MORPHODYNAMIC MODELING OF THE GERMAN BIGHT USING ANN

Tim Berthold[1], Peter Milbradt[2] and Volker Berkhahn[3]

ABSTRACT

In this paper an approach is proposed, how to get a continuous description for sedimentologic measurements. Sedimentologic data has to be taken into account in order to deepen the understanding of morphodynamic processes and to improve simulation and forecasting models. Unfortunately, the data is sparsely distributed over space and time, since the measuring methods are expensive. Due to the small amount of data established interpolation and approximation methods are not suitable in this matter. The approach instead is based on an artificial neural network that is trained by the measured data. Additional information improves the performance of the model.

1. INTRODUCTION

Detailed information on morphodynamic evolution becomes more and more important in coastal engineering. A deepened understanding of morphodynamic processes is essential for coastal protection, shipping and also projects like the installation of offshore wind turbines and the connection to the onshore.

Whereas bathymetric information in terms of the depth (z) of the sea is usually available covering large areas at high spatial resolution, the amount of sedimentological measurements is much less due to the expensive measuring methods. To advance the development of forecasting models and to deepen the understanding of morphodynamic processes, such data must be taken into account. In practice, information on the sediment of the seabed is often modeled in terms of maps. Since the amount of data is not very high, the maps generally base on few measurements compared to the size of the area that they cover. A map gives a snap-shot describing the soil at a certain point in time. Time dependency cannot be captured. Typically, a map is divided into regions, where each region is classified by the configuration of the sediment. The boundaries of the regions lead to discontinuities in the map, which can cause problems in simulation models. As an example, Figure 1 shows a map of the German Bight created by Figge.

[1] Research Assistant, Institute for Computer Science in Civil Engineering, Leibniz University of Hanover, Callinstraße 34, 30167 Hanover, Germany (berthold@bauinf.uni-hannover.de)
[2] Professor, smile consult GmbH, Vahrenwalder Straße 4, 30165 Hannover, Germany (milbradt@smileconsult.de)
[3] Associate Professor, Institute for Computer Science in Civil Engineering, Leibniz University of Hanover, Callinstraße 34, 30167 Hanover, Germany (berkhahn@bauinf.uni-hannover.de)

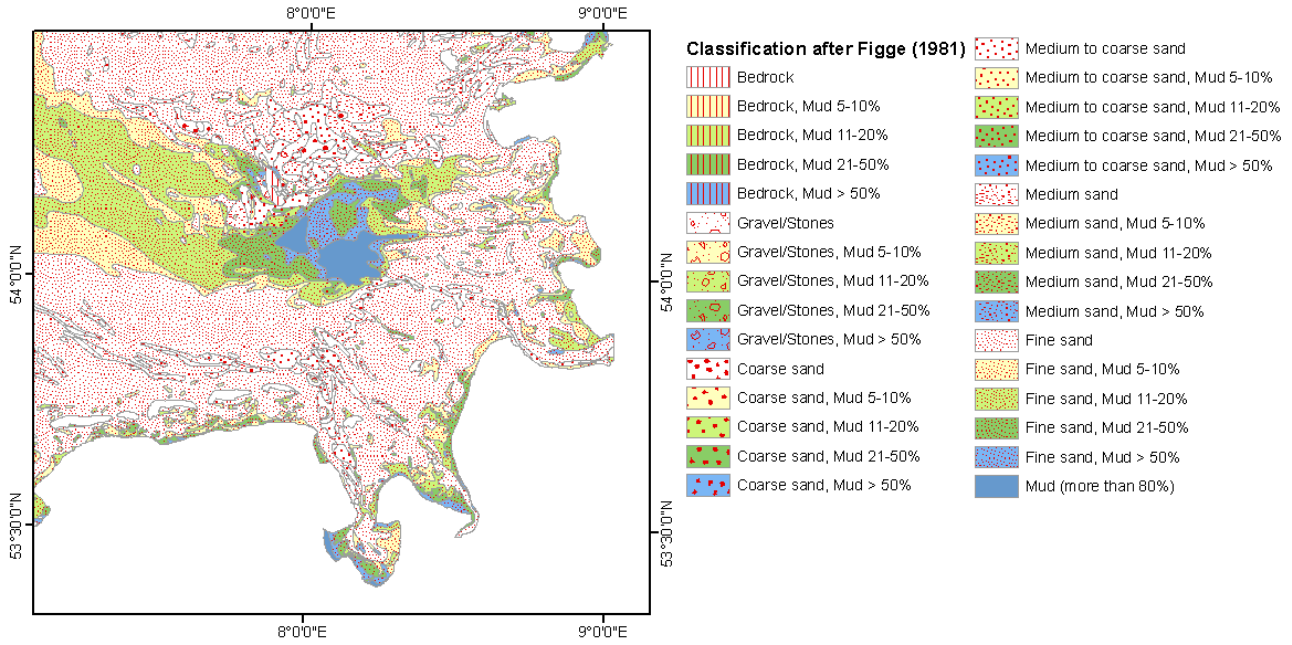

Figure 1 Map of the German Bight describing sedimentological data after classification by Figge (1981).

A consistent digital sedimentological model is desirable. Well-established interpolation and approximation models however, do not lead to an appropriate result due to the low amount of available data.

In this paper, artificial neural networks (ANN) are used to serve as a digital sedimentological model. It is shown how additional data of the bathymetry can be used to set up a model that is able to approximate sedimentological parameters quite well. ANN is a data-based method that is able to "learn" a mapping given by a set of training patterns. During a learning phase, the structure of the network is adapted by a learning rule in an iterative training process. When the training process is finished successfully, the network is able to provide information on the output parameters for any given input parameter within a given range. A more detailed introduction of ANN can be found in the numerous literature.

