Nonlin. Processes Geophys., 14, 211–222, 2007
www.nonlin-processes-geophys.net/14/211/2007/

Nonlinear Processes in Geophysics

Prediction of minimum temperatures in an alpine region by linear and non-linear post-processing of meteorological models

E. Eccel[1], **L. Ghielmi**[1], **P. Granitto**[2], **R. Barbiero**[3], **F. Grazzini**[4], **and D. Cesari**[4]

[1]IASMA Research Centre – Natural Resources Department, Via E. Mach, 1 – 38010 San Michele all'Adige (TN), Italy
[2]Instituto de Física Rosario Conicet UNR Bv. 27 de Febrero 210 bis 2000 Rosario, Argentina
[3]Autonomous Province of Trento – Meteotrentino, Department of Civil Protection, Via Vannetti, 41 – 38100 Trento, Italy
[4]ARPA-SIM Emilia-Romagna, Viale Silvani 6, 40122 Bologna, Italy

Received: 18 December 2006 – Revised: 4 May 2007 – Accepted: 13 May 2007 – Published: 25 May 2007

Abstract. Model Output Statistics (MOS) refers to a method of post-processing the direct outputs of numerical weather prediction (NWP) models in order to reduce the biases introduced by a coarse horizontal resolution. This technique is especially useful in orographically complex regions, where large differences can be found between the NWP elevation model and the true orography. This study carries out a comparison of linear and non-linear MOS methods, aimed at the prediction of minimum temperatures in a fruit-growing region of the Italian Alps, based on the output of two different NWPs (ECMWF T511–L60 and LAMI-3). Temperature, of course, is a particularly important NWP output; among other roles it drives the local frost forecast, which is of great interest to agriculture. The mechanisms of cold air drainage, a distinctive aspect of mountain environments, are often unsatisfactorily captured by global circulation models. The simplest post-processing technique applied in this work was a correction for the mean bias, assessed at individual model grid points. We also implemented a multivariate linear regression on the output at the grid points surrounding the target area, and two non-linear models based on machine learning techniques: Neural Networks and Random Forest. We compare the performance of all these techniques on four different NWP data sets. Downscaling the temperatures clearly improved the temperature forecasts with respect to the raw NWP output, and also with respect to the basic mean bias correction. Multivariate methods generally yielded better results, but the advantage of using non-linear algorithms was small if not negligible. RF, the best performing method, was implemented on ECMWF prognostic output at 06:00 UTC over the 9 grid points surrounding the target area. Mean absolute errors in the prediction of 2 m temperature at 06:00 UTC were approximately 1.2°C, close to the natural variability inside the area itself.

Correspondence to: E. Eccel
(emanuele.eccel@iasma.it)

1 Introduction

Meteorological models are the best prognostic instruments available for operational forecast purposes. The output of such models is only available on a coarse grid, whose nodes are scattered unevenly over the geographical domain. The typical distance between grid points ranges from a few dozen km for general circulation models (GCM) down to a few km for limited area models (LAM). One of the quantities that can be forecasted is the ground level temperature (the temperature at 2 m above the surface), but this prediction is closely tied to the (approximated) topographic position assigned by the model to each grid point.

Large-scale models cannot represent the local topography when the orography is strongly irregular. This is a typical feature of alpine territory, where deviations as large as 1000 m are likely to occur at some grid points (Fig. 1). This effect is particularly evident in the case of deep valleys such as Adige Valley, which is one of the largest in the Alps in terms of both length and depth. In alpine areas, a bias of 4 to 6°C in temperature forecasts is common at grid points where the model elevation is dramatically different from the true elevation.

As a consequence, many numerical weather simulations use downscaling as a post-processing step (Weichert and Bürger, 1998; Schoof and Pryor, 2001; Huth, 2002, 2004; Miksovsky and Raidl, 2005), in order to relate predictions on grid points to real physical sites.

The forecast of spring frosts (or equivalently the prediction of sub-zero nocturnal minimum temperature) is particularly important to agriculture. Unfortunately, it is also a difficult task for meteorological models. During night the thermal profile of the atmosphere strongly depends on meteorological conditions, which can affect the downward flow of cooled air. This flow not only depends on atmospheric stability and cloud cover, but is also influenced by atmospheric circulation on a wider scale and by local orography (André

Published by Copernicus GmbH on behalf of the European Geosciences Union and the American Geophysical Union.

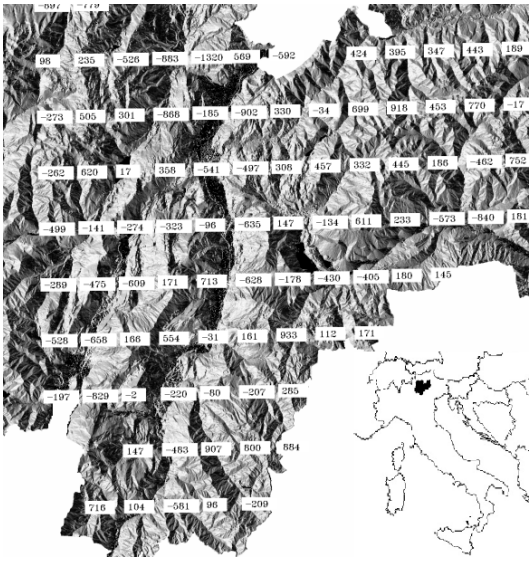

Fig. 1. Differences between the model elevation at LAMI grid points and the true elevation taken from a 10 m resolution DEM.

and Mahrt, 1982; Carlson and Stull, 1986; Gassmann and Mazzeo, 2001).

A good quantitative prediction of minimum nocturnal temperatures at Alpine areas can only be achieved by downscaling the raw (direct) output of numerical models (hereafter DMO). There are several previous approaches to this problem. The simplest are univariate methods as, for example, the application of site-specific offsets (fixed or seasonal) or Kalman filter techniques (Homleid, 1995; Galanis and Anadranistakis, 2002; Anadranistakis et al., 2004; Cane et al., 2004). Although univariate methods have been well tested, multivariate methods have the potential for modelling the influence of both properties of the site and prognostics provided by meteorological models. The use of machine learning (ML) techniques is widespread in meteorological practice, particularly for temperature prediction (Schizas et al., 1991; Abdel-Aal and Elhadidy, 1994; Robinson and Mort, 1997; Arca et al., 1998; Verdes et al., 2000, Basili et al., 2006). Such methods usually create accurate non-linear multivariate models. A potential drawback is that they do not produce understandable relationships between predictors and outputs, which prevent some researchers to adopt these methods.

In this work we apply two different machine learning algorithms to downscale the gridded output of numerical models, to obtain more accurate predictions of minimum nocturnal temperatures in the Adige Valley and compare them to the more traditional multilinear regression model.

