4 Neural network approach

In chapter 2 and especially in section 2. 4, section 2. 5, and section 2. 6 several physical properties of clouds and surfaces have been pointed out. They are in principle utilizable as cloud detection tests. These suggested tests certainly do not claim completeness, but they refer to commonly utilized methods based on assumptions like clouds are high, white, bright or cold, only to enumerate some. Other tests with other channel combinations may also provide important information on the status of cloud coverage. Almost every channel as well as channel combination and therefore almost every cloud detection test is dependent on the viewing angle and in relation to solar channels also on the sun angles. Furthermore, most of these tests are not independent from each other. It can be suggested, that the discrimination between clouds and surfaces utilizing data from the SEVIRI channels forms an under-determined problem in principle, so that every textural, temporal and/or spectral cloud detection technique could produce ambiguous results in some cases. Therefore, a cloud detection output representing cloud covered probabilities instead of boolean values is reasonable. The underlying decision logic should be a diffuse, statistical logic, a so-called “fuzzy” logic. As e.g. described in *Rojas* [1993], artificial neural networks of a multilayer perceptron architecture with

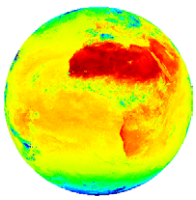
neurons having a sigmoidal output function are suitable for approximating such problems. The learning procedure of the neural network is interpretable as a process of statistical regression.

In the topic of remote sensing, neural networks are utilized as fuzzy classifier for land usage types in the publications of *Foody et al.* [1997] and *Atkinson et al.* [1997] among others. A general overview of the usage of neural networks in remote sensing can be found in *Atkinson and Tatnall* [1997]. Examples of the usage of neural network techniques for cloud detection and/or classification are *Preusker* [1999], *Schröder et al.* [2002], *Bankert* [1994], *Baum et al.* [1997], *Lee et al.* [1990], *Welch et al.* [1992], etc.

4. 1 Network topology

The trained artificial neural networks have all been of a multilayer perceptron architecture holding one hidden layer with 20 and 25 neurons, respectively, depending on the input complexity. The input layer as well as the hidden layer has been supplemented with a bias neuron giving the constant value of one. Figure 4-1 illustrates the applied network topology including data pre- and postprocessing. In this figure, *in* represents the N_i dimensional input vector containing e.g. the SEVIRI viewing geometry, channel brightness temperatures and the BT_{ACSBTE} data. The vector in' is the preprocessed input vector and represents the actual neural network input (section 4. 2). N_h gives the number of hidden neurons in the hidden layer. The weights of the dendrites connecting two adjacent layers are represented by w_{in} and w_{out}. The actual network output out' is converted to the cloud detection output *out* representing cloud covered probabilities (section 4. 3). Neural networks of this type act as universal approximators [*Hornik et al.*, 1989; *Cybenko*, 1989]. The network output out' is calculated from the input vector in' by the "recall function" according to the following equation:

$$out' = \sigma_{out}\left(w_{out\,N_h+1} + \sum_{j=1}^{N_h}\left[w_{out\,j}\ \sigma_{in}\left(w_{in\,N_i+1,j} + \sum_{k=1}^{N_i}\left[in'_k\ w_{in\,k,j}\right]\right)\right]\right) \quad (4\text{-}1)$$