As training patterns, sedimentologic and bathymetric data around the estuary of the river Elbe in the German Bight will be used. The data will be described in Sect. 2. To investigate the potential and the limits of the approach some scenarios will be regarded in Sect. 3, that are subject to the following restrictions:

1. Only a scalar quantity is derived from the sediment data: as a first step, the d50 of the grain-size distribution will be used for approximation in this paper.
2. The approximation of the data is time independent: although the bathymetry is changing significantly over time, this fact will not be considered for now. As a simplification only data will be regarded, that was measured within one year.

2. DATA SET

A characteristic description of sedimentologic data is the grain-size distribution. Typically, a sample of the soil is being collected and then sieved with sieves of different size. In this way the frequency for each grain-size class can be determined, which is summarized in a histogram. Often, the cumulative curve is derived from the histogram as another representation. Because the sample is quantized, the evaluated data is a vector quantity. Figure 2 shows an example for the evaluated data. There are two different scales to describe the grain-size distribution: the phi-scale and the mm-scale. The scales can be transformed from phi to mm and the other way round (see McManus (1988)). In this paper we use the mm-scale, where the grain-size is represented by its diameter in mm.

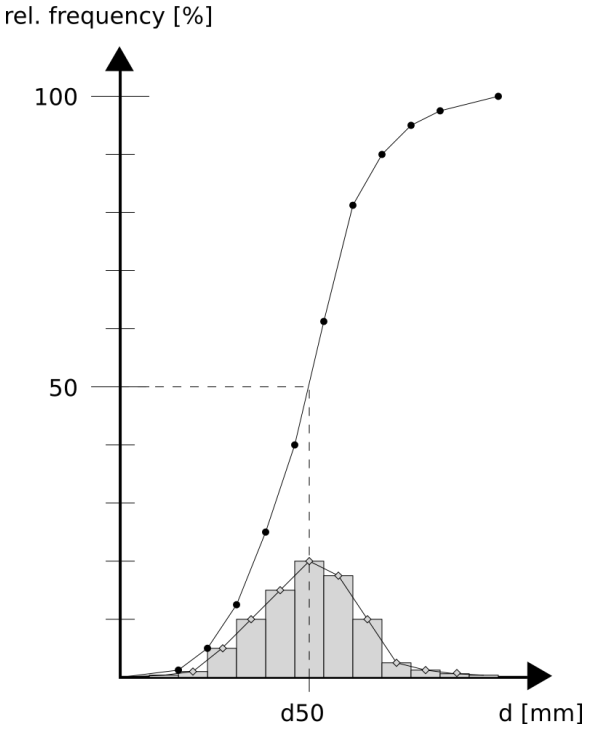

Figure 2 Exemplary representation of evaluated data in terms of a histogram, a grain-size distribution and a cumulative curve.

From the grain-size distribution some characteristic statistical parameters can be derived. The percentile dn specifies the grain diameter for which n % of the found particles are smaller. Hence the d50 describes the diameter, where half of the particles are smaller and the other half is bigger. Other parameters are the sorting coefficient or the variance. All those are scalar quantities.

For the ANN-model, we use sedimentologic data that was measured in the German Bight in about the last 100 years. As the first step is to set up a time independent approximation, the data used as training data must correspond approximately to a snap-shot. This means, that the data of interest has to be recorded nearly at the same time. The problem is that the smaller the interval, the smaller is the resulting data set. On the other hand, the interval must not be too large, because the bathymetry changes over time and the data is not representative anymore. As a compromise we use data that was recorded within the same year. In the year 2005, 473 sediment samples where recorded in the estuary of the river Elbe. For this time a bathymetric model is available also. In Figure 3 the bathymetry and the points of the sediment samples are depicted in the domain of investigation. The colors of the points represent the d50 of the sample. It can be seen, that the spatial resolution of the

sampling points is not very high. For this reason established interpolation or approximation methods are not suitable here.

