Version 03/10/2005, submitted to Atmospheric Research

Sounding–Derived Indices for Neural Network Based Short–Term Thunderstorm and Rainfall Forecasts

Agostino Manzato

agostino.manzato@osmer.fvg.it

Keywords: Neural Network, sounding, thunderstorm, rain.

OSMER – Osservatorio Meteorologico Regionale dell'ARPA Friuli Venezia Giulia

Via Oberdan 18, I–33040 Visco (UD), Italy, tel. +39 0432 934163

Abstract

A neural network based scheme to do a multivariate analysis for forecasting the occurrence and intensity of a meteo event is presented. Many sounding–derived indices are combined together to build a short–term forecast of thunderstorm and rainfall events, in the plain of the Friuli Venezia Giulia region (hereafter FVG, NE Italy).

For thunderstorm forecasting, sounding, lightning strikes and mesonet station data (rain and wind) from April to November of the years 1995–2002 have been used to train and validate the artificial neural network (hereafter ANN), while the 2003 and 2004 data have been used as an independent test sample. Two kind of ANNs have been developed: the first is a "classification model" ANN and is built for forecasting the thunderstorm occurrence. If this first ANN predicts convective activity, then a second ANN, built as a "regression model", is used for forecasting the thunderstorm intensity, as defined in a previous article.

The classification performances are evaluated with the ROC diagram and some indices derived from the Table of Contingency (like KSS, FAR, Odds Ratio). The regression performances are evaluated using the Mean Square Error and the linear cross correlation coefficient R.

A similar approach is applied to the problem of 6 h rainfall forecast in the Friuli Venezia Giulia plain, but in this second case the data cover the period from 1992 to 2004. Also the forecasts of binary events (defined as the occurrence of 5, 20 or 40 mm of maximum rain), made by classification and regression ANNs, were compared. Particular emphasis is given to the sounding–derived indices which are chosen in the first places by the predictor forward selection algorithm.

1 Introduction

The artificial neural networks have been studied since the nineteen sixties (Rosenblatt 1958), but their use for forecasting meteorological events appeared only in the last 15 years. Some "early" references regarding different meteo events could be: Lee et al. (1990) for cloud classification, Frankel et al. (1990) for lightning, Schizas et al. (1991) for minimum temperature, McCann (1992) for thunderstorms, French et al. (1992) or Allen and Le Marshall (1994) for rain, Pasini and Potestà (1995) for visibility and fog, Xiao and Chandrasekar (1996) for snow, Marzban and Stumpf (1996) for tornado, Pankiewicz (1997) for satellite convective cells, Marzban and Witt (2001) for hail. General references to ANN are Masters (1993), Bishop (1996) and Marzban (2002).

The greater advantage in using ANN is their intrinsic non–linearity, which helps in describing complex meteorological events in a better way than linear methods. But this could turn out to be also a drawback, since this intrinsic power permits the ANN to easily fit the database used to train the model. Unfortunately, it is not sure that the good performance obtained by the ANN on the training data will be confirmed also on the new data (generalization ability). To avoid this *overfitting* problem it is crucial to *verify* the ANN, i. e. to divide the original database into training and validation subsets and choose the ANN which has the best performance on the validation dataset (similarly to what done by Navone and Ceccato 1994)[1].

In a previous article (Manzato 2005) a technique to avoid overfitting based on *repeated holdout* is proposed. That means that the database is divided into training and validation subsamples more times, and only the *mean* training and validation errors are considered. In the same article also an algorithm (*forward selection*) to choose a good set of sounding–derived indices as thunderstorm predictors (ANN inputs) was proposed. These two techniques are applied also in the present work. The main differences between this article and the previous one are:

[1]That could lead to overfit the validation dataset, indeed the best way should be to optimize both the training and validation errors. See also Cheng and Titterington 1994.

1. the number of repetitions (bootstrap trials) is 12 instead of 8, while the validation sample is 1/4 of the database (instead of 1/3);
2. the number of random initializations of each ANN has been decreased from 150 to only 50 (so each cycle has $12 \cdot 50 = 600$ runs);
3. the database for thunderstorms has been extended to include the months of October and November;
4. a case is considered as a thunderstorm occurrence if there are at least *three* cloud–to–ground (hereafter C2G) lightning strikes in the target area and not only one, as was before;
5. the same model has also been applied to the rain field (both as classification and as regression model).

In particular point 4 has been included because the forecasters of OSMER have observed some cases with only one lightning without any thunderstorm evidence (even no clouds seen by the OSMER radar located at Fossalon di Grado, GO). So, the one or two C2G lightning in 6 h cases are now not considered as thunderstorm occurrences.

2 Target area and database

The work titled "The 6 Hours Climatology of Thunderstorms and Rainfalls in the Friuli Venezia Giulia Plain" (this same issue, hereafter called CTR) describes in detail the target area (the FVG plain) and the data, including the observation facilities used to collect them. Here it is recalled only that the "cases" studied are the 6 h–long periods starting at the launching time of the Udine–Campoformido radiosounding (WMO code 16044). In fact, the Udine sounding every 6 h has been taken to describe the initial conditions in which the event may develop and the resulting forecast is valid for the subsequent 6 h period.

The thunderstorm database includes all the 6 h periods from April to November of the years 1995–2004 (9210 cases). If there are at least three C2G lightning strikes the case is considered "thundery" (16.3% out of the total). The thunderstorm intensity is then evaluated using the "Calculated Convective Activity in 6 hours" (CALCA6h, Manzato

2003), which takes into account not only the total number of C2G lightning strikes but also the rain and wind observations made by 15 OSMER–ARPA synoptic stations. Figure 3a of CTR shows also the CALCA6h distribution for all the thundery cases.

The rainfall database includes all the 6 h periods of the years 1992–2004 for which the sounding is not missing (17352 cases) and is based only on the maximum rain measured by any of the 15 OSMER–ARPA synoptic stations. The frequency of the "rainy" cases is 23.9% based on a threshold of at least 1 mm in 6 h and becomes 14.3% if the threshold is raised to 5 mm in 6 h. Figure 3b of CTR shows the distribution of MaxRain for all the cases over 1 mm.

3 Thunderstorm classification

The thunderstorm database has been divided into a "total" dataset (1995–2002) and a "test" dataset (2003 and 2004 data). The total dataset has been randomly divided into "training" (3/4) and "validation" subsamples (1/4). The process of dividing differently into training and validation has been repeated in 12 times (bootstrap trials), to obtain a robust statistic.

The predictand variable is the binary thunderstorm occurrence, defined as the presence of at least 3 lightning strikes in the FVG plain during the 6 h period. The predictors investigated are listed in table 1. They are 55 variables: excluding those labeled as *"only rain"*[2], there are 38 sounding indices computed by the SOUND_ANALYS.PY program (Manzato and Morgan 2003), the 6 h difference for 14 of them (not listed), the julian date (JJJ) and hour (HH) of the sounding and, lastly, the activity (thunderstorm occurrence) during the 6 h before (ACTP). With respect to Manzato (2005), two new predictors have been added in this work: the SWISS index (Huntrieser et al. 1997) and the Shear in the lowest 3 km (Shear3).

[2]All the variables labeled as "only rain", except the new SST, already proved to be useless in the thunderstorm multivariate analysis of Manzato (2005).

Since many of these predictors are correlated with each other, they can't all be used together. For this reason, a *forward selection algorithm* has been implemented to choose a proper set of predictors. First is chosen the variable that gives the best fit when taken alone. In the following cycle the one which gives the best fit, when used together with this variable, is added. That is repeated adding a new variable to the set of the already selected ones, as long as the system skill still increases.

An ANN has been used to fit the data. Figure 1 shows a schematic ANN with 4 inputs, one hidden layer with 2 *hidden neurons* and 1 output. The output of a feed–forward ANN with only one hidden layer is computed by the following equations (e. g. Venables and Ripley 2002):

$$y = f_o\left(\beta_o + \sum_{j=1}^{H} \omega_j \cdot h_j\right), \qquad \text{with} \qquad h_j = f_j\left(\beta_j + \sum_{i=1}^{I} \omega_{ji} \cdot x_i\right), \tag{1}$$

where I is the total number of the x_i predictors, H is the number of hidden neurons, ω is the "gain" of each neuron, while β is its "bias" and finally f are the *activation functions*, which here are taken to be the *logistic functions*:

$$f(z) = \frac{1}{1 + e^{-z}} \ . \tag{2}$$

During the forward selection algorithm the number of neurons in the hidden layer has been fixed to $H = \text{Int}\left(\frac{I-1}{2}\right) + 1$, based on previous experience.

The training sample is used to find the ANN coefficients ω and β, using the **R** module called `nnet` (Venables and Ripley 2002). The initial weights are given randomly, and the software tries to optimize them, minimizing an error function. Usually in classification problems this error is taken to be the *cross-entropy error* (CEE):

$$CEE = -\frac{1}{N} \cdot \sum_{n=1}^{N} [t_n \cdot \ln(y_n) + (1 - t_n) \cdot \ln(1 - y_n)], \tag{3}$$

where y_n is the ANN output, t_n is the observed occurrence/non–occurrence target (respectively 1 or 0) and N is the total number of cases used. The cross-entropy error computed on the training sample is then called "training error" (TE). With the weights found, the same CEE is then computed on the validation subsample to estimate the "validation error"

(VE). The weights optimization on the training sample and the verification on the validation sample are repeated for *50* different initial weight sets, while only the ANN with the *lowest VE* is taken, to prevent the training sample overfitting.

Figure 2 shows an example of the 12 TE-VE couples (called TV points) for the first 9 variables selected. The 12 points set on the top–right are obtained using only 1 variable, while those at bottom–left are obtained using together all the variables already selected. At each cycle, the 12 TV points seem to lie along a straight line (with slope of about -2.7), as though lower TE *must* correspond to larger VE and vice versa. A rough explanation of this feature is possible considering that there is a small subset of "bad data", which produce large errors, because they are noise *or* simply because they are very difficult to forecast (like the outliers). If during the training/validation division many of these data are assigned to the training sample, then the ANN will obtain larger TE and lower VE, and vice versa. Of course the presence of some of these "bad" data in the small verification sample can change VE a lot. On the other hand, in the training sample they are a minority, and that explains why TE changes less than VE in the 12 bootstraps, i.e. the points slope. For example, in Manzato (2005) the validation sample was larger (1/3 instead of 1/4) and the slope of the line fitting the bootstrap points was lower (-1.6 instead of -2.7).