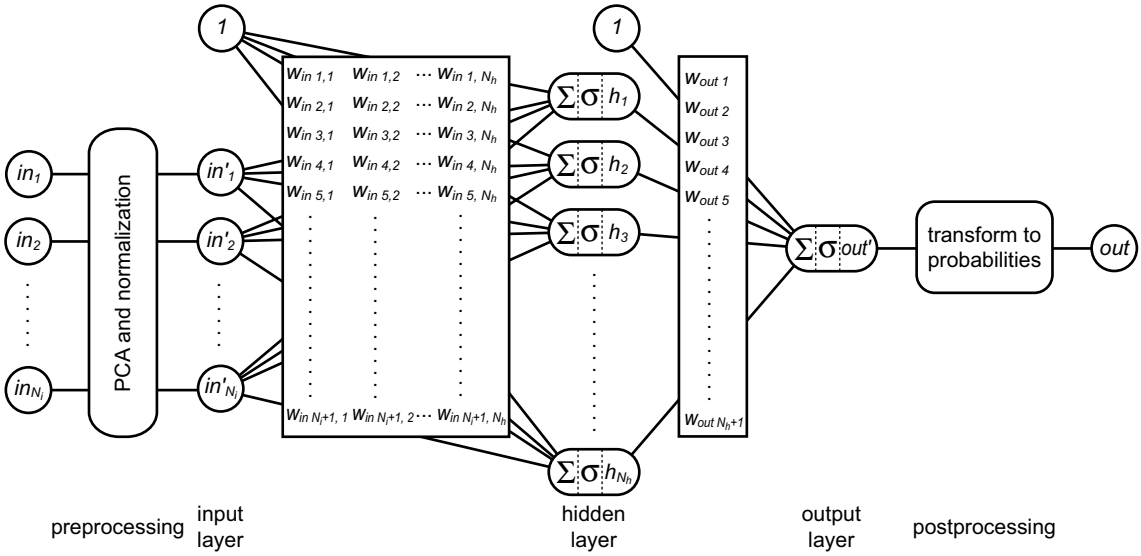


Figure 4-1: General neural network topology of all trained networks including preprocessing and postprocessing.

This equation consists of two consecutive matrix multiplications, each followed by applying a sigmoid function $\sigma_{in/out}$ of the form

$$\sigma_{in/out} = \frac{1}{1 + e^{-c_{in/out}\, x}}, \tag{4-2}$$

where $c_{in/out}$ stands for a constant depending on N_i and N_h, respectively. Equation (4-1) shows, that the number of neurons in the hidden layer N_h limits the potential complexity of the recall function. Predictions on the optimal number of hidden neurons are generally not possible, as this value is strongly problem specific [*Rojas*, 1993]. In this context, best results have been obtained using 20 and 25 hidden neurons, respectively. Due to their possibly more complex decision logic, 25 hidden neurons have been used for those networks utilizing also information from the channels in the visible spectral region. Whereas 20 hidden neurons have been used for the remaining networks.

All networks have been trained with the quick-propagation method which is a refinement of the back-propagation method and which is similar to Newton's method [*Rojas*, 1993]. This kind of training is a form of supervised learning by means of a training dataset. Aim

of the training process is to find those weight matrices w_{in} and w_{out} that minimize the root mean square difference between the output of the recall function out' and the corresponding training dataset output. The resulting recall function can be understood as statistical regression of the training dataset by means of superimposition of sigmoid functions.

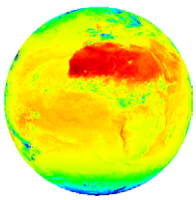
The software environment utilized for creation and training of such neural networks have been developed by Dr. René Preusker at the *Institut für Weltraumwissenschaften*.

4. 2 Training dataset

Twelve different neural networks have been trained. The network identifiers as defined in table 4-1 describe the networks' situation of application. For each combination of **day**/**nig**ht/**twi**light, **lan**d/**sea**, and the availability of reliable $BT_{\boldsymbol{ACS}BTE}$ information one network has been trained. This situation dependent separation is meant for simplifying the learning process of each single network. The sun zenith angle separates day (<85°), twilight (between 85° and 90°), and night (>90°). The sun glint angle has been defined as deviation from the angle of total reflectance. It indicates possible sun glint on sea surfaces when less than 30°. Both networks for sea surfaces under twilight conditions are designed to handle also sun glint cases. Table 4-1 gives an overview of all trained networks, their network identifier, their situation of application, their input vector composition, and their quality regarding to the training dataset (section 4. 3).

The training dataset has been created by manual cloud classification within SEVIRI images. Its quality is decisive for the reachable quality of the recall functions.

A graphical user interface has been developed, providing the user with extremely magnified 25 times 25 pixels measuring image fragments for manualy classifying pixel by pixel. These fragments are randomly chosen in date, time of day, and position. A manual region selection capability was only implemented to ensure that the final dataset also includes seldom special cases like snowy surfaces.



