technische universität
dortmund

Fakultät Physik - Experimentelle Physik V

Data Mining on the Rocks

A measurement of the atmospheric muon neutrino flux using IceCube in the 59-string configuration and a novel data mining based approach to unfolding

Dissertation
zur Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)
vorgelegt von

Dipl. Phys. Tim Ruhe

Dortmund, Mai 2013

CONTENTS

Chapter 1

Introduction

Ever since its postulation by Wolfgang Pauli and its discovery by Reines and Cowan [1, 2] the neutrino has been the most enigmatic of the elementary particles. Although first discovered at a nuclear reactor, many different sources of neutrinos emerged over time.
Solar neutrino experiments [3, 4, 5, 6] were able to detect the flux of neutrinos emerging from nuclear reactions in the sun, which posed a new challenge, the so called "solar neutrino problem", to the neutrino physics community.
This problem was later resolved by the measurement of neutrino oscillations of atmospheric [7] and solar neutrinos [8] by Super-Kamiokande [9]. The measurement of neutrino oscillations further implies, that neutrinos can in fact have non-zero mass.
The current best upper limit of $m_\nu < 2.2$ eV originates from the Mainz neutrino experiment [10]. An upper limit of $m_\nu < 2.5$ eV is reported by the Troitsk group [11]. A further improvement on this limit is expected from the KATRIN experiment [12].
In the scope of astroparticle physics, neutrinos are also of interest [13]. Neutrinos from the type II Supernova 1987A were observed in 1987 [14, 15], being the first and only neutrinos ever observed from outside the solar system to date. Other astrophysical sources, such as Active Galactic Nuclei, are also believed to emit neutrinos [16, 17, 18]. However, no neutrinos from these astrophysical objects have been observed to date. The cosmological background of relic neutrinos is believed to exist, but is yet to be discovered as well [19].
Atmospheric neutrinos [16, 19, 20, 21, 22] form another source of neutrinos, spanning several orders of magnitude in energy. Within current research the atmospheric neutrino spectrum is of interest, as its shape, especially at high energies, is still subject to rather large uncertainties [23]. Moreover, a possible contribution of so-called prompt neutrinos [24] is still to be measured. Atmospheric neutrinos are also interesting in the scope of point source searches as they form the most significant background to these. A detailed understanding of the atmospheric neutrino spectrum is thus crucial for the detection of point sources as well as for the interpretation of their spectra.
All three arguments provide the background for measurements of the atmospheric

neutrino spectrum using large state of the art neutrino telescopes.
Within the analysis presented in this thesis the atmospheric neutrino spectrum was obtained in an energy range from 100 GeV to 1 PeV. This is the highest energy measured for an atmospheric muon neutrino spectrum to date and corresponds to an increase in energy of more than half an order of magnitude compared to previous IceCube measurements [22]. This increase in energy range was achieved by applying state of the art machine learning algorithms in the event selection and by using the unfolding software TRUEE [25], which allows for the use of up to three input variables.

The thesis is organised as follows:

Chapter 2 briefly reviews the theory of atmospheric neutrinos and their role in astroparticle physics. Furthermore, an introduction to the relevant machine learning algorithms used in this thesis is given.

Chapter 3 covers the event selection procedure carried out using state of the art data mining techniques, including an MRMR Feature Selection [26] as well as the training and testing of a Random Forest [27]. Moreover, detailed comparisons of data and Monte Carlo simulation are shown. Finally, the application of the Random Forest on real data including systematic studies is presented.

Chapter 4 explains the application of the unfolding algorithm TRUEE on the neutrino sample obtained in chapter 3. Six different unfolding settings are examined towards their stability and their dependency on changes in the ice model. Special studies on the unfolding result of the highest energy bin are shown. A detailed binwise and statistically reliable estimation of the systematic uncertainties is presented. Finally, a discuission of the unfolded flux spectrum is given, including detailed comparisons to theoretical models and previous experimental results.

Chapter 5 introduces a novel data mining based approach towards the unfolding of smeared distributions. Studies on toy Monte Carlo simulations are shown. A comparison to results obtained with TRUEE is given together with a comparison to other existing unfolding approaches. Furthermore, an outlook on possible real world applications is presented.

In **Chapter 6** an overall summary of the thesis including an outlook is presented.

Chapter 2

Theoretical Background

Within this chapter the theoretical background of the thesis is briefly reviewed. Focus is put on the physics of atmospheric neutrinos, as well as on the machine learning algorithms applied in the event selection.

2.1 Neutrinos as Astronomical Messengers

Even 100 years after the discovery of Cosmic Radiation by Victor Hess [29] the mechanisms by which particles are accelerated up to energies of $\approx 10^{20}\,\mathrm{eV}$, are still unknown. Although high energy cosmic rays are routinely observed at various experimental sites (e.g. the Pierre Auger observatory in Argentina [30]) their origin cannot be assigned to specific regions in the sky. This is due to the fact that the cosmic radiation mainly consists of charged particles. Approximately 79% of the primary nucleons are protons and approximately 79% of the rest are bound in helium nuclei [31]. After acceleration by cosmic sources, the trajectories of these particles are bent by interstellar and intergalactic magnetic fields (see figure 2.1). Thus, upon the point of detection basically all directional information is lost.
The primary component cannot be directly observed with Earth bound detectors, as the particles interact with nuclei in the atmosphere, producing what is known as an extended air shower (EAS) [16, 17].
A small part of the cosmic radiation ($\approx 0.01\%$ [19]), however, consists of high energetic photons. As they carry no electrical charge these photons can travel through space enirely undeflected, making the identification of their sources a relatively easy task. Unfortunately, the Earth's atmosphere is opaque to a large part of the photon energy spectrum. One possible way to circumvent this difficulty, posed by the atmosphere, is the use of sattelite bound devices (e.g. PAMELA [32], AMS [33] and Fermi [34]). However, due to weight considerations the size of detection devices feasible to be used in sattelite experiments is limited, thus also limiting the detectable number of photons per time unit.
Another way to detect high energetic γ radiation is by using the atmosphere itself for detection. Just like the charged component of the cosmic radiation high energy photons interact with nuclei in atmosphere, producing extended air showers.

The topology of these air showers, however, is significantly different from those produced by the interaction of a proton or a nucleus. Nevertheless, a large number of charged particles are produced, which travel faster than the speed of light in air, thus, initiating the emission of Cherenkov light [35]. This very faint and very short light pulse can be observed using dedicated Cherenkov telescopes (e.g. MAGIC [36], HESS [37] and FACT [38]).
Despite the fact that the observation of high energetic photons carries information on the location of their sources, no information on the acceleration mechanism itself can be inferred. Currently hadronic and leptonic acceleration processes are taken into consideration. See references [16, 17, 39] for a detailed introduction to the acceleration processes in cosmic sources.
In case of a hadronic acceleration the emission of high energy photons would be accompanied by the emission of neutrinos. The observation of high energy photons and neutrinos from the same source would, therefore, reveal details on the acceleration mechanism in cosmic sources.
As neutrinos do not carry any electrical charge their trajectories are undeflected by interstellar and intergalactic magnetic fields, allowing for the identification of their source(s). Moreover, due their small cross sections, neutrinos can traverse large amounts of matter without interaction. Compared to the observation of γ-particles this is one of the biggest advantages, as high energy γ-rays can interact with the Cosmic Microwave Background (CMB) (see [40, 41, 42]) as well as with the isotropic radio background and the infrared/optical background [17].
This advantage is at the same time the biggest challenge, as neutrinos can only be observed through leptons created by interactions with matter (see section 2.4). Up to a certain point, however, this can be accounted for by using detection volumes of the order of 1 km^3. In order to achieve these large instrumented volumes the detection medium should be as cheap as possible. Since the Cherenkov effect is utilised in order to detect neutrinos, the second requirement on the detection medium is transparency. These two very basic requirements for the detection medium leave two choices: Water [43, 44] and deep glacial ice [45, 46].
These large scale neutrino telescopes also provide the opportunity for measurements of the atmospheric neutrino spectrum (see section 2.2). Due to their large detection volumes a large number of neutrino events can be collected per time unit, which in turn leads to an extension of the experimentally accessible energy range. Furthermore, atmospheric neutrinos form the dominant background in point source searches. A detailed measurement of the atmospheric neutrino flux is thus interesting upon itself but also a prerequisite for the detection and interpretation of a neutrino flux from cosmic sources.

2.2 The IceCube Detector

IceCube is a state of the art neutrino telescope located at the geographic South Pole. Its 5160 Digital Optical Modules (DOMs), mounted on 86 vertical strings, are arranged in a three dimensional grid at depths between 1450 m and 2450 m,

corresponding to an instrumented volume of $1\,\mathrm{km}^3$. The vertical spacing between DOMs on a string is 17 m, whereas the spatial distance between individual strings is 125 m [47, 48].
The low energy extension DeepCore consists of six strings, deployed in in a ring of 72 m radius, and the seven innermost IceCube strings. Each of the DeepCore strings is equipped with 60 DOMs identical to standard IceCube DOMs, except for new a model of PMT, with a $\approx 30\%$ increased quantum efficiency [47]. On DeepCore strings 50 DOMs are deployed at depths between 2100 m and 2450 m with a DOM to DOM distance of 7 m. These depths provide the clearest ice leading to increased scattering and absorbtion lengths of $l_{scat} = 50\,\mathrm{m}$ and $l_{ab} = 230\,\mathrm{m}$. Thus, compared to AMANDA a larger fraction of unscattered photons can be observed [49].
The remaining DOMs on every DeepCore string are placed at shallower depths between 1750 m and 1850 m above a major dust layer. This way they can be utilised as a veto for extremely vertical background muons. Using DeepCore the energy threshold of the entire IceCube detector can be lowered to 10 GeV [47, 49]. Additionally an air shower array, called IceTop, is located on top of the detetor. Details on DeepCore can be found in [49]. Additional information on IceTop is given in [50, 51].
Two primary detection channels are observed in IceCube. The first one being *track-like* events from charged current ν_μ interactions of the form:

$$\nu_\mu + N \longrightarrow \mu + X. \tag{2.1}$$

The other detection channel are *cascade-like* events, originating from charged current (CC) interaction of ν_e and ν_τ and from neutral current (NC) interactions of all neutrino flavours [48].
When the threshold energy is exceeded, charged leptons as well as cascades induced by NC interactions initiate the coherent superposition of light pulses, due to polarisation effects in the medium. This superposition of light pulses then forms a characteristic light cone (Cherenkov effect [35]), detected by the DOMs, from which the track of the particle can be reconstructed [52]. Since the direction of the charged lepton deviates only marginally from the direction of the high energy neutrino, the direction of the incoming neutrino can be reconstructed as well.
The pointing resolution of IceCube was measured utilising a shadowing effect of the moon [53, 54]. Deficits of 7.6σ and 12.7σ were observed for the 40 and 59-string configuration, respectively. The pointing resolution of the detector was found to be of order 1° [54].
The light pulses, recorded by the IceCube DOMs, are converted into digital waveforms. These digitised waveforms are then read out and time stamped everytime a DOM triggers. A single majority trigger (SMT) is applied for the in-ice-array as well as for IceTop. A coincidence of eight or more hits on nearest or next-to-nearest neighbour DOMs, within a time window of $5\,\mu$s, is required for the in-ice part of the detector [48]. Details on the IceCube data acquisition system are given in [55].

Despite the fact that the detector was completed in December 2010 data was already taken in various previous configurations. The data on which this analysis is based were taken between May 2008 and May 2009 by IceCube in the 59-string configuration, which corresponds to a detector livetime of 346 days.

2.3 Atmospheric Neutrinos

Despite the fact that the atmospheric neutrino spectrum has already been measured by several experiments [20, 56] including AMANDA [21] and IceCube in the 40-string configuration [22], its shape, especially at high energies is subject to rather large uncertainties [23].
Atmospheric neutrinos are produced by the decay of charged mesons originating from interactions of cosmic rays in the Earth's atmosphere. Up to energies of $\approx 100\,\mathrm{TeV}$ [21] their flux is dominated by neutrinos from Pions and Kaons decaying via [18]:

$$\pi^+ \longrightarrow \mu^+ + \nu_\mu, \tag{2.2}$$
$$\pi^- \longrightarrow \mu^- + \bar{\nu}_\mu, \tag{2.3}$$
$$K^+ \longrightarrow \mu^+ + \nu_\mu, \tag{2.4}$$
$$K^- \longrightarrow \mu^- + \bar{\nu}_\mu. \tag{2.5}$$

For energies smaller than several GeV the muons themselves decay before reaching the ground, yielding two ν_μ for every ν_e [18, 24]. Due to the rather large energy threshold of the IceCube detector ($\approx 10\,\mathrm{GeV}$) the contribution of electron neutrinos from muon decay can be negelected. Thus, neutrinos from three body Kaon decays (K_{e3}) become the main source of atmospheric electron neutrinos [18]:

$$K_L \longrightarrow \pi^\pm + e^\pm + \nu_e(\bar{\nu_e}). \tag{2.6}$$

Due to their relatively long lifetime ($2.6033 \times 10^{-8}\,\mathrm{s}$ for $\pi^\pm$ and $1.2386 \times 10^{-8}\,\mathrm{s}$ for $K^\pm$ [31]) Pions and Kaons can lose a certain fraction of their initial energy in collsions before decay. This constant competition between energy loss by collision and decay steepens the spectrum of atmospheric neutrinos by roughly one power compared to the primary cosmic ray spectrum [22], yielding approximately $\frac{dN}{dE} \propto E^{-3.7}$ for $E < 1\,\mathrm{PeV}$ and $\frac{dN}{dE} \propto E^{-4.0}$ for $E \geq 1\,\mathrm{PeV}$ [24].
The flux of conventional atmospheric neutrinos can be described in a simple analytic form [21]:

$$\frac{dN}{dE_\nu d\Omega}(E_\nu, \theta) = A_\nu \left(\frac{E_\nu}{\mathrm{GeV}}\right)^{-\gamma} \left[\frac{1}{1 + 6E_\nu/E_\pi(\theta)} + \frac{0.213}{1 + 1.44E_\nu/E_{K^\pm}(\theta)}\right], \tag{2.7}$$

with $A_\nu = 0.0285\,\mathrm{GeV}^{-1}\mathrm{cm}^{-2}\mathrm{sr}^{-1}$ and $\gamma = 2.69$. Additional models of the conventional component of the atmospheric neutrino spectrum are given in [57, 58, 59]
Despite the isotropic distribution of cosmic rays the spectrum of high energy atmospheric neutrinos is a function of the zenith angle. The energy spectrum of

horizontal neutrino events is flatter compared to vertical tracks. This is due to the fact that the density gradient seen by vertically travelling mesons is much larger than that seen by horizontal ones. As a consequence horizontally travelling mesons have a much higher probability of decaying before losing energy in collisions [21, 22]. At TeV energies the flavour ratio ν_e:ν_μ:ν_τ is approximately 0.05:1:0 [47].
At higher energies a second component starts to contribute to the atmospheric neutrino spectrum. This component, called prompt neutrinos, originates from the decay of charmed mesons. These charmed mesons have rest frame lifetimes of order 10^{-12} s [31], which in turn means that they decay before given any opportunity of losing energy in collisions. Thus, the spectrum of the prompt component is expected to follow the spectrum of the cosmic rays more closely ($\frac{dN}{dE} \propto E^{-2.7}$) [21, 22]. Atmospheric neutrinos from charm decays have not been measured so far but are expected to become dominant above 100 TeV [22]. The exact threshold energy, however, strongly depends on the underlying model [24].

2.4 Muons

2.4.1 Neutrino Induced Muons

Since neutrinos cannot be observed directly, interactions within the detector or the surrounding medium are utilised for their detection. In the case of an atmospheric ν_μ-spectrum, the leptonic partner of the ν_μ, the μ, is detected. The interaction of neutrinos with the surrounding medium is depicted in equation (2.1).
From equation (2.12) one finds that a muon of energy 1 TeV can travel distrances of ≈ 2.4 km in ice. Thus, the effective detection volume of IceCube is enhanced. Furthermore, as the muon range is a function of energy (see eq. (2.12)), the effective detection volume of the detector is a function of energy as well. This energy dependency needs to be taken into account, when converting the observed event rates into neutrino fluxes (see section 4.1).
The calculation of the rate of neutrino induced muons depends on three components: the neutrino flux, the muon production cross section and the muon-energy relation. In general the muon rate is given as [16]:

$$\frac{dN_\mu}{dE_\mu} = \int_{E_\mu}^{\infty} dE_\mu \left(\frac{dN_\nu}{dE_\nu}\right) \left(\frac{dP(E_\nu)}{dE_\mu}\right). \tag{2.8}$$

In equation 2.8 the first factor in the integrand represents the spectrum of neutrinos. In the most general case these neutrinos can be of terrestrial or extraterrestrial origin, respectively. The second factor depicts the probability of a neutrino on a trajectory passing through the detector and producing a muon in the energy interval $[E_\mu, E_\mu + dE_\mu]$ at the detector. This factor, thus, depends on the propagation of muons through matter and on the physics of the neutrino interaction [16].
From equation (2.8) one already finds that the observed rates of atmospheric muons

cannot be directly converted into an atmospheric neutrino spectrum. A considerable amount of smearing, which has to be taken into account, is introduced by the second term of the integrand. Furthermore, the energy of the muon is a quantity, which cannot be observed directly as well. Energy estimators need to be used in order to determine the energy of the neutrino induced muon. However, these energy estimators also suffer from smearing introduced by the limited resolution of the detector. This smearing, due to the $\frac{dP(E_\nu)}{dE_\mu}$ term in equation (2.8) and the limited detector acceptance requires the use of unfolding algorithms for the determination of the atmospheric neutrino spectrum (see section 4.1).
Equation (2.12) can be used to obtain a rough estimate on the rate of muons from atmospheric neutrinos near the vertical. It is given as 1.4×10^{-13} upward events per cm^2 per second per sr [16]. The full calculation of rates, however, is rather complex. A full treatment of this calculation is given in reference [16].

2.4.2 Atmospheric Muons

Atmospheric muons produced in the decay of Pions and Kaons (see equation (2.5)) also enter the detector, due to the range of the muons. In contrast to neutrino induced muons, however, atmospheric muons only enter the detector from above. As the rate of atmospheric muons exceeds the rate of neutrino induced muons by a factor of 10^6, atmospheric muons form the largest source of background in an atmospheric neutrino analysis. The simplest way to reject this source of background is the selection of upward going tracks. By doing so, the background of atmospheric muons can be reduced by roughly three orders of magnitude.
A small fraction of atmospheric muons ($\approx 0.1\%$), however, is falsely reconstructed as upward going, due to insufficient reconstruction of the track. Taking into account the large excess with respect to neutrino induced muons, misreconstructed atmospheric muons still pose a challenge to the event selection in every neutrino analysis. A further reduction of background events depends on the aims of individual analyses and is carried out either by the use of straight cuts or by the utilisation of machine learning algorithms (see section 2.5).

2.4.3 Muon Energy Loss

Two effects contribute to the energy loss. Energy is lost continuously through the ionisation of the surrounding medium and stochastically through radiative processes. These radiative processes involve bremsstrahlung, pair production and photonuclear interactions [31]. Therefore, the muon energy loss can be written as the sum of ionsiation energy loss and stochastic energy loss [31]:

$$-\frac{dE}{dx} = a(E) + b(E)E, \tag{2.9}$$

where ionisation losses are represented by $a(E)$ and stochastic energy losses by $b(E)$. Typical values of $a(E)$ and $b(E)$ in ice where computed in reference [60]:

$$a(E) = 0.259\,\mathrm{GeV\ mwe}^{-1} \tag{2.10}$$

$$b(E) = 3.63 \times 10^{-4}\,\mathrm{mwe}^{-1}. \tag{2.11}$$

From equation (2.9) the mean range of a muon can be estimated to [31]:

$$x_0 \approx \frac{1}{b} \ln\left(1 + \frac{E_0}{E_{crit}}\right). \tag{2.12}$$

$E_{crit} = a/b$ depicts the critical energy at which energy losses due to ionisation equal stochastic energy losses. Energy losses above several hundred GeV are dominated by stochastic processes [31].
Secondaries produced by the muon through bremsstrahlung and pair production, will also suffer from radiative energy losses, generating more secondaries. This results in an electromagnetic cascade. Once the energy of the generated electrons falls below the critical energy, ionisation losses become dominant and the cascade ceases [61].
Detailed simulations were carried out in reference [62] in order to parametrise the behaviour of the cascades. The number of Cherenkov photons can then be calculated as [63]:

$$N_C = L_{\mathrm{eff}}(E_0) n_C, \tag{2.13}$$

where L_{eff} represents the effective track length given as [62]:

$$L_{\mathrm{eff}} = 0.894 \times \frac{E_0}{1\,\mathrm{GeV}} \times 4.889\,\mathrm{m}. \tag{2.14}$$

The factor n_C in equation 2.13 represents the intgrated Frank-Tamm formula.
Energy losses due to photonuclear interactions will result in secondary hadrons, which again interact, producing more secondaries. This will result in a hadronic cascade. In general the yield of Cherenkov photons will be lower when compared to an electromagnetic cascade. The reasons for this can be found in the emission of neutrons, which do not produce Cherenkov photons, in lost energy due to larger hadronic binding energies and in the larger Cherenkov photon emission threshold for charged hadrons [63].
Due to the decay of Pions and Kaons, hadronic cascades can have an electromagnetic component, which increases with energy and does not contribute to a further development of the cascade.
According to [62], the effective track length for hadronic cascades can be parametrised via:

$$L_{\mathrm{eff}} = 0.860 \times \frac{E_0}{1\,\mathrm{GeV}} \times 4.076\,\mathrm{m}. \tag{2.15}$$

2.5 Selected Machine Learning Algorithms

This section introduces the relevant techniques from machine learning and data mining used to achieve the separation of neutrino and background events.

2.5.1 Random Forests and Decision Trees

The Random Forest algorithm, introduced by Breiman in 2001 [27], is a classification tool, which uses an ensemble of relatively weak classifiers (decision trees) in order to obtain a prediction. Compared to Boosted Decision Trees [64, 65] the outcome of every tree is independent from its predecessors. For a binominal classification task the final outcome is obtained by averaging the results of the individual trees:

$$s = \frac{1}{n_{\text{trees}}} \sum_{i=1}^{n_{\text{trees}}} s_i. \tag{2.16}$$

The s_i can be computed in different ways depending on the implementation. Growing of Classification and Regression Trees (CART) is described in the following.
Starting from a *root node* containing all examples of a training set D a split is created producing a number of descending nodes. How many of these new nodes are created again depends on the implementation used and on the problem at hand. Using numeric attributes the creation of a split corresponds to answering the question "$x_i \leq x_{is}$". Such a split will create a hyperplane boundary perpendicular to the coordinate axes [66].
As simple and compact trees with only a few nodes are preferred, the aim of the algorithm is to create descendend nodes that are already as pure as possible. For convenience reasons, however, the impurity of a node is measured rather than the purity [66]. The most popular measure is the entropy impurity given as [66]:

$$i(N) = -\sum_j P(\omega_j) log_2 P(\omega_j), \tag{2.17}$$

where $P(\omega_j)$ denotes the fraction of patterns at node N in class ω_j.
Another definition of impurity, particularly useful for the two category case, can be interpreted as a variance impurity and is given as [66]:

$$i(N) = P(\omega_1)P(\omega_2). \tag{2.18}$$

A generalisation of equation (2.18) can be written as [66]:

$$i(N) = \sum_{i \neq j} P(\omega_i)P(\omega_j) = \frac{1}{2}\left[1 - \sum_j P^2(\omega_j)\right]. \tag{2.19}$$

The measure in equation (2.19) is called the *Gini Impurity*. The *misclassification impurity* is defined as:

$$i(N) = 1 - \max_j P(\omega_j). \tag{2.20}$$

In order to obtain the optimum split at a certain node, however, the decrease in entropy, given as [66]:

$$\Delta i(N) = i(N) - P_L \cdot i(N_L) - (1 - P_L) \cdot i(N_R), \tag{2.21}$$

has to be minimised. Within this notation N_L and N_R are the left and right descendent nodes respectively. The impurities at these left and right descendent nodes are $i(N_L)$ and $i(N_R)$ and P_L is the fraction of examples being placed in N_L after the split. The use of the entropy impurity in equation (2.21) corresponds to an information gain [66]. One should note that the optimisation in equation (2.21) is local and that there is no guarantee that this local optimum corresponds to a global one [66].
In order to create independent trees within a Random Forest some sort of randomness needs to be introduced into the splitting of nodes. In general this is achieved by selecting a certain fraction of available attributes at every node. The optimum split is then determined using this subset rather than all available attributes. Thus, every tree will be different and, even more important, independent from the previous ones.
The successive splitting of nodes is stopped when either no further split can be performed or a certain stopping criterion is reached. The first case corresponds to all examples being described by the same patterns. For the second case several stopping criteria can be considered. The first of such criteria is that the decrease in impurity Δi is smaller than some user defined threshold β [66].
Another possibility to determine when to stop splitting is, when a certain minimal number of examples is present in each of the nodes. This method has benefits comparable to those of the k-nearest-neighbour classifier [67, 68]. The number of terminal nodes is large in regions with a large density of examples and small in the opposite case [66].
After the tree has been fully grown, labels need to be assigned to the terminal nodes. In case of zero impurity the assigned label is simply that of the class present in the node. For the more typical case of positive impurity, however, two different approaches can be used. The first one is a simple majority vote, whereas the second one corresponds to assigning the average of classes present in the node as a label. One should note that the second approach is only feasible for two class problems.

2.5.2 Minimum Redundancy Maximum Relevance Feature Selection

In general a representation of the data in fewer dimensions needs to be found in order to achieve a high performance classification within a reasonable amount of time, while utilising a reasonable amount of resources [69]. Thus, features need to be selected and deselected according to certain criteria. Probably the most intuitive algorithms performing such a Feature Selection are Forward Selection and Backward Elimination.
Both, Forward Selection and Backward Elimination, however, only take the relevance of a feature into account. This approach might fail for the case of high dimensional data with many correlated attributes. This is, for example, the case for IceCube data, where certain quantities, e.g. the zenith angle, have been reconstructed by various reconstruction algorithms.