2 Methods

2.1 Choice of post-processing approach

The two most well-known statistical approaches to the downscaling of numerical outputs are "Perfect Prog" (Perfect Prognosis) and "MOS" (Model Output Statistics) – Wilks (1995). Both methodologies build functional relationships (which may or may not be explicit) between numerical forecasts and observations, but they differ in the way these relationships are inferred. In the Perfect Prog technique, it is assumed that the atmospheric state predicted by the model exactly matches the true atmospheric state (hence the name "perfect prognosis"). Relationships are then established between the model outputs and the observed variables ("predictands") during a training stage. The relationships obtained from past data are modelled over a certain "lead time", which, in the algorithm training stage, had been used as an "analysis" input. This approach demonstrates the importance of calculating individual downscaling offsets; even though the model is considered to be perfect, its low spatial resolution still requires site-specific adjustments. The algorithms linking predictors to predictands are always developed with a reference to an observational dataset, rather than numerical weather prediction (NWP) prognostics.

In the MOS approach, relationships are obtained by using the model's outputs as predictors (e.g., for model run at 00:00 UTC, temperature forecast at different grid points at a lead time of +30 h), and measured quantities as predictands. This approach, however, also takes into account "errors" intrinsic to the model itself. It has the disadvantage of being strictly applicable only over periods of time when the algorithms used in the NWP models are homogeneous. Changes in NWP calculation modules might alter model performance and require a new MOS parameterisation.

Each technique has its pros and cons. Perfect Prog has the disadvantage of not taking into consideration discrepancies between the NWP model and the true atmospheric state (Wilson, 2001). It is stable, however, and does not require adjustment after changes to the NWP model. MOS, on the other hand, takes into account systematic biases in the NWP output. Its major shortcoming is that any change to the model, such as an improvement in spatial resolution or a new parameterisation, may alter the performance of the post-processing. The MOS therefore needs to be checked after any major change. Nonetheless, for the purposes of this paper we chose an MOS approach to allow for the correction of systematic errors in the model. A different MOS was developed for each dataset: two models (ECMWF and LAMI) and two times of operation (12:00 and 00:00 UTC).

2.2 Investigation area and meteorological data

An irregular orography characterises alpine regions, whose mountain massifs are carved by deep valleys. Agricultural

areas in the territory of Trento cover either gentle valley sides or large, flat valley floors; the Adige valley region, in particular, is highly urbanised and its territory has been put to intensive use in fruit-growing. In this context, DMOs are problematic because the "internal" orography of the models is highly simplified and thus generally different from the true orography. This is true not only in terms of the absolute elevation of grid points, but also in terms of accurately representing valley widths and other morphological features. A comparison of differences in grid point elevations in the region between the LAMI topographic model and 10-m Digital Elevation Model (DEM) is reported in Fig. 1).

The target area is the middle reach of the Adige river valley, extending from the meteorological station of S. Michele, close to the northern border of Trentino, to that of Rovereto (Fig. 2). The area covers a fruit- and vine-growing region; south and north of it, apple (the frost-endangered crop) is less common. The selected area contains a total of three meteorological stations, which are situated on the floor of the valley. The Adige valley is large and deep, with a flat bottom formed by alluvia from the Adige River. Figure 2 indicates the geographic positions of all grid points and the meteorological stations. The variation in height over the reach is very low; over a distance of about 40 km (from Rovereto to S. Michele), the altitude rises from 170 to 210 m above sea level. The inclination of transverse terrain in the valley is also negligible. The three stations inside the valley can be considered as lying on generally flat terrain, even though they are surrounded by mountain peaks rising as much as 2000 m above the valley floor. The area reflects a good climatic homogenity, allowing to treat it as a meteorological unit; especially south of Rovereto, minimum temperatures are often higher than in the selected area.

Due to prevailing thermal inversion conditions during frost episodes, the valley bottom of the Adige river is a particularly frost-prone area. Thanks to its low elevation, however, phenological development in the valley is generally more advanced than in the hilly surroundings, resulting in an increased spring frost risk all over the valley bottom area. This reason led us to select the central reach of this valley as a suitable target area for minimum temperature modelling.

The post-processing model is calibrated using the mean hourly 2 m temperature, measured from 06:00 to 07:00 local time (05:00 to 06:00 UTC). This choice is close to the usual time of temperature minima under standard conditions of clear sky and calm wind, and it is also quite close to the standard lead time of 06:00 UTC used in NWP models.

2.3 NWP models

The European Centre for Medium-Range Weather Forecasts (ECMWF) operates a semi-lagrangian global model. The operational model in use through 2006 (T511, L60) carries out a triangular truncation on the spectral components down to a grid size of 0.5° (about 55 km N-S × 39 km E-W at 45° N).

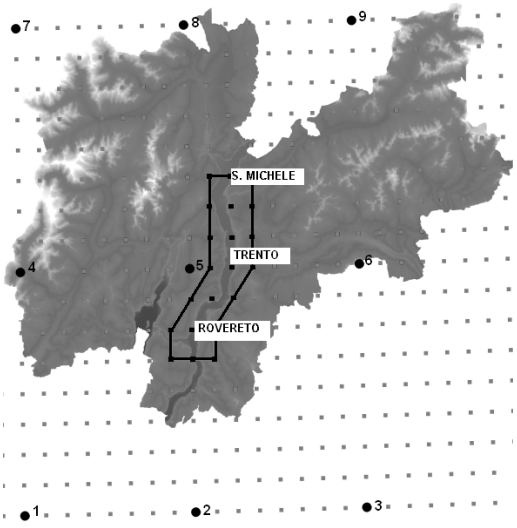

Fig. 2. The ECMWF (large spots, numbered) and LAMI (small spots) grids over the province of Trento. The 21 grid points used for LAMI post-processing fall inside the marked area.

The vertical dimension is discretised into 60 levels. Initial conditions are obtained using the four-dimensional data assimilation scheme "4Dvar" (User Guide to ECMWF products, http://www.ecmwf.int/products/forecasts/guide/). The regional model "Lokal Modell" (LM) is based on the fundamental hydro-thermodynamical equations describing compressible, non-hydrostatic flow in a moist atmosphere. LM does not make any scaling approximations. The equations are written in advective form and solved numerically using the finite difference method, with leapfrog time stepping. The Italian implementation of LM, denoted LAMI (Schättler and Montani, 2005), covers Italy and the surrounding region with a horizontal resolution of 0.0625° (grid size of 7 km) and 35 vertical levels. Since 2003 it has included a continuous assimilation cycle based on the nudging scheme, which includes surface observations, radio soundings, and aircraft reports. As for the physical parameterisation, in 2002 the first-order turbulence scheme was replaced by a new second-order scheme based on turbulent kinetic energy equations. In 2003, the parameterisation of microphysical processes was again updated in order to take into account the ice phase in clouds.