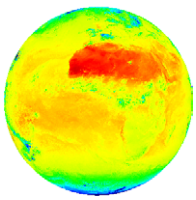
network identifier	situation			network input composition												output quality		
	(l)and / (s)ea	(d)ay / (n)ight / (t)wighlight or sun glint	BT_{ACSBTE} available	BT_{134}	BT_{120}	BT_{108}	BT_{ACSBTE}	BT_{087}	BT_{039}	NIR_{016}	VIS_{008}	VIS_{006}	satellite zenith angle	sun zenith angle	sun glint angle	classifications with confidence > 0.80 [%]	classifications with confidence > 0.90 [%]	classifications with confidence > 0.95 [%]
day_lan_acs_nn	l	d	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	99.8	99.5	99.2
day_lan_nn	l	d	-	✓	✓	✓	-	✓	✓	✓	✓	✓	✓	✓	✓	97.8	96.9	95.1
twi_lan_acs_nn	l	t	✓	✓	✓	✓	✓	✓	✓	-	-	-	✓	✓	✓	97.4	95.4	92.7
twi_lan_nn	l	t	-	✓	✓	✓	-	✓	✓	-	-	-	✓	✓	✓	94.4	91.3	90.0
nig_lan_acs_nn	l	n	✓	✓	✓	✓	✓	✓	✓	-	-	-	✓	-	-	98.6	97.0	95.6
nig_lan_nn	l	n	-	✓	✓	✓	-	✓	✓	-	-	-	✓	-	-	98.1	96.6	93.6
day_sea_acs_nn	s	d	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	99.2	99.1	98.2
day_sea_nn	s	d	-	✓	✓	✓	-	✓	✓	✓	✓	✓	✓	✓	✓	98.8	98.1	95.6
twi_sea_acs_nn	s	t	✓	✓	✓	✓	✓	✓	-	-	-	-	✓	-	-	97.9	96.0	94.5
twi_sea_nn	s	t	-	✓	✓	✓	-	✓	-	-	-	-	✓	-	-	94.3	89.8	81.5
nig_sea_acs_nn	s	n	✓	✓	✓	✓	✓	✓	✓	-	-	-	✓	-	-	98.0	97.3	94.4
nig_sea_nn	s	n	-	✓	✓	✓	-	✓	✓	-	-	-	✓	-	-	96.3	94.7	91.9

Table: 4-1: Overview of all trained neural networks including network identifier, situation of application, input vector composition, and output quality.

The classification software recommends the user to classify at least 80% of each fragment, so that also several “critical” cases may be learned by the neural networks. Histogram equalization, intensity scaling between 5% and 95% percentiles and/or color inversion are applicable on all displayed data for visualization of special features. Additionally, it is possible to switch between two consecutive images to identify clouds by their motion above ground.

#	channel composition	day	night	physical bases	brief description
1	b/w: $-BT_{108}$	✓	✓	section 2. 5. 1 section 3. 1. 2	surface (warm target)→black; cloud (cold target)→white
2	b/w: BT_{ACSBTE}-BT_{108}	✓	✓	section 3. 1. 5 section 3. 3	surface (low diff.)→black; clouds (high diff.)→white
3	b/w: $(BT_{134}$-$BT_{108})$ $-1.5(BT_{120}$-$BT_{108})$	✓	✓	section 2. 5. 3 section 2. 5. 2	surface (low target)→black; cloud (high target)→white
4	b/w: BT_{108}-BT_{087}	✓	✓	section 2. 5. 4 *EUMETSAT* [2004e]	non-xeric surface→gray; fog/low stratus/desert→white; high cloud→black
5	b/w BT_{108}-BT_{039}	-	✓	section 2. 5. 5 *EUMETSAT* [2004e]	surface→gray; fog/low stratus→white; high cloud→black
6	red: BT_{120}-BT_{108} green: BT_{108}-BT_{087} blue: BT_{108}	✓	✓	section 2. 5 *EUMETSAT* [2004e]	desert→white; dust cloud→pink; thin cirrus/contrails→black
7	red: BT_{120}-BT_{108} green: BT_{108}-BT_{039} blue: BT_{108}	-	✓	section 2. 5 *EUMETSAT* [2004e]	surface (in general)→pink; surface (snow)→red; fog/low stratus→green; thin cirrus→black
8	b/w: VIS_{006}	✓	-	section 2. 5. 6	non-snowy and non-xeric surf. (low reflectance)→black; cloud (high reflect.)→white
9	true color composite (NIR_{016}, VIS_{008}, VIS_{006} col. rot., BT_{108} weighted)	✓	-	section 2. 5	clouds→white; non-snowy surface→green/brown; sea surface→black/dark blue
10	red: VIS_{006} green: NIR_{016} blue: IR_{039}	✓	-	section 2. 5 *EUMETSAT* [2004e]	non-snowy surf.→green/blue; snow→red; fog/low stratus→white
11	red: VIS_{006} green: IR_{039} blue: BT_{108}	✓	-	section 2. 5 *Rosenfeld* [2000] *Rosenfeld and Lensky* [1998]	non-snowy surface→blue; snow→red; fog/low stratus→white

Table: 4-2: This table shows the channel compositions implemented in the classification software. Polychromatic compositions are defined by their red, green, and blue components. Also given is the applicability at day- and nighttime, a very brief description, and references to the physical background.