The Minimum Redundancy Maximum Relevance (MRMR) [26, 70] Feature Selection aims to circumvent this problem by taking into account the relevance of the individual attributes as well as their redundancy compared to the attributes already selected in previous iterations. Depending on the implementation and the user's choice the quality criterion is calculated as:

$$Q = R(x, y) - \frac{1}{j} \sum_{x' \ in \ F_j} D(x', x) \tag{2.22}$$

or

$$Q = \frac{R(x, y)}{\frac{1}{j} \sum_{x' \ in \ F_j} D(x', x)}. \tag{2.23}$$

In equations (2.23) and (2.22) R refers to the relevance criterion that is calculated depending on the type of the attribute x and the type of the class y. D refers to the redundancy of the attribute x computed with respect to an already selected attribute x'. To compute the final redundancy of an attribute the pairwise average of the redundancy with respect to all attributes that have already been selected is calculated.
Relevance and redundancy are computed using the correlation in case both attributes are numerical and by utilising the mutual information in case one of the attributes is nominal.

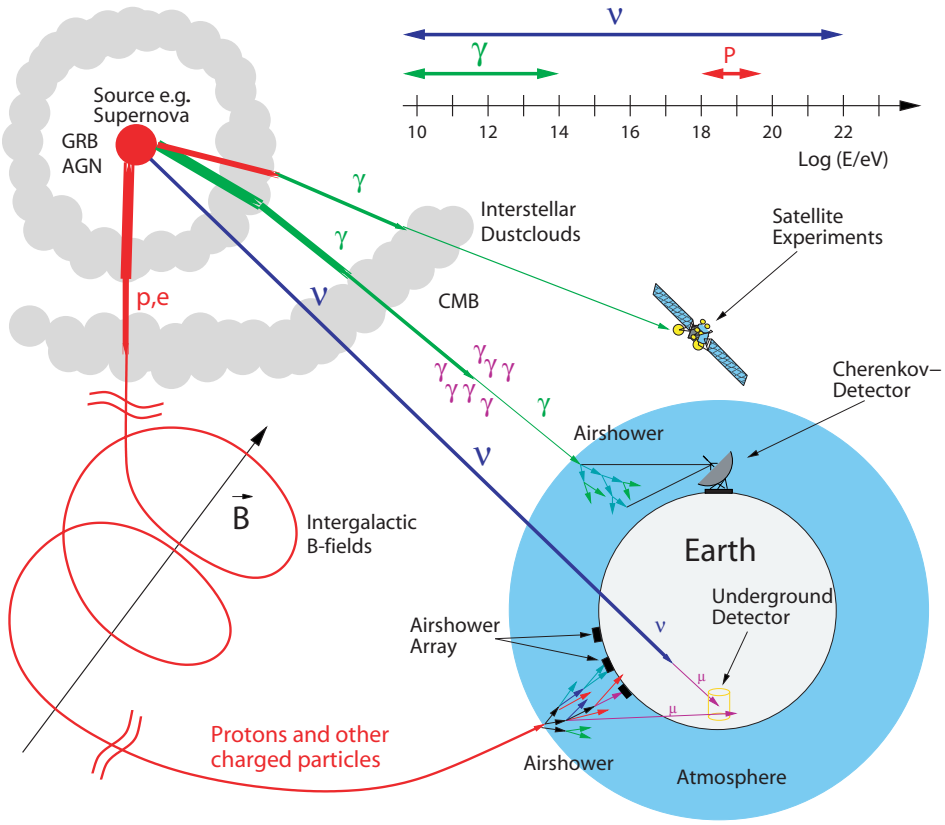


Figure 2.1: Propagation of various astronomical messenger particles from their sources to detectors on Earth. Trajectories of electrically charged particles such as electrons, protons and heavy nuclei get bent thus, having lost all directional information at their point of detection. High energy photons directly point back to their source, but their observation is limited in energy, due to interactions with the Cosmic Microwave Background (CMB). Moreover, their detection can only proceed indirectly by utilising the Cherenkov effect in air. Like photons, neutrinos do not carry any electrical charge and therefore also directly point back to their sources. Due to their small interaction cross sections, they can traverse cosmic distances without absorbtion. On the other hand their small interaction cross section requires huge detection volumes in order to detect a significant number of neutrinos. Figure after [28].

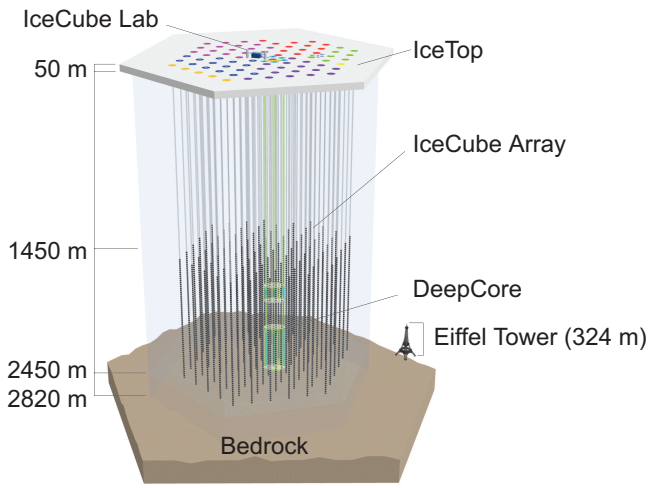


Figure 2.2: The IceCube detector located at the geographic South Pole. 5160 Digital Optical Modules, mounted on 86 strings, are arranged in a three dimensional grid at depths between 1450 m and 2450 m, forming an instrumented volume of 1 km^3. The 6 innermost, more densely instrumented strings are part of the low energy extension DeepCore. The air shower array IceTop consists of water tanks instrumented with photomultipliers and located on top of the IceCube strings.

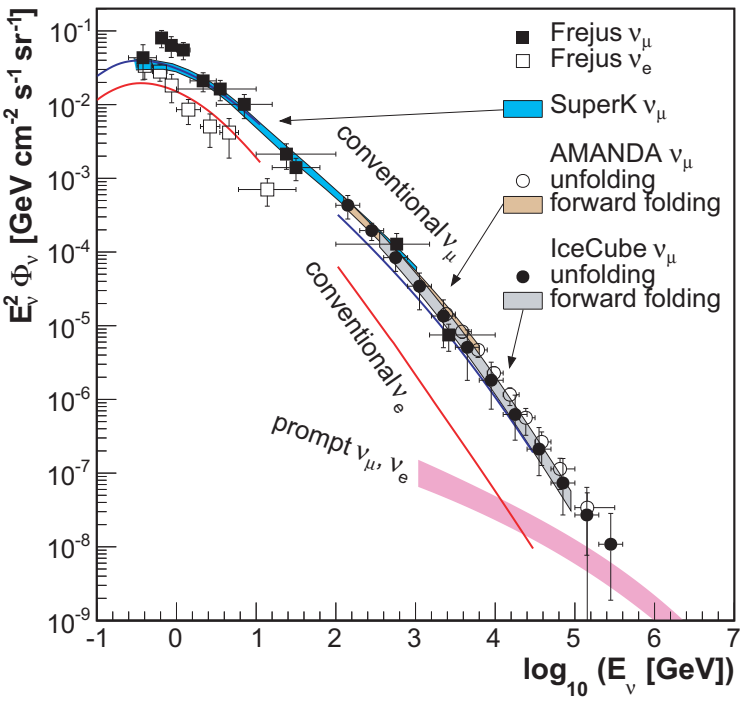


Figure 2.3: Atmospheric Neutrino Flux weighted by E^2 as a function of energy as measured by various experiments. At energies of $\approx 10^6$ GeV the contribution of prompt neutrinos (represented in magenta) originating from the decay of charmed mesons becomes nonnegligible.

Chapter 3

Separation of Signal and Background

Within this chapter the necessary steps in obtaining a high quality neutrino sample are summarised. The procedure, as a whole, can be broken down into three steps, the first one being a preprocessing of data including the application of precuts. A detailed algorithm-based Feature Selection was carried out as a second step. The third step consisted of the training, testing and application of a Random Forest [27].
Monte Carlo simulations were used at various stages of the event selection. The background of atmospheric muons was simulated using the air shower simulation CORSIKA [71], whereas the neutrino generator NUGEN [72] was used for the simulation of signal events.
All machine learning specific tasks were carried out using the data mining environment RAPIDMINER [73, 74]. This chapter focuses on the results of the individual stages of the event selection rather than on the technical side including the implementation of specific processes in RAPIDMINER. Details on the RAPIDMINER processes used for the selection of neutrino events are presented in [75].
All cuts were developed on the so-called "burnsample", which corresponds to 10% of the data or 33.281 days of IceCube in the 59-string configuration (IC-59). All numbers quoted in this chapter correspond to the burnsample, unless stated otherwise.

3.1 Precuts

Starting from a level at which a significant amount of reconstruction algorithms were already run (level 3), two precuts were applied in order to reduce the background and to increase the quality of the retained tracks.
The first cut was applied on the LineFit velocity ($v_{LF} \geq 0.19$) in order to reject poorly reconstructed and cascade-like events (see figure 3.1). The more spherical light pattern initiated by cascades results in smaller values of v_{LF}. Events with small values of v_{LF}, however, need to be rejected, as long well reconstructed tracks lead to a better energy reconstruction, which in turn leads to a better measurement of the energy spectrum.

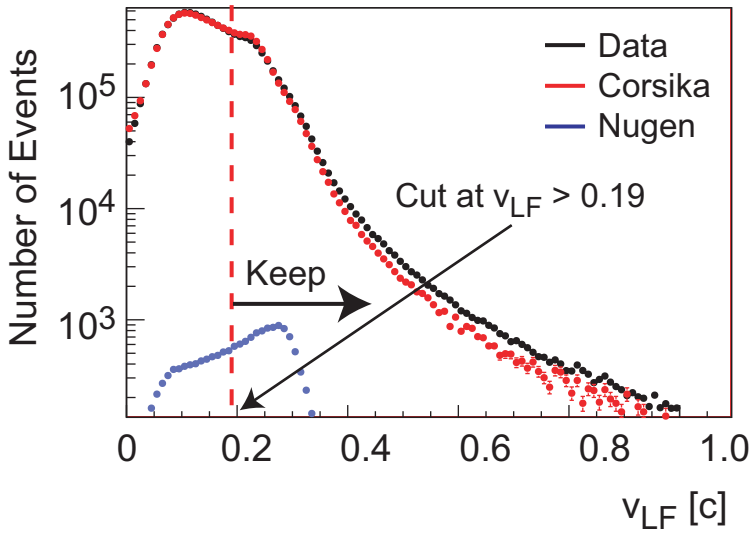


Figure 3.1: Cut on the LineFit velocity v_{LF} excluding events with $v_{LF} < 0.19$. This cut mainly intends to increase the quality of the selected events, as cascades show a spherical event pattern, which in turn leads to small values of v_{LF}.

The second cut was placed on the zenith angle (θ_{MPE}), reconstructed from a Multi-Pulse-Extraction (MPE) Fit. It aimed at significantly reducing the background of atmospheric muons entering the detector from above. A cut at $\theta_{MPE} = 88°$ was chosen in order to retain high energy neutrino events entering the detector close to the horizon (see figure 3.2).
The application of these cuts reduced the number of background events from 19.56×10^6 to 1.68×10^6, while retaining 8 990.87 out of 15 743.1 neutrino induced events. This corresponds to a background rejection of 91.4% at a signal efficiency of 57.1%. Note, that signal efficiency is often referred to as recall, when used in the context of machine learning and data mining. Both terms are used synonymously in this thesis.
In the following the event selection after the application of precuts will be referred to as level 4.

3.2 Feature Selection

3.2.1 Preselection of Attributes

Prior to an algorithm-based Feature Selection a preselection of attributes was carried out in order to reduce the required computing resources. The preselection consisted of four steps, explained in the following:

- **Consistency Check:** Data and Monte Carlo simulations were checked for consistency. Attributes not present in either were excluded. At this stage mainly attributes carrying information on the Monte Carlo simulation e.g. energy and zenith angle of the primary particle, were excluded.

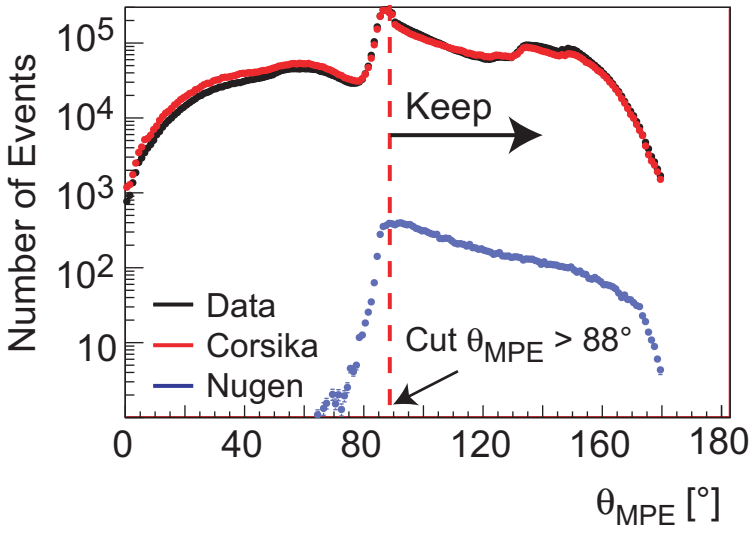


Figure 3.2: Cut on the reconstructed zenith angle (θ_{MPE}). This cut is mainly used to reject downward going atmospheric muon events. The threshold of $\theta_{MPE} = 88°$ was chosen to keep high energy neutrino events originating near the horizon and to slightly increase the field of view.

- **Check for Missing Values:** At this point a check for missing values was performed. Based on studies presented in [76], attributes containing more than 30% missing values were excluded from the selection. Missing values might occur in certain attributes, in case the fit fails to reconstruct certain parameters from a given event pattern.

- **Excluding Sources of Potential Bias:** As a third step attributes that were known to be useless, redundant or a source of potential bias were excluded. This mainly concerned information on time and date as well as sky coordinates.

- **Excluding Highly Correlated Attributes:** Within this step highly correlated attributes ($\rho = 1.0$) were excluded from the selection. Attributes with smaller correlation were not excluded as they were handled in the automated Feature Selection.

Within this preselection of attributes a reduction from $\approx$ 2600 to 477 attributes was achieved, which resulted in a significant amount of saved computing time and memory.

3.2.2 Feature Selection Stability

Automated Feature Selections are sensitive to statistical fluctuations in the example sets they are performed on. Therefore, it has to be ensured that the Feature Selection performs in a similar way on statistically independent subsets of examples, drawn from the same distribution. This is done using the Feature Selec-

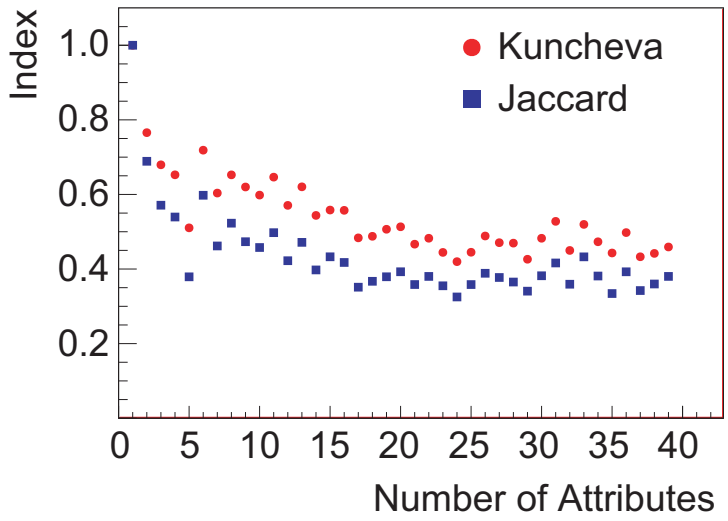


Figure 3.3: Feature Selection stability for a simple Forward Selection. With an increasing number of attributes, a decrease in stability is observed.

tion Stability operator included in the Feature Selection Extension [70] for RapidMiner. The stability of the Feature Selection is investigated by using a procedure similar to a simple Cross Validation [77]. The full set of examples is split into m disjoint subsets. Then the Feature Selections are performed on each of the subsets and the outcome is compared using either the Jaccard index or Kuncheva's index. The Jaccard index is defined as:

$$J = \frac{|A \cap B|}{|A \cup B|}, \tag{3.1}$$

whereas Kuncheva's Index is given as [78]:

$$I_C(A, B) = \frac{rn - k^2}{k(n - k)}. \tag{3.2}$$

In equation (3.2) the parameter k represents the size of the subset, whereas $r = |A \cap B|$ represents the cardinality of the subset. The total number of features availaible is given as n.
Figure 3.3 depicts the stability of a simple forward selection [77] as a function of the selected attributes, by applying the introduced indices. Kuncheva's index is shown in red, whereas the Jaccard index is depicted in blue. One finds that both stability measures decrease with an increasing number of attributes remaining at a constant level of roughly 0.5 for a number of attributes $n_{Att} \geq 20$. This can be interpreted as only 50% of the selected attributes being selected everytime the Forward Selection is run. Thus, the Forward Selection cannot be considered stable in this analysis and was not used as a Feature Selection algorithm accordingly.
Figure 3.4 shows the stability of the Minimum Redundancy Maximum Relevance (MRMR) Feature Selection [26] as a function of the number of attributes considered. The Jaccard index is shown in blue, whereas Kuncheva's index is depicted in

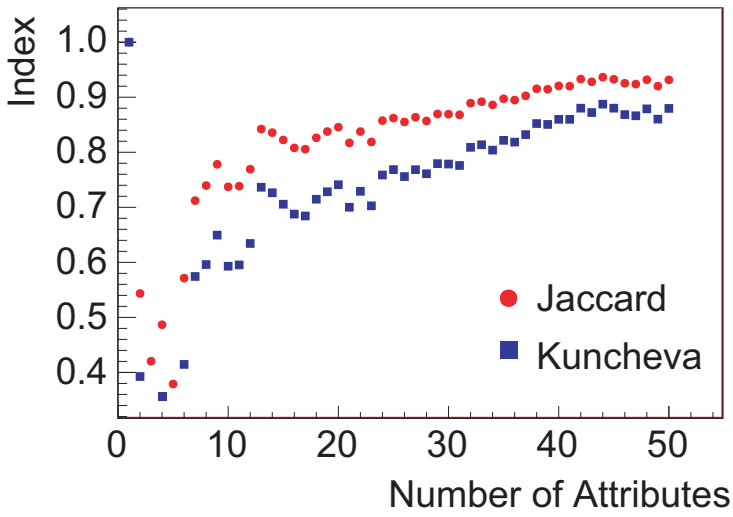


Figure 3.4: Feature Selection stability for an MRMR Feature Selection. The stability increases with an increasing number of selected attributes and starts to saturate for $n_{Att} \geq 20$.

red. One finds that both stability measures increase as the number of attributes increases, going into saturation around $n_{Att} = 20$. It should also be noted that the maximum for the Jaccard and Kuncheva's index is reached if only one attribute is selected. One attribute, however, is not sufficient in order to perform a reliable separation of signal and background using multivariate methods.
Further, it is observed that both stability measures are well above 0.8 in case the number of attributes exceeds 20. The MRMR Feature Selection can thus, be considered stable and was used to extract the attributes used as input parameters for the separation of signal and background using multivariate methods.
Backward Elimination [77], was not considered for Feature Selection, as the computing resources were found to greatly exceed the ones required by MRMR and a Forward Selection.

3.3 Data/MC Comparison

A fair agreement between simulated Monte Carlo events and real data is a prerequisite for a successful application of any machine learning algorithm on real data. As the agreement between simulated neutrino events and real neutrino events cannot be tested before the event selection, the checks focussed on comparing real data to simulated background events. A possible alteration of real data events by the presence of neutrino events can be neglected at this stage, since the fraction of neutrino events in the data sample is well below the 1% level.
Figures 3.5 to 3.7 show the comparison of real data events to simulated background events for the individual attributes obtained from the MRMR Feature Selection. CORSIKA events are depicted in red, whereas real data is shown in black. Si-

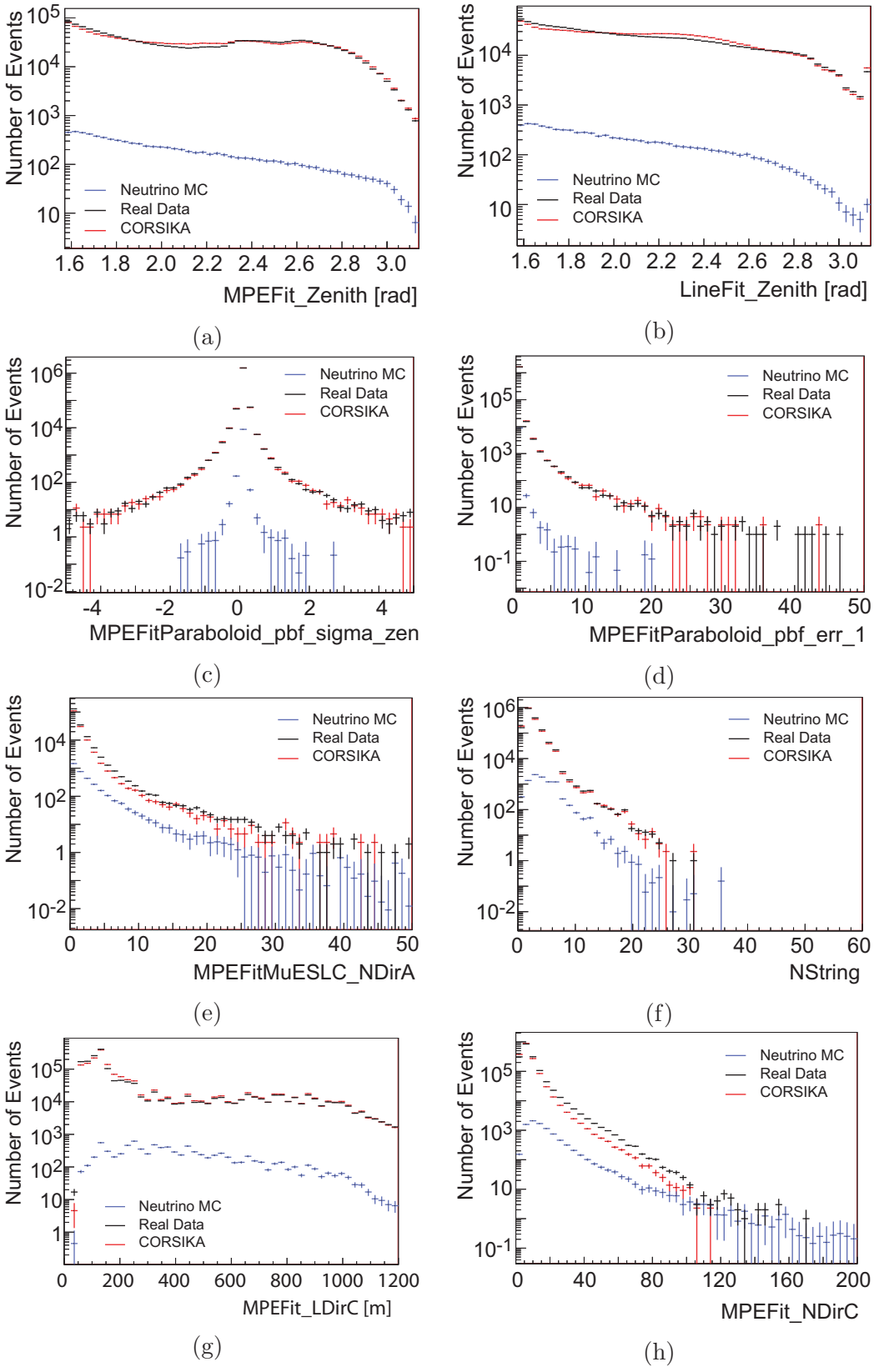


Figure 3.5: Data/Monte Carlo comparison for various attributes at level 4 of the event selection.