As far as investigation periods are concerned, five years (2001–2005) of ECMWF data were used for the 12:00 run (hereafter ECMWF-12). For the 00:00 run (hereafter ECMWF-00), only the 2004–2005 period was used due to the unavailability of previous outputs. We performed a separate evaluation of the 2004–2005 period for ECMWF-12, for purposes of comparison with ECMWF-00. For LAMI's 00:00 run (LAMI-00), the period of available data started in 2002 and ended in 2005. The NWP output quantities used as

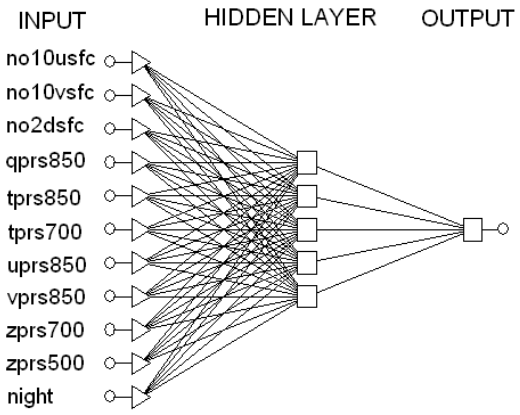

Fig. 3. An MLP network example for the calculation of minimum temperature.

input for the post-processing algorithms are reported in Table 1. The "lead times" for the various runs are +42 h for ECMWF-12, and +30 h for both ECMWF-00 and LAMI-00. Each model thus provides a forecast of hour 06:00 UTC on the day following each forecast issue.

The grid points considered in this analysis (Fig. 2) are close to the central part of Adige Valley. There are 9 grid points for the ECMWF model, and 21 for LAMI.

2.4 MOS algorithms

Post-processing was carried out in four different ways: a simple mean bias correction, a multilinear model, and two machine-learning techniques (a neural network and a random forest). Each of these MOS techniques is described below.

2.4.1 Mean bias correction

We selected the grid points that best predicted temperature for the target area in each model (those with the highest determination coefficient r^2). For the ECMWF grid, point number 5 was chosen as a reference, being the closest to the target area. To obtain an unbiased value, we simply subtracted the average difference between the DMO and measured temperatures from the DMO for this point at 06:00 UTC.

2.4.2 Multilinear regression (MLR)

Multilinear regression techniques are often applied in the post-processing of temperature forecasts from meteorological models. In the simplest case (linear regression), the raw output correction is a function of only one parameter (Woodcock and Southern, 1983; Massie and Rose, 1997). In the more general case, many parameters (predictors) enter into a multilinear model (Sugahara, 2000; Schoof and Pryor, 2001; Casaioli et al., 2003). We applied a backward stepwise multilinear regression, where predictors were selected by the Akaike information criterion (AIC). This selection was implemented using the R package "mass" (Venables and Ripley, 2002). The AIC index is computed as follows:

$$\mathrm{AIC} = 2N - 2n \ln\left(\frac{\mathrm{RSS}}{n}\right) \quad (1)$$

where:
N = number of predictors
n = number of cases (days with prediction and observation)
RSS = sum of squared residuals.

The importance of a single predictor can be determined by comparing this index with the value found after its removal from the pool. Those which produce the greatest change in AIC are retained in the MLR model.

2.4.3 Artificial Neural networks (ANN)

Neural networks have seen extensively use in meteorology over the last decade (Navone and Ceccatto, 1994; Hsieh et al., 1998; Tangang et al., 1998). From their origins as models of human brain function, they have evolved into powerful non-linear statistical models (Bishop, 1995). The most widely used neural network is the "Multilayer Perceptron", or MLP (Rumelhart and McClelland, 1986), whose architecture is shown in Fig. 3. An MLP is formed from layers of individual processing units, which are usually called neurons. Each neuron takes its input from all elements in the previous layer, evaluates a (usually non-linear) function of the inputs, and forwards this result to the next layer. In the simplest case there are only two layers of neurons. The first takes its input values directly from the data, and computes a non-linear transformation. The second layer consists of a single unit, which computes some linear combination of the first layer's outputs. Each connection between two neurons is given a relative weight, which is adjusted during a training phase in order to minimize some measure of error (typically, the mean square error between observed and predicted values). Neural networks are extremely flexible non-linear regressors, but they are prone to overfitting. Even when appropriate methods are implemented to avoid this problem, any error measure calculated on data used in the training stage is expected to be biased. For this reason, only completely independent datasets should be used to evaluate the performance of neural networks. We used a commercial software package (STATISTICA Neural Networks; StatSoft Inc., http://www.statsoft.com) in all the experiments reported in this paper.

2.4.4 Random Forest (RF)

New modelling tools based on ensembles (or groups) of predictors have recently been introduced (Breiman, 1996, 2001; Freund and Schapire, 1995; Granitto et al., 2005; reviews from Ho, 2002, and Tresp, 2001). These have been consistently shown to be more accurate than single-predictor

Table 1. Predictors available to the downscaling algorithms. Some units are modified before the application of post-processing algorithms.