Several different channel combinations give indications of the cloud coverage by highlighting different physical aspects. The choice of suitable channel combinations enables the creation of a training dataset of high quality and is consequently a substantial basis of the resulting algorithm. The compositions of all implemented channel combinations have been chosen according to the investigations in chapter 2 and according to publications like *EUMETSAT* [2004e], *Rosenfeld and Lensky* [1998], and *Rosenfeld* [2000]. An overview of these channel combinations is given in table 4-2. Additionally, figure 4-2 shows some of them to illustrate their specific characteristics applicable for discrimination of clouds from surface. Especially difficult cases like stratus clouds or fog at nighttime, thin cirrus, and snowy surfaces are highlighted.

Another important aspect of the training dataset is the amount of included information, because the whole variability of the input data space has to be covered. In this case, the correlation matrix of the input dataset should have been converged [*Preusker*, 1999].

In order to investigate this convergence, repeatedly two non-intersecting partial quantities have been randomly chosen from the whole manually classified input dataset of almost 90000 input vectors. Depending on the size of the partial quantities (number of elements of random samples), figure 4-3 shows the predictand and the standard deviation of the maximum difference of both correlation matrixes, at each time based on 100 random samples. With roughly 15000 elements, the standard deviation of the maximum difference of correlation coefficients is lowered to less than 0.01. For even larger partial quantities weak convergence on zero may be presumed. The twelve networks of table 4-1 have been trained with datasets of approximately 18000 to 53000 elements. 10% of all manually classified data have been separated from the training dataset and have only been used as test dataset. Its investigation during training gives important information whether the network is still generalizing the problem or if it is only memorizing the training dataset [see also *Geman et al.*, 1992].

The ratio of cloud covered cases per cloud free plus cloud covered cases was set to 0.5 by means of randomly chosen duplication.

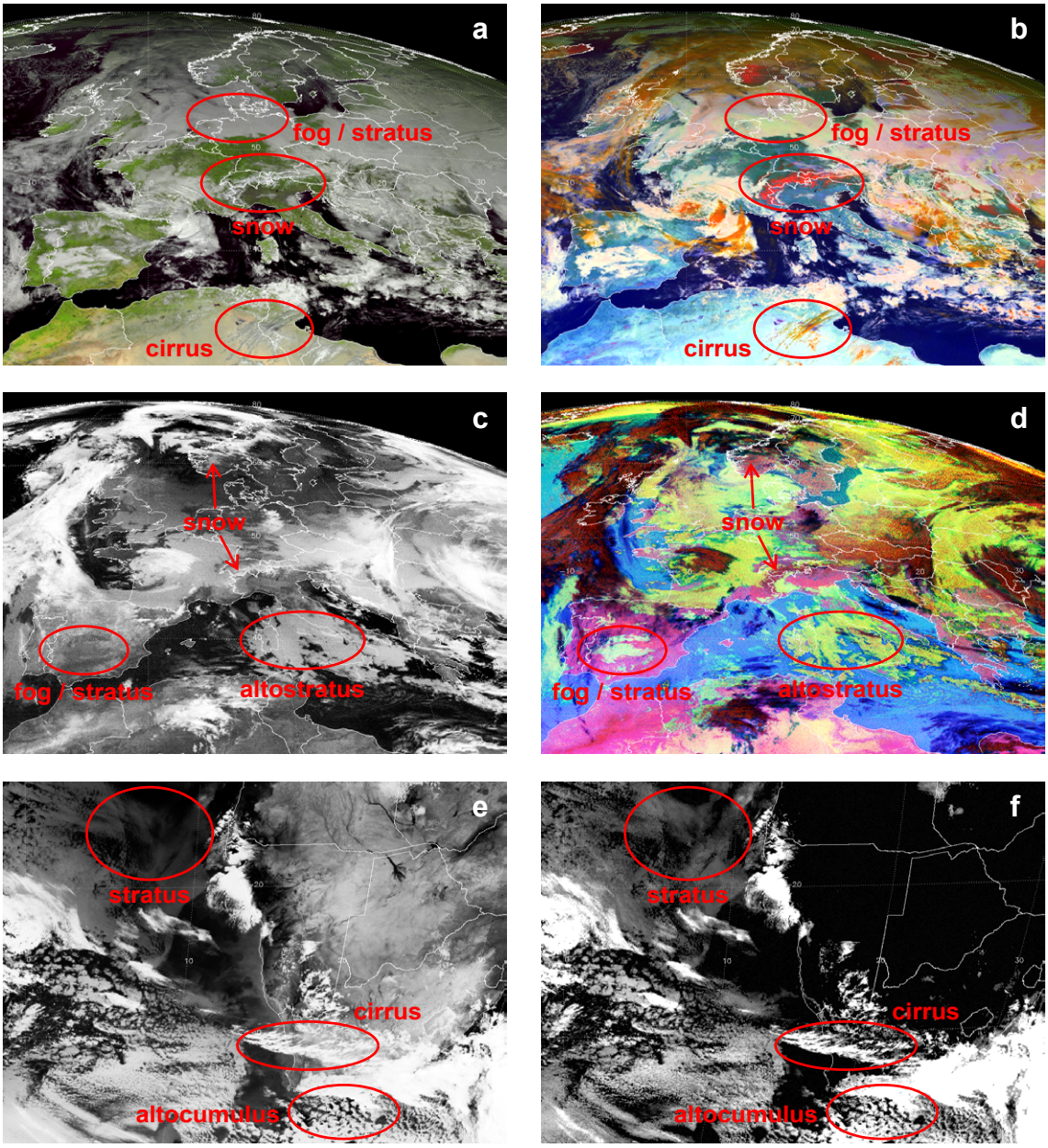
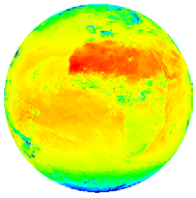