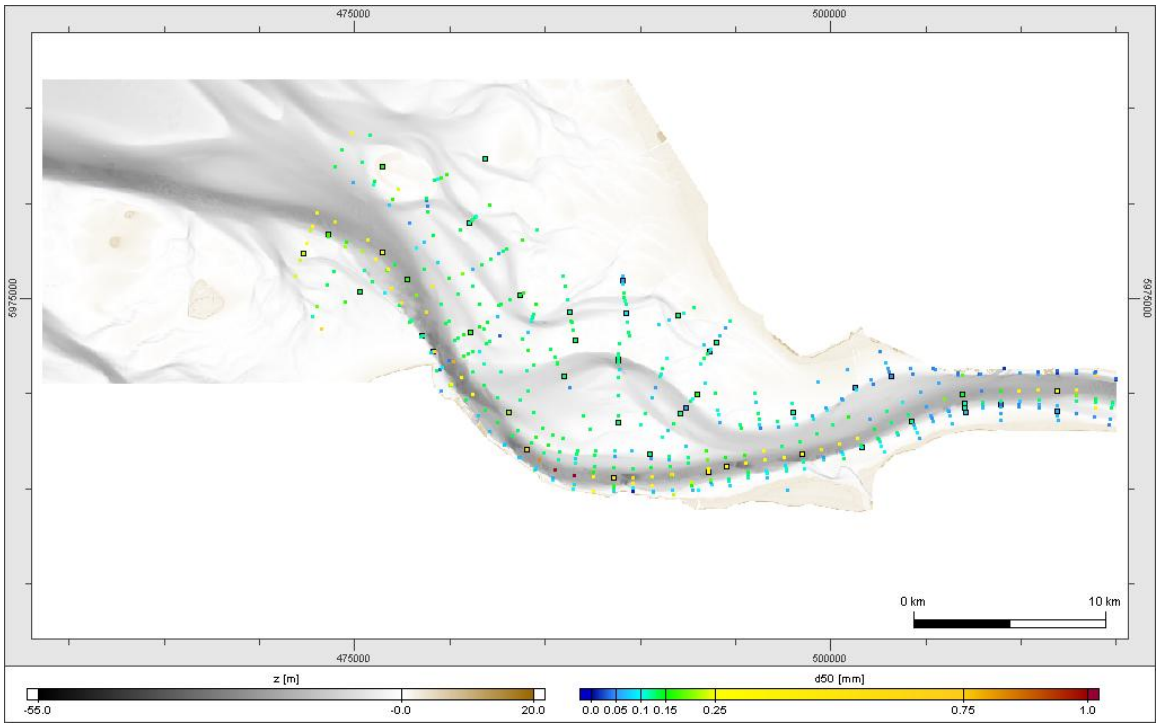

Figure 3 Domain of investigation in the German Bight (estuary of the river Elbe). The bathymetry (background) is the visualization of a bathymetric model of the region on July 1st in 2005. The colored dots represent the locations of the sediment samples that were recorded in the year 2005. The color represents the d50 of each measurement according to the given scale. 45 sediment samples are used as test patterns and are marked by a black rectangle.

3. ANN-APPROXIMATION MODEL

Using ANN as an approximation model for bathymetric data was successfully demonstrated in Berthold and Milbradt (2009). The approach described in this paper is a continuation of this idea. Again we use a feed-forward topology (compare Figure 4) with sigmoidal activation functions that is trained by a supervised learning method (a modification of the backpropagation learning rule using an additional momentum term).

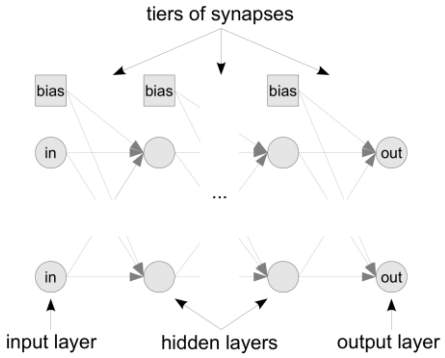

Figure 4 Exemplary illustration of the used feed-forward topology. The size of the input layer depends on the scenario, whereas the output layer consists of 1 neuron representing the d50-parameter. The sizes of the hidden layers will be varied.

The aim is to obtain a consistent approximation model of the d50-parameter for the domain of interest using the data described by Figure 3. For this purpose we will examine the following scenarios varying the input parameters. To evaluate the ability of generalization for the different scenarios, the test set depicted in Figure 3 is used. This data will never be used for training. The test set will be the same for each scenario to achieve a better comparability. Again, there is a problem due to the small size of available data: it is important to have a test set that is disjunct to the training set in order to prevent the model of overfitting. To estimate the ability of generalization, the test set must be representative and not too small. On the other hand, valuable training patterns are not available, if they are used as test patterns. The test patterns were manually picked, trying to obtain a representative test set with patterns of all kind (the variety of d50 and the position of the samples). The size of the test set is about 10 % of the available data.

Since the approximation results depend strongly on the chosen network topology, the sizes of the hidden layers are varied for each scenario. In turn each network topology will be trained 10 times to cover the influence of the randomly chosen initial synapses weights.

3.1 Dependency of the sample position

Using common interpolation and approximation methods the approximated value directly depends on its position (here (x,y)). This will be regarded in the first scenario, where patterns of the form $(x,y) \rightarrow d50$ will be used for training and of course testing. In Berthold, Milbradt and Berkhahn (2010) a geometrical interpretation of particular neurons for the mapping $(x,y) \rightarrow z$ was presented. It was shown that topologies with relatively large hidden layers are useful for the approximation of the bathymetry. Hence, we will regard network topologies with up to 40 neurons in the hidden layers.

Due to the little amount of sediment samples a poor ability of generalization of the model is expected. In the next section the influence of the depth of the bathymetry will be regarded in order to improve the model.

3.2 Dependency of the depth of the sample position

Figure 3 reveals the tendency that coarser particles are found more frequently in channels. This trend can also be seen in the scatter plot in Figure 5, where the d50 is plotted against the depth z of all samples in the year 2005. We do not describe the dependency of the parameters by a classic statistical method here, since we will investigate this issue within a second scenario of the form

$(z) \rightarrow d50$. In the strict sense z is part of the position of the measurement. However, we try to find a two-dimensional approximation and the depth of the sample position is unique, because the sample is always recorded at the soil. Thus z can rather be interpreted as additional information than as part of the position.