Figure 2 says that the most important variable taken alone is the difference of temperature at 500 hPa, which is a classical measure of *potential instability*. DT500 is a refined version of the classical Lifted Index (Galway 1956), which uses a different initial parcel. In some conditions (e. g. without the months of October and November) DT500 is replaced by DTC, which is very similar, but measures the temperature difference at a fixed level of parcel temperature (-15 °C) instead of at a fixed pressure (500 hPa). The best variable to be used together with DT500 is the 6 h–before activity (ACTP), which takes into account the *persistence effect*. After that, follows the maximum CAP strength (defined as the maximum increase with height of the saturated equivalent potential temperature, Θ_{es}), which is a good measure of *convection inhibition*. Next follow two variables related to the *water vapour field*: the mean North–South component of the vapour density flux (VFlux) and the mean relative humidity in the first 500 hPa (MRH).

With MRH the VE starts to decrease less than the TE, because the mean values (unfilled squares) no longer lie on the bisector line. Also, the points start to become more scattered, while the mean errors are no longer spaced regularly as before. That happens because, at this point, there is a greater improvement adding a new hidden neuron (the number to the right of the variable label) than adding a new variable. That is why MLWv (the North–South mean wind in the first 6 km) shows a greater improvement than the "6 h period of the day thunderstorm probability", P(HH), or of the "standard deviation of the balloon vertical velocity between 1 and 6 km of height", VVstd. After that the ninth variable (the bulk Richardson number, BRI) is chosen, the mean VE doesn't decrease anymore. That means that the ANN is starting to saturate or to overfit the training data.

It is also evident that the best way to proceed should be to optimize *contemporaneously* the choice of the I inputs and the number of hidden neurons, H. That means to try all the possible sets of different inputs, each one with many different H. Unfortunately, that is too time consuming (computationally). What has been done here is to use the first 9 variables chosen by the previous forward selection algorithm and then try all the hidden neuron numbers in the range $1 \leq H \leq 8$. Again, that is repeated for 12 bootstraps and the lowest *mean VE* is obtained for H = 6. Finally, to choose the "best model", the total (1995–2002) database has been fitted with an ANN with **9 inputs** and **6 hidden neurons**, while the independent test (2003–2004) database has been used to verify the performance. The *total error* obtained is 0.247 on 7407 cases, while the *test error* is 0.234 on 1759 cases.

To evaluate the classification performance, figure 3 shows the "zoom" (top–left quadrant) *Relative Operating Characteristics* diagram (hereafter ROC, introduced by Swets 1973), obtained for the total and the test databases. The two curves are superposable, meaning that there is no overfitting problem. The vertical segment connecting the ROC curve to the diagonal line POD=POFD is the Kuipers Skill Score (Hanssen and Kuipers 1965, Gandin and Murphy 1992, but first introduced by Peirce 1884), because KSS = POD - POFD. Choosing, on the total database ROC curve, the threshold ($\bar{y} = 0.13$) which maximizes the KSS means to fix a particular *table of contingency*, which summarizes different aspects of the classification performance. As visible in table 2, some interesting

scores computed on the total dataset are: KSS = 0.69, FAR = 0.56 (False Alarm Rate), O = 36 (Odds Ratio, see Stephenson 2000). On the test sample they are almost similar: **KSS = 0.69**, **FAR = 0.60**, O = 32.

There are two main differences between this thunderstorm classification ANN and the corresponding one chosen in Manzato (2005). The first is the choice of DT500 instead of the similar DTC as first variable, while the second is the choice of VFlux (the fourth variable). Both these differences arise when adding the October and November data. A last – relatively unimportant because at the end – difference is the use of VVstd instead of the temperature at the LCL (Tbase). As in the previous work, also here different pre–processing methods have been tested. The results shown have been obtained using the *posterior probability pre–processing* (details in Manzato 2005), which transforms each variable value into its event probability, so that all the different variables become more comparable between them. Doing other runs with the simpler *Z–scores pre–processing* leads to a slightly worse test error (0.239), still obtained using 9 inputs and 6 hidden neurons (maximum KSS=0.68 with FAR=0.62).

4 Thunderstorm regression

In this section an ANN for forecasting the thunderstorm intensity is built, as a regression model for estimating the CALCA6h value (target variable). An ANN for regression problems is similar to those seen in the previous section, but usually the output activation function, f_o in equation 1, is taken to be the *linear function*: $f(x) = x$. Also the error to be minimized is different, since usually it is taken to be the *Mean Square Error* (MSE):

$$MSE = \frac{1}{N} \cdot \sum_{n=1}^{N} (CALCA6h_n - y_n)^2 \,, \tag{4}$$

where the CALCA6h takes into account the number of lightning strikes, the rainfall and the maximum wind gust, as shown in equation 1 of Manzato (2003).

For this regression problem, all the 1995–2002 data with *no lightning strikes* (CALCA6h = 0) were taken out of the database, since they are the majority and can strongly affect the

fitting model (for example changing the ANN mean output). To increase to a maximum the remaining data, it was found useful to use also the data with only one or two lightnings, which increase the CALCA6h $\neq 0$ database number from 1280 up to 1790. This dataset has been divided into training (3/4) and validation (1/4) in 12 different ways (bootstraps). For each bootstrap, only the ANN with the lowest VE, on 50 different initializations with random weights, is taken. That is repeated by the forward selection algorithm for each of the possible input variables. Again, during this phase, H has been fixed to Int $\left(\frac{I-1}{2}\right) + 1$. In this case Z–scores pre–processing gave slightly better results.

As shown in figure 4, the first variable selected is MRH, which, when taken alone, has also the best linear correlation coefficient with CALCA6h: R= 0.34. While the potential instability measured by DT500 is the most important index for the FVG thunderstorm development (occurrence/non–occurrence), the *air saturation* in the mid–low levels measured by MRH is the single variable most correlated with the thunderstorm "intensity" (meaning as production of electric discharges, rain and wind). As in the previous section, ACTP is chosen as second variable. In this case ACTP is not simply 0 or 1, as was in the classification problem, but it is the 6 h–before value of the *continuous* variable CALCA6h. The third variable is the well known *Convective Inhibition* index (CIN), while the Maximum Buoyancy (Morgan and Tuttle 1984, Manzato and Morgan 2003) follows as fourth. Next follow P(HH), the wind shear in the lowest 3 km (Shear3) and the mean buoyancy acceleration of the lowest 250 hPa (b_PBL).

Looking at figure 4 it is possible to see all the 12 bootstrap MSE TV points for the first 7 variables, pre–processed with Z–scores. The large jump of their mean values (unfilled squares) between ACTP and CIN, or between MaxBuo and P(HH), is justified by the increase of H (shown to the right of the labels). Comparing this figure with that of the classification problem (figure 2) it is possible to see that now there is a larger *variability* of the 12 bootstrap errors, with respect to the distance between the mean values. That is probably due to the smaller dataset, which increases the sensitivity to the "bad" data and how these are picked up into training and validation samples.

In this case it is not easy to choose the best ANN architecture. Using ANNs with

respectively 5, 6 and 7 inputs and trying all the hidden neurons in the range $1 \leq H \leq 6$ leads to a minimum mean VE (on 12 bootstraps) respectively for 3, 4 and 3 hidden neurons. If now these 3 network architectures are used to fit the total database (1995–2002 active cases) and then tested on the test sample (2003-2004 active cases) what happens is that the test error is much larger (about 10%) than the total error, while in Manzato (2005) it was even worse. That means that the ANN seems to be *much too complex for the database used*, and so overfits the total dataset. To obtain a test error similar to the total error the ANN complexity *must be reduced* to 5 inputs and only 2 hidden neurons. This ANN obtains a test error of 0.0292, even lower than for the 3 more complex ANNs. Unfortunately, doing things in that way, there is the possibility of overfitting the test sample, which should *never* be used during the choice of the best ANN.

To solve this dilemma another approach is proposed. Indeed, one can look to the single 12 bootstraps and simply choose one of them, because they are selected with the criterion of the lowest *validation* error. In particular the bootstrap 10 of the ANN with **7 inputs** and **3 hidden neurons** had a VE quite low (0.0276) and a *similar* TE (0.0279). Using the ANN made with these weights to predict the total sample (1741 not missing cases) leads to a total error of 0.0278, while on the independent 2003–2004 sample (319 cases) it leads to a test error of 0.0289. The linear cross correlation coefficients are $\mathrm{R} = 0.50$ on the total sample and $\mathbf{R = 0.49}$ on the test sample (p–value $< 2.2\,10^{-16}$). These Rs show an improvement of 47% with respect to the best linear correlation with only one variable (MRH).

Figure 5 shows the scatterplot between the CALCA6h values and the ANN forecast for the thundery cases, with not missing indices, of the test sample. The result is not bad, but has the tendency to underestimate the strongest cases (CALCA6h ≥ 0.8).

5 Rain classification

As explained in CTR, the rainfall database is different from that used for thunderstorms, because it begins in 1992 and uses all the months. Since the events with less than 5 mm

of rain accumulated in 6 h are not considered particularly interesting, this classification problem tries to forecast the 2478 cases (14%) with at least 5 mm of maximum rainfall. The predictors investigated for the rain problem are 66 variables: 38 sounding–derived indices listed in table 1 (excluding those labeled as *"only thunderstorm"*), the 6 h difference for 24 of them, the julian date (JJJ) and hour (HH) of the sounding and the activity (rain occurrence) during the 6 h before (ACTP). Lastly, in this case the Sea Surface Temperature (SST) is added, because in CTR was found to be an important parameter for heavy rain.