Abbreviation	Variable	ECMWF-12	ECMWF-00	LAMI-00
cpsfc	convective precipitation [m]	×	×	
hccsfc	high sky cover $[0-1]$	×		×
lccsfc	low sky cover $[0-1]$	×		×
lhtflsfc	surface latent heat flux $\left[\mathrm{W\,m^{-2}}\right]$			×
mccsfc	medium sky cover $[0-1]$	×		×
mslsfc	atmospheric pressure at sea level [Pa]	×	×	×
night	night length [h]	×	×	×
nlwrssfc	surface net long wave $\left[\mathrm{W\,m^{-2}}\right]$			×
nswrssfc	surface net short wave $\left[\mathrm{W\,m^{-2}}\right]$			×
no10usfc	zonal wind at 10 m $\left[\mathrm{m\,s^{-1}}\right]$	×	×	×
no10vsfc	meridional wind at 10 m $\left[\mathrm{m\,s^{-1}}\right]$	×	×	×
no2dsfc	dew point at 2 m $\left[^{\circ}\mathrm{C}\right]$	×	×	×
no2tsfc	temperature at 2 m $\left[^{\circ}\mathrm{C}\right]$	×	×	×
qprs500	specific humidity at 500 hPa $\left[\mathrm{kg\,kg^{-1}}\right]$	×	×	
qprs700	specific humidity at 700 hPa $\left[\mathrm{kg\,kg^{-1}}\right]$	×	×	
qprs850	specific humidity at 850 hPa $\left[\mathrm{kg\,kg^{-1}}\right]$	×	×	×
shtflsfc	surface sensible heat flux $\left[\mathrm{W\,m^{-2}}\right]$			×
T.db	temp. at 2 m predicted for the day before [°C]	×	×	×
tccsfc	total sky cover $[0--1]$	×	×	×
tprs500	temperature at 500 hPa [°C]	×	×	
tprs700	temperature at 700 hPa [°C]	×	×	
tprs850	temperature at 850 hPa [°C]	×	×	×
tpsfc	total precipitation [m]	×	×	
tsoildpl	soil temperature [°C]			×
uprs500	zonal wind at 500 hPa $\left[\mathrm{m\,s^{-1}}\right]$	×	×	
uprs700	zonal wind at 700 hPa $\left[\mathrm{m\,s^{-1}}\right]$	×	×	
uprs850	zonal wind at 850 hPa $\left[\mathrm{m\,s^{-1}}\right]$	×	×	×
vprs500	meridional wind at 500 hPa $\left[\mathrm{m\,s^{-1}}\right]$	×	×	
vprs700	meridional wind at 700 hPa $\left[\mathrm{m\,s^{-1}}\right]$	×	×	
vprs850	meridional wind at 850 hPa $\left[\mathrm{m\,s^{-1}}\right]$	×	×	×
wprs700	vertical velocity at 700 hPa $\left[\mathrm{Pa\,s^{-1}}\right]$	×		
zprs500	geopotential height at 500 hPa [m]	×	×	
zprs700	geopotential height at 700 hPa [m]	×	×	
zprs850	geopotential height at 850 hPa [m]	×	×	

models (Bauer and Kohavi, 1999). An ensemble is a set of individual regression models which are combined to solve a single problem. There are many possible ensemble construction strategies; each predictor can belong to a different kind of model (MLR, ANN, etc.), be fitted to a different subset of the full dataset, or even differ only in the initial conditions of the fitting procedure. To produce a prediction for new data, first the inputs are evaluated by each regressor and then the results are combined to form a final decision. Usually, a simple weighted average is used to combine the responses of the ensemble.

Among such ensemble techniques, one of the most successful approaches is based on the well-known statistical principle of "bias-variance tradeoff". According to this principle, any statistical method of prediction with low bias (i.e., an intrinsic capacity to accurately model any distribution) also has high variance (i.e., models fitted to several different samples drawn from the same distribution of data tend to be diverse). Interestingly, if a combined regressor is formed using several models with low bias and high variance, then the overall variance can be reduced. The result is a regression method which on average is more accurate than any of its

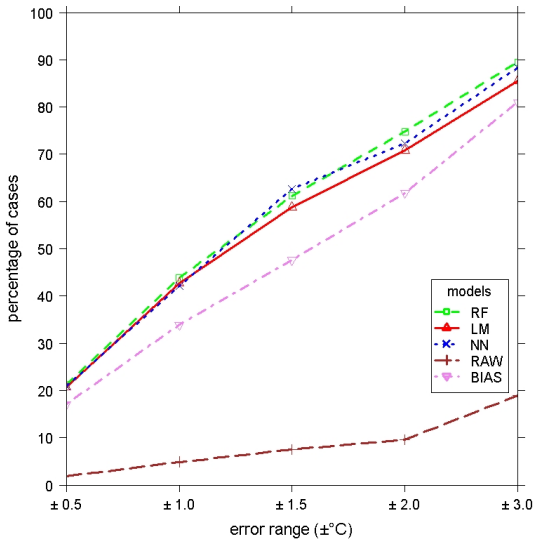

Fig. 4. Distribution of downscaled predictions in five error classes for four different post-processing algorithms. In this figure the algorithms only made use of outputs from the ECMWF grid point closest to the target area (number 5 – see Fig. 2). Legend: RF = Random Forest; LM = Multi-linear model; NN = Artificial Neural Network; RAW = raw (direct) model output; BIAS = bias-corrected model output.

members (Geman et al., 1992). Furthermore, the accuracy of the ensemble grows with the degree of diversity among its individual members. The Random Forest (hereafter RF) method is based on this property.

An RF is formed by growing several regression trees, which individually are very unstable; i.e., a small change in the dataset can result in large changes in the regression model (Breiman, 1996). A regression tree (Breiman et al., 1984) consists of a set of nodes that branch out from a root node. Each node contains a question with several possible answers, each leading either to another node or a "leaf" (a terminal node with an associated prediction). Several automatic, recursive procedures have been developed to build (or grow) regression trees, but most of these are also deterministic. To create diversity in the ensemble, RF fits each tree to a bootstrap replica of the sample data. The bootstrap procedure (Efron and Tibshirani, 1983) creates a sample of the same length as the available dataset, randomly drawn from the original data, with duplication allowed (i.e., each example is picked at random from the full dataset, whether or not it has been picked before). To increase diversity further, only a small random sample of all possible features (predictors) is made available to the fitting algorithm when growing each node of the tree. These two sources of diversity are easy to implement and lead to ensembles with very good prediction performance.

One of the most important features of the RF method is that it limits overfitting, even when the ensemble contains thousands of trees. Its error rate on independent examples converges smoothly to a limiting value as the number of trees approaches infinity. In practice, the RF algorithm has only one free parameter: the number of predictors m made available to each node during tree growth. Breiman (2001) has shown, however, that results are not strongly dependent on this parameter and that the default value of m (the square root of the total number of features) usually gives nearly optimal results. A package implementing RF is available for the statistical software R (Liaw and Wiener, 2005), and is used in this paper.

3 Results

The performance of the post-processing models described above has been assessed in three different aspects:

1. Accuracy; that is, deviation between the predicted minimum temperature and the observed value (hourly average). An error distribution is reported, as the percentage of forecasts with deviations within the following 5 intervals from the measured value: ± 0.5°C, ± 1.0°C, ± 1.5°C, ± 2.0°C and ± 3.0°C.

2. Mean absolute error (MAE); that is, the average of the absolute values of the differences between predicted and measured values.

3. Correlation between predicted and observed temperature, expressed as Pearson's correlation coefficient (r).

In order to compare different downscaling approaches, each post-processing algorithm has been compared to NWP output. In general, the site-specific, unbiased value (see Sect. 2.4.1) for grid point number 5 (which is closest to the target area) can be thought of as a term of comparison to quantitatively assess the improvement achieved with each technique.