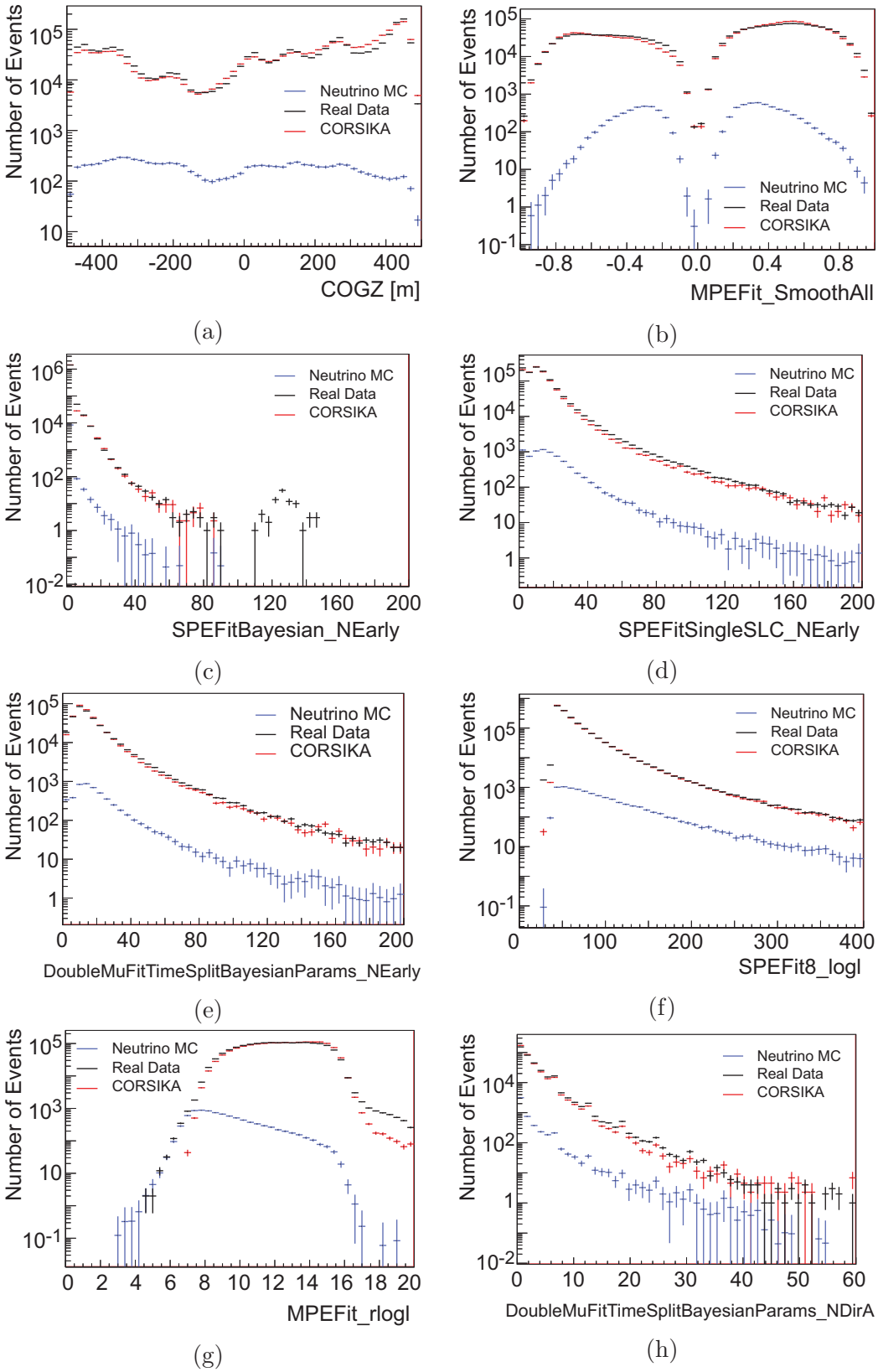


Figure 3.6: Data/Monte Carlo comparison for various attributes at level 4 of the event selection.

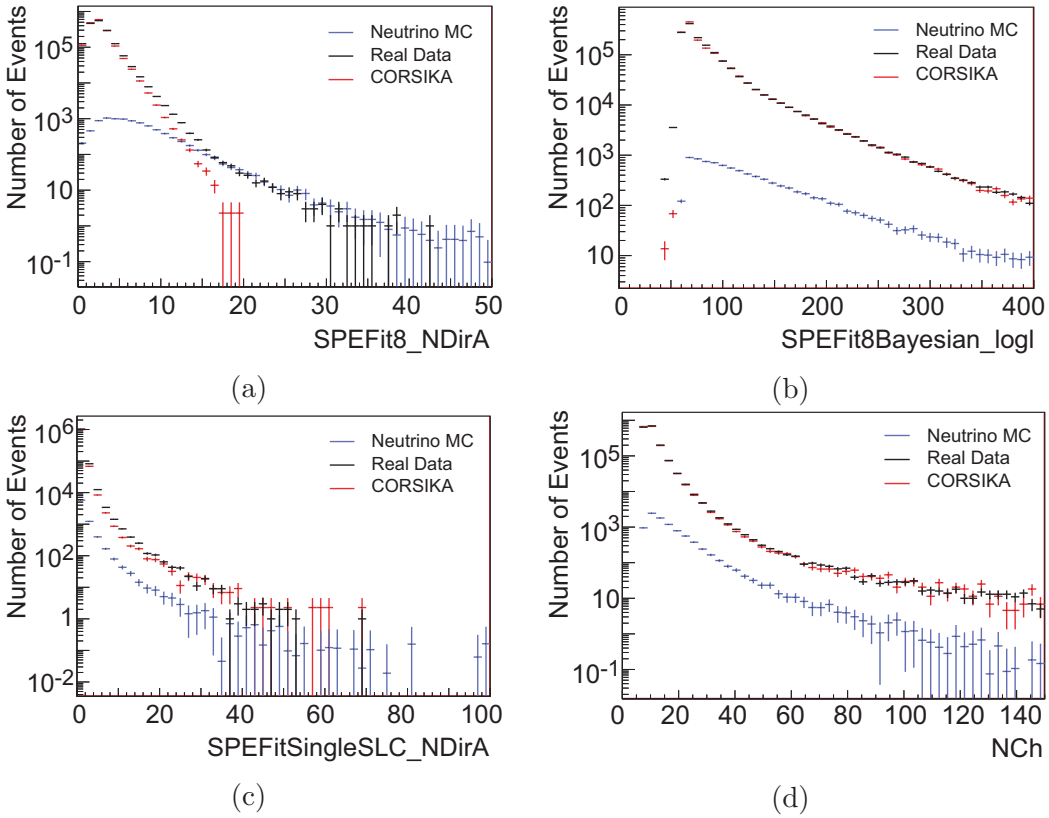


Figure 3.7: Data/Monte Carlo comparison for various attributes at level 4 of the event selection.

mulated neutrino events, in blue, are shown for completeness. All attributes are depicted after the application of precuts, therefore, corresponding to level 4 of the event selection.
In general a good agreement between data and CORSIKA is observed. One should however note, that a complete matchup could not be achieved due to the fact that South Pole ice was used as a detection medium. Like any other natural medium South Pole ice suffers from impurities that need to be taken fully into account in order to model the medium correctly. This modelling of South Pole ice is a particular challenge, as its properties can only be measured indirectly. Details on measurement of South Pole ice properties are given in [79].
In [79] several layers of dust were observed, which change the optical properties of the ice. The largest of these dust layers was found at $\approx -2150\,\mathrm{m}$. It is noteworthy, that the positions of these dust layers strongly influence the event rates at different depths inside the detector. Figure 3.6a shows the center of gravity of the charge collected by the optical modules. In figure 3.6a the point of origin corresponds to the center of the detector ($\approx 2050\,\mathrm{m}$). The different event rates observed for different depths in this plot correspond to the different layers of dust. The smallest event rate is observed at $\mathrm{COGZ} = -100\,\mathrm{m}$, which corresponds to a depth of 2150 m and therefore to the position of the largest dust layer.
The differences in event rate are observed for neutrinos as well. They are, however, less prominent, due to the smaller overall event rate.

3.4 Construction of Additional Attributes

Following an approach presented in [80] three attributes were created in addition to the ones selected in the automated Feature Selection. The attributes were created according to the following equations [80]:

$$\mathrm{Delta\_logl} = \mathrm{SPEFitBayesian\_logl} - \mathrm{SPEFit8\_logl} \tag{3.3}$$

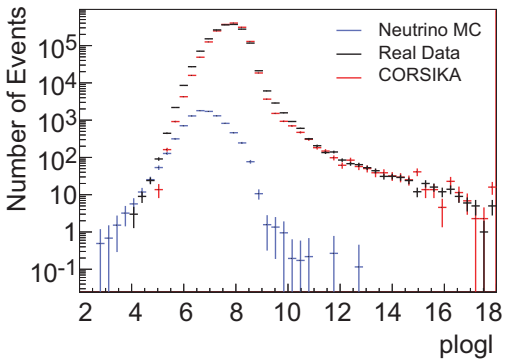
$$\mathrm{plogl} = \frac{\mathrm{MPEFit\_logl}}{\mathrm{NCh} - 2.5} \tag{3.4}$$

$$\mathrm{DeltaZenith} = |\mathrm{LineFit\_Zenith} - \mathrm{MPEFit\_Zenith}| \tag{3.5}$$

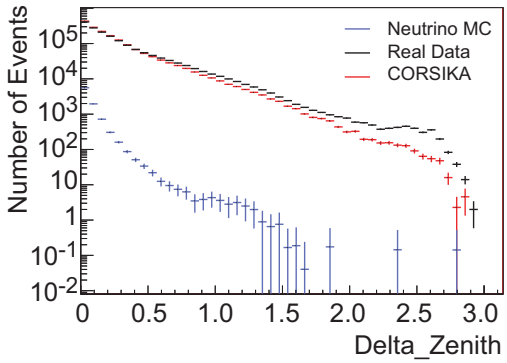
All three attributes were found to be among the 10 attributes showing the highest separation power (see section 3.6), with Delta_logl being by far the strongest attribute in the event selection. The attribute plogl was found to be the fourth strongest, Delta_Zenith was found to be the eighth strongest attribute. It can thus, be concluded that the event selection was significantly improved by adding these additional attributes.

3.5 Random Forest Event Selection

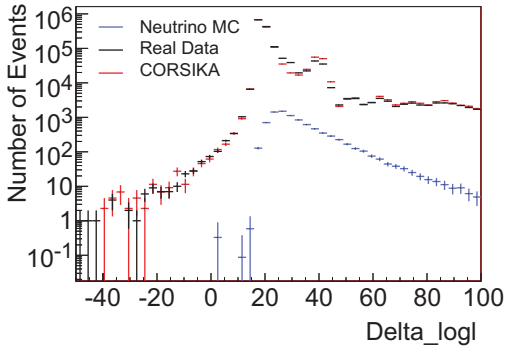
Tree-based machine learning algorithms are well known for their stability and good interpretability. Moreover, very good results were recently obtained using Boosted



(a) Data MC comparison for plogl.



(b) Data MC comparison for Delta_Zenith.



(c) Data MC comparison for Delta_logl.

Figure 3.8: Data/Monte Carlo comparison for the three attributes generated prior to the Random Forest event selection. In general a good agreement between data and MC is observed. The CORSIKA distributions are found to be rather distinct from the neutrino simulation.

Decision Trees for the separation of signal and background events in an atmospheric neutrino analysis with IceCube in the 40-string (IC-40) configuration [22]. In addition, ensembles of trees are in general found to outperform single classification trees. Therfore, an ensemble of trees (Random Forest) was chosen for the separation of signal and background in this analysis.
A Random Forest was chosen over Boosted Decision Trees, in order to achieve an indirect comparison of both algorithms on different detector configurations and analyses. In addition to being rather robust classifiers, Random Forests were found to outperform BDTs and several other machine learning algorithms in reference [81].
The Random Forest from the Weka Extension package for RapidMiner [82] was used, since studies showed a somewhat more stable behaviour, compared to the default implementation of a Random Forest in RapidMiner [83].
The Random Forest performance was evaluated in a 5-fold cross validation using 70 093 signal and 749 921 background events. To avoid overtraining the number of events entering the learner was restricted to 2.7×10^4 per class. The simulated neutrino spectrum used for training the Random Forest is distributed according to a power law of the form $\frac{dN}{dE} \propto E^{-2}$. This spectrum differs significantly from the energy distribution of atmospheric neutrinos ($\frac{dN}{dE} \propto E^{-3.7}$).
There are, however, two reasons for using an E^{-2}-spectrum as an input for the training of a Random Forest. Neutrinos are created following E^{-1} and E^{-2} spectra within the IceCube Monte Carlo chain. Moreover, neutrinos are forced to interact once their track leads them close enough to the detector. The application of both procedures ensures that a large number of neutrino events is simulated for all energies within a reasonable amount of time. Specific neutrino spectra, e.g. an atmospheric neutrino spectrum, can then be obtained by a reweighting of events according to the desired spectrum.
The first argument for using an E^{-2}-spectrum can be made as follows: Due to the power law behaviour of the atmospheric neutrino spectrum the largest part of the events is observed in the low energy region. This analysis, however, aims at measuring the atmospheric neutrino spectrum at high and intermediate energies, where the observed uncertainties are large and the contribution of the prompt component to the overall flux becomes nonnegligible. Thus, in order to avoid a too strong bias towards the low energy region an E^{-2}-spectrum was chosen, as this spectrum ensures a sufficient number of high energy neutrino events to learn on.
Secondly, tests using a reweighted spectrum for the Random Forest training, showed a siginificant drop in the recall of neutrino events.
Atmospheric muon events simulated using the air shower simulation Corsika [71] were used as background examples. The simulated datasets included single, as well as double and triple coincident muons. Coincident muon events are detected in case two or more muons, produced in different air showers, enter the detector from different directions. Coincident muons are not to be confused with muon bundles, where a number of muons, produced in the same air shower, enters the

detector from the same direction. In general, the tracks of these double and triple coincident events cannot be easily reconstructed. Thus, these events might mimic an upgoing neutrino induced muon. The track quality parameters, on which the Random Forest is trained, should be rather poor for most of the coincident events. Hence, the Random Forest is expected to be able to recognise and reject coincident muon events.
The ratio of signal and background events used for the training of the forest was chosen to be $1:1$. Using this ratio might not seem intuitive, since the number of atmospheric muon events exceeds the number of neutrino events by roughly three orders of magnitude, which in turn results in a highly skewed distribution. Using the real signal to background ratio for the training of the Random Forest, on the other hand, would require a large number of background events, and thus exceed the available computing resources. Note, that 90% of all muons have been removed prior to the Random Forest. In addition, there would not be enough neutrino events left at the lower lying nodes in order to perform meaningful splits. In fact, if the real signal to background ratio was used, no neutrino events might be left at all after the first couple of splits. This, however, would render the advantages of tree-based classification algorithms useless.
Tests using different signal to background ratios for training resulted in a better recognition of signal or background, respectively, depending on the majority class. This result is quite intuitive, since the forest has seen more events of the majority class and hence, more patterns for this specific class have been memorised. Recall and background rejection are of almost equal importance for the classification task at hand. Using a ratio of $1:1$ was found to be a good and reliable tradeoff.
The ice model Spice-1 [84], which utilises the IceCube LED calibration system for the determination of ice properties, was used for the production of signal and background Monte Carlo. The nominal DOM efficiency was set to 1.0. The influences of ice model and DOM efficiency on the event selection were studied in section 3.8.

Figure 3.9 depicts the output score of the Random Forest. This output score, called signalness, can be interpreted as the probability of an event to be a signal event. Real data is shown in black, CORSIKA in red and NUGEN in blue. The sum of CORSIKA and NUGEN is depicted in magenta. Two distinct peaks are observed at signalness $s = 1.0$ (signal peak) and $s = 0.0$ (background peak), indicating a well trained forest. The distribution of CORSIKA events has been scaled as the number of data events exceeded the number of available background events by roughly a factor of 2.
A good agreement between the sum of NUGEN and CORSIKA is observed for the high signalness region ($s \geq 0.98$). Between signalness $s = 0.2$ and $s = 0.8$ a slight mismatch of the two histograms is observed, leading to an underestimation of background in this region. Such an underestimation of background for lower signalness values might, however, result in an underestimation of background events in the high signalness region and, thus, in an underestimation of atmospheric muons in the final neutrino sample. One should note, however, that the slope of both

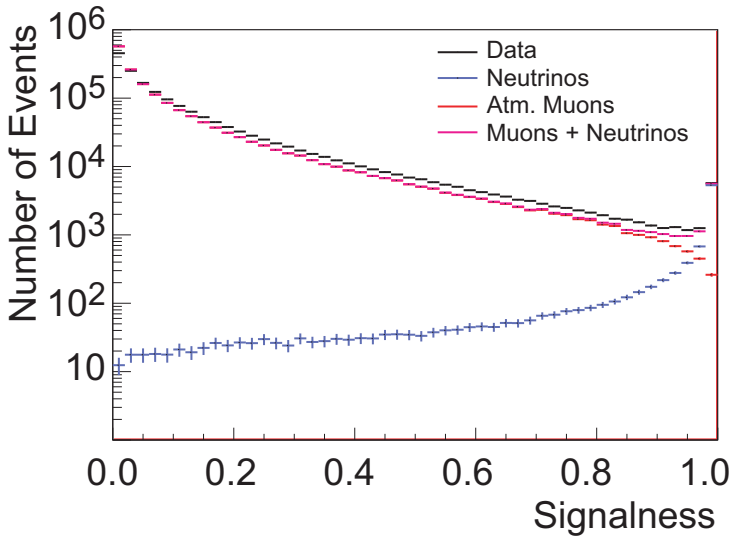


Figure 3.9: Random Forest output score for data (black), CORSIKA (red) and NUGEN (blue). The sum of CORSIKA and NUGEN is depicted in magenta. A good agreement between the sum of CORSIKA and NUGEN is observed in the high signalness region ($s \geq 0.98$). A slight mismatch is observed between signalness $s = 0.2$ and $s = 0.8$. The slope of both graphs, however, is similar.

histograms is similar in the intermediate signalness region, indicating that the mismatch can be resolved through the application of a simple scaling procedure.
From figure 3.9 one further finds that an additional cut needs to be placed on the signalness in order to obtain a high purity neutrino sample. In order to keep the contamination of atmospheric muons in the final neutrino sample as low as possible, only signalness cuts above $s = 0.99$ are considered (see table 3.1).

3.5.1 Background Scaling

Figure 3.10 shows the ratio of data/$(\nu_{MC} + \mu_{MC})$ in the interval from $s = 0.2$ to $s = 0.8$. Within this interval the dominant part of data is still formed by atmospheric muons, allowing for the determination of a scaling factor, that corrects for the data/MC mismatch observed in figure 3.9. The fit to the data points is depicted in red. The scaling factor f was determined to $f = 1.23$.
Figure 3.11 depicts data (black), NUGEN (blue) and CORSIKA (red) as a function of signalness after the application of a scaling factor on the CORSIKA. The sum of simulated muon and neutrino events is shown in magenta. One finds that the data/MC mismatch in the regions of high and medium signalness ($s \geq 0.2$) is resolved by the application of this scaling factor. The enhanced mismatch for $s < 0.2$, however, is not relevant within the scope of this analysis, as all events in this region will be rejected by a cut on the signalness.
Figure 3.12 shows the same plot but on a linear scale. Again an excellent agreement between data and Monte Carlo simulations is observed, showing that no

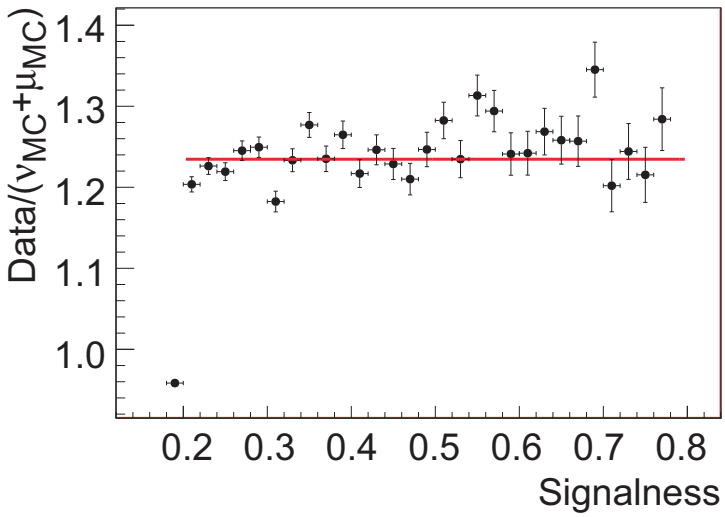


Figure 3.10: Ratio data/$(\nu_{MC}+\mu_{MC})$ as a function of signalness in the intervall $s = 0.2$ to $s = 0.8$. Within this signalness region, data are still dominated by atmospheric muons, allowing for the determination of a constant scaling factor, correcting for the data/MC mismatch. The fit to data points is depicted in red.

mismatches were hidden in the logarithmic plot. Moreover, the very good agreement between data and Monte Carlo with respect to the Random Forest output score (signalness), shows that the forest is well trained and that the application on real data is justified.
Figures 3.13 and 3.14 show the high signalness region, on a logarithmic and a linear scale respectively, after the application of a scaling facor $f = 1.23$ on the CORSIKA events. Real data is depicted in black, NUGEN in blue and CORSIKA in red. The sum of simulated muon and neutrino events is shown in magenta. Again a very good agreement between data and Monte Carlo simulation is observed, indicating a stable and well trained Random Forest.

3.5.2 Statistical Uncertainties of the Event Selection

Statistical fluctuations in the training and test set will influence the estimation of the remaining signal and background events after the application of a cut on the signalness s. To obtain a reliable estimate these statistical fluctuations need to be taken into account.
In this work the statistical uncertainties were obtained using a 5-fold cross validation. The outcomes of the individual iterations of the cross validation were then evaluated for different signalness cuts. For all of these cuts the number of signal and background events was averaged and the statistical uncertainties on those numbers were calculated.
The results of this compuation are summarised in table 3.1. The errorbars, especially on the simulated neutrino events and on the sum of simulated muon and

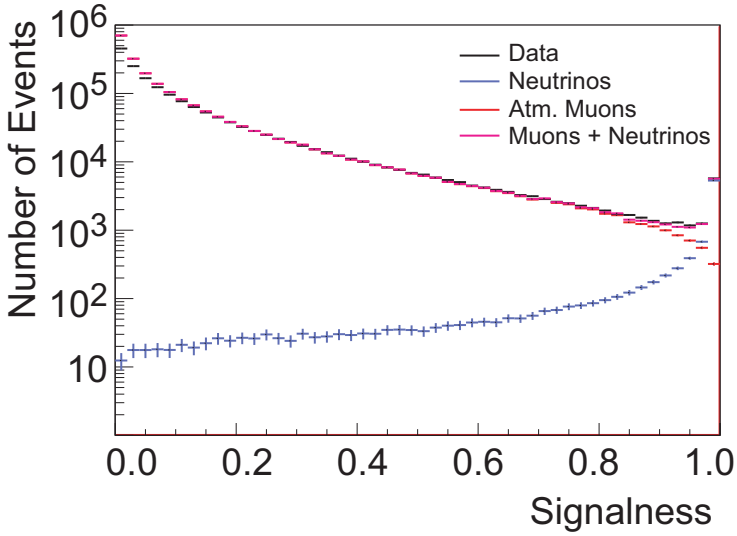


Figure 3.11: Real data (black), NUGEN (blue) and CORSIKA (red) as a function of signalness after the application of a scaling factor on the CORSIKA. The sum of muon and neutrino events is shown in magenta. One finds that the data/MC mismatch is resolved for signalness values exceeding 0.2. The created mismatch for signalness $s < 0.2$ is not relevant within the scope of this analysis, as all events within this region will be rejected by an additional cut on the signalness.

neutrino events are rather small. These small errorbars reflect the very small difference in the performance of the forest, when trained and tested using different subsets of events. The Random Forest can, thus, be considered stable, which again justifies the application of the algorithm on real data.
The errorbars on the remaining background events are larger (up to 91% depending on the cut), but can be explained by statistical effects. Since the number of real data events exceeded the number of available background Monte Carlo by more than a factor of two, the output of the Random Forest, in terms of background events, needed to be scaled in order to obtain a reliable background estimate. Moreover, only very few simulated atmospheric muon events (on average less than 1 per cross validation iteration) survived a signalness cut of $s = 1.0$. Thus, the relative deviation between individual iterations of the cross validations becomes large, leading to a large RMS error on the mean.
Table 3.1 also summarises the anticipated purity of the final neutrino sample for different cuts on the signalness. The purity of the final neutrino sample should be be greater than 95% in order to be able to perform a proper analysis of the sample. The impurities imposed by the remaining atmospheric muons, however, should be kept as small as possible. Moreover, the energy of these remaining muons cannot be estimated with 100% certainty. Thus, in order to avoid a fake signal at high energies in the final neutrino sample, a strict cut on the signalness needs to be chosen.
The final cut on the signalness was chosen to $s = 1.0$, in order to achieve a high

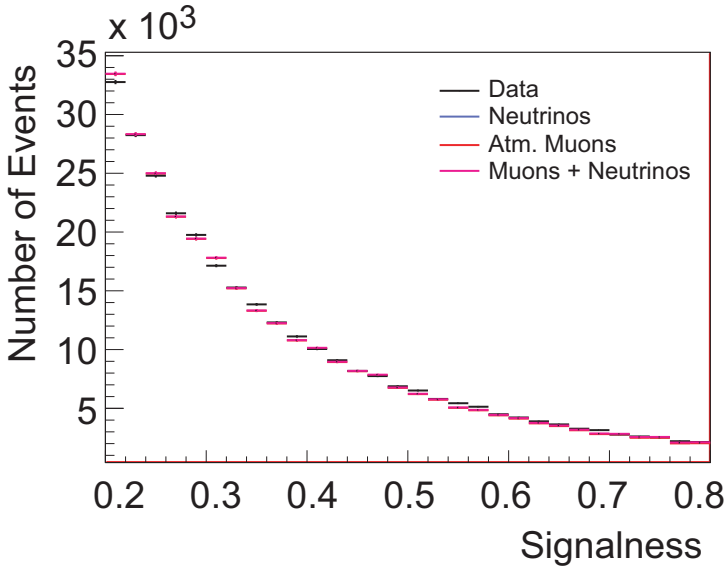


Figure 3.12: Real data (black), NUGEN (blue) and CORSIKA (red) as a function of signalness after the application of a scaling factor on the CORSIKA, shown on a linear scale. The sum of simulated muon and neutrino events is shown in magenta. After the application of a scaling factor an excellent agreement between data and Monte Carlo is observed, even on a linear scale.

quality sample of atmospheric neutrino events with a low background contamination. By applying this final cut on the signalness, the number of background events is reduced from 1.68×10^6 to 11 ± 10. This corresponds to a background rejection of more than five orders of magnitude. At a signal efficiency of 31.9% with respect to level 4 2873 out of 8990.9 neutrino events are retained by the application of the signalness cut. The sample obtained by this final cut, is in the following be referred to as level 5 or neutrino level.