Figure 4-2: Some exemplary channel combinations given by table 4-2 are illustrated in this figure. Europe on February, 16th 2004, 12:00 UTC is shown in **a** and **b**. These images correspond to channel combination 9 and 10 (histogram equalized), respectively. In **b**, thin cirrus is highly visible due to its temperature and phase. Snow is obviously different from fog due to its low reflectance at $1.6\,\mu m$ and $3.9\,\mu m$ (section 2. 5. 5, section 2. 5. 8, and section 2. 5. 6). **c** and **d** correspond to histogram equalized channel combination 3 and 7, respectively and show the same region but at 00:00 UTC. In **c**, most clouds are visible due to their height (section 2. 5. 3), but snow and very low clouds are only distinguishable in **d** due to their different emissivities at $3.9\,\mu m$ (section 2. 5. 5). **e** and **f** show southern Africa on May, 2nd 2004, 00:00 UTC with channel combination 1 (histogram equalized) and 2, respectively. Several cloud features become only visible in **f**, when comparing BT_{108} to BT_{ACSBTE} (section 3. 1. 5 and section 3. 3).



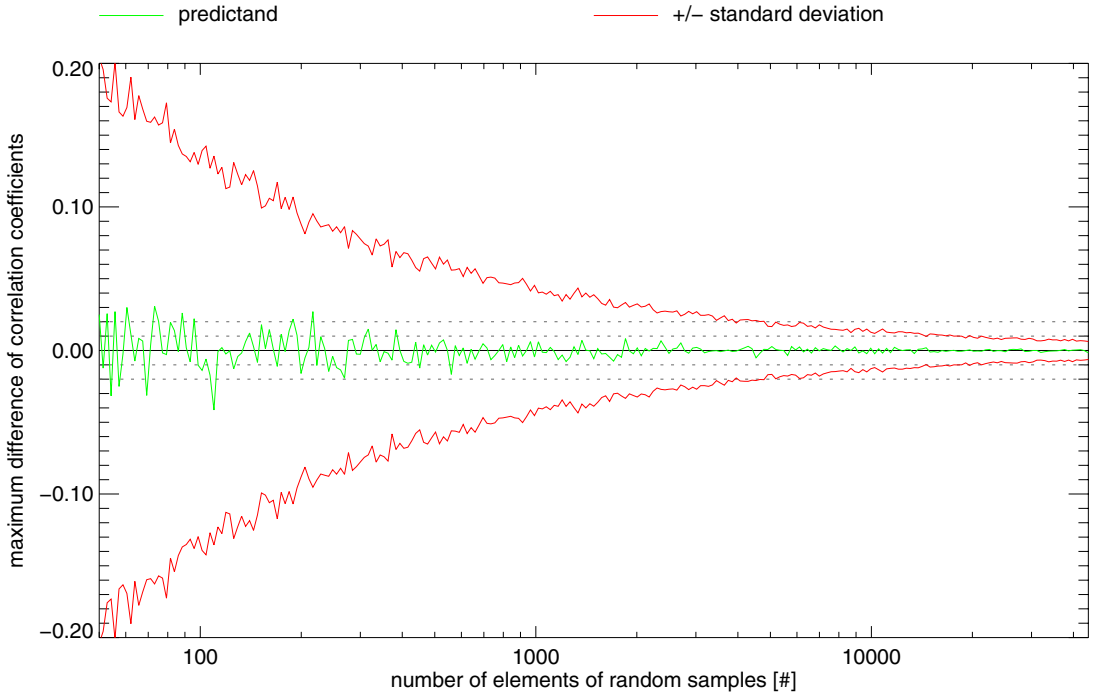


Figure 4-3: The convergence behavior of the correlation matrix, depending on the size of the investigated partial quantities of the input dataset is illustrated in this figure. The predictand and the standard deviation of the maximum difference of correlation coefficients, investigating at each point 100 times two randomly chosen non-intersecting partial quantities can be seen.

In order to simplify the process of training, all input data is preprocessed by principle component analysis (PCA). This preprocessing is applied to each input vector element except the sun and viewing geometry angles.

4. 3 Network output

The efficiency of the resulting recall functions can be determined regarding to the corresponding training and test datasets. It has to be analyzed how well the recall function agrees with the trained output. Figure 4-4 shows the normalized frequency distribution of

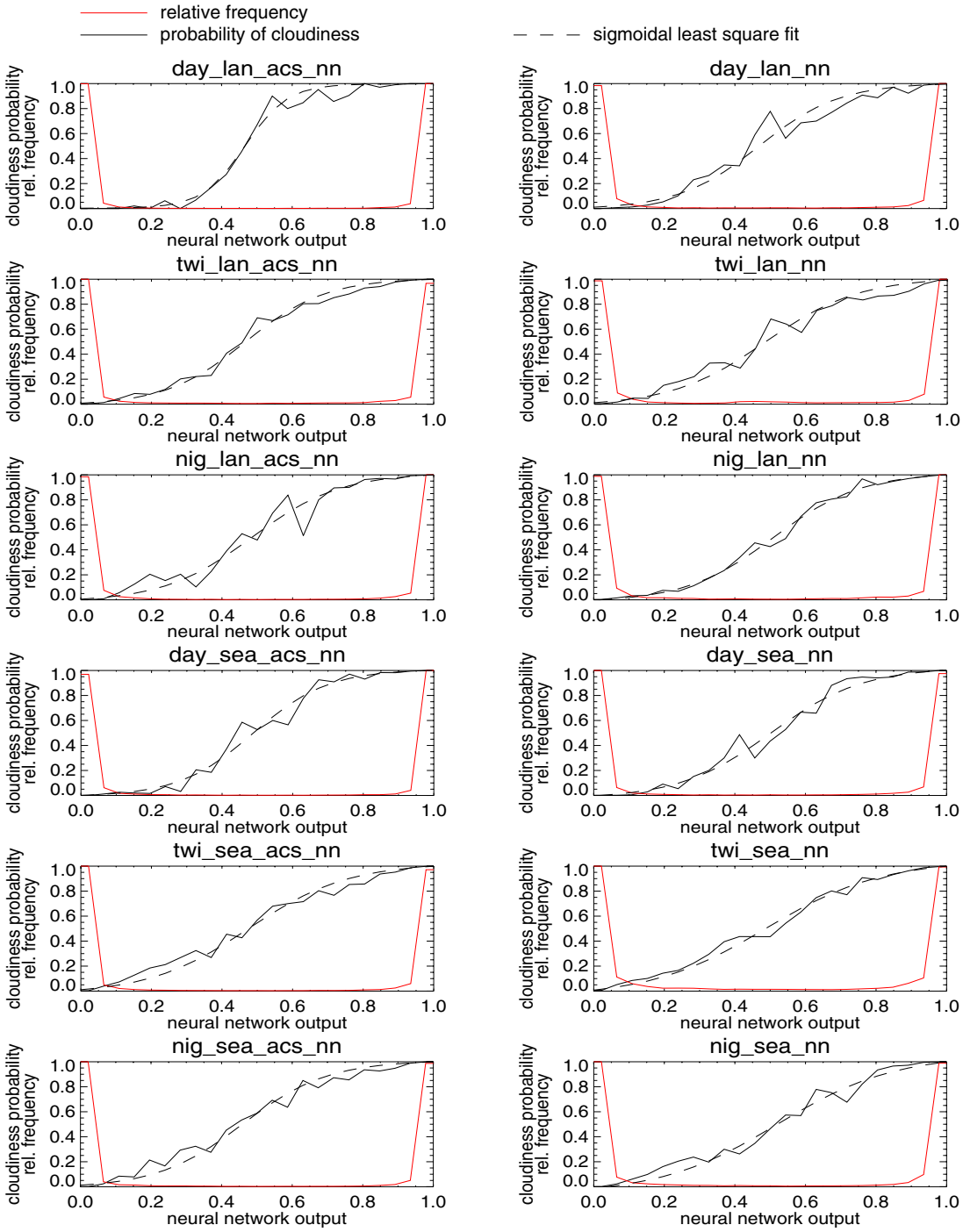
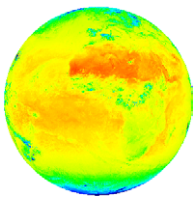


