used. The general overestimation of spatial correlations by all

DOI: 10.1002/joc

Table VI. Skewness coefficients for minimum temperature (top) and maximum temperature (bottom) for various methods. Values greater in absolute sense than 0.4 are in bold.

	Soda	Zugs	Sala	Bamb	Hohe	Vale	Smol	Prag
Observed	−.27	**−.44**	.07	**−1.10**	**−.42**	−.35	**−.52**	**−.81**
pointwise regr.	−.19	−.25	.09	−.25	−.35	−.23	−.21	−.24
4 PCs regr.	−.29	−.25	.01	−.18	−.22	−.08	−.20	−.18
12 PCs regr.	−.26	−.09	.09	−.10	−.15	−.18	−.13	−.08
20 PCs regr.	−.08	−.05	.06	−.08	−.09	−.18	−.11	−.08
NN pointwise	−.10	**−.48**	.03	**−.74**	**−.43**	−.28	**−.56**	**−.80**
NN 20 PCs	.24	–	**.43**	–	−.24	–	–	**−.94**
T-mode, Z5, 4 cl.	−.07	−.26	.11	−.25	−.33	−.16	−.25	−.35
T-mode, Z5, 11 cl.	−.20	−.27	.09	−.27	−.29	−.14	−.22	−.26
T-mode, Z5, 18 cl.	−.18	−.28	.11	−.35	−.29	−.07	−.26	−.33
T-mode, Z0, 4 cl.	−.13	−.24	.08	−.32	−.37	−.14	−.25	−.38
k-means, Z5, 4 cl.	−.15	−.27	.08	−.35	−.33	−.08	−.19	−.31
k-means, Z0, 4 cl.	−.20	−.23	.13	−.30	−.31	−.08	−.26	−.35

	Soda	Zugs	Sala	Bamb	Hohe	Vale	Smol	Prag
Observed	**−.64**	**−.60**	−.16	−.20	−.15	**−.81**	**−.91**	−.29
Pointwise regr.	.02	−.28	−.13	−.14	−.31	**−.47**	−.15	−.16
4 PCs regr.	−.28	−.26	−.24	−.21	−.22	−.13	−.20	−.19
12 PCs regr.	−.15	−.11	−.11	−.13	−.17	−.25	−.10	−.09
20 PCs regr.	.05	−.05	−.21	−.05	−.09	−.34	−.09	−.02
NN pointwise	−.24	**−.60**	.05	−.30	−.19	**−.73**	**−.62**	−.09
NN 20 PCs	.03	–	.13	–	.01	–	–	.16
T-mode, Z5, 4 cl.	−.06	−.32	−.17	−.18	−.25	**−.46**	−.25	−.22
T-mode, Z5, 11 cl.	−.14	**−.42**	−.06	−.14	−.18	**−.44**	−.24	−.13
T-mode, Z5, 18 cl.	−.20	**−.45**	−.07	−.17	−.21	**−.48**	−.27	−.19
T-mode, Z0, 4 cl.	−.11	−.33	−.16	−.20	−.31	**−.41**	−.28	−.19
k-means, Z5, 4 cl.	−.06	−.36	−.20	−.22	−.25	**−.44**	−.29	−.20
k-means, Z0, 4 cl.	−.13	−.37	−.02	−.16	−.29	−.38	−.26	−.17

methods stems mainly from the fact that the downscaling models, unlike the real data, do not explicitly include local effects, unrelated to large-scale forcings. From this point of view, the inflation does not appear to be a sensible way of reproducing the variance in the downscaled series, and adding noise is theoretically preferable instead (von Storch, 1999). However, the persistence is simulated with a fairly acceptable degree of accuracy by several methods, and adding noise would severely deteriorate it, which was already shown for the linear methods by Huth *et al*. (2001). The reason for a larger positive bias in the persistence in pointwise NN models relative to the linear pointwise regression is unclear.