3.6 Attribute Importance

The WEKA Random Forest can not only be utilised for classification but also for checking the importance of the individual attributes. This functionality was used to validate the performance of the MRMR Feature Selection. Within this study all three constructed attributes (see section 3.4) were ranked among the top eight attributes. This again emphasises the importance of these three attributes for the use in the learning process. The outcome of the study is summarised in table 3.2. As not all of the attributes can be discussed in detail only the top five attributes are discussed in this section.
For Delta_logl the importance within the event selection becomes clear when looking at the distributions of signal and background Monte Carlo, respectively (see figure 3.8c). Two clearly distinct peaks are observed for the background MC, whereas only one peak is observed for the signal Monte Carlo. The maxima of

Cut	ν_{MC}	μ_{MC}	$\nu_{\mathrm{MC}} + \mu_{\mathrm{MC}}$	Data Events	Purity [%]
0.980	5392 ± 49	359 ± 87	5751 ± 100	5831	93.84
0.982	5306 ± 40	317 ± 65	5632 ± 76	5691	94.43
0.984	5197 ± 38	261 ± 50	5458 ± 63	5551	95.30
0.986	5079 ± 31	219 ± 36	5298 ± 48	5389	95.94
0.988	4933 ± 36	182 ± 31	5115 ± 48	5216	96.51
0.990	4777 ± 38	149 ± 24	4926 ± 45	5010	97.03
0.992	4596 ± 35	120 ± 31	4716 ± 47	4473	97.32
0.994	4371 ± 35	87 ± 27	4458 ± 44	4507	98.07
0.996	4075 ± 39	45 ± 16	4120 ± 42	4138	98.91
0.998	3642 ± 40	34 ± 19	3676 ± 44	3648	99.07
1.000	2872 ± 47	11 ± 10	2883 ± 48	2765	99.60

Table 3.1: Expected number of simulated muon and neutrino events for various cuts on the signalness, compared to the number of events observed in real data. In general a good agreement between the sum of simulated events and real data events is observed. A complete matchup could not be achieved as certain systematics, such as the modelling of the ice have not been taken into account at this point.

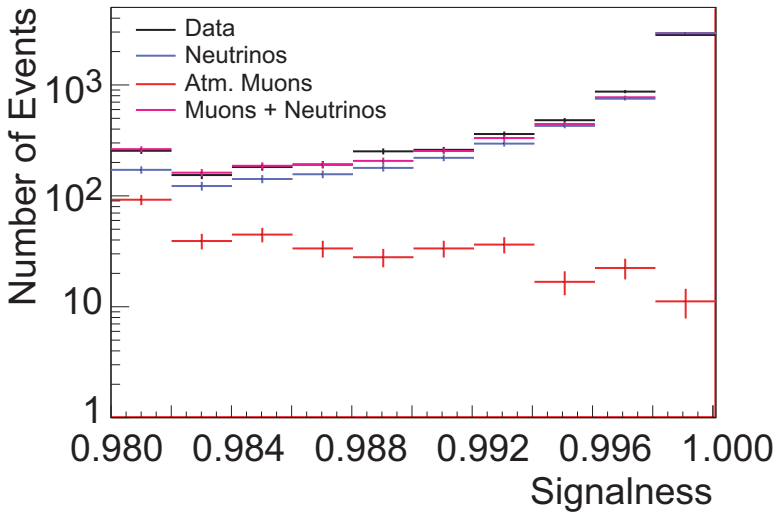


Figure 3.13: Random Forest output score for the high signalness region. Due to the large contamination of the level 4 sample with atmospheric muons a signalness cut within this region needs to be chosen in order to achieve the required purity.

both peaks in the background are clearly shifted with respect to the simulated signal distribution, clearly depicting the importance of this particular attribute for the event selection.
MPEFit_LDirC (see figure 3.5g) shows a strong peak in the background distribution at small track length, possibly originating from poorly reconstructed events. Compared to the background, the signal distribution does not show a strong maximum and in general appears to be more uniformly distributed.
Signal and background distributions were found to be less distinct for LineFit_Zenith (see figure 3.5b). Moreover, no clear maxima can be observed for any of the distributions. The importance of this attribute, thus, most likely originates from synergy effects with other attributes.
For plogl (see figure 3.8a) two distinct maxima can be observed for the signal and background distribution, respectively, emphasising the importance of this attribute.
The signal and background distributions observed for MPEFit_SmoothAll were also found to be clearly distinct. The maxima of both distributions were observed at different positions. Additionally the background distribution appears to be less steeply falling, compared to the distribution of simulated signal events.
In conclusion one finds that the importance of individual attributes can in most cases be attributed to distinct distributions in signal and background. For some attributes, however, synergy effects with other attributes appear to be equally important. Such synergy effects can in general not be spotted by eye. It can thus be concluded that a detailed algorithm based Feature Selection plays a significant role in the separation of signal and background using machine learning techniques.

Attribute	Importance
Delta_logl	11.274
MPEFit_LDirC	7.219
LineFit_Zenith	5.013
plogl	4.845
MPEFit_SmoothAll	4.668
MPEFit_NDirC	4.030
SPEFit8_NDirA	3.914
DeltaZenith	3.788
MPEFit_rlogl	3.779
SPEFit8Bayesian_NLate	3.547
NString	3.180
SPEFit8Bayesian_logl	2.254
MPEFit_Zenith	2.175
SPEFit8_logl	2.103
COGZ	2.084
MPEFit_logl	1.891
MPEFitParaboloid_pbf_sigma_zen	1.520
MPEFitParaboloid_pbf_err_1	1.116
SPEFit8Noisey_NEarly	0.756
DoubleMuFitTimeSplitBayesianParams_NDirA	0.750
NCh	0.661
MPEFitMuESLC_NDirA	0.646
DoubleMuFitGeoSplitBayesianParams_StatusI	0.441
SPEFitSingleSLC_NDirA	0.398
DoubleMuFitTimeSplitBayesianParams_NEarly	0.334
SPEFitSingleSLC_NEarly	0.271
SPEFit8Bayesian_StatusI	0.212
DoubleMuFitTimeSplitBayesian_nmini	-0.005

Table 3.2: Attribute importance as estimated by the Weka Random Forest.

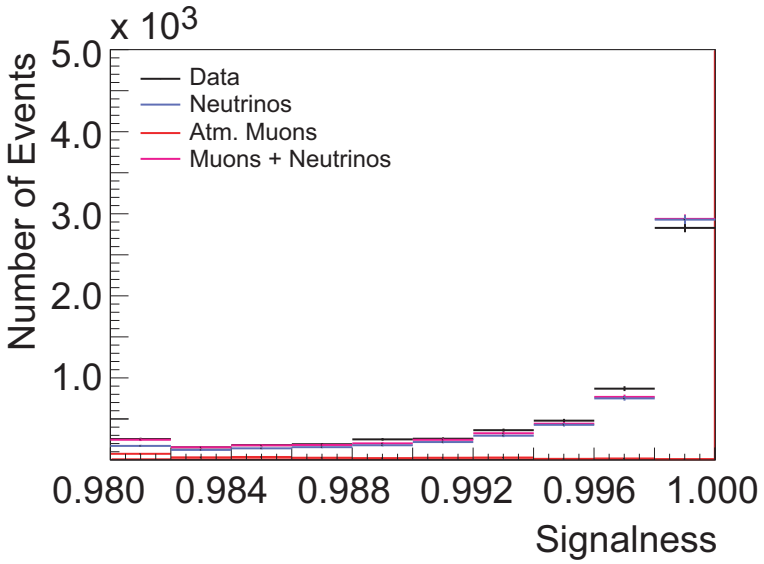


Figure 3.14: Data (black), NUGEN (blue) and CORSIKA (red) as a function of signalness in the high signalness region, depicted on a linear scale. The sum of simulated muon- and neutrino events is shown in magenta. A very good agreement between data and Monte Carlo simulations is observed.

3.7 Data/MC Comparison after Event Selection

A comparison of real data and simulated neutrino events after application of the Random Forest can serve as a cross check on the reliability of the forest. This comparison is depicted in figures 3.15 to 3.17. Data is shown in black, whereas the simulated neutrino events are depicted in red. In order to have enough events to perform a statistically meaningful comparison these checks were carried out on the full year of IceCube in the 59-string configuration (346.3 days).
In general a good agreement between data and Monte Carlo simulation is observed. One should, however, note that the number of data events is $\approx 8\%$ smaller than the number of simulated events. This is due to the fact that an underfluctuation in the event rate was observed with respect to the number of events expected from Monte Carlo simulation. The underfluctuation is discussed in section 3.9. No scaling was applied to any of the plots.
This underfluctuation mainly appears close to the horizon, as can be seen from figures 3.15b and 3.15a. These slight mismatches between data and simulation can be tolerated as no overfluctuations were observed in any of the distributions. Overfluctuations in certain regions of an input variable would indicate a possible underestimation of the background contamination in that region.
Furthermore, it should be noted, that certain features of the distributions are smoothened by the application of the Random Forest. This is particularly visible for the center of gravity in the z-direction COGZ, depicted in figure 3.16a. Compared to the distribution before the application of the forest, shown in figure 3.6a,

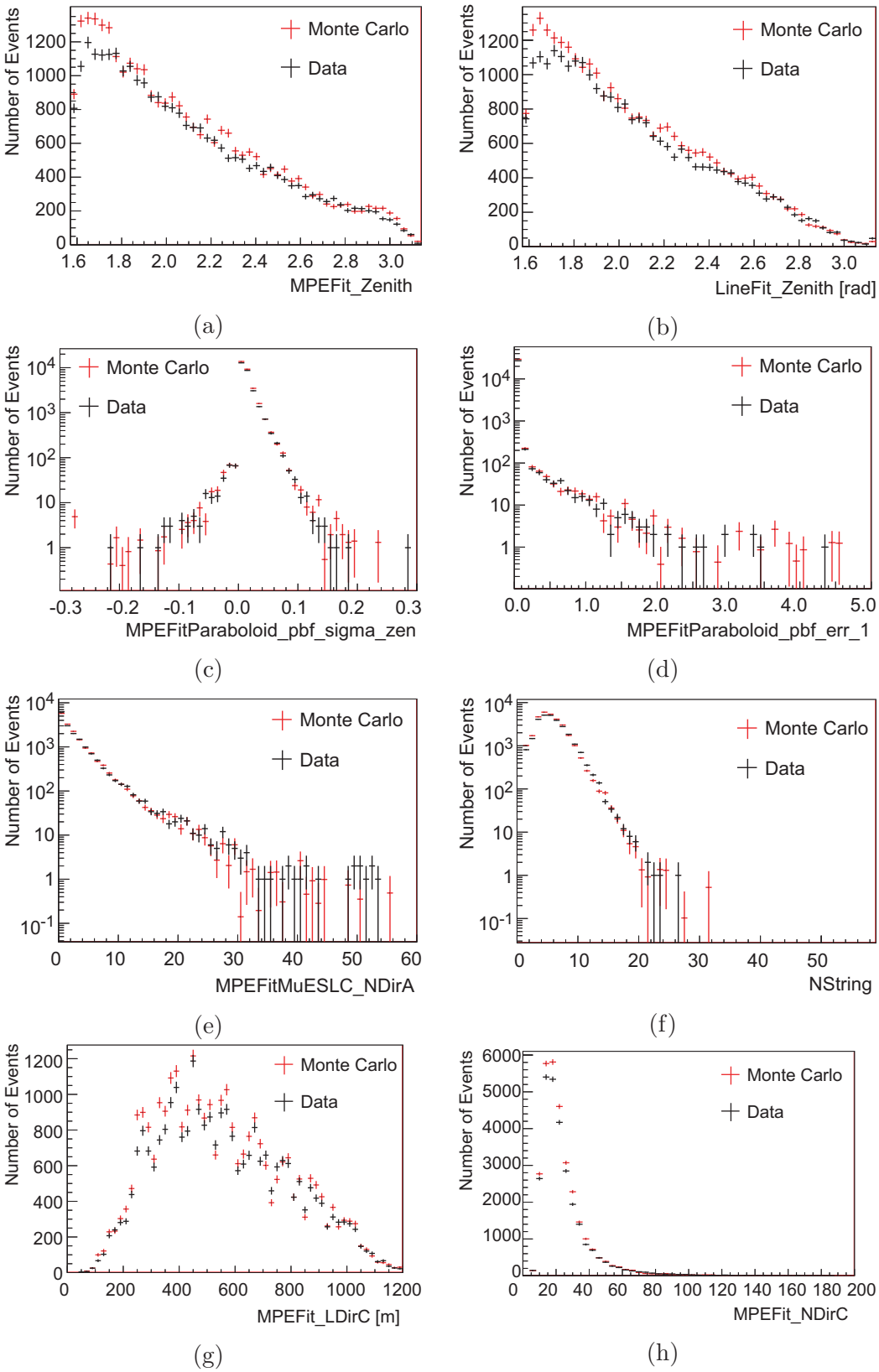


Figure 3.15: Data/Monte Carlo comparison for various attributes at the final neutrino level.

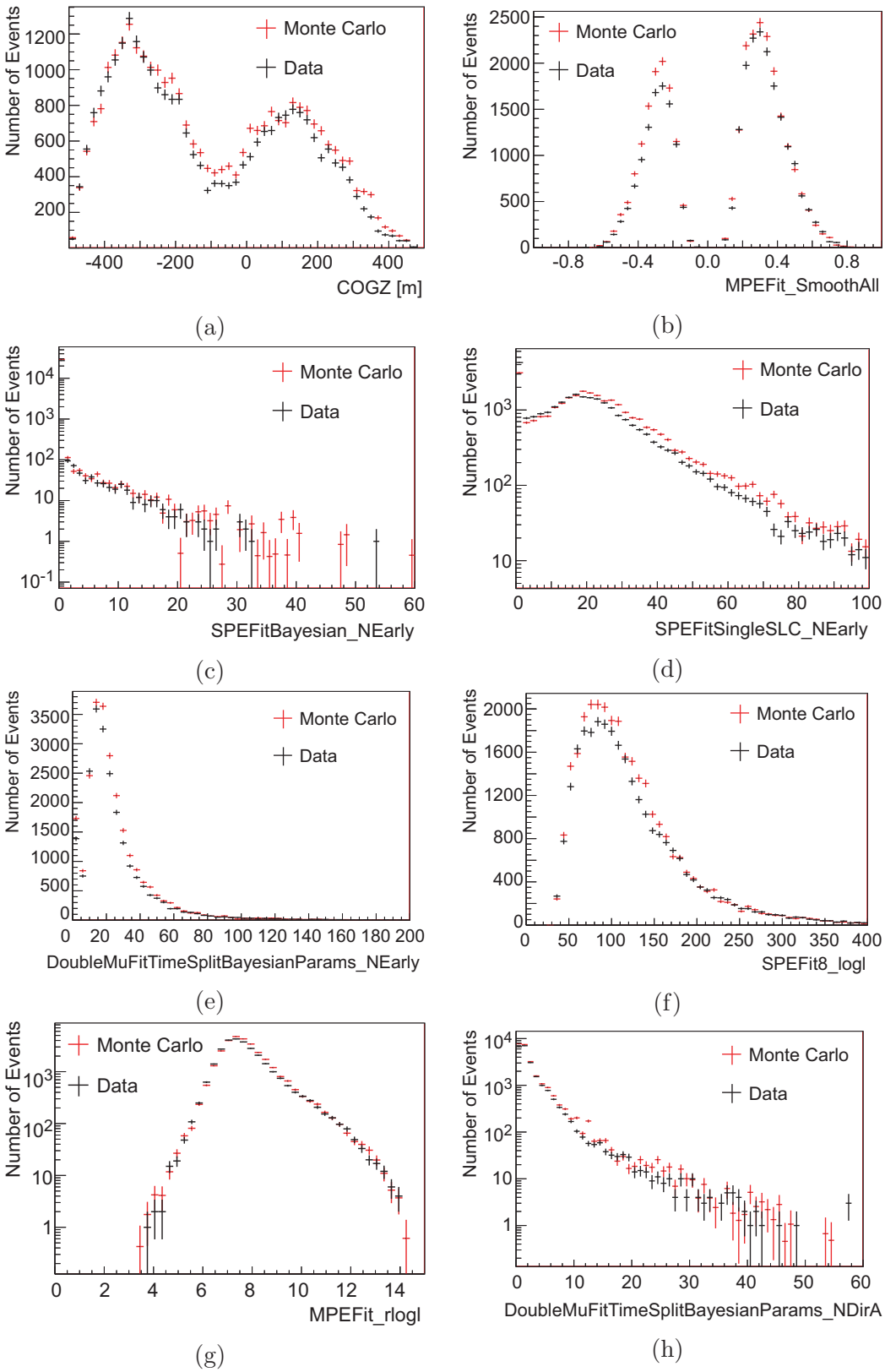


Figure 3.16: Data/Monte Carlo comparison for various attributes at the final neutrino level.

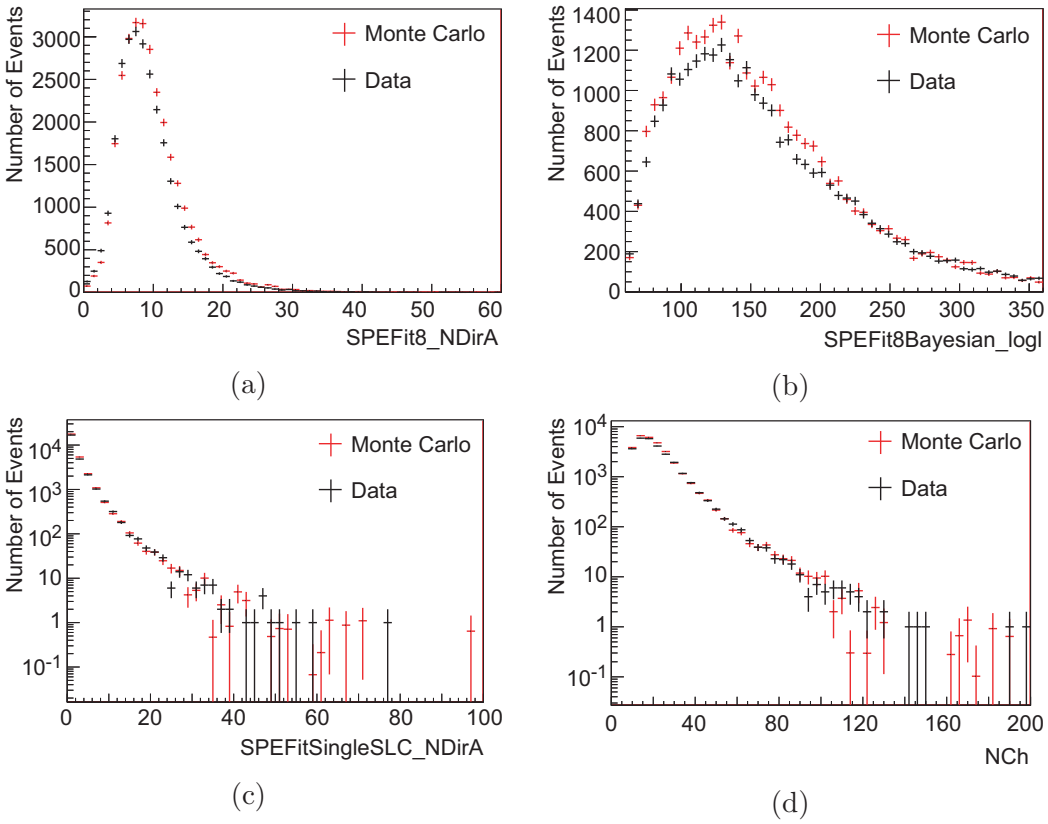


Figure 3.17: Data/Monte Carlo comparison for various attributes at the final neutrino level.

most of the modulations in the event rate are smoothened. The position of the largest dust layer, however, can still be inferred from the distribution.

3.8 Systematic Uncertainties of the Event Selection

As the separation of signal and background depends on the Monte Carlo simulation used for training, it also depends on the ice model and DOM efficiency used to generate these simulations. Differences in these two quantities from reality will thus result in different performances of the forest on real data. Hence, ice model and DOM efficiency need to be taken into account as systematic uncertainties of the event selection.

3.8.1 DOM Efficiency Uncertainties

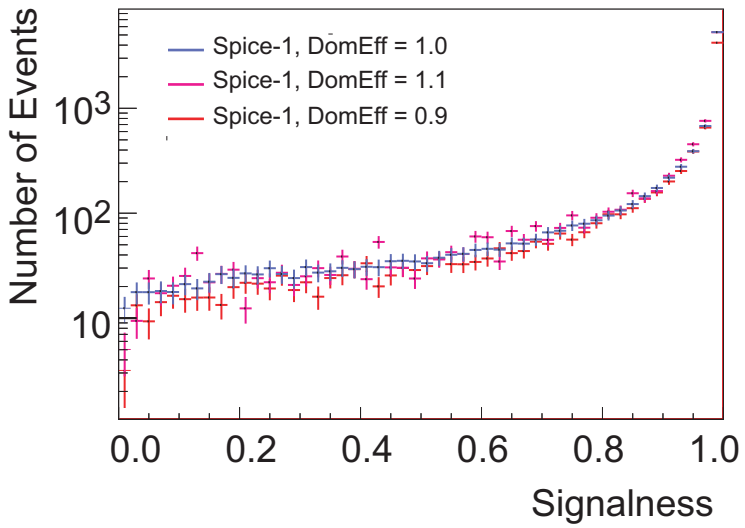


Figure 3.18: The Random Forest output score for different DOM efficiencies. A DOM efficiency of 1.0, used as a standard in this analysis, is shown in blue. DOM efficiencies of 1.1 and 0.9 are shown in red and magenta respectively. It should be noted that a DOM efficiency of 1.1, corresponds to an increase of 10% above the nominal value, as an absolute efficiency of 110% is, of course, unphysical. A good agreement between all three efficiencies is observed. A slight mismatch appears in the very last bin.

Figure 3.18 shows the Random Forest output score for different DOM efficiencies. A DOM efficiency of 1.0 was used as a standard in this analysis and is shown in blue. DOM effiencies of 1.1 and 0.9 are depicted in red and magenta, respectively. It should be noted that all DOM effiencies cited here are relative efficiencies. This means that DOM efficiencies of 1.1 and 0.9 correspond to a 10% increase/decrease with respect to the simulation standard. One should further note that an increase or decrease in DOM efficiency might mimic other systematic

Cut	D.E. = 1.0	D.E. = 1.1	D.E. = 0.9	$\frac{\text{D.E. = 1.1}}{\text{D.E. =1.0}}$	$\frac{\text{D.E.=0.9}}{\text{D.E.=1.0}}$
0.980	5392 ± 49	5393	4273	1.00	0.79
0.982	5306 ± 40	5268	4191	0.99	0.79
0.984	5197 ± 38	5142	4078	0.99	0.78
0.986	5079 ± 31	4979	3960	0.98	0.78
0.988	4933 ± 36	4837	3811	0.98	0.79
0.990	4777 ± 38	4655	3643	0.97	0.76
0.992	4596 ± 35	4429	3446	0.96	0.78
0.994	4371 ± 35	4160	3215	0.95	0.74
0.996	4075 ± 39	3748	2890	0.92	0.71
0.998	3642 ± 40	3194	2448	0.88	0.67
1.000	2872 ± 47	2396	1803	0.82	0.63

Table 3.3: Dependency of the number of retained neutrino events on the DOM efficiency. All results were obtained using Spice-1 at a DOM efficiency of 1.0 were used for the training of the forest.

effects, e.g. interaction cross sections and increased/decreased scattering and absorbtion lengths.
A good agreement in shape is observed for all three graphs, except for the very last bin. A DOM efficiency of 0.9 was found to result in a lower overall event rate which becomes prominent in the last bin. The event rates as well as the systematic uncertainties introduced to the event selection by variation of the DOM efficiency are summarised in table 3.3.
All in all, one finds that the dependency of the event selection on changes in the DOM efficiency is rather strong, resulting in a 18% deviation for a 10% increased DOM efficiency. The deviation for 10% decreased DOM efficiency is even larger ($\approx 40\%$). Due to the large disagreement between data and simulated events generated with a nominal DOM efficiency of 0.9, it can be concluded that the uncertainties derived for this DOM efficiency largely overestimate the actual uncertainty. Moreover, the DOM efficiency was increased by 10% for the simulation standard of IceCube in the 79-string configuration.

3.8.2 Ice Model Uncertainties

Figure 3.19 shows the output score of the Random Forest for different ice models. Spice-1 [84], used as a standard in this analysis, is shown in blue. Spice:Mie [85] and WHAM are depicted in red and black, respectively. A nominal DOM efficiency

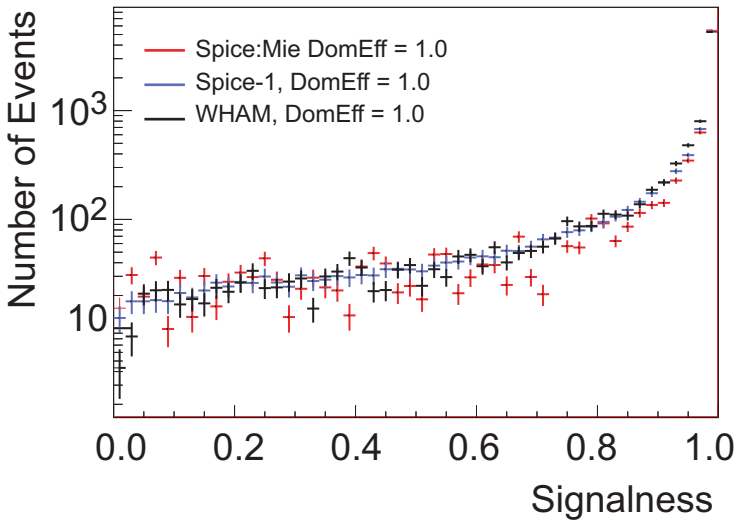


Figure 3.19: Random Forest output score for different ice models. Spice-1 [84], used as a standard in this analysis, is shown in blue. Spice:Mie [85] and WHAM are depicted in red and black, respectively. A nominal DOM efficiency of 1.0 was used for all three ice models. Slight deviations are observed for signalness values $s \leq 0.9$. For the high signalness region all three graphs were found to agree well.

of 1.0 was used for all three ice models. Slight deviations were observed especially for Spice:Mie. For the high signalness region, however, all three ice models were found to agree.
The expected data rates for Spice-1, Spice:Mie and WHAM for various cuts on the signalness are summarised in table 3.4. One finds that the expected event rates for Spice-1 and Spice:Mie agree within 5%. Larger deviations are observed when comparing Spice-1 and WHAM. These deviations are a result of the large mismatches between Spice-1 and WHAM for various attributes used for selection. This becomes most visible for signalness cuts of $s \geq 0.998$ and $s = 1.0$.
Since the mismatches are also observed between WHAM and data, it can be concluded that the deviations between Spice-1 and WHAM overestimate the systematic uncertainties. Thus, the uncertainties introduced by variations of the ice model can be approximated by the difference between Spice-1 and Spice:Mie. Accordingly, the systematic uncertainties introduced through assumptions on the ice is smaller than 5%.