Figure 4-4: Neural network output quality in reference to the training and test dataset. Depending on the network output (in 25 bins) the normalized frequency distribution, the cloud covered probability in reference to the training and test dataset, and the corresponding fit with a sigmoid function are illustrated. All networks using BT_{ACSBTE} are placed on the left side.



network output out' for all trained networks. It also illustrates, depending on the network output, the corresponding cloud covered probability p_{cc} calculated from the number of manually classified cloud free (n_{cf}) and cloud covered (n_{cc}) cases:

$$p_{cc}(out') = \frac{n_{cc}(out')}{n_{cc}(out') + n_{cf}(out')} \tag{4-3}$$

The result of a weighted least square fit of p_{cc} with a sigmoid function σ_{fit} is also illustrated in this figure. By means of the ratio p_{cc} a confidence value c_{onf} can be defined:

$$c_{onf}(out') = \left|p_{cc}(out') - 0.5\right| + 0.5 \tag{4-4}$$

This value specifies the certainty in how far it is possible to assign a cloud coverage status to a given network output.

Ideally, the histograms in figure 4-4 should show the value 1 only for the network outputs 0 and 1, and for all other cases they should show the value 0. The corresponding cloud covered probabilities should be something like a step or logistic (e.g. sigmoid) function through the points (0, 0) and (1, 1). Accordingly, it should ideally be possible to assign a cloud covered status with a confidence value of 1 to every network output.

As assessment of quality of the trained networks, table 4-1 gives the percentage of the training and test datasets which could be classified with a confidence value greater than 0.80, 0.90, and 0.95, respectively. It shall be emphasized that each network using BT_{ACSBTE} produces distinctly better results than the corresponding networks working without that data. The training datasets have been identical except for the BT_{ACSBTE} data. This fact indicates, that the calculation of BT_{ACSBTE} obviously increases the amount of independent information. This effect is best visible at the networks for sun glint and twilight conditions over sea surfaces (*twi_sea_nn* and *twi_sea_acs_nn*), as these networks are the only ones with input vectors containing solely information from the thermal channels and the viewing geometry (table 4-1) to enable usage also under sun glint

conditions. Table 4-1 shows that the average ACSBTE network classifies 95.8% and that the average non-ACSBTE network classifies merely 91.3% of all training and test data with a confidence value greater than 0.95.

To obtain comparable results from all networks and to make these results physically interpretable, a postprocessing was set up, applying the fitted sigmoid function σ_{fit} on the output out' of the recall function:

$$out = \sigma_{fit}(out') \tag{4-5}$$

The final result out of the cloud detection is now interpretable as probability of cloud coverage in reference to the training and test dataset.

A cloud detection result of e.g. 0.98 now represents a cloud covered probability of 0.98 independent of the quality of the utilized network. Consequently, networks with lower quality (table 4-1) should ideally only lead to a higher amount of results with low confidence values.