7. Conclusions

In this study, we examine whether the introduction of non-linearity into statistical downscaling models of daily temperature brings an improvement over the linear models. The non-linearity is introduced in two ways: by artificial neural networks and by a stratification of the dataset by circulation pattern classifications. The comparison is carried out for eight stations in Europe, differing in their geographical and climatic settings, in winter. Several criteria are employed to quantify the performance of downscaling models: the degree of fit between the downscaled and observed series (in terms of correlation coefficients); the shape of statistical distributions, namely the deviations from normality, characterized by their skewness and kurtosis; the persistence of the downscaled series; and a spatial structure of downscaled temperature, quantified by autocorrelations. The results can be summarized in the following items:

- The best linear method is the pointwise regression, which in most criteria outperforms the regression of predictor's principal components. Of the neural network models, the pointwise NN is better than the PC-based NN. Among the classification models, the lowest number of classes, 4, yields consistently better results than the higher numbers of classes, 11–18; the selection of the classification method (k-means clustering or T-mode PCA) and the level where the height patterns are classified (500 or 1000 hPa) has only a minor effect on the performance.
- The best fit with the observations is achieved by the linear method. The application of NNs and stratification of data do not bring any gain in the downscaling skill, with only a few marginal exceptions.
- Neural networks are the only model capable of producing distributions deviating from normality, i.e. with a

DOI: 10.1002/joc

Table VII. Kurtosis coefficients for minimum temperature (top) and maximum temperature (bottom) for various methods. Values less than −0.47 or greater than 0.63 are in bold.

	Soda	Zugs	Sala	Bamb	Hohe	Vale	Smol	Prag
Observed	**−.97**	−.07	−.18	**1.31**	.22	−.44	**−.57**	**.66**
Pointwise regr.	.07	−.10	−.22	−.05	.07	−.03	−.04	−.06
4 PCs regr.	−.16	−.05	−.44	.05	–	−.33	−.19	.02
12 PCs regr.	.09	−.26	−.31	.05	−.02	−.26	−.16	.03
20 PCs regr.	.08	−.34	−.31	−.26	−.26	−.18	−.07	−.27
NN pointwise	−.14	–	.27	**1.83**	.41	.07	.39	**1.61**
NN 20 PCs	.36	–	.32	–	.26	–	–	**3.58**
T-mode, Z5, 4 cl.	−.08	−.07	−.04	.14	.25	−.06	−.12	.40
T-mode, Z5, 11 cl.	.10	−.08	.03	.13	.27	.06	.18	.13
T-mode, Z5, 18 cl.	.06	−.01	.01	.39	.20	.02	.11	.25
T-mode, Z0, 4 cl.	−.16	−.16	−.14	.10	.12	−.09	−.04	.22
k-means, Z5, 4 cl.	.09	−.11	−.02	.44	.17	−.12	−.03	.28
k-means, Z0, 4 cl.	−.05	−.15	−.17	.12	.08	−.13	.03	.25

	Soda	Zugs	Sala	Bamb	Hohe	Vale	Smol	Prag
Observed	−.04	.43	.18	.54	−.20	**.69**	.49	.39
Pointwise regr.	−.01	−.09	.42	.02	−.14	.16	.01	−.08
4 PCs regr.	−.17	−.02	−.17	.03	−.04	−.23	−.22	.04
12 PCs regr.	−.02	−.15	−.04	.01	−.11	−.17	−.07	.04
20 PCs regr.	.03	−.29	.05	−.28	−.30	.02	−.04	−.24
NN pointwise	−.22	.41	.46	**1.34**	−.08	**.69**	**.70**	.18
NN 20 PCs	−.29	–	.39	–	.15	–	–	.24
T-mode, Z5, 4 cl.	−.17	−.04	.53	.15	−.08	.34	.05	.12
T-mode, Z5, 11 cl.	−.03	.18	.22	.17	−.14	.12	.13	.10
T-mode, Z5, 18 cl.	.10	.20	.27	.37	−.08	.38	.08	.25
T-mode, Z0, 4 cl.	–	.01	.37	.06	−.04	.09	.04	.10
k-means, Z5, 4 cl.	.01	.01	.38	.32	−.13	.23	.15	.25
k-means, Z0, 4 cl.	−.05	.02	.48	.19	−.02	.12	.07	.06

non-zero skewness and kurtosis. They reproduce the skewness relatively well, but capture only the sign, not the magnitude of kurtosis.