The ANN technique is the same used in section 3. Figure 6 shows the 12 bootstrap TV points for the first 9 variables (pre–processed using their "5 mm rainfall posterior probabilities") chosen by the forward selection algorithm. As for the thunderstorm intensity (which considered also the rain field) also for the occurrence of at least 5 mm in 6 h, MRH is the best single variable to look at. Interestingly, the Maximum Buoyancy is the second variable chosen, and that is probably because it is a measure of "how difficult it is to lift the air". In fact, the presence of *vertical lifting* is an essential ingredient in the rain production mechanism. Next follow two variables related to the *flux*: the mean North–South component of the wind in the lowest 6 km (MLWv) and the 6 h difference of the mean North–South component of the water vapour flux (D_VFlux). This is the first "D_" (i. e. "difference with the previous case") variable chosen by the forward selection algorithm. Other variables of this kind will be selected by the rain ANNs, meaning that the typical *timescale* of the rainfall is probably larger than that of the thunderstorm events (usually less than 6 h) and also information measured, for example, 9 h before can be very useful.

After that, there are the CAP variable (able to inhibit the lifting) and the rain activity (occurrence/non–occurrence) of the previous 6 h interval (ACTP), i. e. persistence is the sixth choice. There follows a jump, due to the increase of H from 3 to 4, made by the 6 h difference of the balloon vertical velocity standard deviation (D_VVstd), which is related to the *vertical turbulence*. Lastly, the convective inhibition (CIN) and the mean wind speed in the first 6 km (MLWspd) are chosen, even if the mean CEE improvements seem very small. Comparing this figure with the corresponding one for thunderstorms (figure 2), it looks as though the variability of the 12 TV points is similar. The mean VE from the ANN with

[I=1, H=1] to the ANN with [I=7, H=4] decreases by 24% for the thunderstorms and by 23% for the rainfalls.

To choose the best ANN architecture all the hidden neuron numbers in the range $1 \leq H \leq 7$ were tested (on 12 bootstraps), using all the previous 9 inputs. The lowest mean VE is obtained for **9 inputs** and **4 hidden neurons**. This model is then used to fit all the total database, obtaining an ANN with a CEE of 0.206 (on 12742 cases). That ANN gives a test CEE of 0.195 (on 2572 cases) for the test sample. Also figure 7 shows that the two ROC diagrams for the total and test samples are superposable, confirming the performance validity of this classification model.

This performance can be evaluated in table 3, which shows for the test sample a **KSS = 0.72** with a **FAR = 0.59** and an Odds Ratio of 37. Considering that the event probability is a little lower than in the thunderstorms case (14% against 16%), it is possible to say that the rain classifier is slightly better than the thunderstorm classifier. In both cases, almost 90% of the events are detected (POD), but with a high percentage of false alarms. Of course, these results can be changed choosing another threshold, instead of that which maximizes KSS ($\bar{y} = 0.16$). Then it is possible to decrease both POD and FAR (and, obviously, KSS).

6 Rain regression

The target variable for this regression problem is:

$$t_n = \frac{\ln(1 + \mathrm{MaxRain_n})}{5.5}\,, \tag{5}$$

where $\mathrm{MaxRain}_n$ is the maximum rainfall measured during the n 6 h period by any of the 15 stations in the FVG plain. The normalization coefficient of 5.5 assures reduction of the historical maximum (184 mm) to less than one. A similar problem has been studied also by Kuligowski and Barros (1998), who tried to predict the 6 h *winter* rainfalls using the rainfall of the previous case together with some sounding–derived indices (in particular the wind at 700 hPa). The main differences with their work is that here the regression problem

is applied only to the "rainy" cases, while the forecast of rain occurrence is assigned to another ANN. In particular, to maximize the database dimension, it has been found useful to include all the cases with MaxRain $\geq$ 1 mm in 6 h, which are 3392 in the total sample and 570 in the test sample.

Figure 8 shows the first 9 variables (with Z–scores pre–processing) chosen by the rain regression forward selection algorithm. The most important variable is the *water vapour flux*, North–South component, in the lowest 3 km (VFlux). As shown in CTR, it has a linear correlation with MaxRain of R = -0.32 (negative VFlux means southerly wind, i. e. more rainfall). The second variable chosen is again the activity during the previous case (ACTP), i. e. $t_{(n-1)}$. The third variable is the sounding–derived "updraft", which SOUND_ANALYS.PY computes as UpDr=$\sqrt{2\,\mathrm{CAPE}}$ using the CAPE energy accumulated up to the level of -15 °C instead of reaching the parcel Equilibrium Level (details in Manzato 2003). That is the only variable chosen in this problem which is a measure of *convective instability*.

Next follow the mean levels humidity (MRH) and wind (the North–South component, MLWv), while after that there is the Sea Surface Temperature (SST), measured at -2 m a. m. s. l. by the Trieste station. The seventh variable is the 6 h variation of VFlux: considering that VFlux was also the first variable, that means that the 6 h rain intensity is strongly related to the North-South *advection of humid air*. Lastly there are the 6 h variation of the mean East–West wind in the lowest 6 km (D_MLWu) and the CAP strength. Comparing this figure with figure 4 it is possible to find that the decrease of the mean MSE VE between the [I=1, H=1] and the [I=7, H=4] ANNs is 21% for rainfall and only 14% for the CALCA6h case.

To choose the best ANN one can look at figure 9, which shows the 12 TV points for the ANNs with **9 inputs** and $1 \leq H \leq 6$. As best ANN the bootstrap 11 of **H=5** (circled), which has a low VE (0.0180) and similar TE (0.0181), is chosen. Using this ANN to fit the total database leads to a MSE of 0.0181, with R=0.57 on 2854 cases (p–value $< 2.2\,10^{-16}$), while on the 2003–2004 test sample the MSE becomes 0.0184 and **R=0.53** on 525 cases (which means an improvement of 65% with respect to VFlux alone).

Figure 10 shows the scatterplot of the measured rainfall (transformed with equation 5) and the ANN output, y. As was for the thunderstorm intensity, the strongest rainfalls are underestimated. More in general, the standard deviation of the forecast (0.09) is about only half of that of the observed target (0.16). To increase the forecast variance, Karl et al. (1990) proposed the "inflation" method, which just multiplies the forecast anomalies by the ratio of the two standard deviations. Von Storch (1999) claims that this method increases also the forecast MSE and proposed to randomize the unexplained variance with noise. What has been found in this work is that ANN with lower "complexity" (e. g. fewer hidden neurons) can obtain similar MSE and R compared to more complex ANN, but with slightly lower variance.

7 Comparing regression and classification models

Comparing the TV diagrams for the classification and regression ANNs (figure 2, 4, 6 and 8) it seems that the classification ANNs are able to decrease more their validation error before saturating. This can be due to the fact that the regression database is much smaller, because the non–event cases (which are the majority) are taken out. Then, one can think to obtain better results using many categorical classification models instead of one single regression model. For example, instead of a single regression ANN to estimate the maximum rainfall amount, one can use a sequence of classification ANNs to estimate the occurrence of 1, 2, 5... and so on mm of rainfall.

The regression ANN of the previous chapter is here "converted" in classification models for different categories of accumulated rain and their performances are compared to those of true classification models. It is possible to get a classification model from a regression ANN just putting a threshold on the ANN forecast, to dichotomize the output. Figure 11 shows the maximum KSS and the maximum Heidke Skill Score (HSS) for the classification models, obtained from the previous regression ANN, while forecasting the occurrence of 1, 2, 5, 10, 20, 40 and 100 mm of rain. In abscissa there is the logarithm of the rain occurrence category, while in ordinate there are the event probability (i. e. the frequency)

and the maximum KSS and HSS (obtained using different thresholds).

In the range [1, 40] mm the event frequency (dashed line) decreases almost linearly with the logarithm of the rain amount category, meanwhile the maximum KSS (continuous line) increases almost linearly with it (from 0.50 to 0.79). In this case, the same statistical model (the regression ANN) has a maximum KSS which increases inversely with the event frequency, as if it should be "easier" to forecast the rarest event occurrence. Instead, the maximum HSS (dotted line) has a non-linear behavior: it is almost constant (0.46) when the event frequency is in the range [9, 22] %, while it drops dramatically when the frequency decreases at very low values (rare events). That happens because, by definition, the HSS depends strongly on the event frequency.

Figure 12a shows the zoom ROC (on the 1992-2002 dataset) for three of these rain categories: the occurrence of at least 5, or 20 or finally 40 mm of rain. The three thresholds applied to the regression ANN output to maximize KSS are 0.33, 0.36 and 0.41. The ROC curve becomes "higher" increasing the rain amount because the maximum KSS increases (0.61, 0.72 and 0.79 respectively) and KSS = POD - POFD.

Figure 12b shows the ROC for the same three classes of rain but obtained by three different classification ANNs: for the 5 mm of rain it is the same ANN presented in chapter 5, while for the 20 and 40 mm they are two ANNs made using 8 inputs and 3 hidden neurons or 7 inputs and 3 hidden neurons respectively (using simple Z–scores pre–processing). The variables used in these last two models are listed in table 4, where all the ANN inputs are ranked for an easy comparison. Also in this case the maximum KSS increases (0.70, 0.74 and 0.78) as the event frequency decreases, but the differences between the three classification models are not so large as was for the three classifiers obtained by the same regression model. Table 4 shows how also the mean Cross Entropy Error improves with the event rarity.

The 5 mm "true" classification model is much better than the 5 mm classification made by the regression ANN. The 20 mm classification model is just a little better than the regression ANN, while the two 40 mm models reach the same performance level (the two ROCs are superposable). So, above a certain rain amount (in this case between 20 and

40 mm) it is not worth building a sequence of different classification models instead of one single regression model. Probably, for forecasting extreme events, the regression models based on MSE are preferable to classification methods.

8 Conclusions

In this work six ANNs have been found: four for classification of events (6 h periods with at least 3 C2G lightning strikes and 6 h periods with at least 5, 20 or 40 mm of maximum rainfall) and two for intensity regression (CALCA6h and maximum rainfall in 6 h). In operational use, these ANNs are used "in series": if the classification ANN forecasts event occurrence[3], then the regression ANN is used to estimate its intensity (which is shown in the forecaster room), otherwise the output is set to 0. The thunderstorm ANN–based forecast has been used with profit by OSMER forecasters since 2001, while the rainfall forecast was implemented in 2005.

ANNs proved to be a powerful statistical method for performing a multivariate analysis, but special care must be used to prevent overfitting. In particular it has been shown how a repeated holdout method, based on 12 bootstraps, can give good results. It also has been shown how fitting all the total database, in order to build the final ANN, can be very dangerous when the variability of the single bootstraps is large. In this case (regression ANNs) it is better to use always a criterion based on the validation samples, as choosing one single bootstrap.