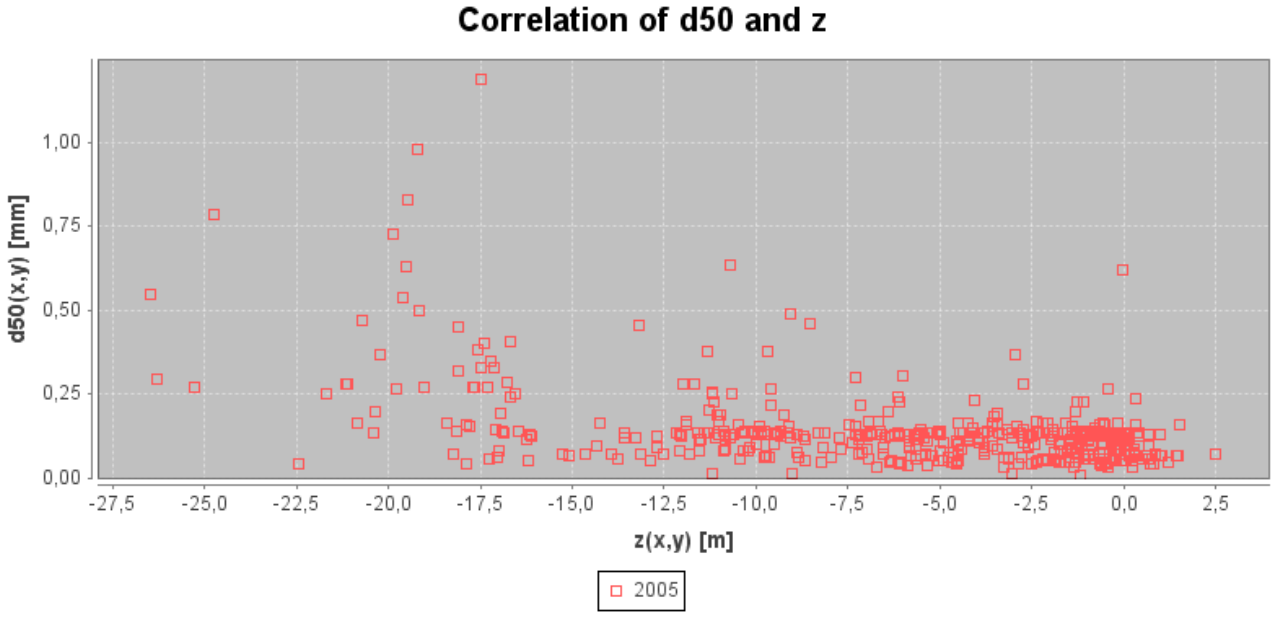

Figure 5 Scatter plot of the grain-size (d50) of the samples over the depth (z) of the position where the samples where recorded. A small tendency can be seen that coarser particles occur more frequently at deeper locations.

3.3 Combining the position and the depth

In the third scenario the input parameters of scenario 1 and 2 will be combined, since the grain-size is dependent of both, its position and the depth of the sea at that position. According to this the mapping that will be trained is $(x,y,z) \rightarrow d50$.

3.4 Dependency of the bathymetric environment

As mentioned in Section 3.2 bigger particles are found more frequently in channels. A deeper z-value is an indicator for a channel at that position. But a channel cannot be identified until the information of its environment is taken into account. This will be regarded in the fourth scenario. Since the bathymetric information is available for any position by the bathymetric model within its bounds it is possible to retrieve the depth in the environment of the sediment sample. In this scenario we additionally use the depth at 8 positions on a sphere around the sampling point as depicted in Figure 6. The sphere is defined using the maximum metric. The positions of the environment are like the positions of neighbours defined in the Moore neighbourhood on a regular quadratic grid.

By choosing a too big radius the essential information of the bathymetric environment cannot be detected in the same way as if the radius was too small. For the selection of the radius, the bathymetry was analyzed and it was found out, that the width of the channels at the soil is about 200 m to 500 m. This scenario is therefore run twice with a radius of 150 m and 300 m each. The mapping is $(z,z^{NE},z^{E},z^{SE},z^{S},z^{SW},z^{W},z^{NW},z^{N}) \rightarrow d50$ but will briefly be referred to as $(z,z^{moore(r)}) \rightarrow d50$.

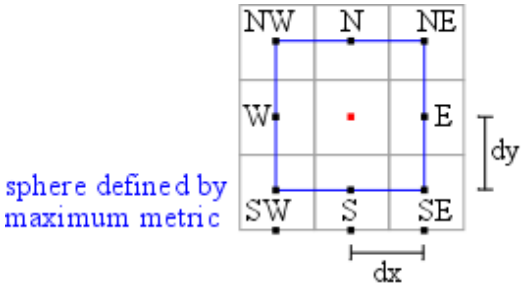

Figure 6 Definition of the additional positions (black dots) in the environment of the sampling position (red dot). The black dots are positioned on a sphere (marked blue) with the sampling position as its center and radius $r = dx = dy$ using the maximum metric. The positions are placed like on a quadratic regular grid using a Moore neighbourhood.

3.5 Combining the position and the surrounding parameters

Analog to Section 3.3 the explicit position (x,y) of the sampling point and the bathymetric information of its environment $(z,z^{moore(r)})$ from Section 3.4 will be combined in the last scenario. The d50 is approximated by the mapping $(x,y,z,z^{moore(r)}) \rightarrow d50$ here.

4. EVALUATION OF THE MODEL

The evaluation of the approximation model is a difficult task. The only possibility is to measure the performance of the model with regard to the observed values by comparing the approximated output values to the measured ones. Since the aim is to obtain a continuous approximation model for the whole region around the measurements the accuracy of the model for any position in this region would have to be evaluated ideally. Of course this is not possible at all and if it was there would be no need for an approximation model. Usually the concept of the test set is used to address this matter: the observed data is being divided into two disjunct sets, the training set, which is used as training data and the test set, which is used to measure the performance of the model with respect to unknown patterns.

Now, there are some challenges regarding the evaluation. First of all, it is clear that the selection of the test set fundamentally influences the quality of the evaluation of the generalization performance. A good performance of the model determined with regard to an uncharacteristic test set is not a confidential one. And even if the test set was found to be representative, in the strict sense statements can only be given for the used test set. Secondly, it seems to be insufficient to investigate the performance concerning the test set only, because the approximation of the patterns in the training set should be good as well for a continuous approximation. Thirdly, one of the biggest challenges is to define an adequate performance measure. The most commonly used performance measure in literature is the root mean square error (RMSE), which is defined in Equation 1, where the number of patterns is denoted by n, the observed output of the i-th pattern by y_i and the approximated output of the i-th pattern by $\hat{y}_i$.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (1)$$

The RMSE can be used to compare the approximation quality of analog models. This is done in Figure 7, where the minimum approximation error (RMSE) of each scenario pertaining to the test set is shown for each network topology. Since we have 10 calculations for each network topology of each scenario, the best performance results in the depicted range (min/max denoted by the dashed lines). The continuous lines denote the means of the minimum values.