- The 1-day lag autocorrelations are best reproduced by the linear method.
- The interstation correlations are overestimated by all the methods, the classifications being closest to observations.

To summarize, the non-linearity leads to an improvement only in the shapes of statistical distributions for the neural networks and in the spatial structure for the classifications. Taken all results together, the pointwise linear regression can be regarded the best performing method of all those examined.

The reason why the non-linear methods tend to be inferior to the linear ones, especially in the fit to observations, can be seen in several effects. First, there might be a problem with the non-linearity of relations between predictors and predictands. The relationships between the predictors and daily surface temperature may be intrinsically linear, or close to linear, or the non-linearity may have a character of noise and may be too complex to be described by quite a simple architecture of the NNs (Mikšovský and Raidl, 2005). Second, there are too many parameters in the NN models to be determined, which leads to a larger uncertainty in their estimation. Third, in the classification models, the loss in skill due to smaller sample size outweighs the gain due to potentially better-fitting models within individual classes. Also, the preprocessing of data (standardization of anomalies, use of PCs, selection from all gridpoints) that was proved to be suitable for the linear methods, may be sub-optimal for the NNs.

The linear downscaling is the best of the examined methods for downscaling daily extreme temperature. Although the family of the examined methods is relatively broad and includes a variety of different linear methods and combinations of predictors (cf. Huth, 2002), as well as NNs and different classification approaches, there are, potentially, plenty of other methods not included in our comparisons. We think there still is a potential for an improvement of the NN downscaling method: the number of neurons in the input layer (i.e. the number of predictors) must be sufficiently low in order to reduce the uncertainty in the estimates of the parameters and to keep them stable, but the few predictors must represent the relevant signal. The two requirements act against each other and it is not clear, *a priori*, if an effort in searching for such a method would be successful. We have demonstrated that taking the principal components as the predictors is not a

DOI: 10.1002/joc

Table VIII. One day lag autocorrelations (×1000) for maximum temperature. Values within +/−0.040 from the observations are in bold; values below (above) this range are in italics (in light print). In the last column, the mean absolute error of the 1-day lag correlations (×1000) is shown for selected models.

	Soda	Zugs	Sala	Bamb	Hohe	Vale	Smol	Prag	MAE
Observed	**717**	**737**	**737**	**825**	**778**	**707**	**810**	**850**	0.0
Pointwise regr.	**748**	793	**769**	**822**	**804**	**746**	**795**	**827**	28.1
4 PCs regr.	920	921	914	928	925	927	923	931	–
12 PCs regr.	902	896	888	922	903	904	904	921	–
20 PCs regr.	846	855	891	887	859	846	898	892	–
NN pointwise	825	855	**756**	**796**	**783**	767	**840**	**838**	47.6
NN 20 PCs	837	–	855	–	835	–	–	**869**	–
T-mode, Z5, 4 cl.	**736**	782	**737**	**781**	**786**	**730**	*750*	*786*	32.9
T-mode, Z5, 11 cl.	*672*	761	*605*	*698*	**764**	**689**	*712*	*709*	–
T-mode, Z5, 18 cl.	*653*	749	*558*	*641*	**743**	*662*	*660*	*666*	–
T-mode, Z0, 4 cl.	**730**	790	**727**	*777*	**795**	**731**	*764*	*765*	37.0
k-means, Z5, 4 cl.	**730**	780	**721**	*779*	**789**	**712**	**781**	*786*	28.4
k-means, Z0, 4 cl.	**744**	784	**721**	*766*	**795**	**713**	768	*778*	35.7

Table IX. Interstation correlations (×100) with Prague for maximum temperature. Three lowest (highest) values for each station are printed in bold (italics). In the last column, the mean absolute error of interstation correlations is shown.