A last word on the forward selection algorithm: it does not assure finding the absolute *best set* of inputs, but just a good one. All the possible sets should be tested to find the very best one, but that is too much time consuming. Moreover, the last variables selected, when ANNs are near to overfitting or to saturating, can change rerunning the algorithm, since at the end the VE differences between the "best" variables become very small.

Even if the sounding–derived indices computed by the SOUND_ANALYS.PY program

[3]For rain occurrence it is considered the 5 mm classification ANN.

are a multitude, the ones which were *more frequently* chosen in the first five places of table 4 are only a few. They are: **MRH**, **ACTP**, **VFlux**, **MaxBuo**, **CAP** and **MLWv** (plus KI, D_SWISS and D_VVstd in the classification of 20 and 40 mm of rain). More in general, the variables which proved to be *more useful* in this multivariate study can be grouped in four categories: 1) *potential instability* indices: like DT500 (a refined Lifted Index), KI, MaxBuo and UpDr; 2) *convective inhibition* indices: like CIN and CAP; 3) *mean wind and humidity* indices in the mid–low levels (in the lowest 3–6 km): like MLWv, VFlux and moreover MRH; 4) the *persistence effect* (ACTP), which was chosen as second variable for 3 times.

Finally it must be stressed that these results can be strongly due to the particular FVG orography, so it should be very interesting to try a similar approach in different regions. Future development of this work will be focused on model output pseudo–soundings (i. e. to their derived pseudo–indices), to improve the time forecast validity, and also to other forecast fields.

Acknowledgments

The author thanks his wife and dr Griffith Morgan for English improvement and for encouragement, and the anonymous reviewers for the suggestions. Last but not least, he thanks the "GNU generation" for having provided such good – and free! – tools as R, Python, LaTeX, Emacs, as well as Linux itself.

References

Allen, G. and J. F. Le Marshall, 1994. An evaluation of neural networks and discriminant analysis methods for application in operational rain forecasting. *Australian Meteorological Magazine* **43**, 17–28.

Bishop, C. M., 1996: *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford, 482 pp.

Boyden, C. J., 1963. A simple instability index for use as a synoptic parameter. *Meteorol. Mag.* **92**, 198–210.

Cheng, B. and D. M. Titterington, 1994. Neural Networks: A Review from a Statistical Perspective. *Stat. Sci.* **9**, 2–54.

Flueck, J. A., 1987. A study of some measures of forecast verification. Preprints *10th Conf. Probability and Statistics in Atmospheric Sciences*, Edmonton, Alberta, Amer. Meteor. Soc., 69–73.

Frankel, D. S., Schiller, I., Draper, J. S., and A. A. Barnes, 1990. Investigation of the prediction of lightning strikes using neural networks. Preprints 16th Conference on Severe Local Storms, October 22-26, Kananaskis Provincial Park, Alberta, Canada, Amer. Meteor. Soc., Boston, 7–11.

French, M. N., Krajewski, W. F., and R. R. Cuykendall, 1992. Rainfall Forecasting in Space and Time Using Neural Network. *J. Hydrology* **137**, 1–31.

Galway, J. G., 1956. The lifted index as a predictor of latent instability. *Bull. Amer. Meteor. Soc.* **37**, 528–529.

Gandin, L. S., and A. H. Murphy, 1992. Equitable skill scores for categorical forecasts. *Mon. Wea. Rev.* **120**, 361–370.

Hanssen, A. W., and W. J. A. Kuipers, 1965. On the relationship between the frequency of rain and various meteorological parameters. *Koninklijk Nederlands Meteorologist Institua Meded. Verhand* **81**, 2–15.

Huntrieser, H., Schiesser, H. H., Schmid, W. and A. Waldvogel, 1997. Comparison of Traditional and Newly Developed Thunderstorm Indices for Switzerland. *Wea. Forecasting* **12**, 108–125.

Karl, T. R., Wang, W. C., Schlesinger, M. E., Knight, R. W., and D. Portman, 1990. A method of relating general circulation model simulated climate to observed local climate. Part I: Seasonal statistics. *J. Climate*, **3**, 1053–1079.

Kuligowski, R. J., and A. P. Barros, 1998. Experiments in Short–term Precipitation Forecasting Using Artificial Neural Networks. *Mon. Wea. Rev.* **126**, 470–482.

Lee, J., Wegner, R. C., Sengputa, S. K., and R. M. Welch, 1990. A neural network approach to cloud classification. *IEEE Transactions on Geoscience and Remote Sensing* **28**, 846–855.

Manzato, A., 2003. A climatology of instability indices derived from Friuli Venezia Giulia soundings, using three different methods. *Atmos. Res.* **67–68**, 417–454.

———, 2005. The use of sounding derived indices for a neural network short–term thunderstorm forecast. Accepted for *Wea. Forecasting* **20**, 896–917.

———, and G. M. Morgan, 2003. Evaluating the sounding instability with the Lifted Parcel Theory. *Atmos. Res.* **67–68**, 455–473.

Marzban, C., 2002. Neural Network Short Course, Amer. Meteor. Soc., 2002. Available at http://www.nhn.ou.edu/~marzban.

———, and G. J. Stumpf, 1996. A neural network for tornado prediction based on Doppler radar-derived attributes. *J. Appl. Meteor.* **35**, 617–626.

———, and A. Witt, 2001. A Bayesian Neural Network for Hail Size Prediction. *Wea. Forecasting* **16**, 600–610.

Masters, T., 1993. *Practical Neural Network Recipes in C++.* Academic Press, 493 pp.

McCann, D. W., 1992. A neural network short-term forecast of significant thunderstorms. *Wea. Forecasting* **7**, 525–534.

Morgan, G. M. Jr., and J. Tuttle, 1984. Some experimental techniques for the study of the evolution of atmospheric thermodynamic instability. Paper presented at the Second International Conference on Hailstorms and Hail Suppression (Proceedings), Printing Office of the Hydrometeorological Service, Sofia, pp. 192–196.

Navone, H. D., and H. A. Ceccatto, 1994. Predicting Indian Monsoon Rainfall: A Neural Network Approach. *Climate Dyn.* **10**, 305–312.

Pankiewicz, G. S., 1997. Neural network classification of convective airmasses for a flood forecasting system. *International Journal of Remote Sensing* **18**, 887–898.

Pasini, A., and S. Potestà, 1995. Short–range visibility forecast by means of neural–network modelling: a case study. *Nuovo Cimento* **18C**, 505–516.

Peirce, C. S., 1884. The numerical measure of the success of predictions. *Science* **4**, 453–454.

Rosenblatt, F., 1958. The Perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review* **65**, 386–408.

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

Stephenson, D. B., 2000. Use of the "odds ratio" for diagnosing forecast skill. *Wea. Forecasting* **15**, 221–232.

Swets, J. A., 1973. The relative operating characteristic in psychology. *Science* **182**, 900–1000.

Venables, W. N., and B. D. Ripley, 2002. *Modern Applied Statistics with S-PLUS*. 4th ed. Springer-Verlag, 495 pp.

von Storch, H., 1999: On the use of "inflation" in downscaling. *J. Climate* **12**, 3505–3506.

Wilks, D. S., 1995. *Statistical Methods in the Atmospheric Sciences*. Academic Press, 467 pp.

Xiao, R. and V. Chandrasekar, 1996. Multiparameter Radar Snowfall estimation using Neural Network Techniques. Proceedings IGARSS 1996, *Geoscience and Remote Sensing Symposium*, 566–568.

TABLE CAPTIONS

Table 1: *The list of all the variables investigated as predictors. Apart from the SWISS index (Huntrieser et al. 1997), the meaning of all these variables has been clarified in Manzato (2003) and Manzato (2005). MUP means "Most Unstable Parcel" and is 30 hPa high.*

Table 2.
2 a: *The Table of Contingency definition for the "total" and "test" thunderstorm samples.*
2 b: *Some performance indices derived by the previous table, following Wilks (1995) and Flueck (1987) for the Probability Of False Detection. The threshold used maximizes the Kuipers Skill Score.*

Table 3: *Same as table 2, but for the 5 mm in 6 h rainfall classification problem.*

Table 4: *The first 9 "best" variables chosen by the forward selection algorithm for the six ANNs presented, together with their associated mean Validation Errors (CEE for classification and MSE for regression). P() means "posterior probability" and Z() means Z–scores pre–processing. The Hidden Neuron number was given by* $H = \mathrm{Int}\left(\frac{I-1}{2}\right) + 1$.

FIGURE CAPTIONS

figure 1: *A schematic example of a generic feed–forward artificial neural network with 4 input "neurons", 2 neurons in the hidden layer and 1 output neuron. For this kind of ANN the total number of weights is* $[(I+2)\cdot H + 1]$, *i. e. 13 in this example.*

figure 2: *All the 12 TV points for the first 9 "best variables" chosen by the thunderstorm classification forward selection algorithm. The set of 12 TV points for each variable is represented alternatively by 2 symbols (circles and triangles) and 3 filling gray scales (dark, medium and light). The unfilled squares, connected by a dashed line, show the mean of the 12 TV point set. The value of H used in each cycle is shown to the right of the variable label.*

figure 3: *The "zoom" ROC diagram (top–left quadrant) showing the Probability of Detection vs. the Probability of False Detection for the thunderstorm ANN using 9 inputs and 6 hidden neurons. The continuous line is computed on the total database (1995–2002), while the dashed line is corresponding to the test database (2003–2004).*

figure 4: *Same as figure 2, but for the first 7 variables of the thunderstorm regression forward selection algorithm. Here the variability of the 12 bootstrap set is larger than before, when compared with the decrease of the mean TV point of each cycle.*

figure 5: *Observed thunderstorm intensity (CALCA6h) vs. the regression 7–I and 3–H ANN output, y, computed forecasting all the active cases of the test sample (2003–2004). Line is the linear correlation between points (R = 0.49).*

figure 6: *Same as figure 2, but for the first 9 variables of the rainfall classification forward selection algorithm. Here the variability of the 12 bootstrap set is similar to that of the thunderstorm classification case.*

figure 7: *Same as figure 3, but for the rainfall classification ANN using 9–I and 4–H. The total database is 1992–2002, while the test sample is still 2003–2004.*

figure 8: *Same as figure 4, but for the first 9 variables of the rainfall regression forward selection algorithm. The bootstrap variability is similar to that of figure 4.*

figure 9: *All the 12 TV points for the ANNs using 9 inputs and a number of hidden neurons in the range* $1 \leq H \leq 6$ *for the rainfall regression case. The circled triangle shows how a single bootstrap with H=5 can obtain a very low TE and VE (about 0.182).*

figure 10: *Same as figure 5, but for the rain intensity case, using a 9–I and 5–H ANN, for all the active cases of the test sample (2003–2004). The line is the linear correlation between points (R = 0.53).*

figure 11: *The performance of the rain regression ANN when used as a classifier of different categories of rainfall occurrence. The event probability (dashed line) decreases with the logarithm of the amount of rain, while the maximum KSS (continuous line with scale on the right) increases. Lastly, the maximum HSS (dotted, same scale as the event probability) drops for the rare event cases.*

figure 12.

