Standard back-propagation artificial neural networks for cloud liquid water path retrieval from AMSU-B data

J. M. Palmer *,F. Romano*, V. Cuomo*
* Institute of Methodologies for Environmental Analysis, National Council Research, Potenza, Italy

Artificial neural networks (ANNs) have many variants, which can have very different behaviours. The success of ANNs to produce good results for a wide variety of problems when little is known about the search space has lead them to become of interest to many scientific disciplines. Ideally if a problem is tested for the first time with an ANN methodology then this methodology should be standard. However this may be problematic for a number of reasons. It is difficult to know the best configuration of parameters for the learning algorithm. Results from individual runs can be irregular. There may be a very large amount of training data making training slow. These problems often cause researches to diverge from the standard back propagation method.

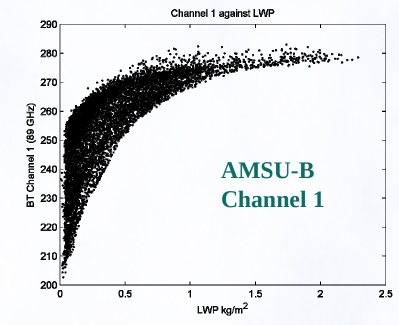
The objective of this study is to test ANN methodology for the problem of cloud liquid water path (LWP_C) derivation, using the advanced microwave sounding unit B (AMSU-B) microwave brightness temperatures. The vertically integrated cloud liquid water, also known as LWP_C plays a key role in the study of global atmospheric water circulation and the evolution of clouds. The ability to derive LWP_C accurately and across large areas therefore means better atmospheric models can be built and tested. Simulated AMSU-B and LWP_C data is fitted using linear, polynomial and standard ANN methods. The ANN method performed the best and gave an average RMS error for between 0.06 and 0.02 kgm^{-2} dependent on the environment.

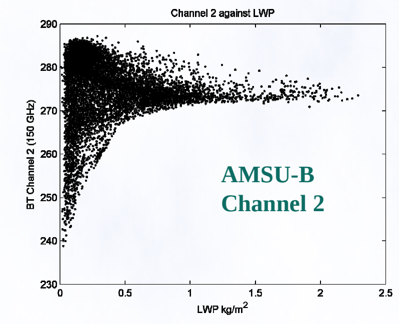
Simulated AMSU-B data

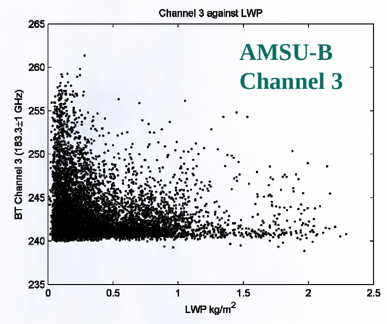
Data used to train any methodology that builds a model from data should be comprehensive. ***The data should express the various facets of the problem*** and its complexity, it should also be numerous enough to be split into training, validation and test sets [8]. Fortunately using simulated data allows a large and detailed data set to be constructed.

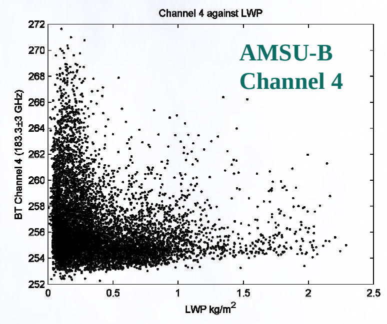
The AMSU-B passive microwave channels simulated using ECMWF [11] atmospheric profiles and with the RTTOV code. The data is split into 4 sets representing different problems dependent on the surface type and season, these are **land-summer (LS)**, **land-winter (LW)**, **ocean-summer (OS)** and **ocean-winter (OW)**. The ECMWF cloud liquid water profiles (*clw*) at 60 atmospheric levels, have been integrated to calculate the LWP_C,

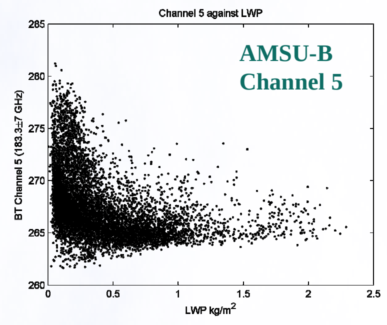
$$LWP_C = \frac{1}{g}\int clw \cdot dp$$