	Soda	Zugs	Sala	Bamb	Hohe	Vale	Smol	MAE
Observed	**3**	**45**	**26**	**89**	**64**	**30**	**42**	–
Pointwise regr.	5	62	41	95	80	40	56	11.4
4 PCs regr.	*21*	*87*	*74*	*98*	*94*	*71*	*82*	32.6
12 PCs regr.	*14*	*74*	*53*	*97*	*88*	*57*	*68*	21.7
20 PCs regr.	6	*74*	*48*	*97*	*88*	*49*	*65*	18.3
NN pointwise	8	**46**	43	90	81	43	58	10.0
NN 20 PCs	**-3**	–	44	–	*88*	–	–	–
T-mode, Z5, 4 cl.	**2**	58	38	91	76	39	49	8.0
T-mode, Z5, 11 cl.	4	53	**31**	**85**	**70**	**32**	**45**	4.1
T-mode, Z5, 18 cl.	4	**51**	**26**	**80**	**66**	**32**	**42**	2.9
T-mode, Z0, 4 cl.	5	59	39	91	76	35	50	8.0
k-means, Z5, 4 cl.	4	58	35	91	76	37	51	7.6
k-means, Z0, 4 cl.	*7*	60	39	91	77	35	50	8.6

way in the right direction as they miss some important information.

Since the examined stations cover a wide range of climates in Europe, from subpolar to Mediterranean, from maritime to continental, as well as from lowlands to mountain tops, we believe that the results can be generalized at least to whole northern midlatitudes.

Acknowledgement

This research was supported by the Grant Agency of the Czech Republic, project 205/02/0871.

References

Bartzokas A, Metaxas DA. 1996. Northern Hemisphere gross circulation types. Climatic change and temperature distribution. *Meteorologische Zeitschrift* **5**: 99–109.

Brinkmann WAR. 1999. Application of non-hierarchically clustered circulation components to surface weather conditions: Lake Superior basin winter temperatures. *Theoretical and Applied Climatology* **63**: 41–56.

Cassou C, Terray L, Hurrell JW, Deser C. 2004. North Atlantic winter climate regimes: Spatial asymmetry, stationarity with time, and oceanic forcing. *Journal of Climate* **17**: 1055–1068.

Cavazos T. 1997. Downscaling large-scale circulation to local winter rainfall in north-eastern Mexico. *International Journal of Climatology* **17**: 1069–1082.

Cavazos T. 1999. Large-scale circulation anomalies conducive to extreme precipitation events and derivation of daily rainfall in northeastern Mexico and southeastern Texas. *Journal of Climate* **12**: 1506–1523.

Cavazos T. 2000. Using self-organizing maps to investigate extreme climate events: An application to wintertime precipitation in the Balkans. *Journal of Climate* **13**: 1718–1732.

Compagnucci RH, Salles MA. 1997. Surface pressure patterns during the year over southern South America. *International Journal of Climatology* **17**: 635–653.

Crane RG, Hewitson BC. 1998. Doubled CO_2 precipitation changes for the Susquehanna basin: Down-scaling from the GENESIS general circulation model. *International Journal of Climatology* **18**: 65–76.

Easterling DR. 1999. Development of regional climate scenarios using a downscaling approach. *Climatic Change* **41**: 615–634.

Enke W, Spekat A. 1997. Downscaling climate model outputs into local and regional weather elements by classification and regression. *Climate Research* **8**: 195–207.

Esteban P, Jones PD, Martín-Vide M, Mases M. 2005. Atmospheric circulation patterns related to heavy snowfall days in Andorra, Pyrenees. *International Journal of Climatology* **25**: 319–329.