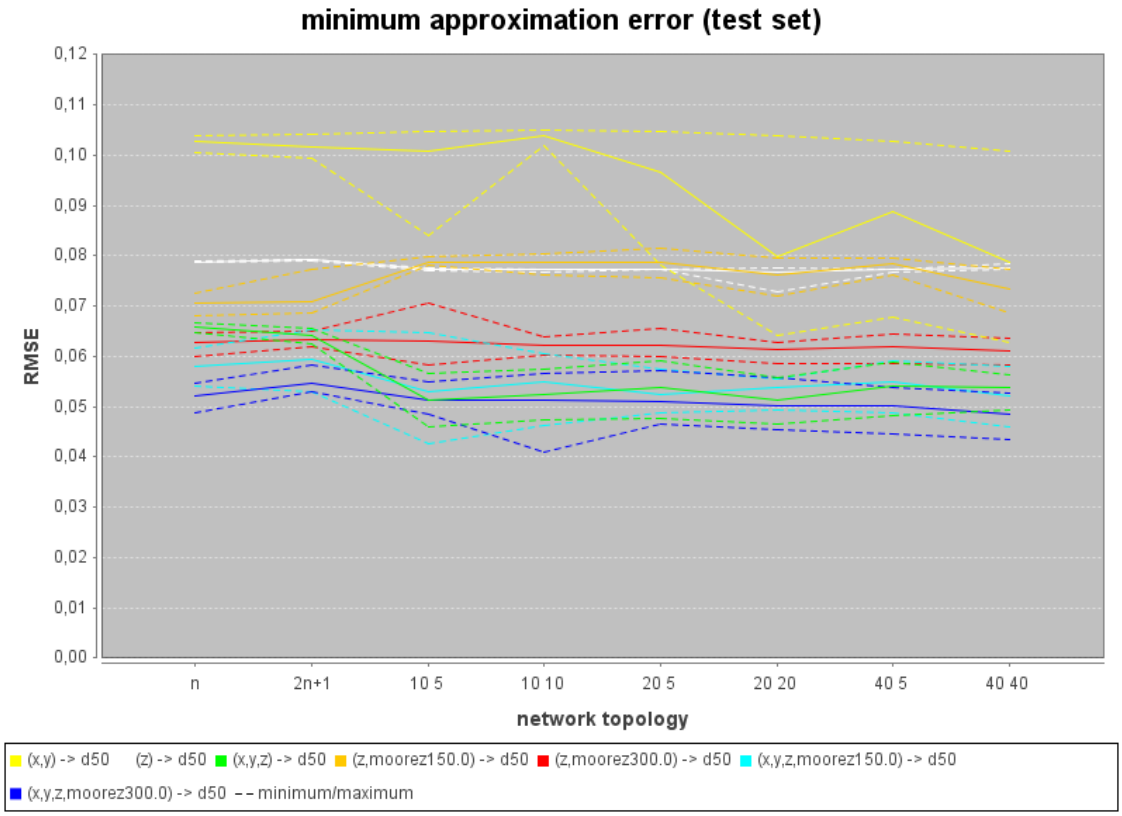

Figure 7 Evaluation of the minimum approximation error (RMSE) relating to the test set for each scenario. 10 ANN where trained for each network topology (of each scenario) starting with different synapse weights. The mean of the 10 calculations is marked by the continuous line, the dashed lines denote the min/max.

First of all, it can be seen that a strong dependency of the chosen network topology exists for scenario $(x,y) \rightarrow d50$, where the generalization performances of the other scenarios do not change very much. Further, it turns out that scenario $(x,y) \rightarrow d50$ does not perform very well in contrast to the others as it was expected. This may be caused by the rarely distributed training patterns. As for the commonly used approximation methods, the information of the position alone is not sufficient to produce a good approximation. The generalization performance of scenario $(z) \rightarrow d50$ is slightly worse. The network is not able to perform very well, since the input data is ambivalent as was shown in Figure 5. But it is interesting to see that all the scenarios, which use a combination of the position and the depth as input parameters ($(x,y,z) \rightarrow d50$, $(x,y,z,z^{moore(150m)}) \rightarrow d50$ and $(x,y,z,z^{moore(300m)}) \rightarrow d50$) perform much better than the other scenarios. In the following *one particular* ANN-model of each scenario will be used for further investigations. Therefore those ANN-models where chosen, which had the best RMSE-value regarding the test set of each scenario (this corresponds to the minimum value of the lines in Figure 7 for each color).