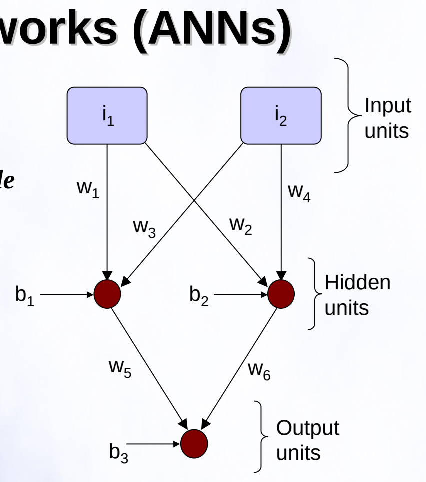


Here the LS simulated brightness temperatures are shown against cloud liquid water path.

Artificial Neural Networks (ANNs)

The term 'artificial neural network' is more historical than descriptive and in fact refers to the original inspiration for their development [6]. ***An ANN is a multivariable function in continuous space***. It is more concisely depicted graphically but each output can also be written as a function of the ANN inputs.

Changing the free parameters of an ANN changes the function it produces. ANN free parameters are called **biases** and **weights**.



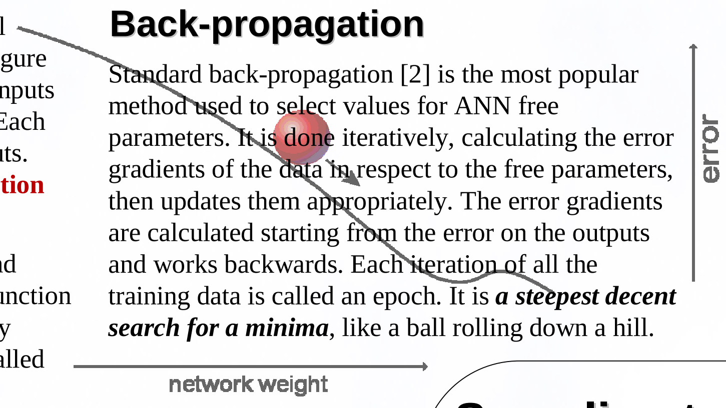
A standard feed forward artificial neural network is shown in the figure to the left and has 3 layers. The inputs biases and weights are labelled. Each unit outputs a function of its inputs. This function is called the **activation function**.

Selecting values for the biases and weights is done to fit the ANN function to data. However this is not easily done… the standard method is called **back-propagation**.

$$Output = f(b_3 + w_5 \cdot f(b_1 + w_1 \cdot i_1 + w_3 \cdot i_2) + w_6 \cdot f(b_2 + w_2 \cdot i_1 + w_4 \cdot i_2))$$

Back-propagation

Standard back-propagation [2] is the most popular method used to select values for ANN free parameters. It is done iteratively, calculating the error gradients of the data in respect to the free parameters, then updates them appropriately. The error gradients are calculated starting from the error on the outputs and works backwards. Each iteration of all the training data is called an epoch. It is ***a steepest decent search for a minima***, like a ball rolling down a hill.

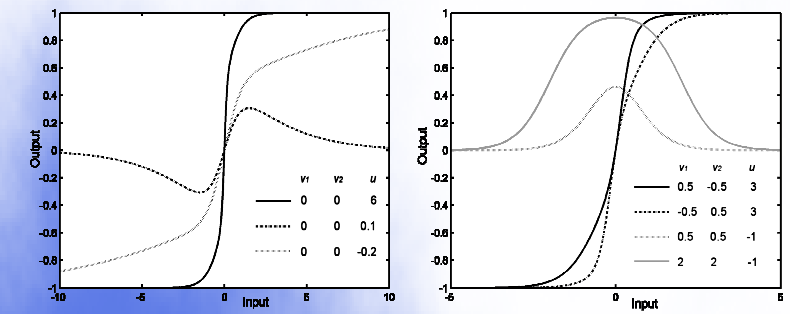


Parameterised activation function

The activation function is the building block of the ANN, typically the sigmoid or hyperbolic tangent functions are used.