12 a: the ROC diagrams obtained by the rain regression *ANN for three categories of rainfall (at least 5, 20 or 40 mm accumulated in 6 h).*

12 b: the ROC diagrams obtained by three different classification *ANNs for the same three categories of rainfall as in 12a.*

TABLES

Table 1.

acronym	**variable**
ACTP	ACTivity of the Previous case
b_PBL	Mean buoyancy acceleration of the lowest 250 hPa
BOY	Boyden Index (Boyden 1963)
BRI	Bulk Richardson Number *(only thunderstorm)*
CAP	Maximum cap (as Θ_{es} difference)
CAPE	Convective Available Potential Energy
CIN	Convective Inhibition
DT500	Difference of temperature at 500 hPa ($T_{env} - T_{parc}$)
DTC	Core difference of temperature (parcel at -15 °C)
EHI	Energy–Helicity index
h_MUP	MUP height *(only thunderstorm)*
HH	Synoptic hour of the sounding
JJJ	Julian date of sounding
KI	K index
LCL	Lifting condensation level
LFC	Level of free convection height *(only thunderstorm)*
LI	Lifted Index *(only rain)*
LLWspd	Low–level (0.5 km) wind speed
LLWu	U component of low–level wind (lowest 0.5 km) *(only rain)*
LLWv	V component of low–level wind (lowest 0.5 km)
LRH	Mean relative humidity in the low levels (lowest 250 hPa)
MEL	Melting Level (parcel at 0 °C)
MLWspd	Mid–level (6 km) wind speed
MLWu	U component of mid–level wind (lowest 6 km) *(only rain)*
MLWv	V component of mid–level wind (lowest 6 km)
MRH	Mean relative humidity in medium levels (lowest 500 hPa)
MaxBuo	Maximum Buoyancy (Morgan and Tuttle 1984)
Mix	MUP mixing ratio
PBL	Planetary Boundary Layer estimated height
PWC	Precipitable water content of cloud
PWE	Precipitable water content of environment
Rel_Hel	Relative Helicity *(only rain)*
Shear	Wind shear in the lowest 12 km *(only thunderstorm)*
Shear3	Wind shear in the lowest 3 km
ShowI	Showalter Index
SST	Sea Surface Temperature (at -2 m) *(only rain)*
SWEAT	Severe WEAther Threat
SWISS	Stability and Windshear Index for Storms in Switzerland
Tbase	Cloud base (LCL) temperature
Thetae	MUP equivalent potential temperature
Trop	Tropopause height *(only thunderstorm)*
UpDr	"Core updraft" (parcel at -15 °C)
VFlux	Mean water vapour v horizontal flux
|VFlux|	Absolute value of VFlux
VV	Radiosonde ascensional vertical velocity *(only rain)*
VVstd	Standard deviation of radiosonde vertical velocity
WBZ	Wet bulb zero height

Table 2 a.

$y > 0.2$	**Observed Yes**	**Observed No**	**Total sample**		**Test sample**	
Forecast Yes	a = event	b = false	a = 1054	b = 1358	a = 211	b = 315
Forecast No	c = miss	d = nonevent	c = 106	d = 4889	c = 25	d = 1208

Table 2 b.

Variable	**Symbol**	**Definition**	**Total sample**	**Test sample**
Bias forecast/real events	BIAS	$\frac{a+b}{a+c}$	2.08	2.23
Probability Of Detection	POD	$\frac{a}{a+c}$	0.91	0.89
Probability Of False Detection	POFD	$\frac{b}{b+d}$	0.22	0.21
False Alarm Rate	FAR	$\frac{b}{a+b}$	0.56	0.60
Hit rate	HIT	$\frac{a+d}{a+b+c+d}$	0.80	0.81
Heidke Skill Score	HSS	$\frac{2\cdot(ad-bc)}{(a+c)\cdot(c+d)+(a+b)\cdot(b+d)}$	0.48	0.45
Kuipers Skill Score	KSS	$\frac{(ad-bc)}{(a+c)\cdot(b+d)}$	0.69	0.69
KSS standard error	E(KSS)	$\frac{N^2-4(a+b)(c+d)KSS^2}{4N(a+b)(c+d)}$	0.01	0.02
Odds ratio	O	$\frac{ad}{bc}$	36	32

Table 3 a.

$y > 0.16$	**Observed Yes**	**Observed No**	**Total sample**		**Test sample**	
Forecast Yes	a = event	b = false	a = 1415	b = 1986	a = 216	b = 394
Forecast No	c = miss	d = nonevent	c = 190	d = 9151	c = 23	d = 1616

Table 3 b.

Variable	**Symbol**	**Definition**	**Total sample**	**Test sample**
Bias forecast/real events	BIAS	$\frac{a+b}{a+c}$	1.93	2.13
Probability Of Detection	POD	$\frac{a}{a+c}$	0.86	0.88
Probability Of False Detection	POFD	$\frac{b}{b+d}$	0.16	0.16
False Alarm Rate	FAR	$\frac{b}{a+b}$	0.56	0.59
Hit rate	HIT	$\frac{a+d}{a+b+c+d}$	0.85	0.84
Heidke Skill Score	HSS	$\frac{2\cdot(ad-bc)}{(a+c)\cdot(c+d)+(a+b)\cdot(b+d)}$	0.50	0.48
Kuipers Skill Score	KSS	$\frac{(ad-bc)}{(a+c)\cdot(b+d)}$	0.70	0.72
KSS standard error	E(KSS)	$\frac{N^2-4(a+b)(c+d)KSS^2}{4N(a+b)(c+d)}$	0.01	0.02
Odds ratio	O	$\frac{ad}{bc}$	32	37

Table 4.

Input rank	Thunderstorm classification	Thunderstorm regression	rain ≥ 5 mm classification	rain ≥ 20 mm classification	rain ≥ 40 mm classification	rain regression
1	P(DT500)	Z(MRH)	P(MRH)	Z(VFlux)	Z(|VFlux|)	Z(VFlux)
	0.352	0.0335	0.278	0.140	0.0474	0.0242
2	P(ACTP)	Z(ACTP)	P(MaxBuo)	Z(KI)	Z(KI)	Z(ACTP)
	0.321	0.0324	0.254	0.125	0.0426	0.0227
3	P(CAP)	Z(CIN)	P(MLWv)	Z(D_VVstd)	Z(D_VVstd)	Z(UpDr)
	0.296	0.0309	0.240	0.114	0.0392	0.0212
4	P(VFlux)	Z(MaxBuo)	P(D_Vflux)	Z(ACTP)	Z(D_SWISS)	Z(MRH)
	0.285	0.0302	0.231	0.111	0.0371	0.0205
5	P(MRH)	P(HH)	P(CAP)	Z(D_SWISS)	Z(MLWu)	Z(MLWv)
	0.278	0.0294	0.221	0.107	0.0365	0.0199
6	P(HH)	Z(Shear3)	P(ACTP)	Z(CAP)	Z(LRH)	Z(SST)
	0.273	0.0290	0.218	0.104	0.0358	0.0196)
7	P(MLWv)	Z(b_PBL)	P(D_VVstd)	Z(MLWv)	P(HH)	Z(D_VFlux)
	0.267	0.0288	0.213	0.102	0.0357	0.0192
8	P(VVstd)	–	P(CIN)	Z(MaxBuo)	Z(D_MLWu)	Z(D_MLWu)
	0.264	–	0.212	0.100	0.0357	0.0190
9	P(BRI)	–	P(MLWspd)	–	–	Z(CAP)
	0.262	–	0.210	–	–	0.0189