3.8.3 Total Systematic Uncertainties of the Event Selection

Summarising the results obtained in the application of the Random Forest on simulation produced with different ice models and DOM efficiencies, one finds that the total systematic uncertainty on the event selection is 19%. This value is fully dominated by the estimated 18% uncertainty originating from changes in DOM efficiency. The event rate obtained on real data can, therefore, be expected to vary within 19% from the event rate estimated on Monte Carlo simulation.

Cut	Spice-1	Spice:Mie	WHAM	$\frac{\text{Spice:Mie}}{\text{Spice-1}}$	$\frac{\text{WHAM}}{\text{Spice-1}}$
0.980	5392 ± 49	5541	5356	1.028	0.993
0.982	5306 ± 40	5440	5233	1.025	0.986
0.984	5197 ± 38	5352	5101	1.030	0.982
0.986	5079 ± 31	5221	4944	1.028	0.973
0.988	4933 ± 36	5097	4771	1.033	0.967
0.990	4777 ± 38	4966	4598	1.040	0.963
0.992	4596 ± 35	4790	4374	1.042	0.952
0.994	4371 ± 35	4580	4087	1.048	0.928
0.996	4075 ± 39	4272	3705	1.048	0.909
0.998	3642 ± 40	3792	3181	1.041	0.873
1.000	2872 ± 47	2885	2310	1.005	0.804

Table 3.4: Dependency of the number of retained neutrino events on the ice model. All results were obtained using Spice-1 at a DOM efficiency of 1.0 for the training of the forest.

Changes in the rate of atmospheric muon events surviving the forest cannot be tested directly. As only a handful of events survives the cuts, a large number of background events is required as input to the overall event selection procedure. The production of these background events would require further time and computing resources. Moreover, one finds that the atmospheric muon rate is already dominated by the statistical uncertainties, due to the small number of events surviving the final signalness cut. Assuming a dependency on systematics comparable to neutrino events, one finds that the total errorbar on the number of atmospheric muons is still dominated by the statistical uncertainty. Even a 100% systematic uncertainty would not become visible in the event rate on real data, due to the much larger rate of neutrino events.

3.9 Application to full year of IC59

The application of the individual event selection steps, developed on 10% of the data yielded 27 771 neutrino events in 346.3 days of livetime, corresponding to 80 neutrino events per day. Compared to the 29 884 neutrino events expected from simulation, this corresponds to an underfluctuation of approximately 8%. This underfluctuation is found to be well within the range expected from the studies on the event selection systematics.

3.10 Comparison to other Results

The IC-59 point source (PS) analysis [86] as well as the IC-40 atmospheric neutrino analysis [22, 80] both utilised Boosted Decision Trees for the selection of upward going neutrino events. A comparison to the event selection at hand, therefore, provides the opportunity of an indirect comparison of Boosted Decision Trees and Random Forests on a real world problem. Such a comparison is of particular interest in the scope of machine learning.
The challenge in comparing the event selection at hand to the IC-59 point source event selection, lies in the different purities of the individual samples at final level. Using the Random Forest a purity well above 99% was achieved, whereas the IC-59 point source sample was found to have a contamination of atmospheric muons on the order of 5%. One should note, however, that a purity of 95% is fully sufficient for point source searches.
Therefore, in order to achieve a fair and reliable comparison of the individual event selections, the signalness cut was loosened to $s \geq 0.984$, resulting in an estimated purity of 95.3%. Comparing the number of neutrino events at this cut level, and therefore at a comparable purity level, one finds that 55 000 neutrino events are selected using the Random Forest preceded by a detailed Feature Selection, whereas $\approx 35\,856$ neutrino events are selected using BDTs. This corresponds to an increase in event rate of more than 50%.
However, as individual event selections might perform differently on different en-

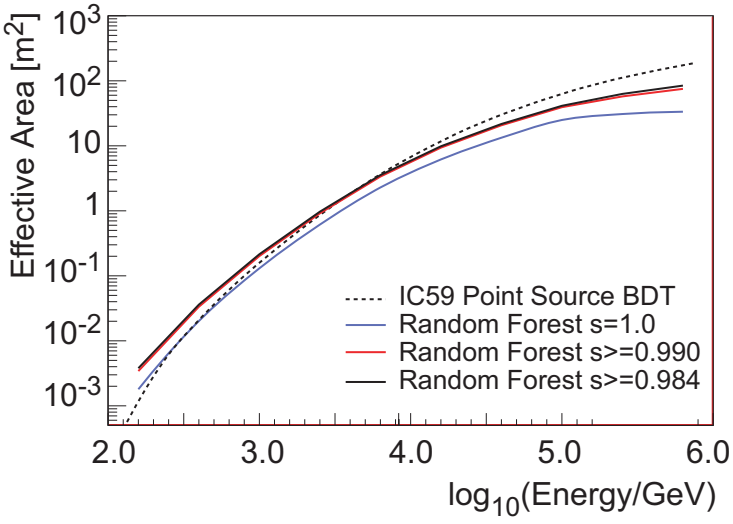


Figure 3.20: Effective area as a function of neutrino energy for three different signalness cuts, as well as for the IC-59 point source (PS) event selection. The event selection using the Random Forest is found to be about twice as effective as the IC-59 Point source event selection. With increasing energy, however, the point source event selection outperforms the Random Forest event selection.

ergy scales, the energy dependence of the samples needs to be taken into account as well. This can be done by comparing the effective areas of the individual event selections.

A comparison of the effective areas obtained at different signalness cuts for the analysis at hand as well as the effective area of the IC59 point source analysis is depicted in figure 3.20. One finds that the Random Forest event selection in general outperforms the BDTs at low energies, whereas an almost equal performance is obtained at intermediate energies. For high energies, however, the performance of BDTs is found to be better.

The event selection used in reference [80] also utilised Boosted Decision Trees. Two BDTs were trained each outperforming the other in a specific energy region. An E^{-1} neutrinos spectrum was chosen as input for the training of the BDTs [80]. Events were selected when passing either BDT [80].

The final event sample in reference [80] consisted of 20 496 upward going neutrino events in 359 days of detector uptime. This corresponds to 57.1 neutrino events per day. Due to an unexpected excess of events near the horizontal, events with zenith angles between 90° and 97° had to be excluded from the event selection. This additional cut resulted in a final number of neutrino events of 17 682, corresponding to 49.3 neutrino events per day [22, 80].

Comparing the numbers of the final neutrino sample, to the event selection at hand (80 neutrino events per day) one finds that an increase of $\approx 62\%$ was achieved with respect to the IC-40 event selection. Part of this increase in event rate, however, is due to the larger detector volume (59 in contrast to 40 strings), which results in an approximate increase of 50% in the trigger rate. Taking into account the im-

proved trigger rate between the individual detector configurations, one finds that an increase of $\approx 8\%$ was achieved for the event selection at hand. Unfortunately a reliable comparison of the effective areas for both event selections cannot be given, as the effective area derived in the IC-40 analysis suffers from a bug in the IceCube neutrino generaor NUGEN.

3.11 Summary and Conclusion

Level	atm. μ	ν_μ
level 3	19.56×10^6	15 743.1
level 4	1.68×10^6	8990.9
level 5	11 ± 10	2872

Table 3.5: Overview over the individual analysis levels.

In total a background rejection better than 99.9999% was achieved with respect to level 3, at a signal efficiency of 18.2%, calculated with respect to level 3 as well. One should note that this corresponds to a rejection of background events of approximately five orders of magnitude, while keeping approximately one fifth of the neutrino events.
The largest part of the background (91.4%) was removed through the application of precuts, while these precuts retained 57.1% of the neutrino events. Remaining background events were removed by the application of state of the art machine learning techniques, namely an MRMR Feature Selection and a Random Forest. Although the absolute number of events removed by the forest is small compared to those removed by the precuts, one should note that these background events were significantly harder to reject.
A signal efficiency of 31.9% was obtained for the machine learning part of the event selection alone.
In summary the event selection presented here was found to show a high performance at an equally high stability. Especially the combination of a powerful Feature Selection and a stable high performance classifier, was found to yield very positive results in terms of signal efficiency and background rejection. Since the loss of neutrino events between level 3 and 4 is rather large, the application of precuts still leaves room for improvements. These improvements can be expected, due to the more homogeneous detector geometry of the full IceCube detector and due to improved track reconstruction algorithms.
The individual steps of the event selection are summarised in table 3.5. Note, that all event numbers correspond to the burnsample on which the event selection was developed (33.281 d).

The application of the event selection to the full year of IC-59 was carried out successfully.

Chapter 4

Energy Spectrum Unfolding

4.1 Introduction to Unfolding

Since the atmospheric neutrino spectrum cannot be accessed directly, it needs to be inferred from the measured rate of neutrino induced muons. Due to stochastic processes involved in muon production and propagation, however, the energy of the observed muon cannot be directly converted into the energy of the neutrino (see equation (2.8)). Moreover, the muon energy distribution cannot be measured itself but needs to be estimated from energy proxies, which are smeared due to the limited resolution of the detector. In general, this problem is described by the Fredholm equation of first kind [87]:

$$g(y) = \int_a^b A(x,y) f(x)\, dx, \tag{4.1}$$

where the distribution of the quantity of interest $f(x)$ is connected to the measured quantity $g(y)$ by the response function of the detector $A(x,y)$. In order so to solve the inverse problem equation (4.1) can be converted into a matrix equation:

$$\vec{f}(x) = \underline{A}(x,y)\vec{g}(y), \tag{4.2}$$

where the response matrix $\underline{A}(x,y)$ needs to be determined. As no direct energy calibration can be used in IceCube, $\underline{A}(x,y)$ is obtained from Monte Carlo simulations. In the most naive assumption, equation (4.2) can be solved by inverting the transition matrix [87]. However, as $\underline{A}(x,y)$ is in general not a square matrix, the inverse might not exist. Moreover, a simple inversion of the matrix might result in an oscillating solution. Equation (4.2) can then be solved by minimizing either a least squares [88] or a log-likelihood expression [87]. Additionally, a regularisation term is introduced to avoid an oscillating behaviour of the solution [87, 88].
Within this work the unfolding software TRUEE [25, 89], which is based on the $\mathcal{RUN}$ algorithm [90], was used for the reconstruction of the spectrum.
More detailed information on regularised unfolding can be found in references [25,

87, 89, 91]. Alternative approaches for solving equation (4.2) are reviewed in references [88] and [92]. A detailed introduction to regularisation is given in [87] and [88].

4.2 Input Variables and Unfolding Settings

The quality of the unfolding strongly depends on the choice of the input variables and the parameter settings for the unfolding algorithm. In general, the choice of the unfolding parameters cannot be considered independent from the input variables. Therefore, both, the set of input variables and the unfolding parameters, need to be determined in an optimisation procedure.
Four energy dependent quantities (LDirC, NDirC, dEdX and NCh) were combined to three distinct sets of input variables as TRUEE is able to handle as many as three input variables [25]. Other combinations of input variables were used as well. These sets are not discussed, as no stable unfolding settings were found for these sets of input parameters. The quality of the four unfolding parameters towards the use in an unfolding (energy dependence, Data/Monte Carlo agreement) will be addressed further down in this section. A summary of the individual input sets can be found in table 4.1.
The most important settings in TRUEE are the number of bins (*nbins*) of the unfolded distribution and the number of knots *nknots*. The strength of the regularisation can be adjusted by the number of degrees of freedom *ndf* [25]. Further information on the individual unfolding parameters and their specific settings can be found in [25, 89, 91].
For the unfolding at hand, the unfolding parameters were chosen to $nbins = 10$ and $nknots = 17$. Given the sets of input parameters two stable settings ($ndf = 5$ and $ndf = 6$) were found for the number of degrees of freedom, where $ndf = 5$ corresponds to a somewhat stronger regularisation. The unfolding settings were chosen in an optimisation procedure and are summarised in table 4.2.

Set	Var. 1	Var. 2	Var3
1	LDirC	NDirC	—
2	LDirC	NDirC	NCh
3	LDirC	NDirC	dEdX

Table 4.1: Summary of the sets of input variables used for unfolding.

Setting	*nbins*	*nknots*	*ndf*
1	10	17	5
2	10	17	6

Table 4.2: Summary of the unfolding parameters.

4.2.1 Description of Unfolding Variables

In this section a short description of the variables considered as input for the unfolding is given.

Number of Direct Photons (NDirC)

Photons detected without being scattered by impurities in the ice are called direct. In IceCube the time residual t_{res}, defined as [93]

$$t_{res} = t_{hit} - t_{geo}, \tag{4.3}$$

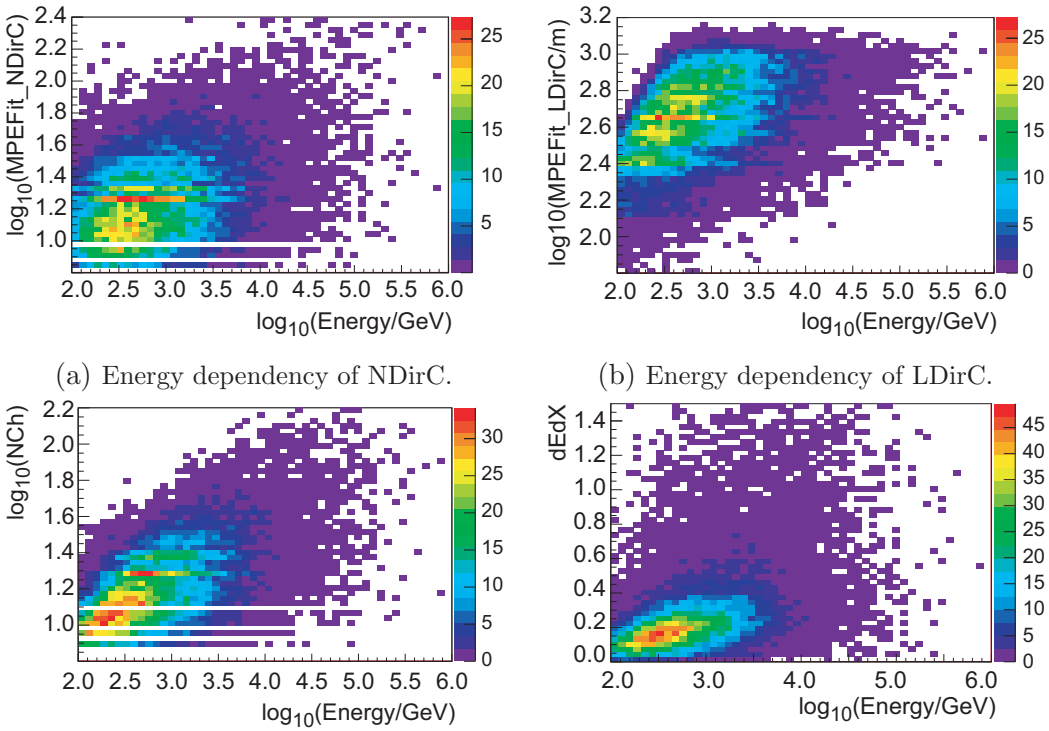
is used to distinguish between direct and non-direct photons. In equation (4.3) t_{hit} represents the actual hit time, whereas t_{geo} depicts the arrival time of an unscattered photon. Both times are computed with respect to the reconstructed particle track. Several time windows indicated by capital letters from A to F are used to define direct photons of different types. Photons reaching a DOM in the interval from $t_{res} = -15\,\text{ns}$ to $t_{res} = 75\,\text{ns}$ are considered as direct of type C [63, 89].
A good correlation between the number of direct photons and the energy of the incoming neutrino is observed, which makes NDirC well suited as an input variable for unfolding.

Direct Length (LDirC)

The direct length, LDirC, can be used as an estimator for the length of the track inside the detector, which itself can serve as an estimator for the energy of the neutrino (see equation (2.12)). The direct length of type C is obtained by projecting all DOMs registering direct photons within time window C onto the reconstructed muon track.

Number of Channels (NCh)

The number of channels (NCh) depicts the number of DOMs detecting photons in an event [63, 89]. NCh is expected to correlate well with energy, since high energy muons will emit more photons, leading to more hit DOMs and a longer track length (see equation (2.13)). This parameter is, thus, also expected to correlate with the track length LDirC.



(a) Energy dependency of NDirC.

(b) Energy dependency of LDirC.

(c) Energy dependency of NCh.

(d) Energy dependency of dEdX.

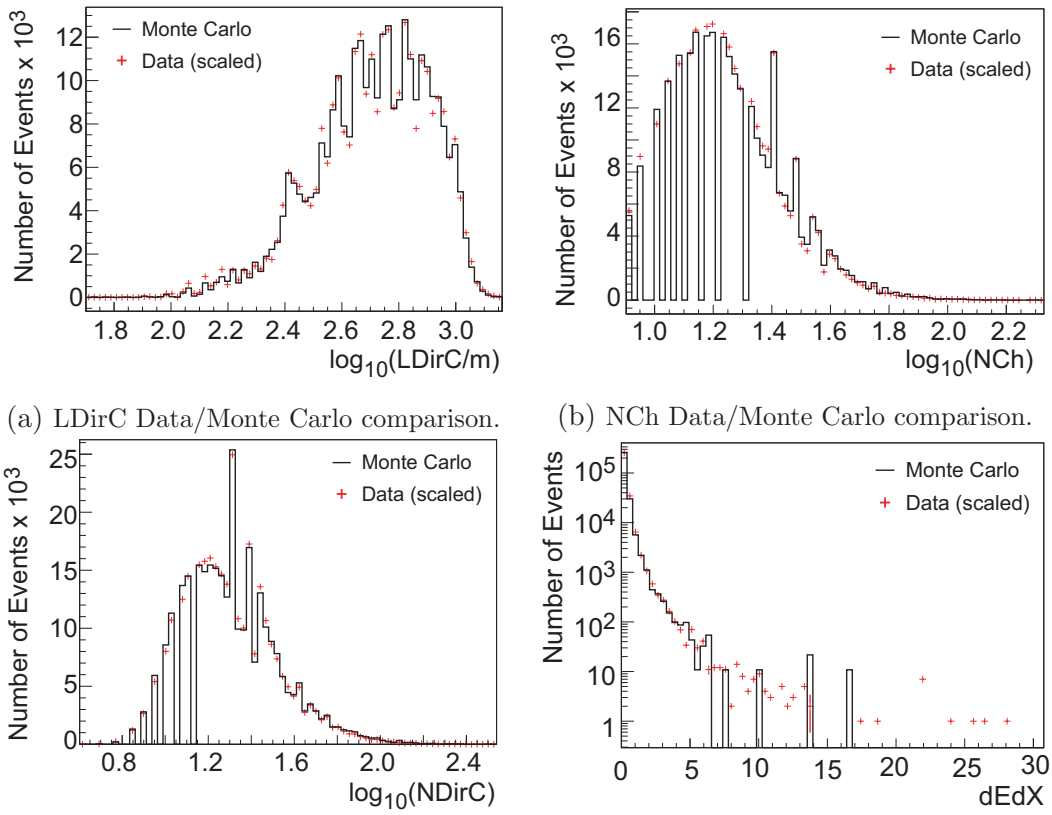
Figure 4.1: Energy dependency of the four variables used as input for the unfolding. A good correlation with energy is observed for each of the variables, indicating that it can be used in an unfolding algorithm. Note, that dEdX is depicted on a linear scale, whereas the other three variables are depicted on a logarithmic scale.

Differential Energy Loss (dEdX)

Since the differential energy loss dEdX of the neutrino induced muon depends strongly on the muon energy (see equation (2.9), it can be used as an estimator for the energy of the muon and, thus, also as an estimator for the energy of the incoming neutrino.

4.2.2 Energy Dependency

Figure 4.1 depicts the energy dependency of the four variables considered as input for the unfolding. A good correlation with energy is observed, indicating that all variables carry valuable information that can be exploited in the unfolding process. Note, that dEdX is depicted on a linear scale, as it enters the unfolding in a linear fashion as well. The three remaining variables (LDirC, NDirC, NCh) enter the unfolding logarithmically and are therefore depicted on a logarithmic scale. The energy dependency of other energy correlated variables can be found in reference [89].



(a) LDirC Data/Monte Carlo comparison.

(b) NCh Data/Monte Carlo comparison.

(c) NDirC Data/Monte Carlo comparison.

(d) dEdX Data/Monte Carlo comparison.

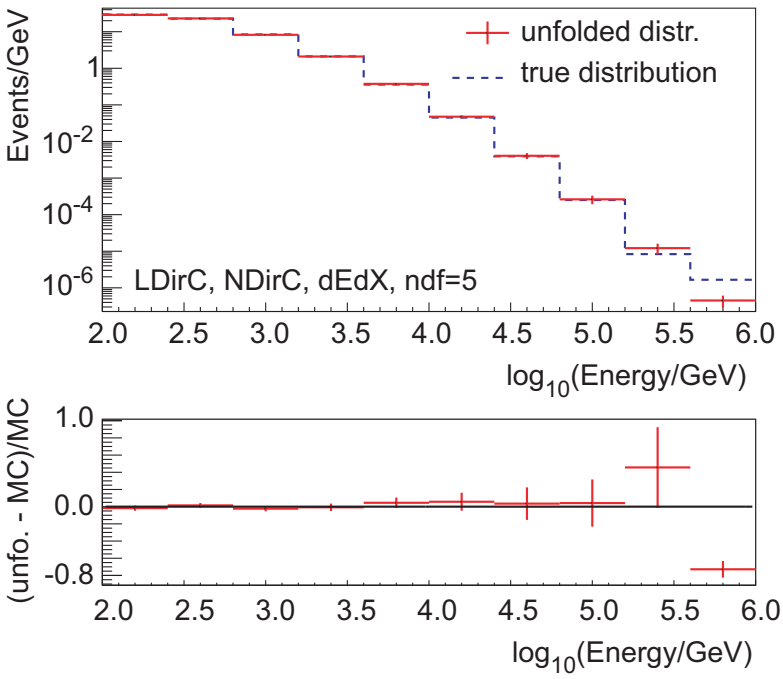
Figure 4.2: Data/Monte Carlo comparison for all variables considered as input for the unfolding. Spice-1 with a nominal DOM Efficiency of 1.0 was used for simulation. A good agreement is observed for all four variables.

4.2.3 Data/Monte Carlo Comparisons

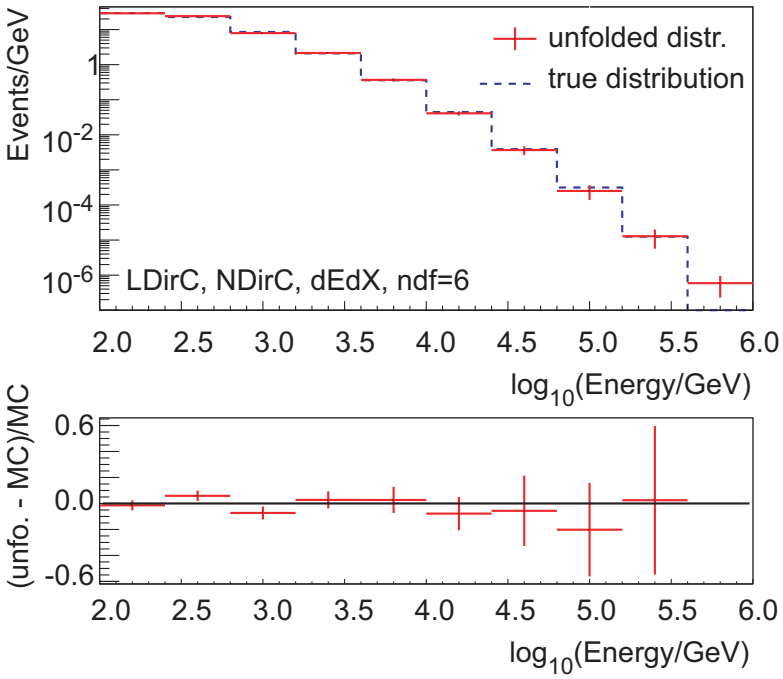
In addition to a reasonably good correlation with energy the variables considered as input for the unfolding should show a good agreement between data and Monte Carlo, as mismatches might affect the unfolding in an undesirable manner.
Figure 4.2 shows the comparison between data and Monte Carlo simulation for all variables considered as input for the unfolding. A good agreement between data and Monte Carlo simulation is observed for all variables, indicating that all four are well suited for the use in an unfolding procedure.

4.2.4 Test Mode Results

Figure 4.3 depicts selected test mode results using LDirC, NDirC and dEdX as input variables for the unfolding. In general, a good agreement between the true and the unfolded distribution is observed. No strong oscillations were found, indicating stable unfolding settings. The solution for $ndf = 6$ appears to be slightly more oscillating compared to the solution for $ndf = 5$. The discrepancies with respect to oscillations, however, are marginal between the two settings and caused



(a) LDirC, NDirC and dEdX. $ndf = 5$.



(b) LDirC, NDirC and dEdX. $ndf = 6$.

Figure 4.3: Selected test mode results using LDirC, NDirC and dEdX as input parameters for the unfolding. A good agreement between the unfolding result and the true distribution is observed.

by the difference in regularisation strength.

Selected test mode results using only LDirC and NDirC as input parameters for the unfolding are shown in figure 4.4. The true and the unfolded distribution were found to agree well. Furthermore, no oscillations in the solution were observed, indicating stable and reliable unfolding settings. Again, the statistical error bars were found to not cover the discrepancy between the two distributions for the highest energy bin.