DOI: 10.1002/joc

Gardner MW, Dorling SR. 1998. Artificial networks (the multilayer perceptron)–a review of applications in the atmospheric sciences. *Atmospheric Environment* **32**: 2627–2636.
Gong X, Richman MB. 1995. On the application of cluster analysis to growing season precipitation data in North America east of the Rockies. *Journal of Climate* **8**: 897–931.
Goodess CM, Jones PD. 2002. Links between circulation and changes in the characteristics of Iberian rainfall. *International Journal of Climatology* **22**: 1593–1615.
Hewitson BC, Crane RG. 1992. Large scale atmospheric controls on local precipitation in central Mexico. *Geophysical Reserarch Letters* **19**: 1835–1838.
Hsieh WW, Tang B. 1998. Applying neural network models to prediction and data analysis in meteorology and oceanography. *Bulletin of the American Meteorological Society* **79**: 1855–1870.
Huth R. 1996a. Properties of the circulation classification scheme based on the rotated principal component analysis. *Meteorology and Atmospheric Physics* **59**: 217–233.
Huth R. 1996b. An intercomparison of computer-assisted circulation classification methods. *International Journal of Climatology* **16**: 893–922.
Huth R. 1997. Continental-scale circulation in the UKHI GCM. *Journal of Climate* **10**: 1545–1561.
Huth R. 1999. Statistical downscaling in central Europe: evaluation of methods and potential predictors. *Climate Research* **13**: 91–101.
Huth R. 2002. Statistical downscaling of daily temperature in central Europe. *Journal of Climate* **15**: 1731–1742.
Huth R, Kyselý J, Dubrovský M. 2001. Time structure of observed, GCM-simulated, downscaled, and stochastically generated daily temperature series. *Journal of Climate* **14**: 4047–4061.
Huth R, Kyselý J, Dubrovský M. 2003. Simulation of surface air temperature by GCMs, statistical downscaling and weather generator: higher-order statistical moments. *Studia Geophysica et Geodaetica* **47**: 203–216.
Jacobeit J, Wanner H, Luterbacher J, Beck C, Philipp A, Sturm K. 2003. Atmospheric circulation variability in the North-Atlantic-European area since the mid-seventeenth century. *Climate Dynamics* **20**: 341–352.
Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, Iredell M, Saha S, White G, Woollen J, Zhu Y, Chelliah M, Ebisuzaki W, Higgins W, Janowiak J, Mo KC, Ropelewski C, Wang J, Leetmaa A, Reynolds R, Jenne R, Joseph D. 1996. The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society* **77**: 437–471.
Karl TR, Wang WC, Schlesinger ME, Knight RW, Portman D. 1990. A method of relating general circulation model simulated climate to the observed local climate. Part I: Seasonal statistics. *Journal of Climate* **3**: 1053–1079.
Klein WH, Walsh JE. 1983. A comparison of pointwise screening and empirical orthogonal functions in specifying monthly surface temperature from 700 mb data. *Monthly Weather Review* **111**: 669–673.
Klein Tank AMG, Wijngaard JB, Können GP, Böhm R, Demarée G, Gocheva A, Mileta M, Pashiardis S, Hejkrlik L, Kern-Hansen C, Heino R, Bessemoulin P, Müller-Westermeier G, Tzanakou M, Szalai S, Pálsdóttir T, Fitzgerald D, Rubin S, Capaldo M, Maugeri M, Leitass A, Bukantis A, Aberfield R, Engelen AFV van, Forland E, Mietus M, Coehlo F, Mares C, Razuvaev V, Nieplova E, Cegnar T, Antonio López J, Dahlström B, Moberg A, Kirchhofer W, Ceylan A, Pachaliuk O, Alexander LV, Petrovic P. 2002. Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. *International Journal of Climatology* **22**: 1441–1453.
Li X, Sailor D. 2000. Application of tree-structured regression for regional precipitation prediction using general circulation model output. *Climate Research* **16**: 17–30.
Luterbacher J, Rickli R, Xoplaki E, Tinguely C, Beck C, Pfister C, Wanner H. 2001. The Late Maunder Minimum (1675–1715)–a key period for studying decadal scale climatic change in Europe. *Climatic Change* **49**: 441–462.