FIGURES

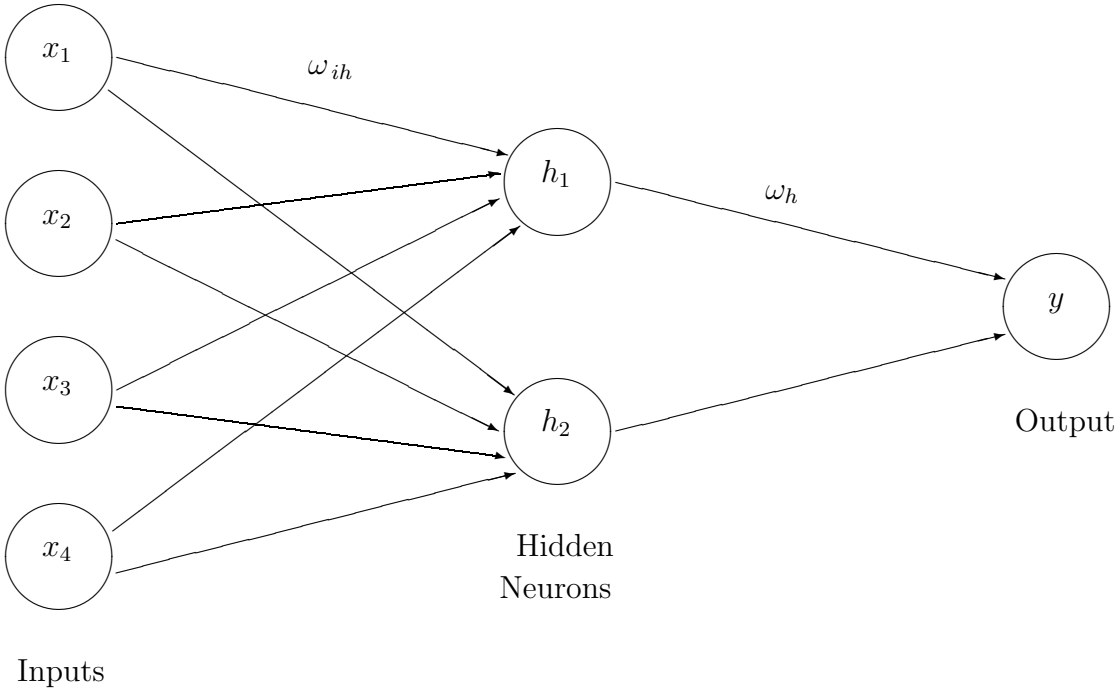

fig. 1

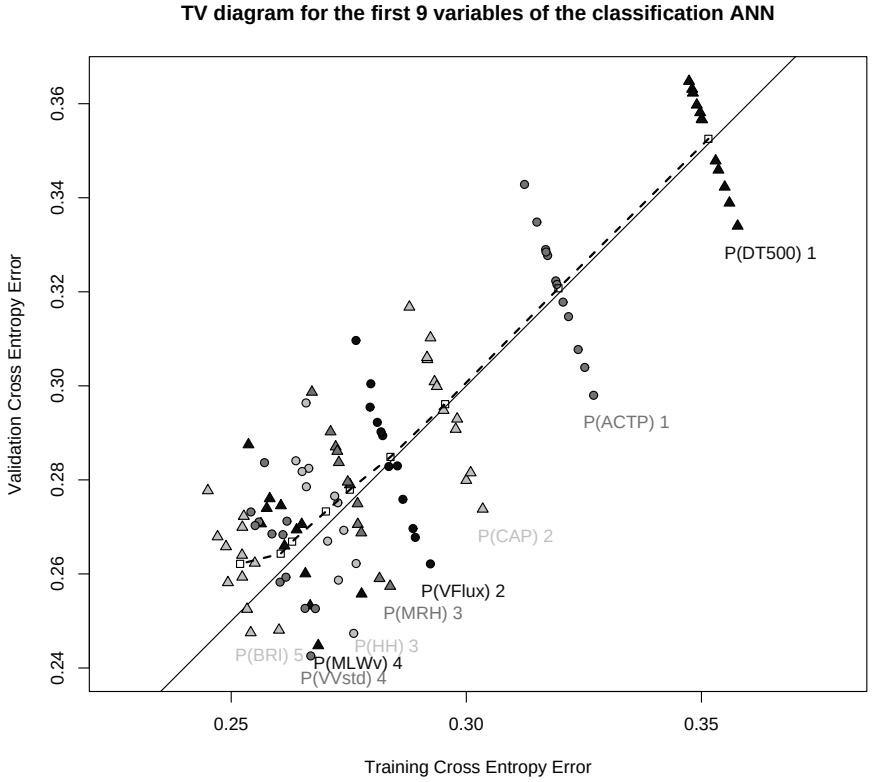

fig. 2

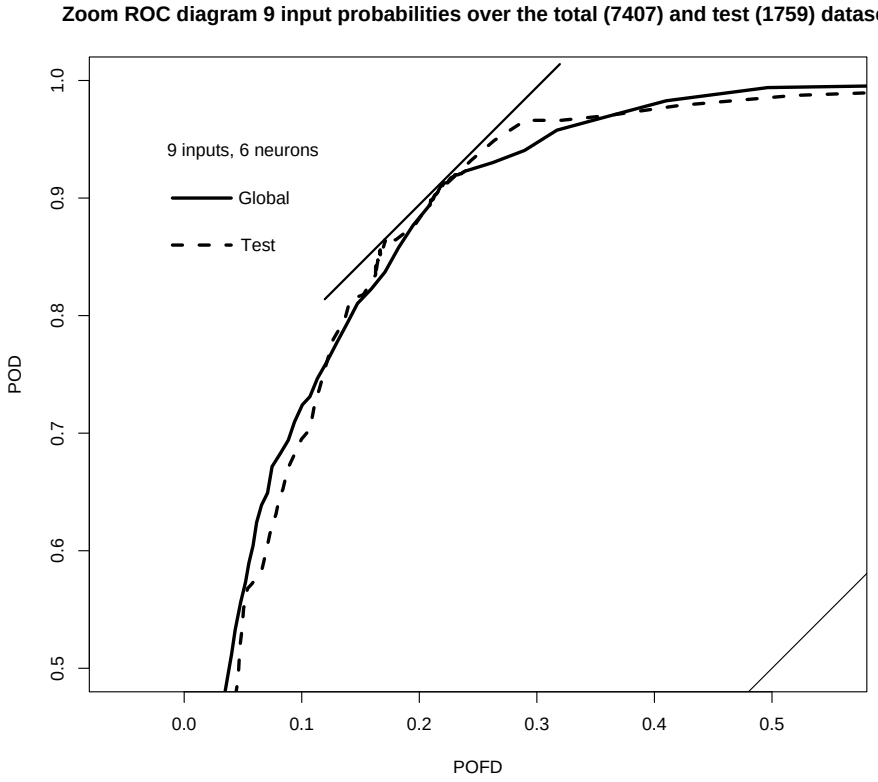
Zoom ROC diagram 9 input probabilities over the total (7407) and test (1759) dataset
9 inputs, 6 neurons
Global
Test
POD
POFD
1.0
0.9
0.8
0.7
0.6
0.5
0.0
0.1
0.2
0.3
0.4
0.5

fig. 3

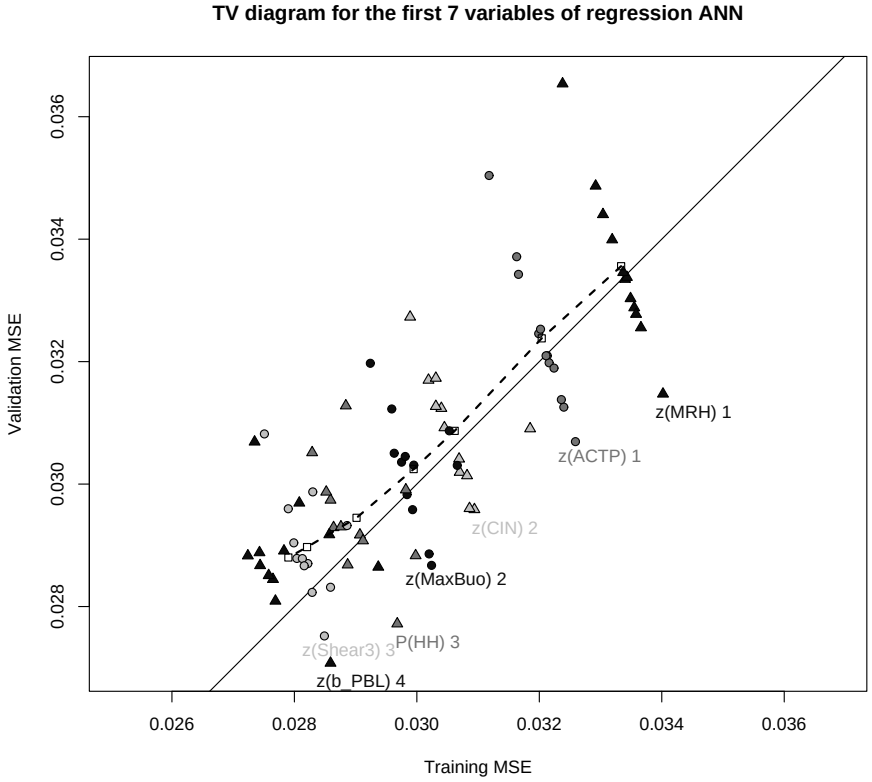
TV diagram for the first 7 variables of regression ANN
Validation MSE
Training MSE
0.036
0.034
0.032
0.030
0.028
0.026
0.028
0.030
0.032
0.034
0.036
z(MRH) 1
z(ACTP) 1
z(CIN) 2
z(MaxBuo) 2
P(HH) 3
z(Shear3) 3
z(b_PBL) 4

fig. 4

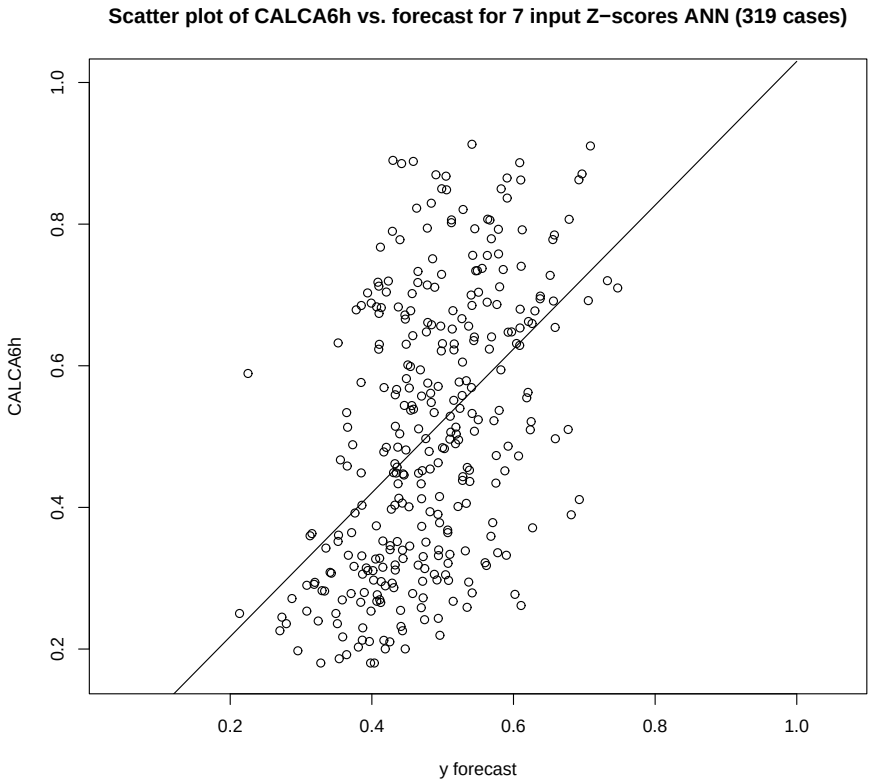
Scatter plot of CALCA6h vs. forecast for 7 input Z−scores ANN (319 cases)
CALCA6h
y forecast
0.2
0.4
0.6
0.8
1.0