3.1 Comparison of post-processing methods

These are the results of the four post-processing algorithms, as applied to one single grid point (number 5) of the ECMWF-00 prediction. The goal of this preliminary analysis is simply to identify the downscaling model with the best overall performance. The most complete version of the algorithm considers all grid points that have the potential to influence the temperature forecast. The results are summarized in Fig. 4, which reports the distribution of errors in classes, and in Table 2, which includes additional statistics. A remarkable improvement is attained with a simple mean bias correction determined from a single, well correlated grid point. Nevertheless, further improvement can be obtained with the multi-parameter models, even though the difference between

Table 2. Performances of different post-processing models, applied to ECMWF-12 (2001-2005). MAE = mean absolute error. r = Pearson's correlation coefficient. RF = random forest. LM = multi-linear model. NN = neural network. DMO = Direct (raw) Model Output. BIAS = mean-bias-corrected output. RF - BIAS = RF improvement with respect to the performance of BIAS.

	RF	LM	NN	DMO	BIAS	RF - BIAS
Error distribution in classes $[\pm°C]$						
0.5	21.3	20.7	20.9	1.9	17.1	4.2
1.0	43.9	42.7	42.1	4.9	33.9	10.0
2.0	74.8	70.8	72.3	9.6	61.8	13.0
3.0	89.5	85.5	88.3	19.0	81.0	8.5
MAE [°C]	1.87	2.10	1.88	6.94	2.47	0.60
r	0.971	0.964	0.972	0.949	0.949	0.022

the three approaches tested is often negligible. The degree of improvement, measured in terms of accuracy, is on the order of a few percent in each error class. RF post-processing gives an average improvement of 9.9 percentage points over every error class compared to the simple average bias correction (column "BIAS" in Table 2), with a maximum improvement of 13 percentage points in the error class ± 2.0°C. The mean absolute error (MAE) is reduced from 2.47°C to 1.87°C by RF post-processing, and the correlation coefficient (r) rises from 0.95 to 0.97.

3.2 Choice of predictors for multi-parameter models

The previous section assessed the performance of models when applied to a single grid point. Because all the grid points surrounding the target area are potentially useful, a subset of the most useful predictors must be selected. The R implementation of RF method is particularly robust in selecting the most influential variables. Since it also yielded the best results on a single grid point, for the RF all grid points are included in the pool of potentially influencing variables. All the following analysis refers now to the RF algorithm. We considered the following set of predictors (divided into four categories): i) All of the NWP output variables evaluated at 06:00 UTC (27 for ECMWF-12, 24 for ECMWF-00, 19 for LAMI-00), for each grid point (9 for ECMWF, 21 for LAMI). ii) The average temperature from 05 to 06 UTC on the previous day (i.e., the average of the three temperatures recorded at each meteorological station). iii) The temperature prediction errors at each grid point, on each of the three previous days for each grid point. iv) The length of the night.

In total, there were 272 possible predictors for ECMWF-12, 245 for ECMWF-00, and 464 for LAMI. The inclusion of forecast errors into the predictor space means that we also considered the recent performance history of the forecast to be influential.

To evaluate the relative influence of these variables, a sensitivity analysis has been carried out on the predictors of grid point number 5. The results of this analysis for other grid points (not given here), were not particularly different. Figures 5 and 6 (for the single-point and nine-point models, respectively) show the relative importance of each variable, ordered according to their influence on the result. As expected, the temperature forecast at 06:00 UTC (no2tsfc) is the most important predictor in both cases. Other especially important variables include the night length, meridional (parallel to the main valley) wind velocity, 2 m dew point temperature, and temperature and humidity in the lowest layer (850 hPa). Remember that these variables are not those that most directly affect the minimum temperature, but rather those that best correct the systematic errors of the model. When the variables related to all grid points are considered together, it can be seen (Fig. 6) that the temperature prognosis of other grid

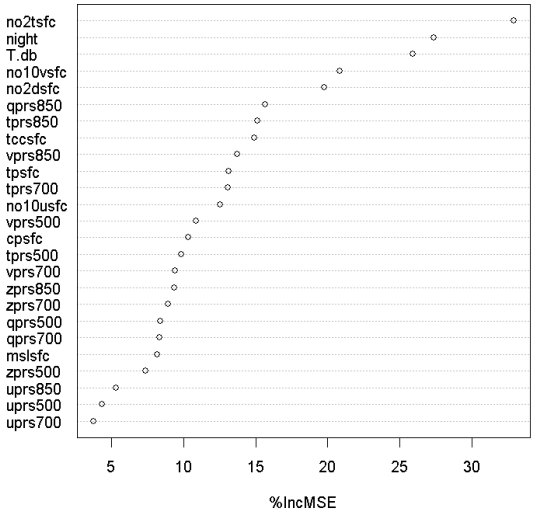

Fig. 5. Sensitivity analysis for the RF model. Single point model (using data only from grid point number 5). The x-axis represents the normalised increase in Mean Standard Error when the variable is removed from the pool of model inputs. For the meaning of the abbreviations, see Table 1.

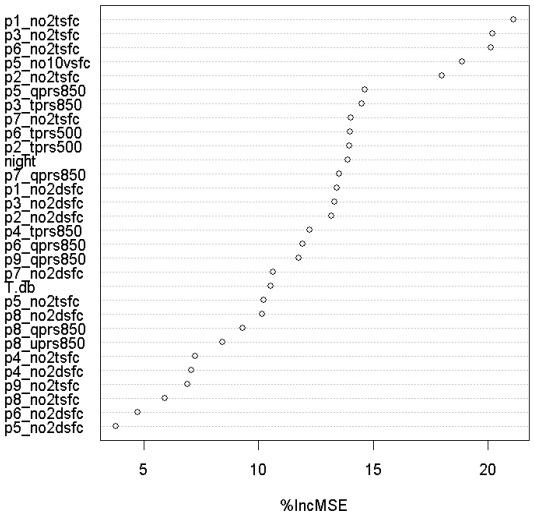

Fig. 6. Same as Fig. 5, for the nine-point model.

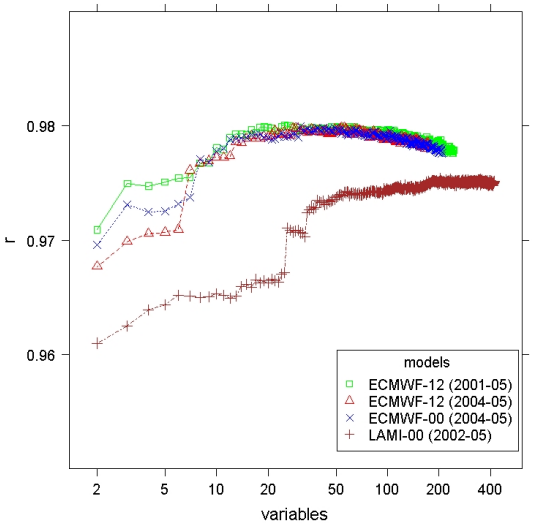

Fig. 7. Correlation coefficient between predicted and observed values, as a function of the number of predictors used (RF algorithm).

points at 06:00 UTC is also important. In other words, many of the grid points surrounding the target area independently contribute to improving the temperature forecast in the Adige valley. More space will be devoted to this issue in Sect. 4.