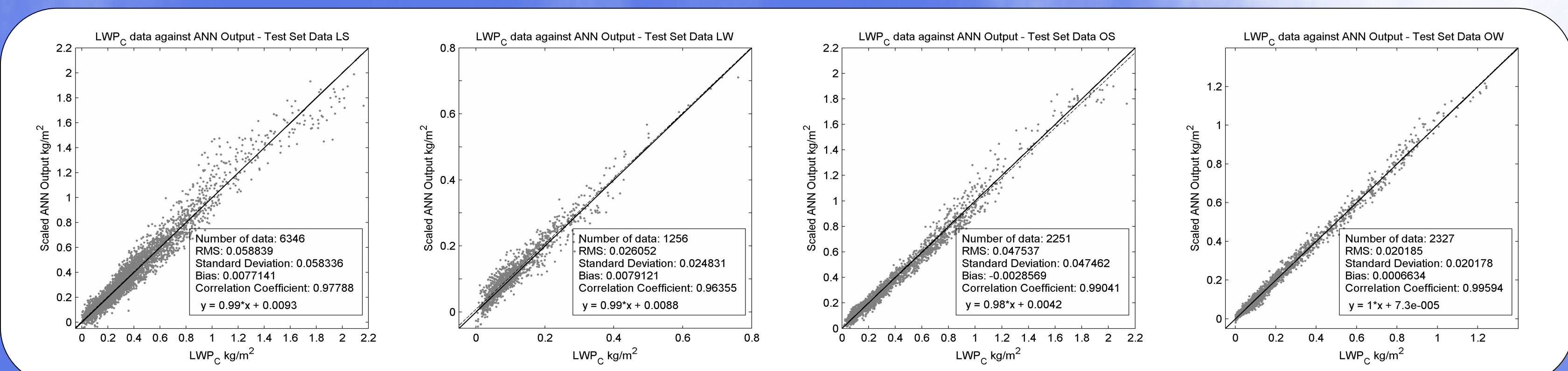
A parameterised activation function has been developed which can represent a high number of possibilities such that can be easily tested and referenced. The figures show a selection of configurations. The function is as follows, *a* and *b* control the scaling and are both normally set to 1.

$$y = \frac{a}{2}\left\{\tanh\left(bx + v_1\right) + \tanh\left(ubx + v_2\right)\right\}$$



ANN Results

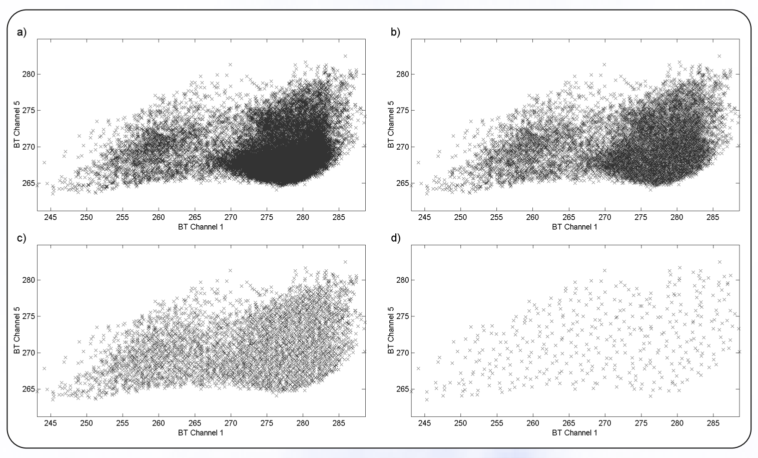
The ANN architecture used was **5-20-5-1** for all networks, this notation references the number of units in each layer from input to output. The hyperbolic tangent activation function is used for all output units, and also for the hidden units in the LW case. The hidden units in the LS, OS and OW cases use the parameterised activation function with v_1=0.5, v_2=-0.5 and u=3. A linear fit of the results is also shown .



Sampling training and validation data

Back-propagation is slow, particularly when there is a large amount of training data. Sampling training and validation data sets is done to reduce their size. This can be done because ***the form of the data is more important than just the quantity***.

Within this study a simple method has been implemented to simple data sets. Each dimension of the data is split into *S* parts, for *d* dimensions, this makes S^d subspaces. Various statistical measures can then be used to select *n* points from each of the subspaces. The plots give a simple example of this. One point is taken at random from each of the subspaces considering only 2 dimensions of the simulated LS data set. a) shows all 25384 data points b) *S*=200 and so 8457 points are selected, c) *S*=100, selecting 3755 points, d) *S*=25 selecting 398 points.