Figure 4.5 shows selected test mode results using LDirC, NDirC and NCh as input variables for the unfolding. A good agreement is observed between the true distribution and the unfolding result. No oscillating behaviour was observed for either of the solutions, indicating stable unfolding settings. It was found that the discrepancy between the true and the unfolded distribution in the very last bin is not covered by the statistical uncertainty.

In summary, one finds that the true and the unfolded distributions agree well for all six unfolding settings presented here. As no signs of oscillating solutions could be observed, all settings are potentially stable. Moreover, it is found that the statistical error for the very last bin does not cover the discrepancy between the true and the unfolded distribution. Whether or not this poses a potential problem to the unfolding cannot be determined from the test mode results alone.

4.2.5 Pull Mode Results

The stability and performance of the individual unfolding settings and sets of input variables was validated in a bootstrapping procedure. Within this bootstrapping 500 test unfoldings, called pulls, were carried out treating 10% (corresponding to 30 000 Monte Carlo events) of the simulated events as pseudodata. For each pull the deviation of the unfolding result from the true distribution was calculated binwise and in units of the statistical uncertainty.

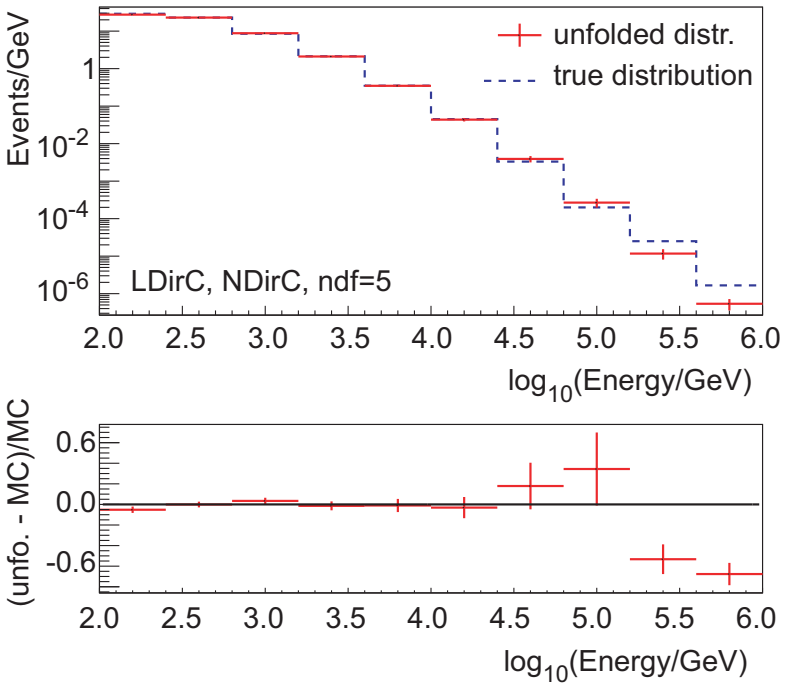
Figure 4.6 depicts the results of the pull mode. For all settings and all sets of input parameters, considered for the use in the final unfolding, an extremely stable performance is observed in the first seven bins. The deviations are found to increase with energy. This behaviour, however, can be attributed to the small statistics in these bins, caused by the steeply falling spectrum of atmospheric neutrinos.

In general, it appears that smaller deviations are observed for the settings using $ndf = 6$, especially in the highest energy bins. This is misleading and due to the fact, that settings with less regularisation are found to produce larger statistical uncertainties. This behaviour is further illustrated in tables C.1, C.2 and C.3.

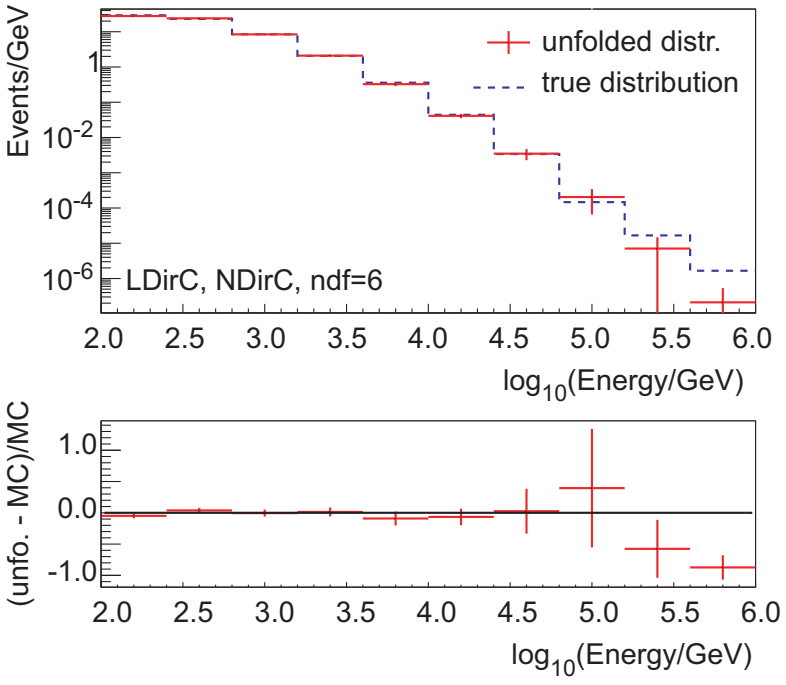
The rather large discrepancies observed for bin 10 are discussed in the next section. The pull distribution derived for individual bins are summarised in appendix B.

4.2.6 Discussion on Bin 10

Figure 4.7 depicts the pull mode distributions of bin 10 for all six stable unfolding settings. Three clearly distinct populations of unfolding results were observed for

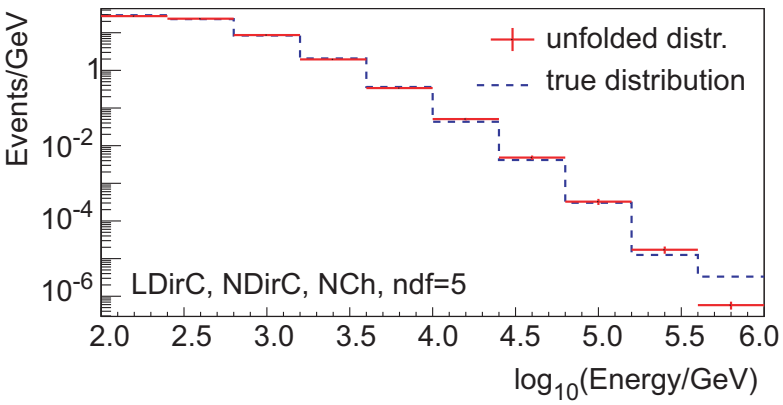


(a) LDirC and NDirC. $ndf = 5$.

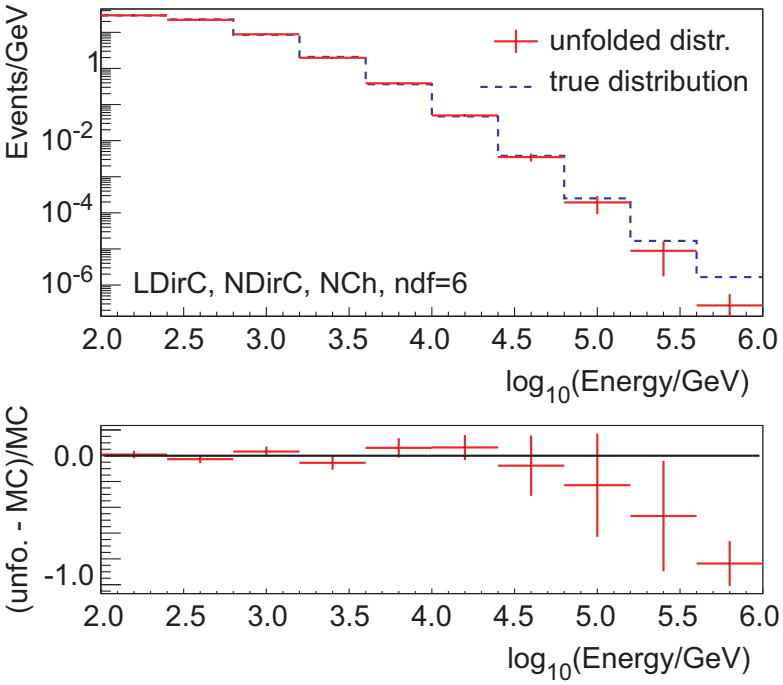


(b) LDirC and NDirC. $ndf = 6$.

Figure 4.4: Selected test mode results using LDirC and NDirC as input variables for the unfolding. A good agreement between the unfolding result and the true distribution is observed.

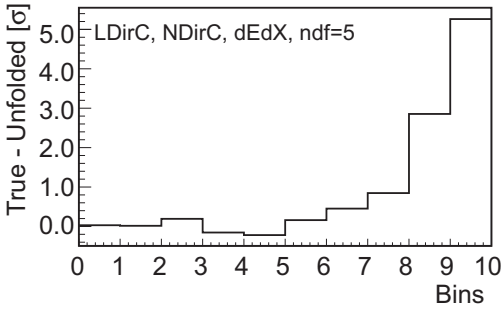


(a) LDirC, NDirC and NCh. $ndf = 5$.

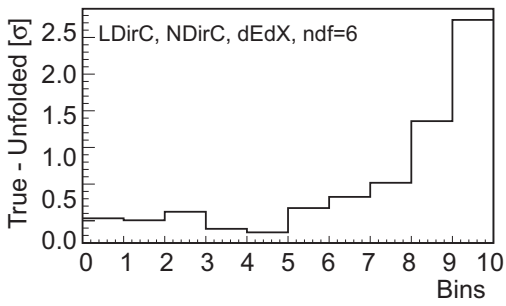


(b) LDirC, NDirC and NCh. $ndf = 6$.

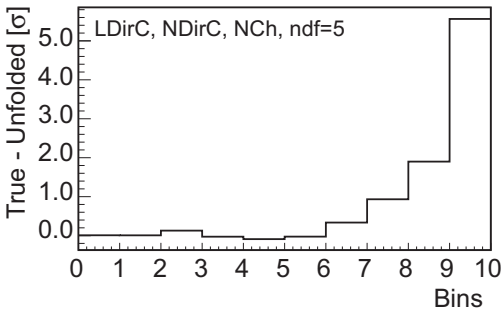
Figure 4.5: Selected test mode results using LDirC, NDirC and NCh as input variables for the unfolding. In general a good agreement between the unfolding result and the true distribution is observed.



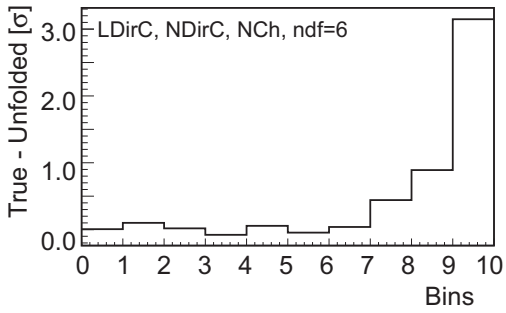
(a) Pull mode results using LDirC, NDirC and dEdX. The number of degrees of freedom was set to 5.



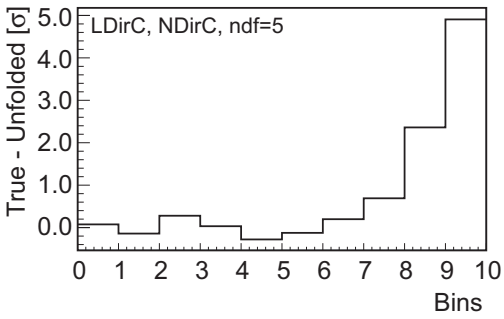
(b) Pull mode results using LDirC, NDirC and dEdX. The number of degrees of freedom was set to 6.



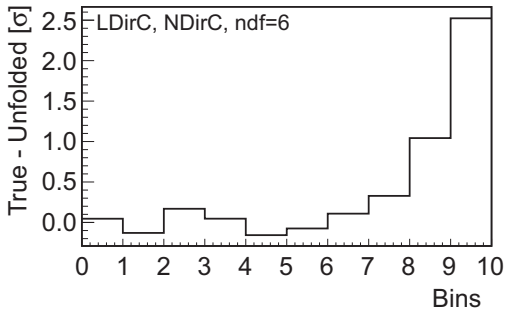
(c) Pull mode results using LDirC, NDirC and NCh. The number of degrees of freedom was set to 5.



(d) Pull mode results using LDirC, NDirC and NCh. The number of degrees of freedom was set to 6.



(e) Pull mode results usings LDirC and NDirC only. The number of degrees of freedom was set to 5.



(f) Pull mode results using LDirC and NDirC only. The number of degrees of freedom was set to 6.

Figure 4.6: Pull mode results for all unfolding settings and all sets of input parameters considered for the use in the final unfolding. For all settings an extremely stable performance is observed in the first seven bins. Deviations are found to increase for higher energy bins, due to the limited statistics, caused by the steeply falling spectrum of atmospheric neutrinos.

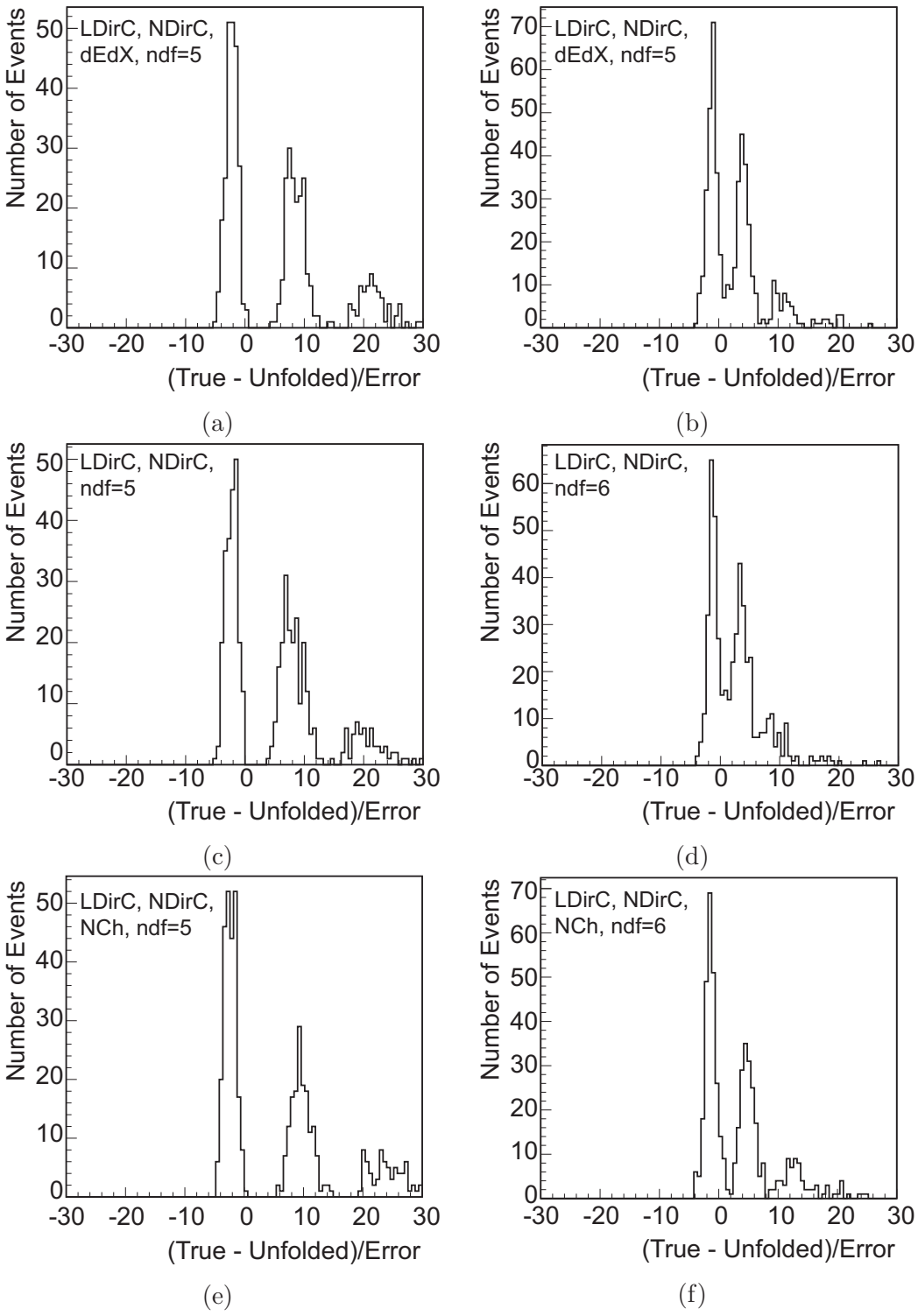


Figure 4.7: Pull mode distributions of bin 10 for all six stable unfolding settings. Three clearly distinct populations are observed for all settings, corresponding to 0, 1 or 2 events in the true distribution (from left to right). Only a fraction of an event, typically 0.5 events, is unfolded for the highest energy bin. Since the statistical uncertainty estimated by TRUEE is not large enough to cover the deviation to either 0 or 1 events, distinct populations are observed.

all settings, corresponding to 0, 1 or 2 events in the true distribution. Such large deviations indicate unstable unfolding settings. For the unfolding at hand, however, the situation is somewhat different.
Due to the sampling of events performed in the pull mode and the steeply falling spectrum, either 0 or 1 events will be found in the last bin of the true distribution. A sampling of 2 or more events is only observed in rather rare cases. The unfolding on the other hand reconstructs a fraction of an event for the last bin. Typically bin contents around 0.5 events are observed. As the statistical error estimated by TRUEE is smaller than the deviation between the true and the unfolded distribution, the discrepancy is not covered.
From figure 4.7 the individual populations appear to have a larger overlap for the $ndf = 6$ settings. This is misleading as the obtained statistical uncertainty is found to be somewhat larger for $ndf = 6$ compared to $ndf = 5$.
Nevertheless, an additional conclusion can be drawn from the pull mode distributions of the highest energy bin. In case there is no event present in the true distribution, the unfolding tends to overestimate the bin content. However, if one or more events are present in the last bin, the unfolding tends to underestimate the true bin content. As the true content of the last bin is not known on data, no clear statement can be made about a possible contribution of prompt neutrinos to the flux in the highest energy bin.
Nevertheless, the atmospheric ν_μ flux can be obtained for the highest energy bin, thus, extending the energy range up to 1 PeV.
Additionally, the question arises whether the statistical uncertainties of the very last bin should be increased due to the findings in this section. In the sections 4.3 and 4.5 the systematic uncertainties of the last bin were found to increase above the 100% limit. Thus, the flux obtained for the highest energy bin is compatible with zero. Therefore, no scaling of the statistical uncertainty was applied to the highest energy bin.

4.2.7 Performance on Real Data

Figure 4.8 shows the results of the unfolding settings using $ndf = 5$, when applied to real data. All three sets of input variables are found to agree within their statistical uncertainties. The set utilising LDirC, NDirC and dEdX, however, is found to be somewhat steeper compared to the other two variable sets. As the last three bins already suffer from a possible underestimation, this behaviour is rather undesirable.
The real data results obtained using $ndf = 6$ are depicted in figure 4.9. All three variable sets are found to agree within their statistical uncertainties for the first couple of bins. For the last three bins an oscillating behaviour is observed for the input set using LDirC, NDirC and dEdX. This oscillating solution results in negative bin contents for the last three bins. The negative bin contents are not shown in figure 4.9. As negative bin contents are unphysical, the setting using this particular set of input parameters (LDirC, NDirC, dEdX) and $ndf = 6$ is no

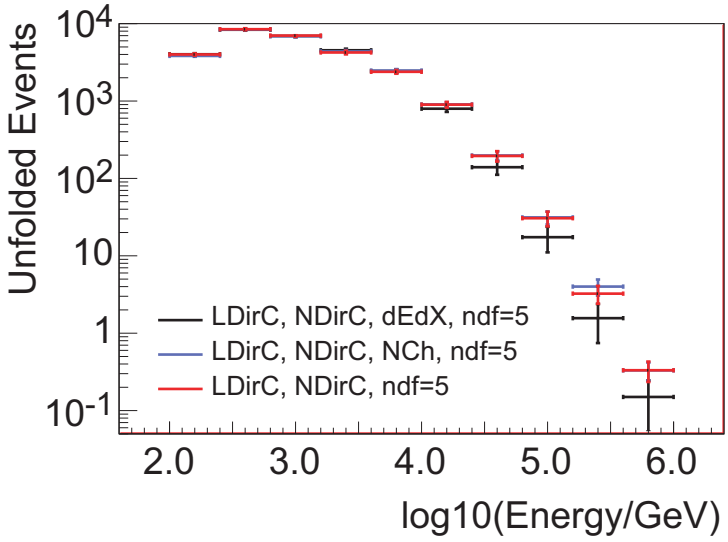


Figure 4.8: Real data results obtained using $ndf = 5$. The results obtained using LDirC, NDirC and dEdX are shown in black, whereas the results obtained with LDirC, NDirC and NCh are depicted in blue. The results utilising LDirC and NDirC only are shown in red. All three unfolding results are found to agree within the estimated statistical uncertainties, indicating stable and reliable unfolding settings.

longer considered for the final unfolding.
Moreover, all input sets used in combination with $ndf = 6$ tend to produce significantly larger statistical errors, compared to those using $ndf = 5$. Thus, the settings using $ndf = 6$ are no longer considered for the final unfolding either.
The unfolding results obtained on real data are summarised in table C.4 and table C.5.

4.2.8 Unfolding Cross Checks

A check comparing the distributions of unused variables on real data to reweighted MC, was carried out in addition to the various unfolding checks performed on Monte Carlo simulations. Within this approach the MC is reweighted according to the unfolding result and compared to real data [25, 91].
The outcome of this particular unfolding check is depicted in figure 4.10. Data is shown in black, whereas the reweighted Monte Carlo is depicted in red. A good agreement between the two distributions is observed for all four parameters. This is another indicator for the stable and trustworthy performance of the unfolding on real data.

4.3 Ice Model Dependency

The sensitivity of the combinations of input variables on the modeling of the ice was studied using the pull mode in TRUEE. For all three sets of input variables 500

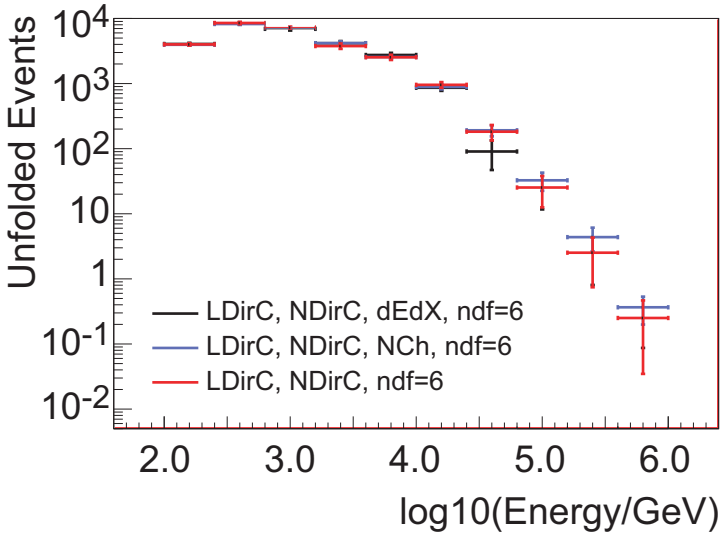
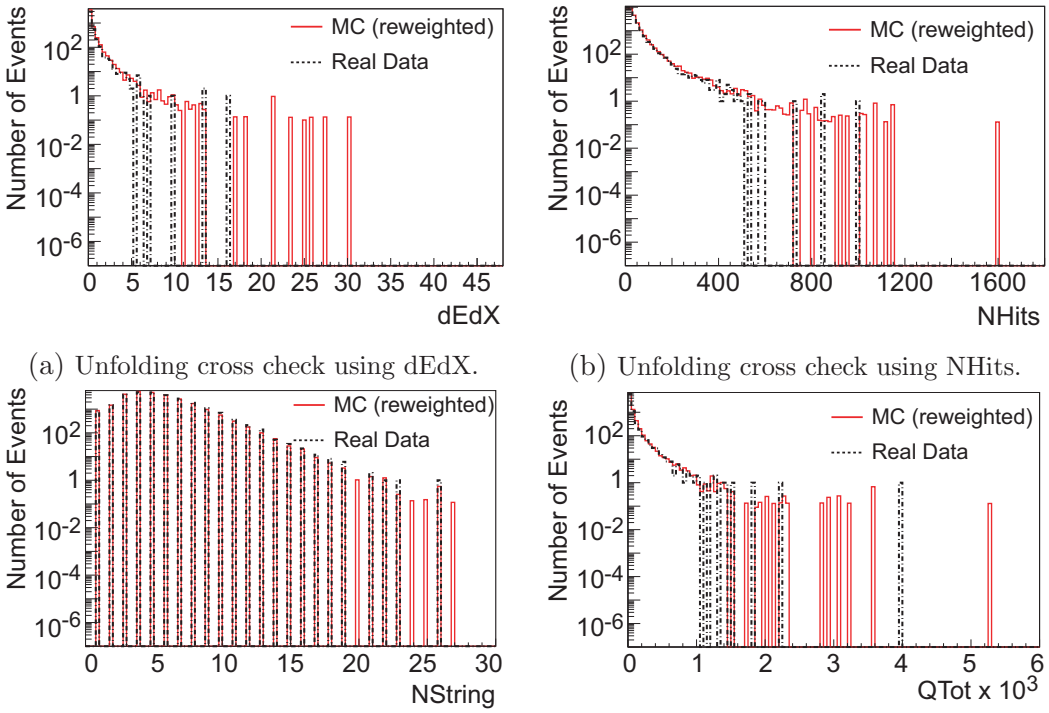


Figure 4.9: Real data results obtained using the different sets of input variables and $ndf = 6$. The results using LDirC, NDirC and dEdX are shown in black, whereas the results utilising LDirC, NDirC and NCh are depicted in blue. The results obtained using LDirC and NDirC only are shown in red. All real data results are found to agree within the estimated statistical uncertainties. Negative bin contents, caused by an oscillating solution, were observed when using LDirC, NDirC and dEdX. Thus, this setting is no longer considered for the final unfolding. The negative bin contents are not shown in the plot.

test unfoldings were performed on 10% of the Monte Carlo set, corresponding to 30 000 events. In either case Spice-1 [84] Monte Carlo was used to build the matrix, whereas Spice:Mie [85] Monte Carlo was treated as pseudodata. The deviation of the unfolded distribution from the true distribution was then computed in units of the statistical uncertainty estimated by the algorithm. Note, however, that a large deviation of order $20\,\sigma$ in a low energy bin (large statistics, small statistical uncertainty) might correspond to a smaller relative error than a $2\,\sigma$ deviation in a high energy bin (small statistics, large statistical uncertainty). To illustrate this behaviour, the deviations were converted into relative uncertainty.
Figure 4.11 shows the dependency of all sets of input variables considered for the final unfolding on the modeling of the ice. The ice model dependency of the unfolding settings used in [89] is shown for completeness. A clear dependency on the ice model is observed for all three sets of unfolding variables. The set using LDirC, NDirC and NCh, however, was found to have the smallest ice dependency of all three sets considered for unfolding. In particular no relative errors larger than $\approx 70\%$ were observed. Note, that deviations larger than 100% were observed for the two remaining sets of input variables. Moreover, the fluctuations between neighbouring bins were found to be significantly smaller using this set of variables. In comparison to the unfolding presented in [89], the ice dependency of the variable set is found to be much smaller, which results in a much larger reliability of the unfolded spectrum. One should note, however, that the ice model uncertainty in



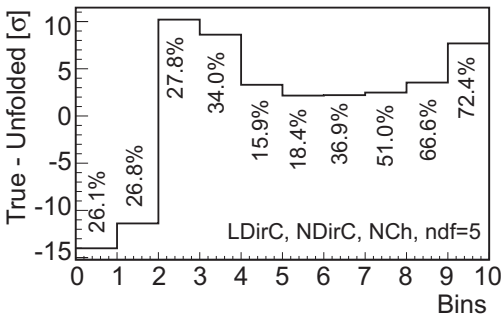
(a) Unfolding cross check using dEdX.