While the RMSE can be used to compare the approximation quality of analog models, the meaning of particular RMSE-values is not obvious. Hence, the performance of the scenarios was determined using some other measures with regard to the training and the test set each. Table 1 gives an overview of the different performance measures applied to the given ANN-models of the described scenarios. The relative deviation concerning a particular pattern i specifies the deviation of the approximated output $\hat{y}_i$ to the expected output y_i in relation to the magnitude of the expected output. The mean relative deviation (MRD) is defined in Equation (2).

$$MRD = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{|y_i|} \tag{2}$$

Table 1 Performance of the approximation models of the different scenarios regarding different performance measures. The values where determined for one particular ANN-model of each scenario.

scenario	RMSE [-]		MRD [rel]		p5 [mm]		p95 [mm]	
	training	test	training	test	training	test	training	test
x,y	0.0874	0.0627	0.4053	0.3049	-0.1414	-0.1094	0.1155	0.1208
z	0.0986	0.0730	0.4417	0.3392	-0.1372	-0.1210	0.0914	0.0812
x,y,z	0.0860	0.0459	0.3881	0.2569	-0.1239	-0.0556	0.0966	0.1006
$z,z^{moore(150m)}$	0.0945	0.0681	0.4234	0.3322	-0.1287	-0.1139	0.0751	0.0808
$z,z^{moore(300m)}$	0.0890	0.0583	0.4078	0.3054	-0.1537	-0.0592	0.0635	0.1149
$x,y,z,z^{moore(150m)}$	0.0791	0.0426	0.3719	0.2628	-0.1224	-0.0718	0.0763	0.0722
$x,y,z,z^{moore(300m)}$	0.0867	0.0411	0.3845	0.2673	-0.1445	-0.0598	0.0680	0.0603

As further performance parameters the 5th (p_5) and the 95th (p_{95}) percentile of the difference $\hat{y}_i - y_i$ will be regarded. The p_{95} specifies the value, below which 95% of the differences may be found. The p_5 is defined likewise. The range $[p_5, p_{95}]$ then defines the range of the deviation of the ANN-model, in which 90% of the observations fall into.

It turns out that again the scenarios that use the position as well as the depth as input parameters perform nearly the same in terms of the MRD: the average of the relative deviation regarding the training set is about 38% each, whereas the approximation of the test set differs about 26% related to the expected values on average. Regarding the 5th and 95th percentile the ANN-model of scenario $(x,y,z,z^{moore(300m)}) \rightarrow d50$ performs best concerning the test set: 90% of the produced output values do not differ more than 0.0603mm from the real output value, while that are 0.0722mm for scenario $(x,y,z,z^{moore(150m)}) \rightarrow d50$ and only about 0.1mm for the model of scenario $(x,y,z) \rightarrow d50$.

Figures 8 to 14 in the appendix show the results of the d50-approximation and more detailed information of the performance in terms of a performance scatter plot for each scenario. It can be seen that the approximation of scenario $(x,y) \rightarrow d50$ seems to be very coarse, while the approximation of the scenarios that solely use the depths as input parameters are very similar to the bathymetric structure. The combination of the position and the depth seem to improve the model.

5. CONCLUSION AND OUTLOOK

In this paper an continuous approximation model for sedimentologic parameters was introduced, that is based on an artificial neural network. Measuring the approximation performance of such a model in particular and of approximation models in general is a big challenge. The performances of the ANN-models of the considered scenarios have been investigated with respect to different measures. The estimation of the approximation performance strongly depends on the used measure and on the data that is used for testing. Some more work has to be done to get suitable measures and test sets in order to determine an adequate approximation performance.

Altogether the presented ANN-approximation model seems to produce quite good results in terms of the approximation of the d50 parameter of the given grain-size measurements. An acceptable continuous approximation was achieved. Particularly the combination of the position of the measurement and the depth of the sea at that position seem to improve the results of the model. Future works should focus on the approximation of other sedimentologic parameters, like the grain-size distribution as a discretized function.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the direct and indirect support from the partners of the AufMod-Project (see Heyer (2010)). Thanks also go to Chris Schiermeyer, Nils Rinke and Simon Berkhahn, who helped out with the evaluation of the results and the figures.

REFERENCES

Berthold, T. and Milbradt, P. (2009) “Artificial Neuronal Networks in Environmental Engineering: Theory and Applications”, Proceedings of the 18th International Conference on the Applications of Computer Science and Mathematics in Architecture and Civil Engineering (IKM 2009), Ed. Gürlebeck, K. and Könke, C., pp. 1611-4086.

Berthold, T., Milbradt, P. and Berkhahn, V. (2010) “Determination of Network Topology for ANN-Bathymetric Models”, Proceedings of the 9th International Conference on Hydro-Science and Engineering (ICHE 2010), Ed. Sundary, V., Srinivasan, K., Murali, K. and Sudheer, K.P., pp. 1650-1661.

Figge, K. (1981) Sedimentverteilung in der Deutschen Bucht, Deutsches Hydrographisches Institut, Karte Nr. 2900 (mit Begleitheft)

Heyer, H. and Schrottke, K. (2010) 1. Statusbericht AufMod (03KIS082-03KIS088) Gemeinsamer Statusbericht für das Gesamtprojekt mit Beiträgen aus allen 7 Teilprojekten.

APPENDIX

The results of the ANN-models of the different scenarios are summarized here in terms of performance scatter plots and the approximation of the d50-parameter for the produced output. For the d50-approximation plots the color scale from Figure 3 was used. The black dots mark the position of the test patterns.