Benchmark Testing

Data Set	S	Training Set			Validation set			Test set		Mean epochs	
		№	Mean	SD	№	Mean	SD	Mean	SD	Last	Val
cancer1	-	350	2.93	0.43	175	1.88	0.33	1.6	1.15	202	137
cancer2	-	350	1.94	0.04	175	1.61	0.05	3.16	0.04	199	81
cancer3	-	350	1.50	0.07	175	2.82	0.06	2.56	0.09	201	89
diabetes1	-	384	15.36	0.75	192	16.42	0.44	18.4	1.94	140	62
diabetes2	-	384	16.3	1.53	192	19.06	0.99	20.33	2.63	132	25
diabetes3	-	384	15.09	1.14	192	18.46	0.64	17.46	1.73	140	51
glass1	-	107	6.99	1.88	54	9.35	0.3	9.42	0.37	230	106
glass2	-	107	6.62	0.89	54	10.1	0.32	10.45	0.35	161	36
glass3	-	107	6.77	1.42	54	9.1	0.24	10.76	0.48	262	144
building1	-	2104	0.13	0	1052	0.81	0.02	0.69	0.02	414	341
building2	-	2104	0.30	0.02	1052	0.34	0.02	0.31	0.02	225	225
building3	-	2104	0.32	0.02	1052	0.31	0.02	0.31	0.02	234	234
cancer1	5	187	4.26	1.25	89	3.41	0.21	**1.42**	0.16	260	149
cancer2	5	177	3.17	0.16	97	3.11	0.08	3.36	0.12	192	83
cancer3	5	167	3.43	2.52	100	5.31	1.41	3.32	2.28	163	53
diabetes1	4	263	19	1.18	164	17.42	0.51	19.09	2.58	190	78
diabetes2	4	256	18.64	0.54	143	20.85	0.53	**18.79**	1.17	119	17
diabetes3	4	248	17.97	1.38	154	21.01	0.73	19.12	2.33	195	51
glass1	4	70	5.67	2.28	44	9.21	0.55	9.68	0.51	298	177
glass2	4	62	6.18	0.77	45	10.69	0.34	10.59	0.53	176	45
glass3	4	66	5.67	1.18	43	9.02	0.24	10.77	0.57	208	81
building1	2	400	0.18	0.02	290	0.85	0.02	0.72	0.03	280	280
building2	2	356	0.41	0.01	274	0.45	0.01	0.35	0.01	180	180
building3	2	350	0.44	0.01	294	0.42	0.01	0.36	0.01	192	192

Benchmark testing is done on all the methodology used in the study. This is not just to test the functionality and implementation of the ANN, but it is also used to test how well the extra methodology extends to problems and data already documented. Here is an example of benchmark testing for the sampling algorithm using PROBEN1 [8] benchmark problems. Firstly each of the problems is tested using their full data sets. In each case the number of data in the reduced sets is shown. The errors shown are the ANN errors[8], they are unit less and each represents 10 runs of the ANN. The standard deviation is labelled SD. Note that the test sets are always the same.

These results are better than expected, the PROBEN1 data sets are already small and it was expected that most of the data was needed to express the problems. ***The test was to see if a smaller data set could be constructed for an experimental phase of work, without sacrificing the behaviour of the ANN and causing a large increase in the test error***. Surprisingly for two of the problems the test error decreased. Notice also that the glass problems still performed well for such small data sets. The larger data sets (building) showed the largest reduction in data and again gave good results.

Method comparison with least squares solution

The least squares fit solution is done using ***matrix methods***. Firstly only AMSU-B channels 1 and 2 are used because they show the most correlation with the LWP_C. Then secondly all 5 channels are used.

Both ***linear and quadratic models*** are tested giving 4 functions as shown below. The 5 AMSU-B channels are labelled C_1 to C_5 and the free parameters are the *a* coefficients. The quadratic forms are written with *b* coefficients because they are inside the squared term. It is however, the *a* coefficients that are fitted, these are the coefficients after the expansion. Notice that the least squares fit is linear in the coefficients but not necessarily in the terms.
The functions are,