McGinnis DL. 2000. Synoptic controls on upper Colorado river basin snowfall. *International Journal of Climatology* **20**: 131–149.
Michaelsen J. 1987. Cross-validation in statistical climate forecast models. *Journal of Climate and Applied Meteorology* **26**: 1589–1600.
Mikšovský J, Raidl A. 2005. Testing the performance of three nonlinear methods of time series analysis for prediction and downscaling of European daily temperatures. *Nonlinear Processes in Geophysics* **12**: 979–991.
Mpelasoka FS, Mullan AR, Heerdegen RG. 2001. New Zealand climate change information derived by multivariate statistical and artificial neural network approaches. *International Journal of Climatology* **21**: 1415–1433.
Müller GV, Compagnucci R, Nuñez MN, Salles A. 2003. Surface circulation associated with frost in the wet pampas. *International Journal of Climatology* **23**: 943–961.
O'Lenic EA, Livezey RE. 1988. Practical considerations in the use of rotated principal component analysis (RPCA) in diagnostic studies of upper-air height fields. *Monthly Weather Review* **116**: 1682–1689.
Olsson J, Uvo CB, Jinno K. 2001. Statistical atmospheric downscaling of short-term extreme rainfall by neural networks. *Physics and Chemistry of the Earth B* **26**: 695–700.
Richman MB. 1986. Rotation of principal components. *Journal of Climatology* **6**: 293–335.
Romero R, Sumner G, Ramis C, Genovés A. 1999. A classification of the atmospheric circulation patterns producing significant daily rainfall in the Spanish Mediterranean area. *International Journal of Climatology* **19**: 765–785.
Santos JA, Corte-Real J, Leite SM. 2005. Weather regimes and their connection to the winter rainfall in Portugal. *International Journal of Climatology* **25**: 33–50.
Saunders IR, Byrne JM. 1996. Generating regional precipitation from observed and GCM synoptic-scale pressure fields, southern Alberta, Canada. *Climate Research* **6**: 237–249.
Saunders IR, Byrne JM. 1999. Using synoptic surface and geopotential height fields for generating grid-scale precipitation. *International Journal of Climatology* **19**: 1165–1176.
Schoof JT, Pryor SC. 2001. Downscaling temperature and precipitation: A comparison of regression-based methods and artificial neural networks. *International Journal of Climatology* **21**: 773–790.
Solman SA, Nuñez MN. 1999. Local estimates of global climate change: A statistical downscaling approach. *International Journal of Climatology* **19**: 835–861.
Solman SA, Menéndez CG. 2003. Weather regimes in the South American sector and neighbouring oceans during winter. *Climate Dynamics* **21**: 91–104.
Terray L, Demory M-E, Déqué M, de Coetlogon G, Maisonnave E. 2004. Simulation of late-twenty-first-century changes in wintertime atmospheric circulation over Europe due to anthropogenic causes. *Journal of Climate* **17**: 4630–4635.
Thode HC Jr. 2002. *Testing for Normality*. Marcel Dekker: New York, Basel; 479.
Trigo RM, Palutikof JP. 1999. Simulation of daily temperatures for climate change scenarios over Portugal: a neural network model approach. *Climate Research* **13**: 45–59.
Trigo RM, Palutikof JP. 2001. Precipitation scenarios over Iberia: A comparison between direct GCM output and different downscaling techniques. *Journal of Climate* **14**: 4422–4446.
von Storch H. 1999. On the use of "inflation" in statistical downscaling. *Journal of Climate* **12**: 3505–3506.
Weichert A, Bürger G. 1998. Linear versus nonlinear techniques in downscaling. *Climate Research* **10**: 83–93.
Wilby RL, Wigley TML, Conway D, Jones PD, Hewitson BC, Main J, Wilks DS. 1998. Statistical downscaling of general circulation model output: A comparison of methods. *Water Resources Research* **34**: 2995–3008.
Zorita E, von Storch H. 1999. The analog method as a simple statistical downscaling technique: Comparison with more complicated methods. *Journal of Climate* **12**: 2474–2489.

DOI: 10.1002/joc