3.3 Comparison among meteorological models

The four different NWPs (model and run time) are now compared after RF post-processing. Table 3 reports their statistical performance measures. The effect of an increase in the number of predictors can be appraised in Figs. 7 and 8, which, respectively, show the correlation coefficient r and the "out-of-bag" (OOB) error as a function of the number of predictor variables. The OOB error is an internal error estimator of the RF, measured on an independent test set, which remains "out of the bag" of data used in the forest development. The OOB error is the standard prediction error, referred to the measured values, when the algorithm is applied to the OOB data set.

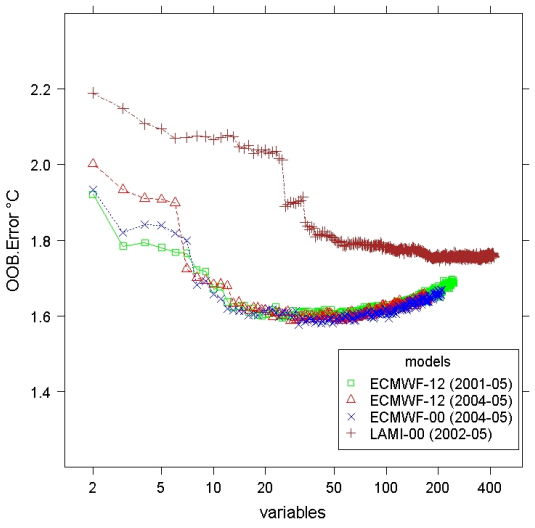

Fig. 8. "Out-of-bag" error between predicted and observed values, as a function of the number of predictors used (RF algorithm).

In general, there is remarkable improvement when the number of variables increases from a few to a few dozen, but no further improvement is observed beyond this point. Rather, when more than 100 variables are used there is a slight decrease in performance. This behaviour is typical of statistical modelling, where a large number of input variables gives the model increased flexibility but usually results in some degree of overfitting. The variable selection process itself can also produce overfitting (Ambroise and McLachlan, 2002), but this effect would be the same for all NWP models and thus not change their relative performance.

When the number of variables is sufficiently high (as in operational applications), there is no appreciable difference among the three ECMWF runs. The ECMWF-00 model, for which only two years of data were available, apparently yields results just as good as the five-year ECMWF-12 model. Nevertheless, in the case of LAMI, attention must be paid to the different meaning of the total number of predictors; in LAMI it is much higher because there are more grid points (21, compared to ECMWF's 9) and each of them is multiplied by the number of meteorological predictors; for this reason, the total number of predictors in LAMI model cannot be compared to the corresponding number for ECMWF. Downscaling with LAMI performs worse than with any ECMWF run, considering the results in terms of OOB error and the other evaluation parameters.

4 Discussion

The outcome of this investigation confirms the advantage of MOS analysis on the DMO. All the methods tested here clearly improve the raw NWP predictions. The best MOS approach (a nine-point RF) yields a mean average error (MAE) of 1.2°C. This is similar to results found by other authors using non-linear methods (Marzban, 2003; Casaioli et al., 2003; Boi, 2004). Nevertheless, it must be recognised that the rather high correlation (r=0.95) between the DMO and measured values is itself a good starting point for designing a MOS procedure. On the other hand, the literature has shown that the advantage of non-linear post-processing algorithms is not universal. Non-linear methods have produced only questionable results, for example, in the prediction of sea surface temperatures (SST) from sea level pressure and SST time series (Tang et al., 2000). The relative advantages of different non-linear methods have also been investigated by Trigo and Palutikof (1999) and Miksovsky and Raidl (2005). These authors also found little difference in the performance of various methods, also in the case of temperature post-processing.

In the context of NWPs, it is useful to analyse which variables can be considered most responsible for the systematic error in the raw output. Indeed, these are the only sources of error that can be reduced by statistical post-processing. A mean daily temperature downscaling can be successfully carried out in geographically regular areas, generally with few variables (Huth, 2002). For minimum temperatures in an alpine region, however, there are several quantities that affect the dynamics of cooling and downvalley air flow. Atmospheric temperatures were among the most influential factors in the NWP prediction, mainly at the 850, 700, and 500 hPa atmospheric layers. The latter layers may seem too high to directly affect temperature at the ground when its elevation is limited to a few hundred meters. However, these temperature data are made available to the algorithm to model the temperature lapse rate in the free atmosphere, which is a measure of atmospheric stability and therefore of the ability of the lower atmosphere to set up a stable layer close to ground and to foster thermal inversion. Indeed, the intrinsic difference in elevation between the topographic models inside NWPs and the real altitude of corresponding sites, is a major source of errors due to the application of an inaccurate lapse rate to the simulated atmosphere. The bigger the error in lapse rate, and the difference in elevation, the larger this kind of error.

The role of wind is outstanding in temperature prediction, especially at night, preventing the formation of a shallow thermal inversion layer close to ground when its speed exceeds some very low threshold value. As for the wind speed at ground level, it can be seen that the meridional component is more important than the zonal component. This can be easily explained by the north-south orientation of Adige Valley, which strongly inhibits transverse air flow. Hence, the predicted (large-scale) zonal circulation is less correlated with the actual flow in the valley than the meridional circulation. In other words, a better parameterisation of wind in the valley increases the quality of temperature prediction, and the better predictability, in the case of a north-south oriented valley, especially comes from the meridional component of wind velocity.

Cloudiness, which normally plays an important role in the radiative balance at ground level, has a significant effect only for total cover. It is not, however, in a position of great importance. This would show ECMWF's good parameterisation for the process of backward reflection of longwave radiation from clouds, making any systematic errors related to the degree of sky cover negligible.

The night length must be considered separately; it is not a meteorological variable, yet it occupies a high position in the ranks of predictor importance. This can be probably explained by considering the basic distinction in temperatures recorded before and after sunrise. For part of the year hour 06:00 UTC (corresponding to 07:00 local time) occurs after sunrise, when the temperature has already experienced a sharp rise; for the rest of the year, however, particularly in winter, at that time temperatures are still close to the minimum night values. Night length is thus capable of triggering a discontinuity along the year in the modelling of early morning temperature evolution.