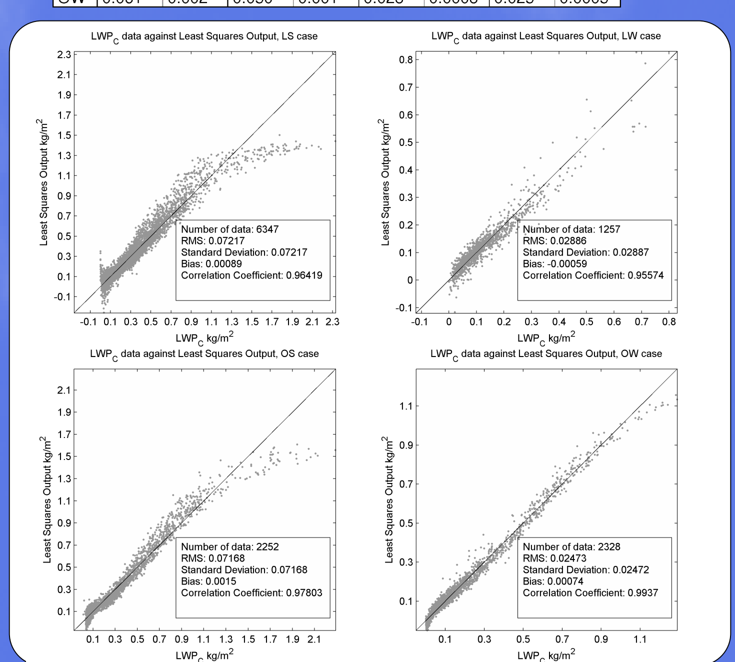
$$f(C_1, C_2) = a_1 + a_2C_1 + a_3C_2$$
$$f(C_1, C_2, C_3, C_4, C_5) = a_1 + a_2C_1 + a_3C_2 + a_4C_3 + a_5C_4 + a_6C_5$$
$$f(C_1, C_2) = (b_1 + b_2C_1 + b_3C_2)^2$$
$$f(C_1, C_2, C_3, C_4, C_5) = (b_1 + b_2C_1 + b_3C_2 + b_4C_3 + b_5C_4 + b_6C_5)^2$$

And the coefficients are calculated,

$$\underline{a} = \left(\underline{\underline{K}}^T\,\underline{\underline{K}}\right)^{-1}\underline{\underline{K}}^T\,\underline{lwp_C}$$

The matrix K is constructed with the function terms as columns and data as rows. The following table shows these results in kgm^{-2}. Using all 5 channels with the quadratic model did the best. These results are plotted for all cases of surface and season.

Data Set	Linear (C_1,C_2)		Linear (C_1,C_2,C_3,C_4,C_5)		Quadratic (C_1,C_2)		Quadratic (C_1,C_2,C_3,C_4,C_5)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
LS	0.156	0.003	0.139	0.003	0.113	0.002	0.071	0.002
LW	0.044	0.001	0.041	0.001	0.040	0.001	0.029	0.001
OS	0.156	0.005	0.134	0.004	0.099	0.003	0.072	0.003
OW	0.061	0.002	0.050	0.001	0.028	0.0008	0.025	0.0005



The ANN performs better than the least squares fit. The least squares solutions are still very good, but ***notice the ends of the fit***; where the LWP_C is very low the fit is much more inaccurate and produces more negative outputs than for the ANN. Equivalently where the LWP_C is high the fit tends to produce underestimates. Two very positive points about these results concerning the least squares fit, are that the ***ANN has many more free parameters,*** and that ***the matrix fitting method is quick***. Therefore a possible extension would be to use an intelligent methodology (discreet optimisation variant), to select/construct the terms for the fit.

References

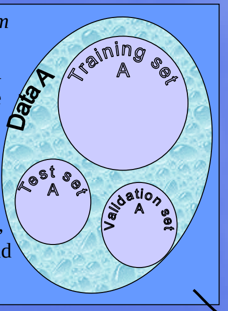
[1] Scott E. Fahlman. *An Empirical Study of Learning Speed in Back-Propagation Networks*. Technical report CMU-CS-88-162, September 1988
[2] R. Fletcher and M. J. D. Powell. *A rapidly convergent descent method for minimization*. Computer Journal, 1963.
[3] H. Hauschildt and A. Macke. *A Neural Network based algorithm for the retrieval of LWP from AMSU measurements*. Institute for Marine Research, Kiel, Germany, 2002.
[3] Thomas Jung, Eberhard Ruprecht, Friedrich Wagner. *Determination of Cloud Liquid Water Path over the Oceans from Special Sensor Microwave/Imager (SSM/I) Data Using Neural Networks*. Journal Of Applied Meteorology, Volume 37, December 1997
[4] David Landgrebe. *On Information Extraction Principles for Hyperspectral Data, A White Paper*. School of Electrical & Computer Engineering, Purdue University, July 1997.
[5] C. Mallet, E. Moreau, L. Casagrande and C. Klapisz. *Determination of integrated cloud liquid water path and total precipitable water from SSM/I data using a neural network algorithm*. International Journal of Remote Sensing, Volume 23, 2002.
[6] M. L. Minsky, and S.Papert. *Perceptrons*. MIT Press, Cambridge, MA, and London, England, 1969.
[7] Lutz Prechelt. *Early Stopping – but when?* Universitat Karlsruhe, Germany, 1997.
[8] Lutz Prechelt. *PROBEN1 – A Set of Neural Network Benchmark Problems and Benchmarking Rules*. Technical Report 21/94, September 30, 1994.
[9] www.class.noaa.gov Comprehensive Large Array-data Stewardship System (CLASS). NOAA.
[10] www.arm.gov The Atmospheric Radiation Measurement (ARM) Program.
[11] www.ecmwf.int European Centre for Medium-Range Weather Forecasts.