fig. 5

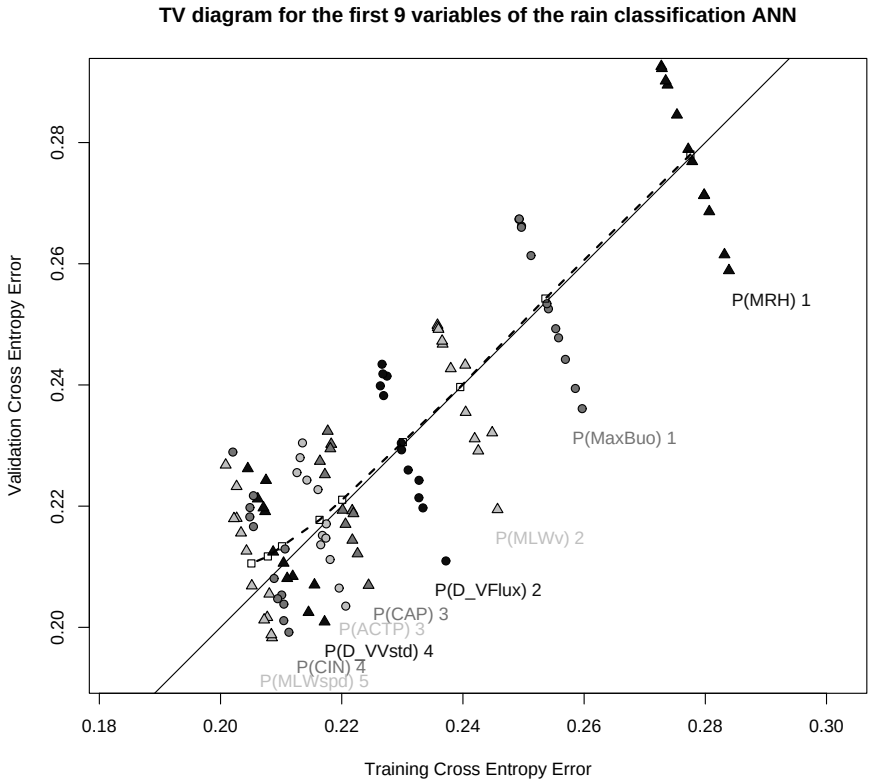
TV diagram for the first 9 variables of the rain classification ANN
Validation Cross Entropy Error
Training Cross Entropy Error
0.18
0.20
0.22
0.24
0.26
0.28
0.30
P(MRH) 1
P(MaxBuo) 1
P(MLWv) 2
P(D_VFlux) 2
P(CAP) 3
P(ACTP) 3
P(D_VVstd) 4
P(CIN) 4
P(MLWspd) 5

fig. 6

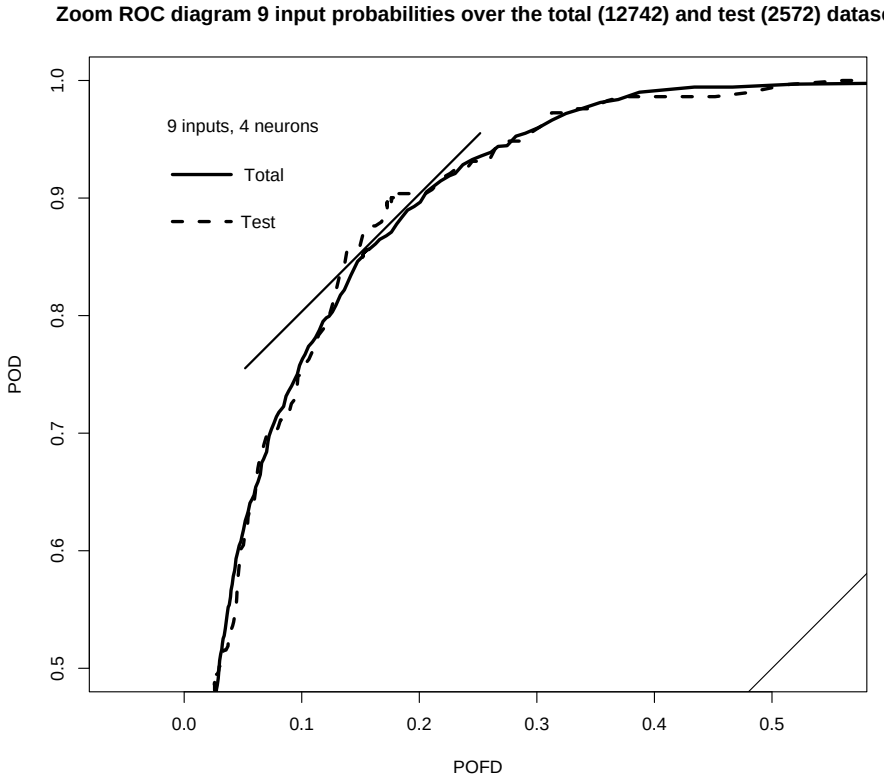
Zoom ROC diagram 9 input probabilities over the total (12742) and test (2572) dataset
9 inputs, 4 neurons
Total
Test
POD
POFD
1.0
0.9
0.8
0.7
0.6
0.5
0.0
0.1
0.2
0.3
0.4
0.5

fig. 7

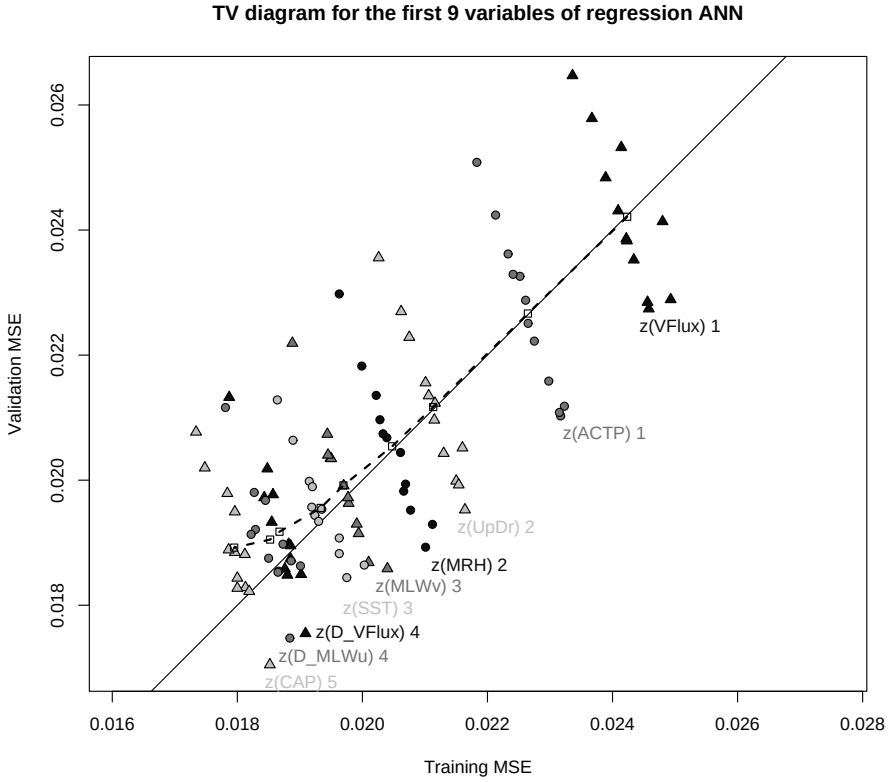
TV diagram for the first 9 variables of regression ANN
Validation MSE
Training MSE
0.026
0.024
0.022
0.020
0.018
0.016
0.018
0.020
0.022
0.024
0.026
0.028
z(VFlux) 1
z(ACTP) 1
z(UpDr) 2
z(MRH) 2
z(MLWv) 3
z(SST) 3
z(D_VFlux) 4
z(D_MLWu) 4
z(CAP) 5

fig. 8

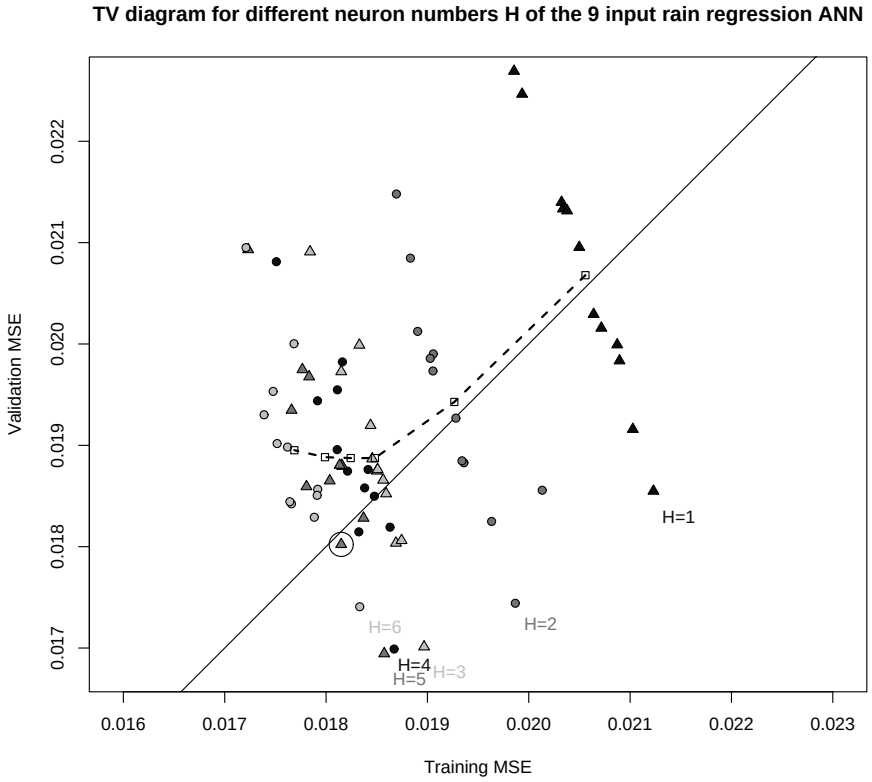
TV diagram for different neuron numbers H of the 9 input rain regression ANN
Validation MSE
Training MSE
0.016
0.017
0.018
0.019
0.020
0.021
0.022
0.023
H=1
H=2
H=3
H=4
H=5
H=6