The prediction errors of the previous three days also are not meteorological in nature, but are potentially useful predictors. Such variables could account for long-term biases in the prediction, correcting systematic offsets. Though such sources of error often exist in NWP outputs, these were not among the most influential predictors as determined by our sensitivity analysis (i.e., the ranks of "importance" in the RF method as shown in Figs. 5 and 6). The reason is to be sought in the procedure of the set-up of the "random forest": the choice of training data sets in the RF is made at random, as is the choice of out-of-bag data. This random selection breaks the continuity of the time series, preventing the identification of temporary, self-consistent biases. The prediction errors of previous days thus leave no trace on the RF algorithm, which is designed to work with general data. This is perhaps the most important shortcoming in a MOS approach. Kalman filtering, applied post-downscaling, could probably make up for the inability of "static" post-processing models – as are all those considered in this research – in coping with time-related biases.

Having applied the most reliable method (RF) to four different NWP outputs, we have observed that there is no appreciable improvement in the ECMWF-00 model compared to the ECMWF-12 model. Moreover, applying the RF correction is less effective on the LAMI-00 data than on the ECMWF data. This is in spite of the former model's higher spatial resolution, which enhances the representation of geographic features. One possible explanation is that even though the 21 grid points considered more closely follow the geography of Adige Valley, they cover a much smaller area

Table 3. The performance of RF on different NWP models. MAE = mean absolute error. r = Pearson's correlation coefficient.

	ECMWF-12 2001–2005	ECMWF-12 2004–2005	ECMWF-00 2004–2005	LAMI-00 2002–2005
Error distribution in classes $[\pm°C]$				
0.5	29.4	27.6	31.3	24.0
1.0	52.1	50.4	53.5	46.9
1.5	68.9	69.9	71.9	64.8
2.0	81.3	81.2	81.8	76.8
3.0	94.1	93.6	93.1	90.9
MAE [°C]	1.20	1.22	1.17	1.36
r	0.980	0.980	0.980	0.975

than the nine grid points of the ECMWF. In other words, information at more than one grid point might be partially redundant, while the true improvement could come from the increase in the real spatial domain. This could enable the downscaling models to capture atmospheric features that typically show at the β-meso scale (20 to 200 km). For example, the pressure gradient in the direction W-E (linked to northerly winds in the Adige Valley) cannot be satisfactorily represented when the horizontal domain is too narrow, as is the case of the small target area of LAMI grid. The related, potentially undersampled phenomena include, e.g., local patterns of wind field and sky clearing. For the same reason, the smaller domain may also decrease the downscaling models resilience to NWP-based errors in the time development of phenomena.

5 Conclusions

NWP output is generally available on a fixed grid rather than at given desired sites. This work has demonstrated the advantages of post-processing NWP data by successfully downscaling the DMO of two models (ECMWF and LAMI) for the minimum temperature. The general purpose of this application was forecasting the minimum spring temperature within a frost-prone region of agricultural interest (the middle Adige Valley, in the Italian Alps). Several different approaches were compared: a simple correction by the mean bias, multilinear regression analysis, and two machine-learning methodologies (a neural network and the "Random Forest" method).

Results show that the model output's accuracy improves after downscaling, particularly in non-linear models. The algorithm that yielded the best results (Random Forest) can be easily automated to process the model output and produce an improved minimum temperature prediction for the following day. Such an application has been working on a server in IASMA since spring 2006. It uses data from both ECMWF-12 and ECMWF-00, and the results are made available daily for end users on the IASMA website.

The residual error (expressed as MAE) is probably as low as possible after downscaling, yet it is still greater than 1°C (1.2°C for the nine-point RF algorithm). The very good agreement among multi-parameter algorithms (especially the non-linear ones) shows that there is a technical limit to the improvement that MOS methods can obtain. Post-processing algorithms can calibrate and correct systematic errors, provided that relationships exist between these errors and output variables. It is not, however, effective in reducing errors with some other origin. This would apply, for example, to errors in the prediction of night sky cloud cover or wind speed. Failures in such model outputs can obviously be ascribed to unsatisfactory knowledge of initial atmospheric conditions, to the discretisation of the atmospheric, or even to mathematical simplifications in the physics of atmospheric processes. Such causes generate errors that are only detectable a posteriori, and thus cannot be classified and corrected within post-processing algorithms. It is likely that a good fraction of the total error in minimum temperature predictions can be ascribed to such reasons. Further improvements in forecasting thus have to be sought in the NWP itself, rather than in complex statistical downscaling.

Finally, it is worth mentioning that the present ECMWF model release, which has been available since January 2006 (T799 L91), has doubled its resolution with respect to the previous one (0.25° instead of 0.50°). Given the limited period available, it was not possible to formulate algorithms on data from this latest release. Nevertheless, the RF model can be applied to the outputs of both ECMWF-12 and ECMWF-00 in the present high-resolution releases since the position of grid points in the previous version remains unchanged. A few years from now, when the model output archives are long enough, it will be possible to repeat this process and compare the results to those obtained with older releases. This will allow us to assess any NWP-related improvements in the minimum temperature prediction.

Acknowledgements. This work was performed as part of the "GEPRI" project, funded by the Autonomous Administration of the Province of Trento (PAT). Thanks to F. Zottele for his help.

Edited by: A. Provenzale
Reviewed by: two referees

References

Abdel-Aal, R. E. and Elhadidy, M. A.: A machine-learning approach to modelling and forecasting the minimum temperature at Dharan, Saudi Arabia, Energy, 19(7), 739–749, 1994.

Anadreanistakis, M., Lagouvardos, K., Kotroni, V., and Elefteriadis, H.: Correcting temperature and humidity forecasts using Kalman filtering: Potential for agricultural protection in Northern Greece, Atmos. Res., 71(3), 115–125, 2004.

André, J. C. and Mahrt, L.: The nocturnal surface inversion and influence of clear-air radiative cooling, J. Atmos. Sci., 39, 864–878, 1982.

Arca, B., Benincasa, F., De Vincenzi, M., and Fasano, G.: A neural model to predict the daily minimum of air temperature, 7th International Congress For Computer Technology in Agriculture, Computer technology in agricultural management and risk prevention. Florence, Italy, 15–18 November 1998, Proceedings, 485–493, 1998.

Basili, P., Bonafoni, S., and Biondi, R.: Analisi e previsione di temperature minime e di gelate sul bacino del Trasimeno, Rivista Italiana di Agrometeorologia, 2006(1), 46–50, 2006.

Bauer, E. and Kohavi, R.: An empirical comparison of voting classification algorithms: Bagging, Boosting, and variants, Machine Learning, 36(1–2), 105–139, 1999.

Bishop, C. M.: Neural Networks for Pattern Recognition. Oxford University Press, 475 pp., 1995.

Boi, P.: A statistical method for forecasting extreme daily temperature using ECMWF 2-m temperatures and ground station measurements, Meteorol. Appl., 11, 245–251, 2004.