(b) Unfolding cross check using NHits.

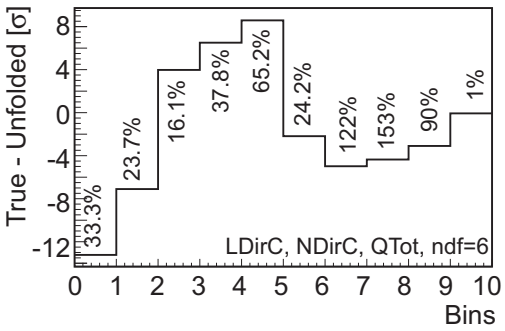
(c) Unfolding cross check using NString.

(d) Unfolding cross check using QTot.

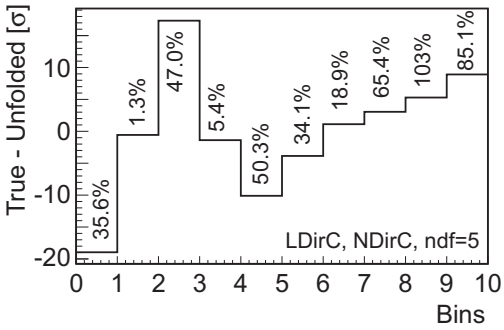
Figure 4.10: Unfolding cross check carried out in TRUEE. Unused energy dependent variables are reweighted according to the unfolding result. The good agreement between data and Monte Carlo indicates stable and reliable unfolding settings.



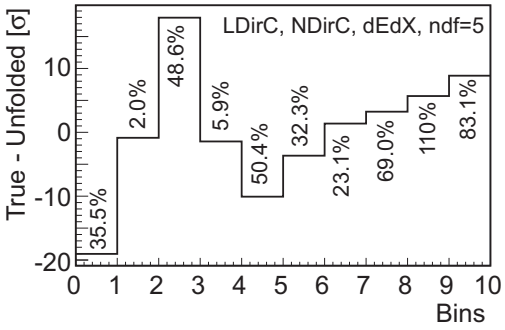
(a) Ice model dependency of the variable combination LDirC, NDirC and NCh.



(b) Ice model dependency of the unfolding presented in [89].



(c) Ice model dependency of the variable combination LDirC and NDirC.



(d) Ice model dependency of the variable combination LDirC, NDirC and dEdX.

Figure 4.11: Ice model dependency of the three sets of input variables considered in the analysis at hand. The ice model dependency of the input variables and settings used in [89] is shown for completeness. A clear dependency on the ice model is observed for all sets of input variables. The unfolding using LDirC, NDirC and NCh, however, was found to have the smallest ice dependency of all input sets. In particular, no relative uncertainties above 100% were observed for this very set of unfolding variables.

reference [89] was accessed differently, as the possibility of performing pulls on different Monte Carlo sets was implemented after the analysis presented in [89] had been finished.
The results of the ice model studies are summarised in table C.6.

4.4 Sensitivity to Unfolding Parameters

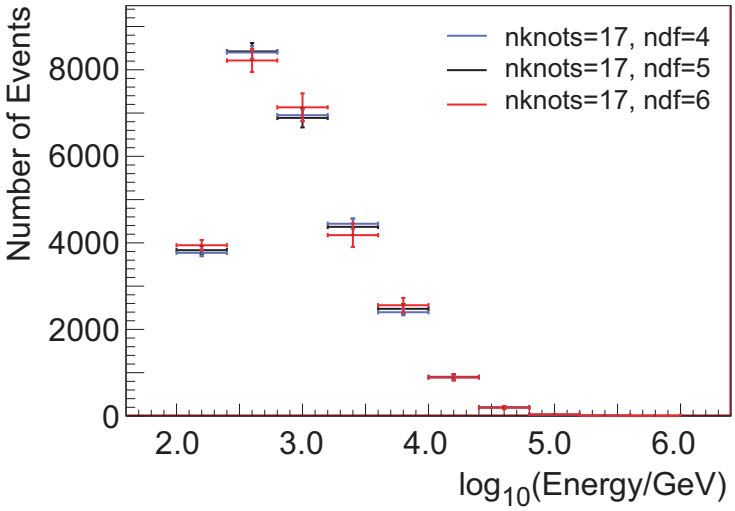


Figure 4.12: Dependency of the unfolding result on the settings used for the number of degrees of freedom. The number of knots was kept constant during the investigation, whereas *ndf* was set to $ndf = 4$, $ndf = 5$ and $ndf = 6$ respectively. The unfolding results are depicted on a linear scale for better visibility. All results were found to agree within the estimated statistical uncertainties.

Studies on the dependency of the unfolding result on specific settings of the parameters *ndf* and *nknots* were carried out. Both unfolding parameters were varied by ± 1 around the value used for the determination of the final neutrino spectrum. For *ndf* this corresponds to using $ndf = 4$, $ndf = 5$ and $ndf = 6$, respectively. The number of knots was set to $nknots = 16$, $nknots = 17$ and $nknots = 18$. All results were obtained using LDirC, NDirC and NCh as input parameters for the unfolding.
Figure 4.12 shows the unfolding results obtained on real data for different settings of the number of degrees of freedom, which corresponds to different settings of the regularisation strength. The number of knots was kept constant. One finds that the slight differences obtained for the individual settings are compatible with each other and well inside the statistical uncertainties. In contrast to other unfolding results presented in this work, a linear scale was used on the y-axis of figure 4.12 in order to display the tiny differences more clearly.
Figure 4.13 shows the dependency of the unfolding results obtained on real data on the setting of the number of knots, while keeping the number of degrees of

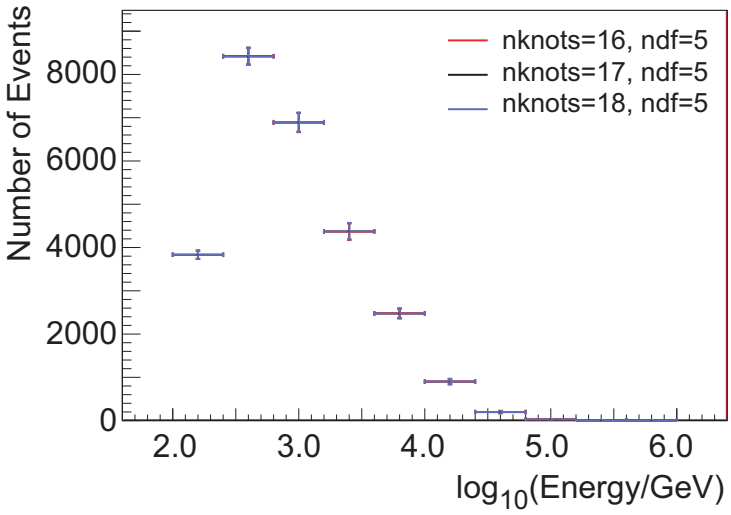


Figure 4.13: Dependency of the unfolding on changes in the number of knots. The number of degrees of freedom was kept constant at $ndf = 5$ throughout the study, whereas *nknots* was varied by ± 1 around the value used to obtain the final neutrino spectrum. The y-axis is depicted on a linear scale to increase visibility. All results were found to agree within the estimated statistical uncertainties.

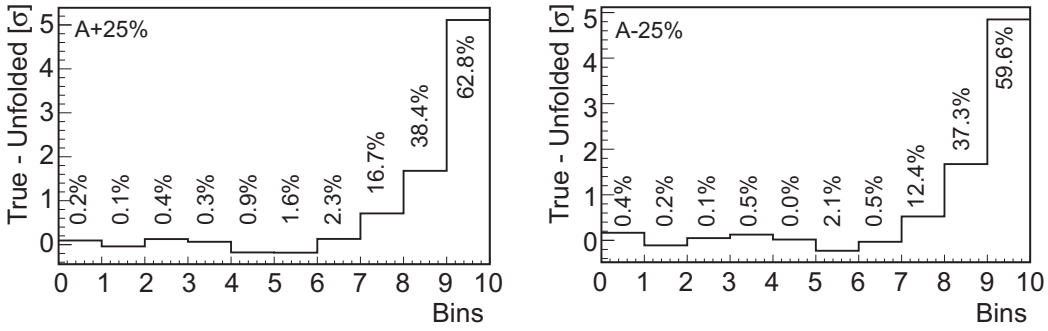
freedom constant. Again a linear scale is used in order to display the differences more clearly. One finds that the observed discrepancies are small and well covered by the estimated statistical uncertainties.
In summary, a weak dependency of the real data result to changes in the unfolding settings is observed. This dependency is, however, due to the differences in regularisation strength. Moreover, all observed deviations were found to be well inside the estimated statistical uncertainties.

4.5 Other Systematic Uncertainties

Apart from the modelling of the ice, the main sources of systematic uncertainties entering the unfolding are the DOM efficiency, the interaction cross sections of muons as well as the normalisation and the spectral index of the atmospheric neutrino spectrum. All uncertainties are addressed separately in this section in a way similar to the procedure outlined in section 4.3.
All errorbars are computed in units of the statistical uncertainty. They can, however, easily be converted into relative errors by a simple scaling, as the statistical uncertainties in individual pulls only deviate from the statistical errors obtained on data on the 1-2% level.

4.5.1 Normalisation Uncertainties

Figure 4.14 shows the systematic errors estimated due to the uncertainty of the normalisation of the atmospheric neutrino flux. The normalisation of the atmospheric



(a) Errorbars due to uncertainties in the normalisation of the atmospheric neutrino flux, obtained for a 25% increased flux normalisation.

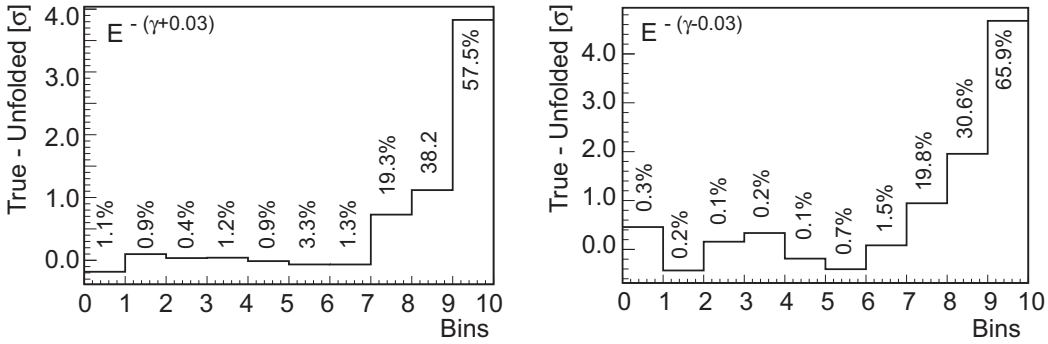
(b) Errorbars due to uncertainties in the normalisation of the atmospheric neutrino flux, obtained for a 25% decreased flux normalisation.

Figure 4.14: Systematic errors due to uncertainties in the normalisation of the atmospheric neutrino flux, estimated for a 25% increased and decreased flux normalisation, respectively. In general, a rather weak dependency on changes in the normalisation is observed for the first 7 bins. The uncertainties are found to increase towards the higher energy bins.

neutrino spectrum was varied by ±25% with respect to the spectrum used for the determination of the unfolding matrix. All uncertainties were computed in units of the statistical error and converted into a relative uncertainty in a simple scaling procedure.
One finds that the uncertainty introduced by the normalisation of the atmospheric neutrino spectrum is small, especially in the low and intermediate energy bins. Deviations below the 3% level were observed for the first seven bins. The discrepancies to the true distribution of events increases for the highest energy bins, which is in part due to the limited statistics in this energy region. However, in order to obtain a conservative estimate for the systematic uncertainties all errorbars derived for the normalisation uncertainty were fully taken into account.
Furthermore, the relative deviations in the plus and minus direction were found to be of comparable size. Accordingly, only the deviations in the plus direction are used for the computation of the final systematic uncertainty. The errorbars, introduced by the normalisation of the atmospheric neutrino spectrum, are assumed to be symmetric.
Moreover, one finds that the normalisation of the atmospheric neutrino flux is retained in the unfolding.
The results obtained for the different normalisations of the atmospheric flux are summarised in table C.7.

4.5.2 Spectral Index Uncertainties

The systematic errors due to uncertainties in the spectral index were obtained by varying the spectral index by ±0.03 in Monte Carlo simulations, according to



(a) Errorbars derived, due to a steepening of the spectrum from $E^{-\gamma}$ to $E^{-(\gamma+0.03)}$.

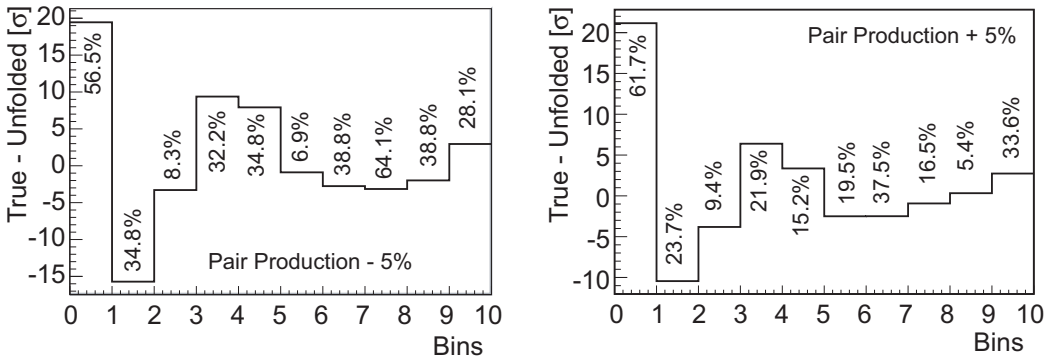
(b) Errorbars derived, due to a flattening of the spectrum from $E^{-\gamma}$ to $E^{-(\gamma-0.03)}$.

Figure 4.15: Uncertainties derived for changes in the spectral index γ. The uncertainties were estimated by varying γ by ± 0.03, which corresponds to a steepening and a flattening of the spectrum respectively.

reference [80]. The simulated events were then used in 500 test unfoldings and treated as pseudodata. The pull mode results are depicted in figure 4.15.
One finds that the deviations are well below the 5% level for the first eight bins. In fact, it should be noted that discrepancies of less than 1% were observed in four different bins. Two bins were found to show average deviations only slightly above the 1% level. The discrepancies were found to increase with energy, leading to an uncertainty on the order of 60% in the very last bin. This large deviation can in part be attributed to the small statistics in the two highest energy bins, but is fully taken into account to achieve a conservative estimate of the systematic uncertainty.
The small uncertainties estimated especially for the low energy region of the spectrum clearly indicate that the spectral index is preserved throughout the entire unfolding process. Such a behaviour is highly desirable in order to obtain a reliable atmospheric neutrino spectrum with small systematic uncertainties. Since the errorbars were found to be comparable when changing the spectral index towards steeper and flatter spectra, respectively, the spectral index uncertainties are assumed to be symmetric. Only the deviations obtained for $E^{-(\gamma-0.03)}$ are used for the final calculation of errors. A summary of the errorbars derived for individual bins is given in table C.8.

4.5.3 Cross Section Uncertainties

Systematic uncertainties introduced by the pair production cross section were estimated by using Monte Carlo sets produced with a 5% increased and decreased pair production cross section respectively. As less or more secondary particles are produced for a decreased or increased pair production cross section, the amount of Cherenkov light produced by individual neutrino induced muons changes accordingly. Changes in the pair production cross section, thus, serve as examples



(a) Systematic uncertainties derived for a 5% decreased pair production cross dection.

(b) Systematic uncertainties derived for a 5% increased pair production cross section.

Figure 4.16: Systematic uncertainties obtained for a 5% decreased and increased pair production cross section, respectively. The first bin is found to be particularly sensitive to changes in this parameter.

for the detection of different amounts of light in the detector. It should be noted, that changes in the amount of light can also be produced by other processes, e.g. changes in the cross sections for bremsstrahlung and ionisation. Despite the fact that these processes vary on different energy scales, changes in the pair production cross section can serve as an estimate. Moreover, using σ_{pp} did not require the production of additional systematic Monte Carlo sets, which is rather time consuming and resource intense.
The results are depicted in figure 4.16. Especially the lowest energy bin is found to be very sensitive to changes in this parameter as deviations on the order of $20\,\sigma$ (60%) are observed. This rather large uncertainty can be understood from the fact that the first bin is close to the energy threshold in IceCube. Changing the amount of light in the detector, will thus result in a shift of the threshold towards the lower and higher energies, respectively. It is worth noting that both an increase and a decrease of the pair production cross section result in an underestimation of events in the first bin. A smaller sensitivity to this parameter is observed for the remaining bins.
The shape of the pull distributions further implies an oscillating behaviour, which can possibly be accounted for by stronger regularisation. As the regularisation chosen for this very unfolding is already quite strong ($ndf = 5$), an even stronger regularisation is not considered.

4.5.4 Uncertainties in the DOM Efficiency

The systematics of the unfolding introduced through uncertainties in the DOM efficiency were estimated using the pull mode in TRUEE. Monte Carlo simulations with a 10% increased and decreased DOM efficiency, respectively, were treated as pseudodata. 500 test unfoldings were performed on these pseudodata and the

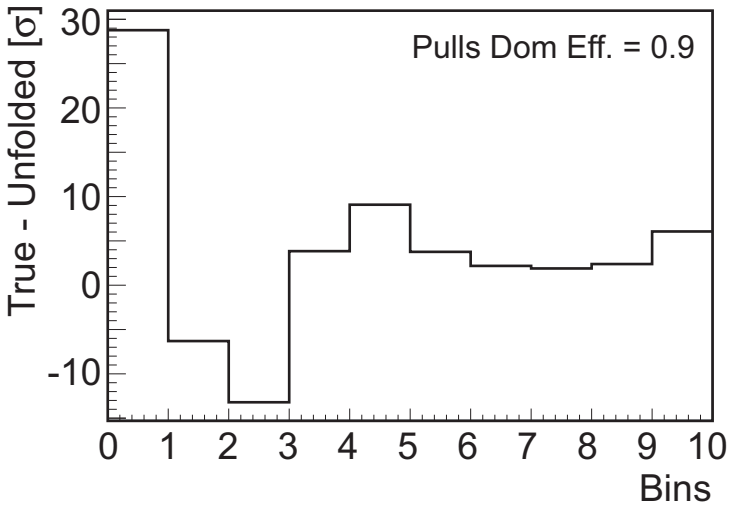


Figure 4.17: Pull mode results using Monte Carlo simulation with a nominal DOM efficiency of 0.9 as pseudodata. Large discrepancies, especially in the first three bins are observed, indicating that the unfolding behaves very sensitive towards changes in this parameter.

deviation from the true Monte Carlo distribution was calculated binwise and in units of the statistical error. The 10% increase and decrease in the DOM efficiency correspond to nominal DOM efficiencies of 1.1 and 0.9 calculated with respect to the standard in the IC-59 simulations. These nominal efficiencies are not to be confused with real DOM efficiencies, where a value of 1.1 would of course be unphysical.
The pull mode results are depicted in figures 4.17 and 4.18. In figure 4.17 large deviations and fluctuations are observed especially in the first three bins. Large deviations in the first two bins, as well as in bins 4 and 5 are observed in figure 4.18. This behaviour indicates that the unfolding behaves very sensitive towards changes in this very parameter. The largest uncertainties are observed for the first bin. Note, that the deviation of the first bin exceeds $30\,\sigma$ in figure 4.18 and is therefore not shown.

The actual size of the deviation becomes understandable when taking into account the test mode results (see figure 4.19). Selected test mode results obtained using Monte Carlo simulation with a DOM efficiency of 0.9 and 1.1, respectively, as pseudodata are shown. The true distribution is shown for comparison. Large deviations of the unfolding result from the true distribution are observed, especially for the first three bins. Due to the logarithmic scale the discrepancies might appear smaller but their actual size can be clearly inferred from the ratio plots. For the first bin for example, deviations of approximately a factor of 8 are observed.
Furthermore, the unfolding result appear to show an oscillating behaviour, which becomes most visible from the ratio plots. Such a behaviour is common for many unfolding procedures. In general, however, it is suppressed by the use of regularisation. As the regularisation for the unfolding at hand is already quite strong, a

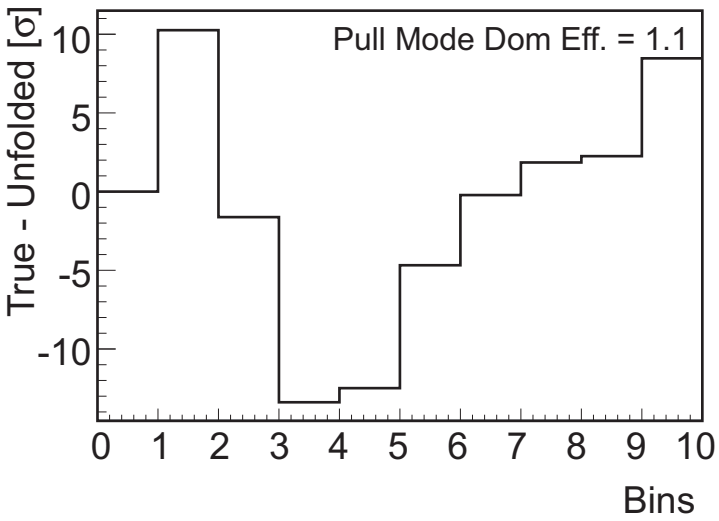
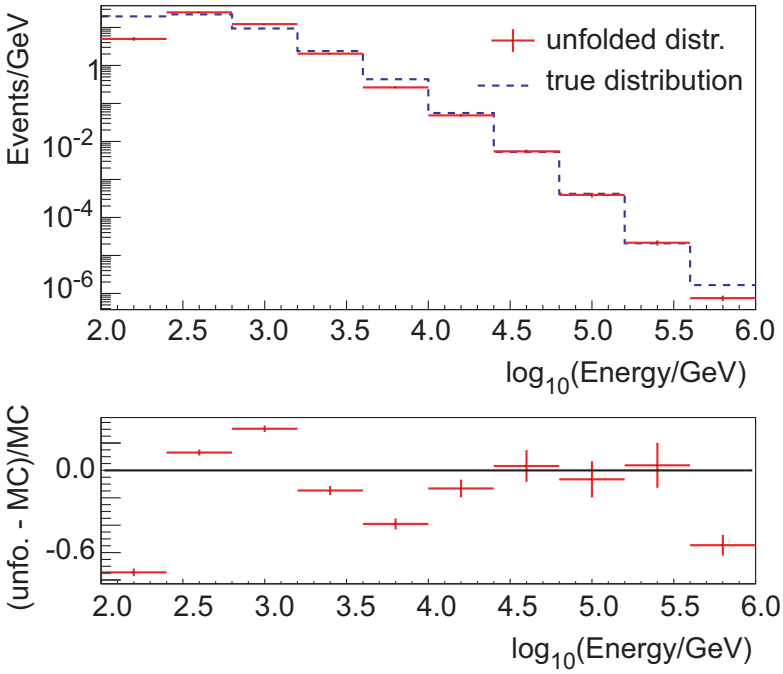
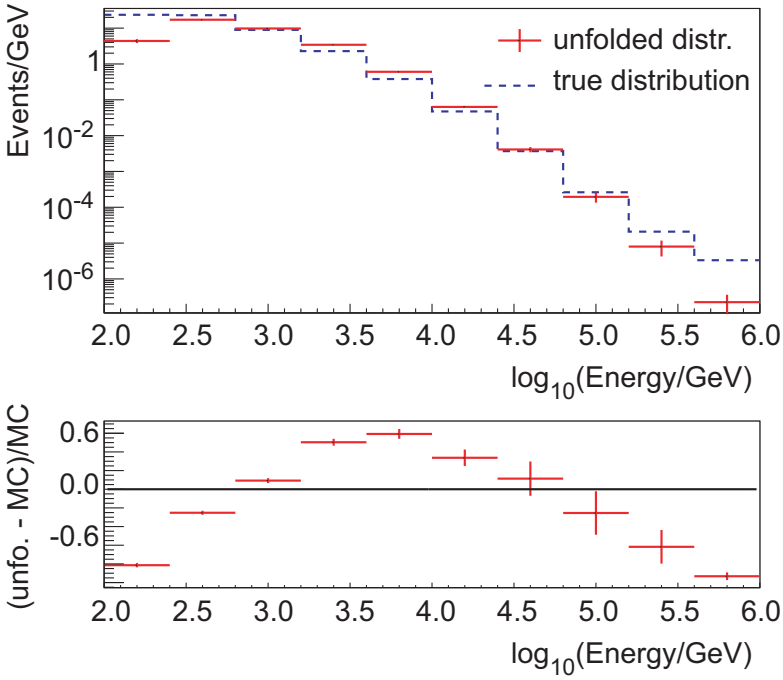


Figure 4.18: Pull mode results using Monte Carlo simulation with a nominal DOM efficiency of 1.1 as pseudodata. Large discrepancies are observed, especially in the first two bins as well as in bins 4 and 5, indicating that the unfolding behaves very sensitive towards changes in this parameter. Note that the devitation in the first bin is $> 30\,\sigma$ and therefore no longer plotted in TRUEE.

further increase in regularisation might introduce an unnecessary bias.
An increase in regularisation can nonetheless account for the observed oscillating behaviour of the unfolded distribution. Using the maximum possible regularisation ($ndf = 2$) appears to smoothen the oscillations (see figure 4.21). The use of maximum regularisation, however, cannot be considered as the solution to this problem since it would result in a very strong bias and an unnecessary smoothening of possible features.
In addition to the pull modes the unfolding was carried out on real data in order to further investigate the dependency of the unfolding on the efficiency of the DOMs. Since the DOM efficiency cannot be varied on data, different nominal DOM efficiencies were used for the determination of the response matrix.
The unfolding results are depicted in figure 4.20. As expected, discrepancies between the results obtained using different matrices are observed. These discrepancies, however, are found to be much smaller than the discrepancies observed in the pull mode results. Especially the results for the first bin should be highlighted. A deviation of 15% is observed between the results obtained with a DOM efficiency of 1.0 and those obtained with a DOM efficiency of 1.1 in the first bin. Furthermore, the results obtained using a DOM efficiency of 1.0 and 1.1 for the determination of the matrix are found to differ only marginally. Somewhat larger discrepancies on the order of 20% to 40% are observed for the bins at intermediate energies.
Taking into account the pull mode results as well as the results obtained on real data, one finds that the influence of the DOM efficiency appears to be much larger in the pull modes. It can, thus, be concluded that an estimation of systematics due to the DOM efficiency using the pull mode would largely overestimate the



(a) Selected test mode result using Monte Carlo simulation with a nominal DOM efficiency of 0.9 as pseudodata.