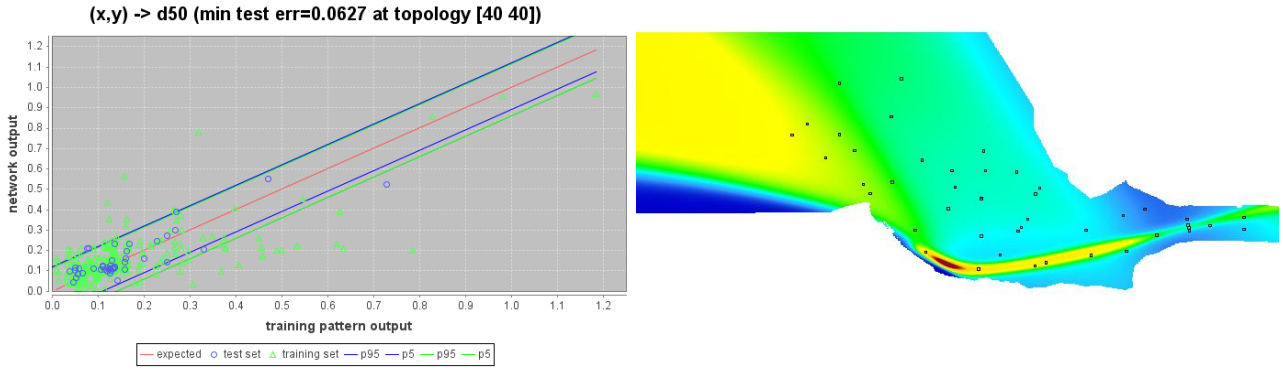

Figure 8 Results for the approximation model for scenario $(x,y) \rightarrow d50$.

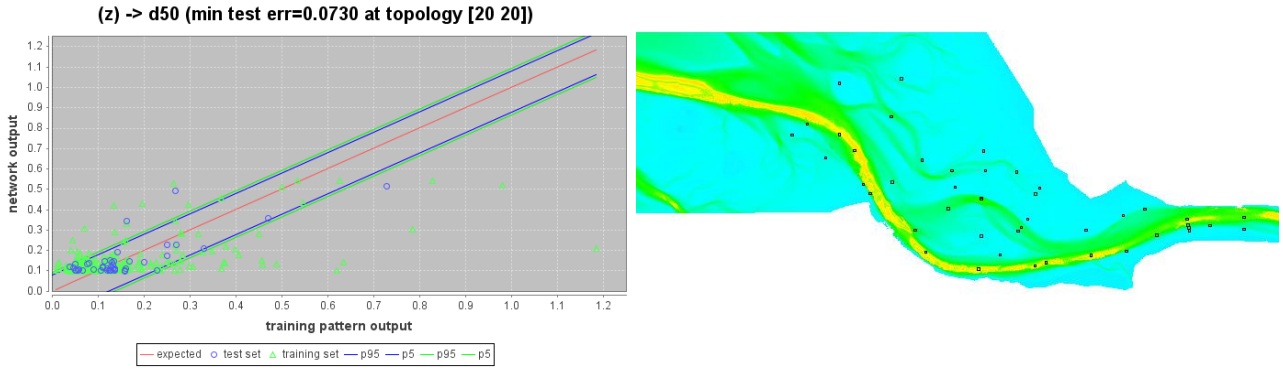

Figure 9 Results for the approximation model for scenario $(z) \rightarrow d50$.

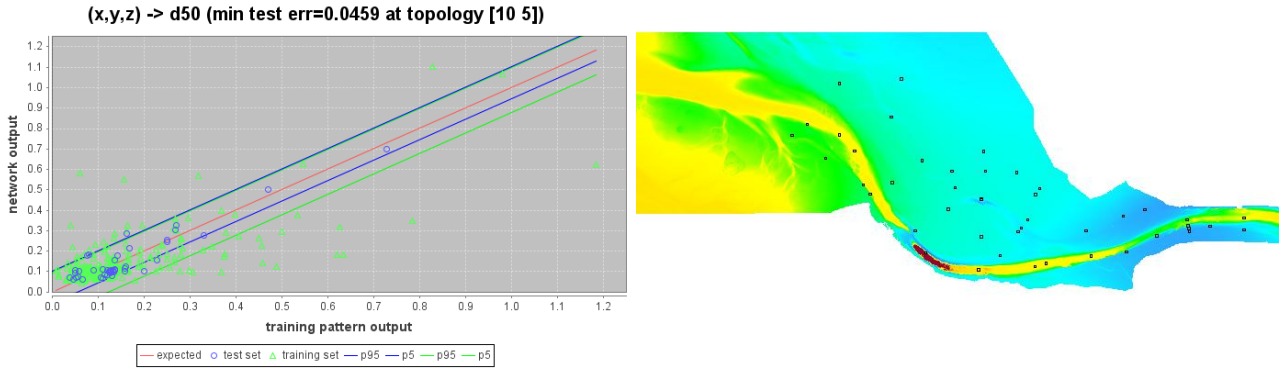

Figure 10 Results for the approximation model for scenario $(x,y,z) \rightarrow d50$.

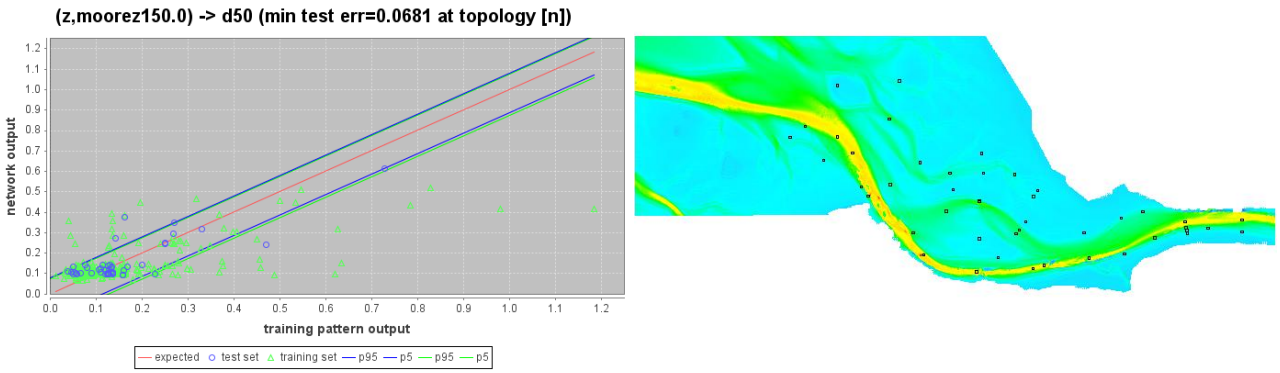

Figure 11 Results for the approximation model for scenario $(z, z^{moor(150m)}) \rightarrow d50$.