Breiman, L.: Bagging predictors, Machine Learning, 26(2), 123–140, 1996.

Breiman, L.: Random Forests, Machine Learning, 45(1), 5–32, 2001.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J.: Classification and regression trees, Belmont, Wadsworth, 368 pp., 1984.

Cane, D., Milelli, M., and Gandini, D.: Improvement of the meteorological parameters forecasts for the XX Olympic winter games venue, Geophys. Res. Abstr., 6, 03637, 2004.

Carlson M. A. and Stull, R. B.: Subsidence in the nocturnal boundary layer, J. Clim. Appl. Meteorol., 25, 1088–1099, 1986.

Casaioli, M., Mantovani, R., Proietti Scorzoni, F., Puca, S., Speranza, A., and Tirozzi, B.: Linear and nonlinear post-processing of numerically forecasted surface temperature, Nonlin. Processes Geophys., 10, 373–383, 2003, http://www.nonlin-processes-geophys.net/10/373/2003/.

Efron, B. and Tibshirani, R. J.: An introduction to the bootstrap, New York: Chapman & Hall, 456 pp., 1983.

Freund, Y. and Schapire, R. E.: Experiments with a New Boosting Algorithm, in: Proceedings of the Thirteenth International Conference on Machine Learning, edited by: Kaufmann, M., 148–156, 1996.

Galanis, G. and Anadreanistakis, M.: A one-dimensional Kalman filter for the correction of near surface temperature forecasts, Meteorol. Appl., 9(4), 437–441, 2002.

Gassmann, M. I. and Mazzeo, N. A.: Nocturnal stable boundary layer height model and ist application, Atmos. Res., 57(4), 247–259, 2001.

Geman, S., Bienenstock, E., and Doursat, R.: Neural Networks and the Bias/Variance Dilemma, Neural Computation, 4, 1–58, 1992.

Granitto, P., Verdes, P., and Ceccatto, H. A.: Neural Networks Ensembles: Evaluation of Aggregation algorithms, Artificial Intelligence, 163, 139–162, 2005.

Ho, T. K.: Multiple Classifier Combination: Lessons and Next Steps, in: Hybrid Methods in Pattern Recognition, edited by: Kandel, A. and Bunke, H., World Scientific, 171–198, 2002.

Homleid, M.: Diurnal corrections of short-term surface temperature forecasts using the Kalman filter, Weather and Forecasting, 10(4), 689–707, 1995.

Hsieh, W. W. and Tang, B.: Applying Neural Network Models to Prediction and Data Analysis in Meteorology and Oceanography, Bull. Am. Meteorol. Soc., 79(9), 1855–1870, 1998.

Huth, R.: Statistical downscaling of daily temperature in Central Europe, J. Climate, 15, 1731–1742, 2002.

Huth, R.: Sensitivity of local daily temperature change estimates to the selections of downscaling models and predictors, J. Climate, 17, 640–652, 2004.

Liaw, A. and Wiener, M.: Breiman and Cutler's random forests for classification and regression, R-Package "randomForest": http://stat-www.berkeley.edu/users/breiman/RandomForests, 2005.

Marzban, C.: Neural networks for postprocessing model output: ARPS, Mon. Wea. Rev., 131, 1103–1111, 2003.

Massie, D. R. and Rose, M. A.: Predicting daily maximum temperatures using linear regression and Eta geopotential thickness forecasts, Weather and Forecasting, 12(4), 799–807, 1997.

Miksovsky, J. and Raidl, A.: Testing the performance of three nonlinear methods of time series analysis for prediction and downscaling of European daily temperatures, Nonlin. Proceses Geophys., 12, 979–991, 2005.

Navone, H. D. and Ceccatto, H. A.: Predicting Indian monsoon rainfall: a neural network approach, Clim. Dyn., 10(6–7), 305–312, 1994.

Robinson, C. and Mort, C.: A neural network system for the protection of citrus crop from frost damage, Computers and Electronics in Agriculture, 16, 177–187, 1997.

Schizas, C. N., Michaelides, S., Pattichis, C. S., and Livesay, R. R.: Artificial networks in forecasting minimum temperature, Institution of Electrical Engineers, Publ. No. 349, 112–114, 1991.

Schoof, J. T. and Pryor, S. C.: Downscaling temperature and precipitation: A comparison of regression-based methods and artificial neural networks, Int. J. Climatol., 21, 773–790, 2001.

Schättler U. and Montani, A. (Eds.): Operational Implementations, COSMO Newsletter No.5, Chapter 4. DWD, Offenbach am Main, Germany, available at http://www.cosmo-model.org/, 2005.

Sugahara, S.: Uma experiência com modelo estatístico (MOS) para a previsão da temperatura mínima diária do ar, Brazilian Journal of Geophysics, 18(1), 3–12, 2000.

Tang, B., Hsieh, W. W., Monahan, A. H., and Tangang, F.: Skill comparisons between Neural Networks and Canonical Correlation Analysis in predicting the equatorial Pacific sea surface tem-

perature, J. Climate, 13, 287–293, 2000.
Tangang, F. T., Tang, B., Monahan, A. H., and Hsieh, W. W.: Forecasting ENSO Events: A Neural NetworkExtended EOF Approach, J. Climate, 11(1), 29–41, 1998.
Tresp, V.: Committee machines, in: Handbook for Neural Network Signal Processing, edited by: Hu, Y. H. and Hwang, J. N., CRC Press, pp. 5.1–5.14, 2001.
Trigo, R. M. and Palutikof, J. P.: Simulation of daily temperatures for climate change scenarios over Portugal: a neural network model approach, Clim. Res., 13, 45–59, 1999.
Venables, W. N. and Ripley, B. D.: Modern Applied Statistics with S, Fourth Edition, Springer, New York, 495 pp., 2002.
Verdes, P. F., Granitto, P. M., Navone, H. D., and Ceccatto, H. A.: Frost Prediction with Machine Learning Techniques, Proceedings of the VI Argentine Congress on Computer Science, pp. 1423–1433, 2000.
Weichert, A. and Bürger, G.: Linear versus nonlinear techniques in downscaling, Clim. Res., 10, 83–93, 1998.
Wilks, D. S.: Statistical methods in the atmospheric sciences, Academic Press, 467 pp., 1995.
Wilson, L.: Statistical interpretation methods applied to ensemble forecast, Proceedings of the WMO Workshop on Ensemble Prediction, Beijing, 2001.
Woodcock, F. and Sputhern, B.: The use of linear regression to improve official temperature forecasts (Adelaide Brisbane Canberra Hobart Melbourne Perth Sydney), Australian Meteorological Magazine, 31(1), 57–62, 1983.