fig. 9

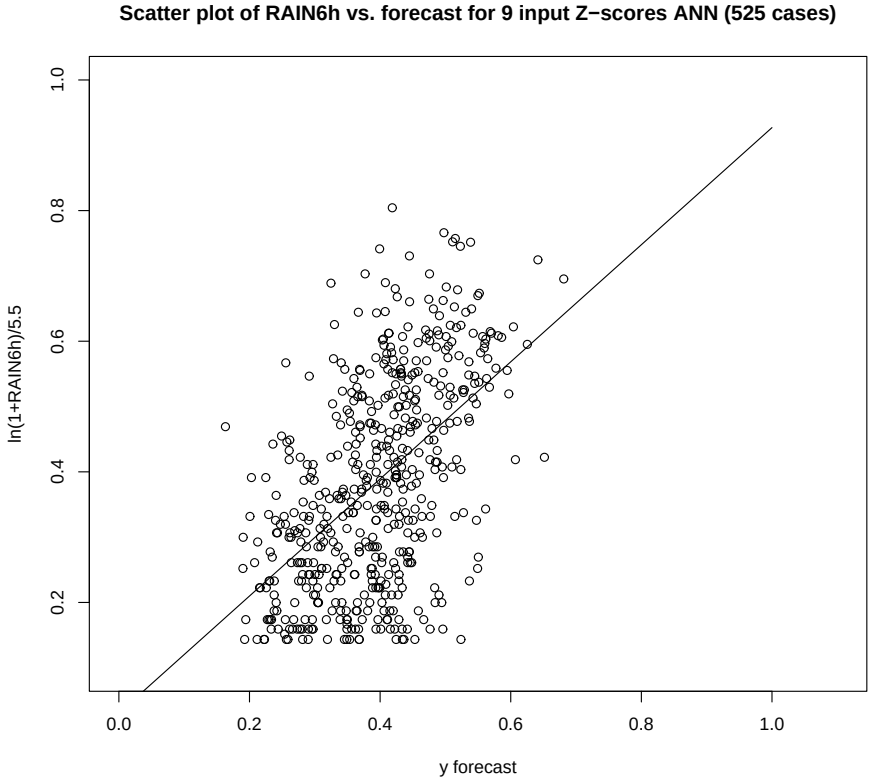
Scatter plot of RAIN6h vs. forecast for 9 input Z−scores ANN (525 cases)
ln(1+RAIN6h)/5.5
y forecast
0.0
0.2
0.4
0.6
0.8
1.0

fig. 10

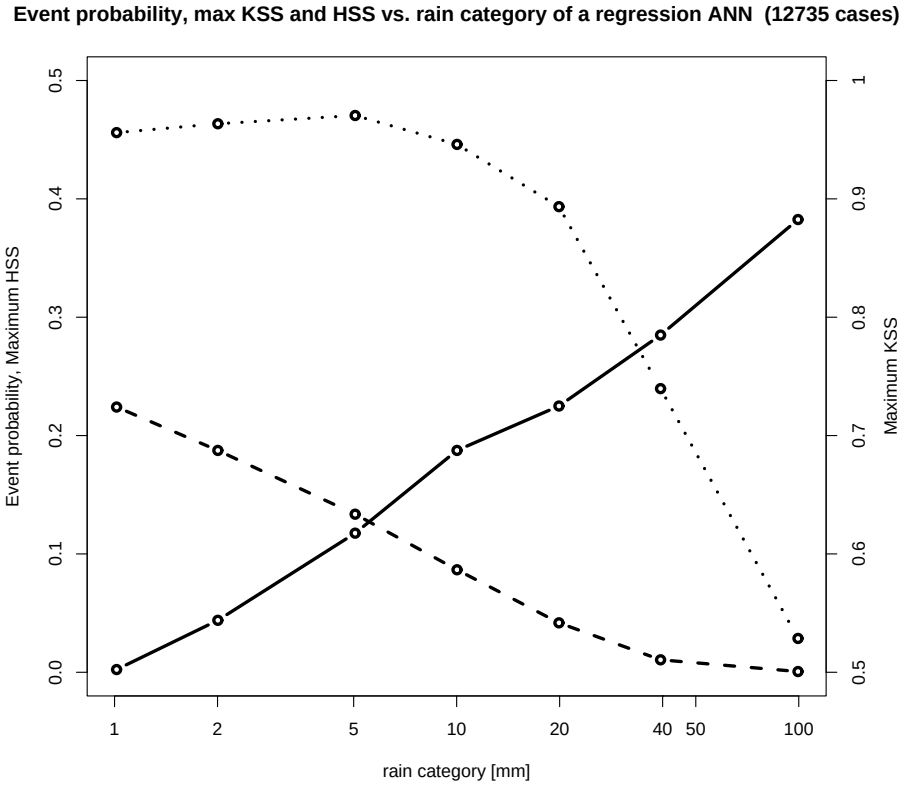

fig. 11

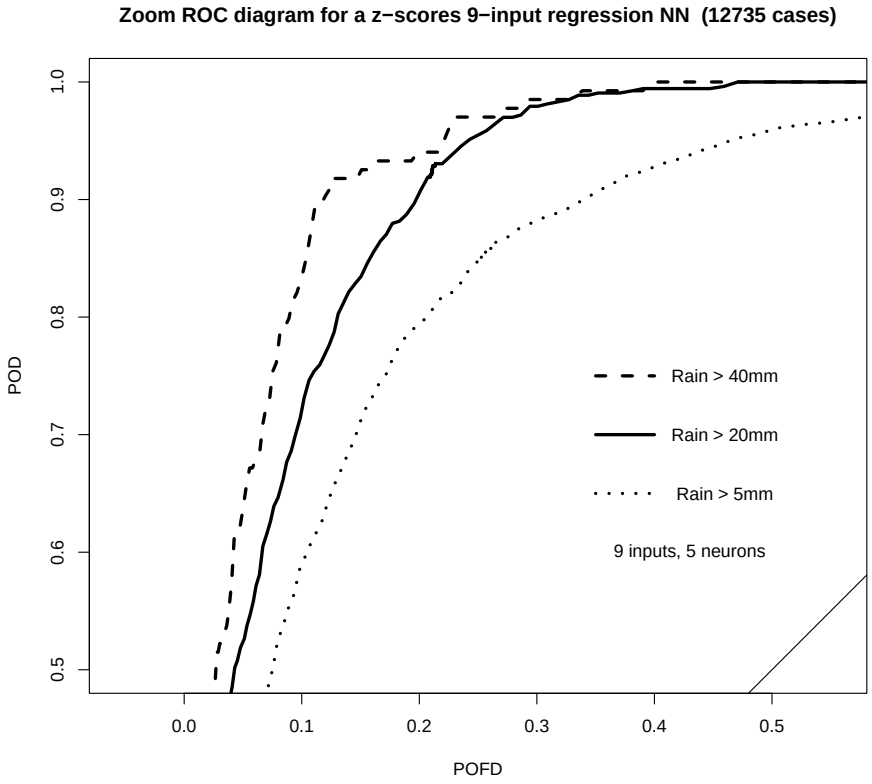

fig. 12a

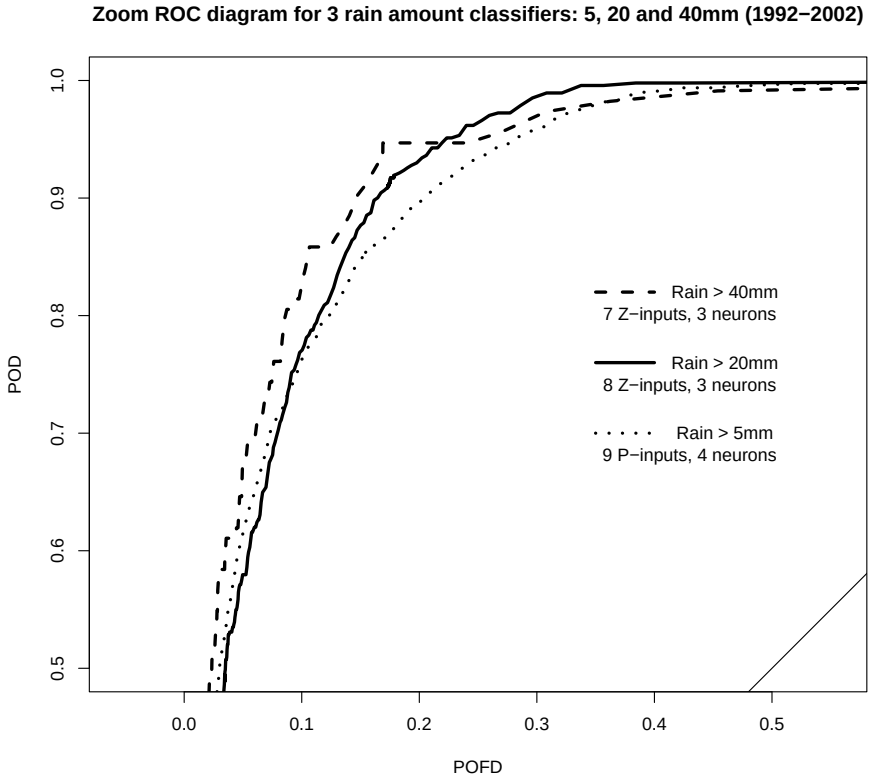
Zoom ROC diagram for 3 rain amount classifiers: 5, 20 and 40mm (1992−2002)
POD
POFD
0.5
0.6
0.7
0.8
0.9
1.0
0.0
0.1
0.2
0.3
0.4
0.5
Rain > 40mm
7 Z−inputs, 3 neurons
Rain > 20mm
8 Z−inputs, 3 neurons
Rain > 5mm
9 P−inputs, 4 neurons

fig. 12b