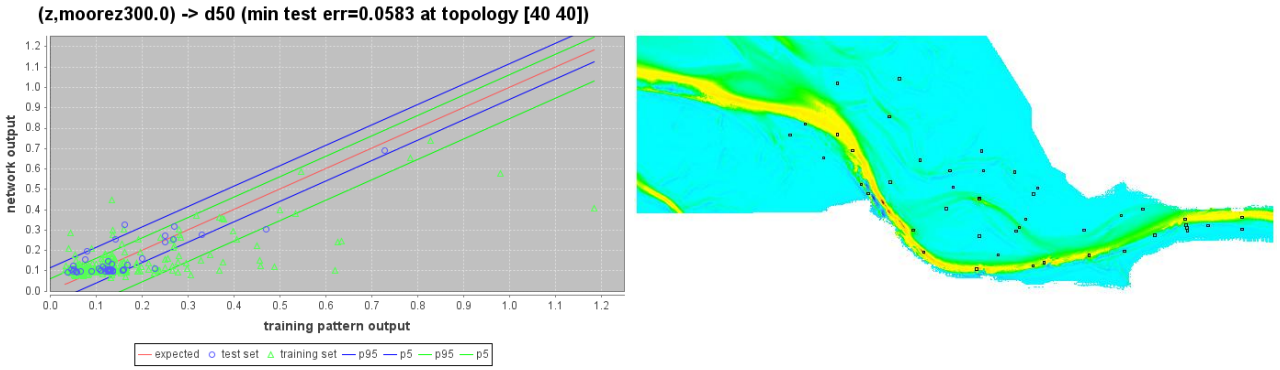

Figure 12 Results for the approximation model for scenario $(z, z^{moor(300m)}) \rightarrow d50$.

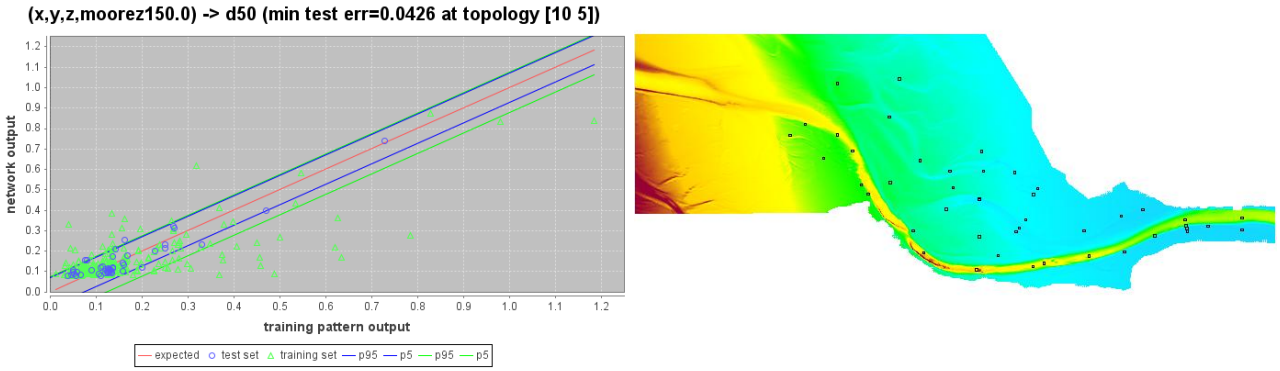

Figure 13 Results for the approximation model for scenario $(x, y, z, z^{moor(150m)}) \rightarrow d50$.

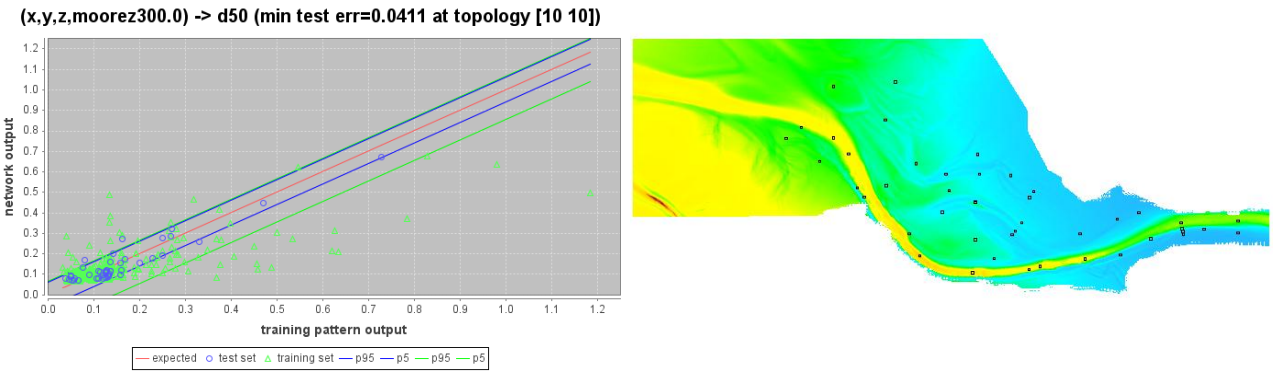

Figure 14 Results for the approximation model for scenario $(x,y,z,z^{moore(300m)}) \rightarrow d50$.