Real Data?

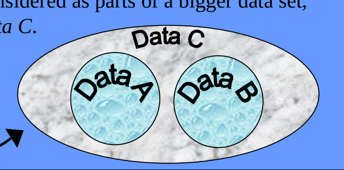
The introduction of real data into the study highlights some important points about learning from data...

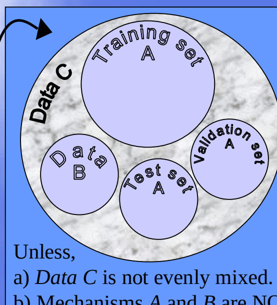
Can an ANN trained with simulated data be tested with real data?

Imagine *mechanism A* is used to create the data set, *data A* and we want to use this data to learn a model for *mechanism A*. We first split the data into *Training set A*, *Validation set A* and *Test set A*.



Now consider a *mechanism B*, which creates *data B*. If mechanisms *A* and *B* are in fact the same, then *data A* and *data B* can be considered as parts of a bigger data set, *data C*.





If a model is trained with *Training set A*, and validated with *Validation set A* such that the best trained state of the model can be selected, then testing the model with any other part of *data C*, would be equivalent…

Unless,
a) *Data C* is not evenly mixed.
b) Mechanisms *A* and *B* are NOT the same.

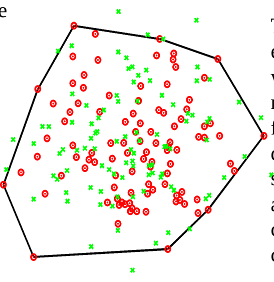
A set of real data has been constructed by co-locating data from the NOAA CLASS web site [9] with ground station data from the ARM web site [10]. A range of LWP_C has been selected from the Southern Great Plains stations Central and Hillsboro. In total the final data set contains 99 data points.

However this data set is not comprehensive enough to learn a model to represent the mechanism that created it. Some simple tests are shown below to demonstrate this.

Test of comprehensive data 1: The convex hull test

The convex hull of a data set the tessellating boundary described by the data set.

The data set is first randomly split into 2 equally sized subsets, *a* and *b*. Both these parts should represent the same function and should be similar. The test is to see what fraction of subset *b* lays within the convex hull of *a*..



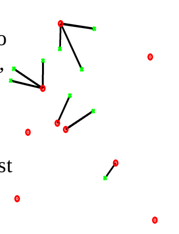
The figure shows an example for a dataset with 2 dimensions made of 200 points from a normal distribution. This is separated into subsets *a* and *b*. 87% of subset *b* can be found in the convex hull of subset *a*.

This test is run 50 times for both the real data and the simulated data (LW case). The test is done using the 5 dimensions of the brightness temperatures.

The Simulated data mean result is 92% with a standard deviation of 0.6%,
The Real data mean result is 37% with a standard deviation of 8%.

Test of comprehensive data 2: The closest point test

Again the data set is first randomly split into 2 equally sized subsets, *a* and *b*. The test is a measure of the mean distance for each point in subset *b* to the closest point in subset *a*.



This is also run 50 times for both the real data and the simulated data (LW case). The test is done using the 5 dimensions of the brightness temperatures. These distances will be measured in Kelvin.

The Simulated data mean result is 0.8K with a standard deviation of 0.007K,
The Real data mean result is 7.4K with a standard deviation of 1.2K.

What happens when we use this real data set to train an ANN?

After the real data set has been split into training, validation and test sets, the small amount of training data can be fitted very well by the ANN with a RMS error of around 0.015kgm^{-2}. This however is the error on the training data and does not consider the validation or test sets. It has been demonstrated that it would be very difficult to split this data set into subsets that are comprehensive enough to represent the same model. When all the data sets are used an the ANN is set to the point of best validation we see this is true because the test set error is around 0.25kgm^{-2}.