(b) Selected test mode result using Monte Carlo simulation with a nominal DOM efficiency of 1.1 as pseudodata.

Figure 4.19: Selected test mode results using Monte Carlo simulation with nominal DOM efficiencies of 0.9 and 1.1 as pseudodata. The true distribution of events is shown for comparison. Large deviations and fluctuations are observed for the unfolded distribution. Moreover, a certain shift of events towards the center of the distribution is observed for the pulls carried out on events generated with a DOM efficiency of 1.1.

effect, especially in the first few bins, where the statistics are largest.
Moreover, the effect of an increased and decreased DOM efficiency might mimic other sources of systematic uncertainties, e.g. an increase or decrease in the muon interaction cross section. On the other hand, changes in the DOM efficiency might be mimicked as well. Thus, by fully accounting for the systematics due to DOM efficiencies and cross sections, one runs the risk of double or triple counting the same uncertainty. A cancellation of effects is possible as well. Since many of these sources of systematics can only be studied in Monte Carlo simulation, it is not possible to fully disentangle the contributions of individual sources on real data. Therefore, the estimation of systematics due to uncertainties in the DOM efficiency must be handled with care in order to not largely overestimate the final size of the errorbars.
In summary, the unfoldings in the pull mode were found to be significantly more sensitive to changes in the DOM efficiency than the unfolding results on real data. This might in turn lead to a significant overestimation of the systematic error, which would not reflect the quality of the data and the overall stability of the unfolding. In addition, an increased and decreased DOM efficiency could possibly be mimicked by other sources of systematics, which might result in considering the same source of uncertainty twice. Therefore, the systematic errors, due to changes in the DOM efficiency are not considered in the final calculation of systematic uncertainties, as their influence might already be covered by the cross section errors. This further allows for a consistent treatment of all systematics using the pull mode in TRUEE. In order to avoid an underestimation of systematics through the exclusion of the DOM efficiency error, dedicated cross checks were carried out to verify the size of the overall systematic uncertainty. These cross checks are described in section 4.5.5.

4.5.5 Cross Checks on the Systematic Uncertainty

Following an approach presented in [89], the detector was divided into two parts according to COGZ in order to verify the obtained systematic uncertainties. In contrast to [89], buffer zones were introduced to avoid events being randomly counted for the one or the other subdetector, due to small variations in the ice or in the reconstructed parameters.
By dividing the detector according to COGZ, no changes in the spectra are expected after acceptance correction (in contrast to dividing the detector according to θ_{Zenith}). Discrepancies are, however, expected due to differences in the ice and other sources of systematics. The comparison of the subdetector spectra to the full detector spectrum can, therefore, be utilised to check the size of the overall uncertainties.
Within this work the detector was divided into an inner and an outer part in a first attempt. The inner part contained all events for which COGZ $\leq 225\,\mathrm{m}$ and COGZ $\geq -225\,\mathrm{m}$. The outer part contained all events for which COGZ $\geq 275\,\mathrm{m}$ or COGZ $\leq -275\,\mathrm{m}$. Accordingly, two buffer zones of 50 m were created.

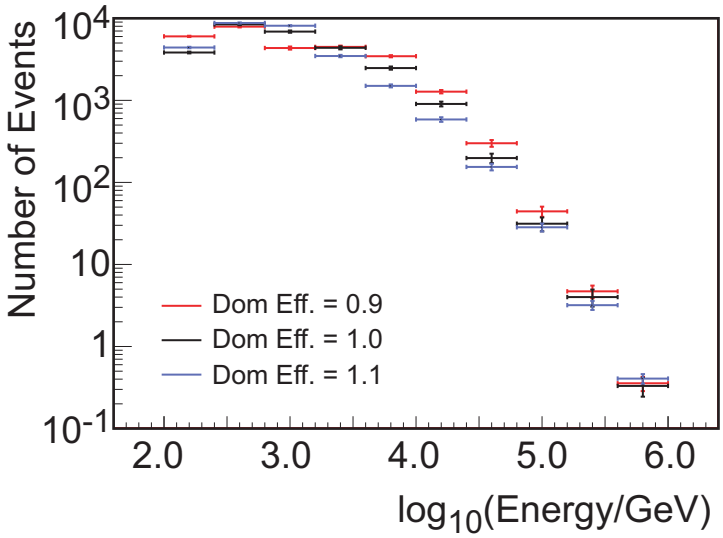
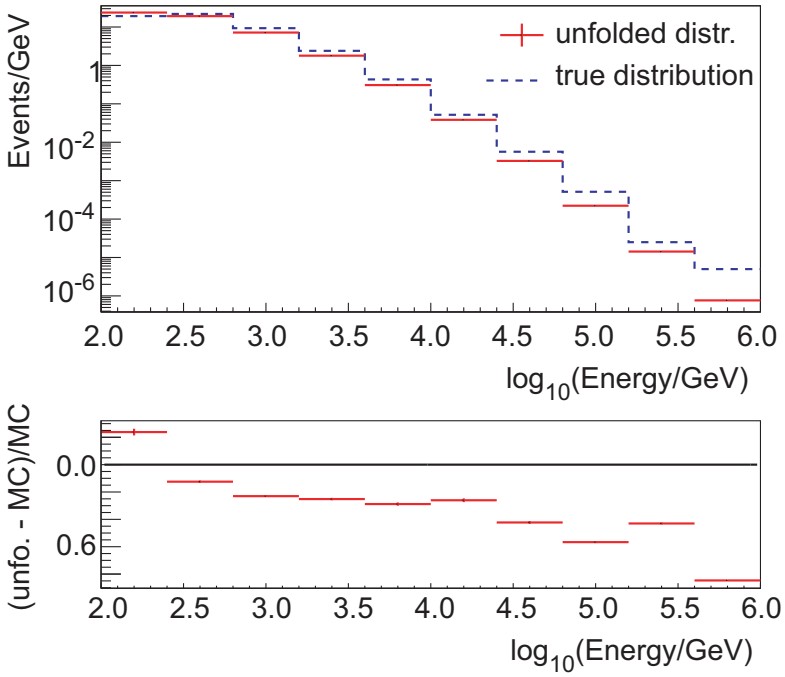
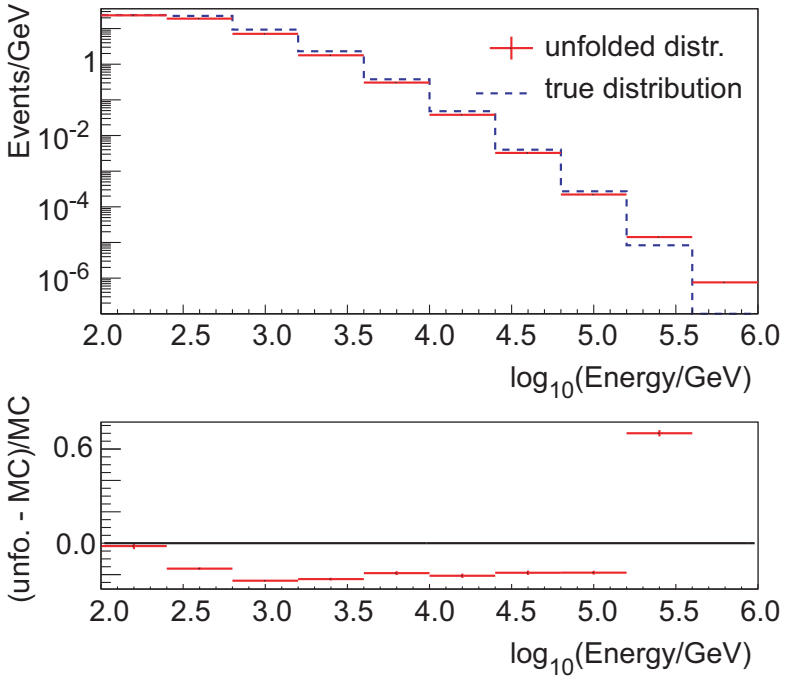


Figure 4.20: Unfolding results obtained on real data using different nominal DOM efficiencies for the determination of the unfolding matrix. The changes in the matrix are found to produce deviations between the individual unfolding results. These deviations, however, are found to be much smaller compared to those observed in the pull mode results.

By dividing the detector into an inner and an outer part, one aims at maximizing the ice model uncertainties. The inner part of the detector is located around the large dust layer, whereas the outer part contains the deep clear ice [79].
Both subdetectors were unfolded using Monte Carlo simulation divided in the same way as real data. An acceptance correction was obtained by using simulation divided in the same way as real data. This produre ensures, that small and implicit mismatches, e.g. data/Monte Carlo mismatches are accounted for as well.
The cross check results comparing the spectra of the inner and outer part of IceCube to the spectrum obtained using the full detector are depicted in figure 4.22. The errorbars of the full detector contain the combined systematic and statistical uncertainties, whereas the errorbars shown for the subdetectors contain the statistical uncertainty only. The x-positions of the two subdetectors were slightly shifted to increase the visibility of the data points.
For most of the data points the full detector spectrum is found to agree with the spectra obtained with the two subdetectors within its errorbars. Deviations are observed for bins 5 and 6, for which the smallest ice model uncertainties and the smallest errorbars were obtained.
Figure 4.23 shows a second cross check, dividing the detector into a top and a bottom part. The bottom part contains all events for which COGZ $\leq -50\,\mathrm{m}$, whereas the top part contains all events for which COGZ $\geq 50\,\mathrm{m}$. The errorbars shown for the full detector spectrum contain the combined statistical and systematic uncertainties. Only statistical errors are shown for both of the subdetectors. Again, the x-values of the subdetectors were slightly shifted in order to increase



(a) Test mode result obtained with a DOM efficiency of 0.9 at $ndf = 2$.



(b) Test mode result obtained with a DOM efficiency of 1.1 at $ndf = 2$.

Figure 4.21: Selected test mode results obtained using different DOM efficiencies at the maximum possible regularisation ($ndf = 2$). Compared to $ndf = 5$ a somewhat better overall agreement is observed. Especially the result for the first bin is found to improve significantly.

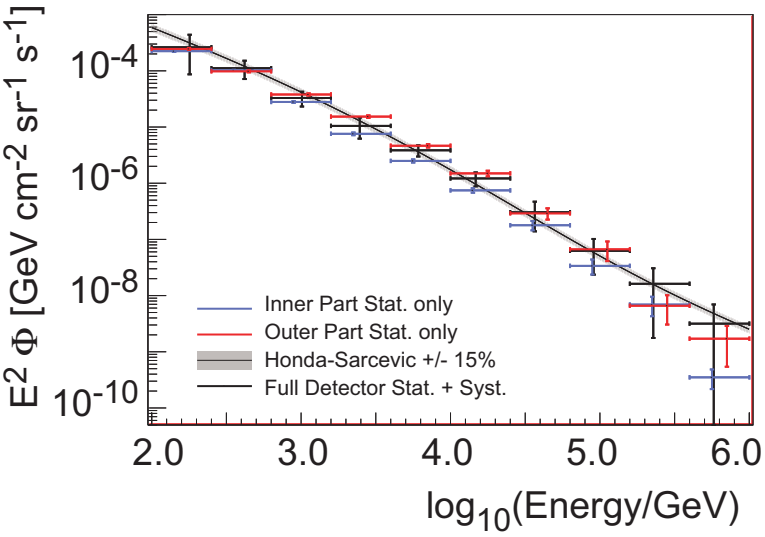


Figure 4.22: Unfolded spectra of the inner- and outer part of the detector, compared to the spectrum obtained using the full detector. The errorbars of the full detector spectrum represent the statistical and the systematic uncertainty, whereas the errorbars for both subdetectors are purely statistical. The subdetector spectra agree with the full detector spectrum for most of the data points. Deviations are observed for bins 5 and 6.

the overall visibility.
The full detector spectrum is found to agree with the subdetector spectra within its errorbars for most of the bins. Deviations are observed for bins 5 and 6. Taking into account that deviations outside the estimated errorbars were observed for these bins in two cross checks, one concludes that the systematic uncertainties are possibly estimated too small for the two bins.

In order to achieve a more conservative estimate of the systematic uncertainties in these bins, the ice model uncertainties for both bins were scaled by a factor of 2. Therefore, the relative ice model error for bin 5 increases from 15.9% to 31.8%. For bin 6 the relative ice model uncertainty is increased from 18.4% to 36.8%. By this scaling the relative size of the ice model uncertainty becomes comparable to the ice model induced errors of the two neighbouring bins (bin 4 and bin 7). The total systematic uncertainty for both bins increases accordingly, from 22.5% to 35.5% in bin 5 and from 27.9% to 42.3% in bin 6.
Using a more conservative estimate is justified by the fact that the errorbars were still found to decrease compared to the IC-40 measurement [22, 80].
The cross checks using the scaled ice model uncertainties are depicted in figures 4.24 and 4.25. The full detector spectrum agrees with the subdetector spectra within its errorbars for all bins. Thus, it can be concluded that the size of the systematic uncertainties has been estimated correctly.

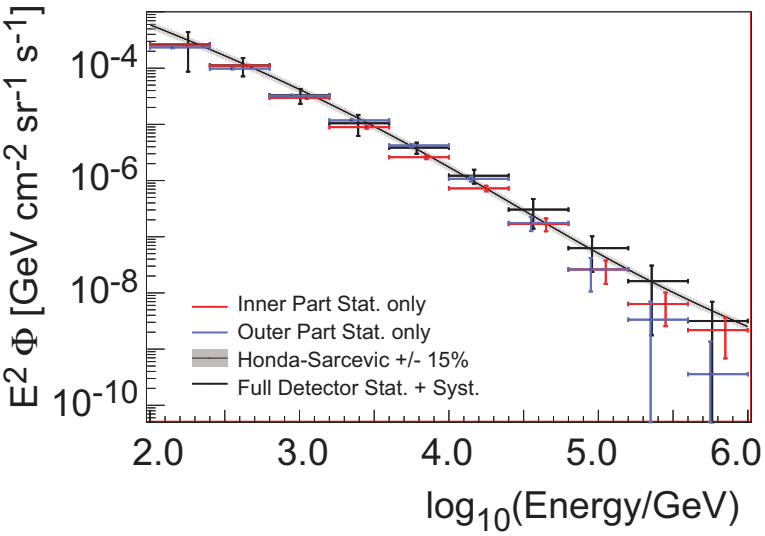


Figure 4.23: Unfolded spectra of the top- and bottom part of the detector, compared to the spectrum obtained using the full detector. The errorbars shown for the full detector contain the combined statistical and systematic uncertainties, whereas the errorbars for the two subdetectors only contain the statistical uncertainty. Within the errorbars a good agreement between the two subdetector spectra and the full detector spectrum is observed. Deviations are observed for bins 5 and 6.

4.6 Acceptance Correction

In order to reconstruct the differential flux of atmospheric neutrinos, the unfolded distribution needs to be corrected for the acceptance of the detector. The dependency of the flux on detector properties, such as solid angle and livetime, is given according to the following equation [80]:

$$N_{Events} = \int dt \int d\Omega \int dE\, \Phi(E,\theta)\, A_{eff}(E,\theta). \tag{4.4}$$

Note that $\Phi(E,\theta)$ in equation (4.4) depicts the differential rather than the integral flux. Since the effective area of the detector is energy dependent (see (2.12)), this area has to be calculated individually for every energy bin. Carrying out the integration in time and solid angle one finds:

$$N_{Events} = T_{Live}\, \Delta\Omega \int \Phi(E,\theta)\, A_{eff}(E,\theta)\, dE. \tag{4.5}$$

Making the transition from the continous to the discrete case:

$$N_{Events} = T_{Live}\, \Delta\Omega \sum \Phi(E,\theta)\, A_{eff}(E,\theta)\, \Delta E. \tag{4.6}$$

Thus, one finds that the relation for an individual bin is given as:

$$\Phi(E,\theta) = \frac{N_{Events}}{\Delta E\, \Delta\Omega\, T_{Live}\, A_{Eff}}. \tag{4.7}$$

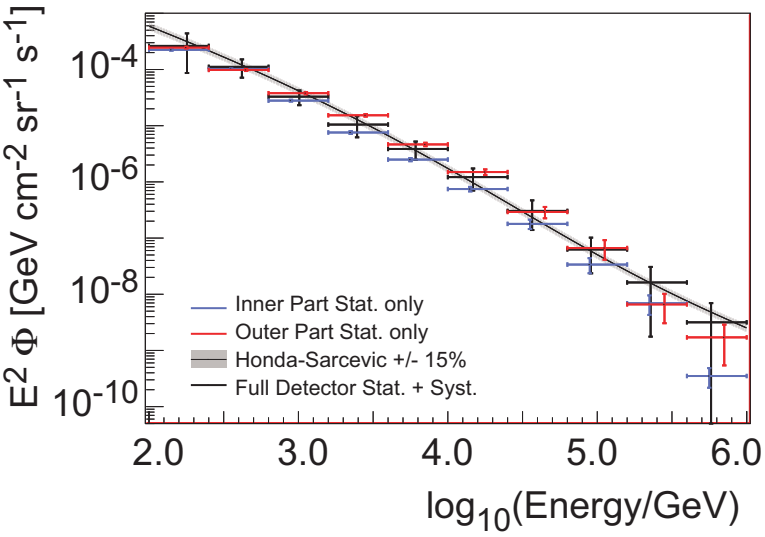


Figure 4.24: Unfolded spectra of the inner and outer part of the detector in comparison to the spectrum obtained by using the full detector. The errorbars for the full detector spectrum contain the combined statistical and systematic uncertainty, whereas the uncertainties shown for both subdetectors are purely statistical. The ice model uncertainties for bins 5 and 6 were scaled by a factor of 2 to account for the deviations observed in these bins. After scaling, the subdetector spetra agree with the full detector spectrum within their errorbars.

As the effective area is not known on real data, it needs to be determined from Monte Carlo simulation:

$$A_{eff}(E) = \frac{N_{Events,MC}}{\Phi(E,\theta)\,\Delta E\,\Delta\Omega\,T_{Live}}, \tag{4.8}$$

with the number of MC events calculated via:

$$N_{Events,MC} = \Phi_{Diff} \times \sum_i OneWeight \times \frac{T_{Live}}{N_{Gen}}, \tag{4.9}$$

where N_{Gen} represents the total number of generated events, which is given as the number of generated events per file, multiplied by the number of files. $OneWeight$ is an event weight assigned to all events during the generation of Monte Carlo. It accounts for neutrino properties, such as cross section as well as for geometrical properties of the detector.

4.7 Final Results

Figure 4.26 depicts the acceptance corrected atmospheric ν_μ energy spectrum. The errobars include the statistical as well as the systematic uncertainty. Two theoretical flux models, Bartol [59] and Honda-Sarcevic [57, 94], are shown for comparison. Note, that the Bartol model does not account for a possible flux of neutrinos from

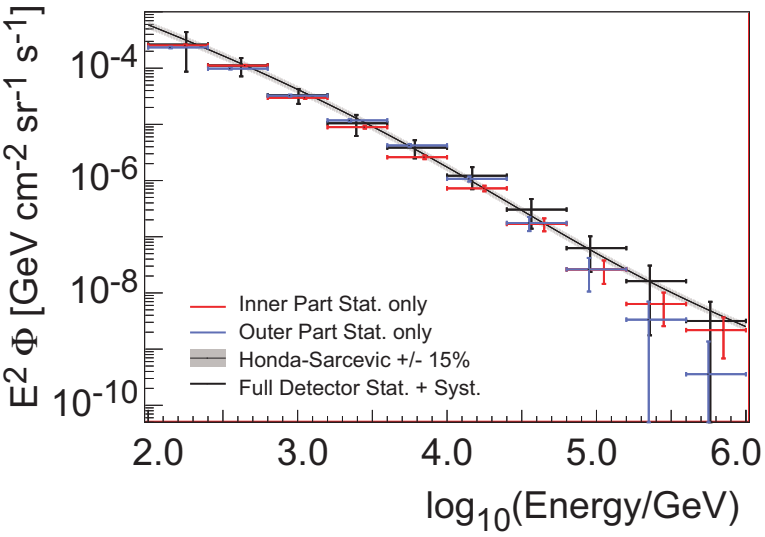


Figure 4.25: Unfolded spectra of the top and bottom part of the detector, compared to the spectrum obtained by using the full detector. The errorbars shown for the full detector contain the statistical as well as the systematic uncertainty, whereas the errorbars for the two subdetectors are purely statistical. The ice model uncertainties for bins 5 and 6 were scaled by a factor of 2 to account for the deviations observed in these bins. After scaling the subdetector spectra agree with the full detector spectrunm within their errorbars for all data points.

the decay of charmed mesons. An uncertainty of 15% is assumed for both theoretical models.
A good agreement between the obtained spectrum and the theoretical models is observed. A relatively large errorbar is observed in the first bin, which is due to the sensitivity of the unfolding towards changes in the muon interaction cross section. In part, this large errorbar also accounts for the sensitivity of the unfolding towards changes in the DOM efficiency, which is not explicitly included in the calculation of the final errors (see section 4.5.4 for details).
It should be noted that the measurement is particularly accurate in the energy region between $10^{2.5}$ GeV and $10^{4.5}$ GeV, as can be seen from the small errorbars in this very region. The errorbars are found to increase with energy, due to smaller statistics and larger systematic uncertainties in the high energy bins. No statement on a possible contribution of prompt neutrinos can be inferred from the spectrum, due to the large errorbars in the two highest energy bins. The flux in the last bin is found to be compatible with zero within the estimated uncertainties.
It should, however, be pointed out that the neutrino sample used in this unfolding provided the very first measurement of the atmospheric neutrino flux in the energy range from 400 TeV to 1 PeV. It is noteworthy that this extension of the ν_μ energy spectrum has been achieved with an increase of $\approx 60\%$ in statistics compared to [22, 80]. Due to the power law dependency on energy, the extension of the spectrum cannot be attributed to the increased statistics alone. It was rather achieved by an improved event selection and the utilisation of the improved unfol-

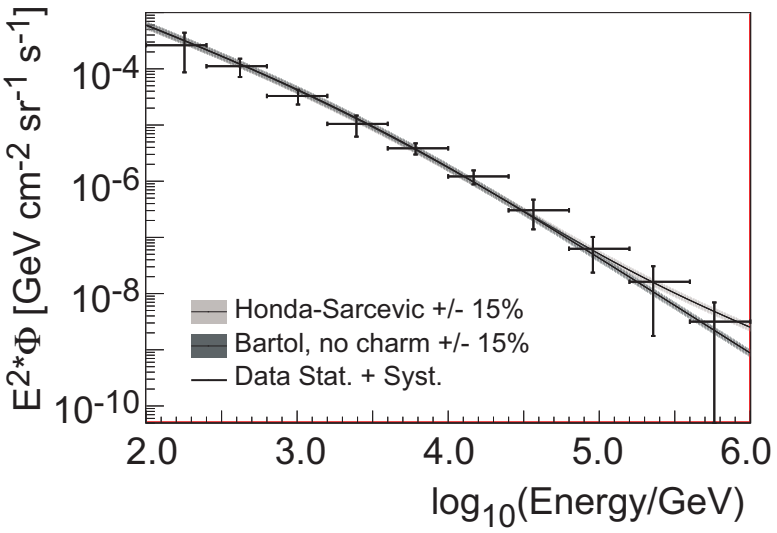


Figure 4.26: Final unfolding result including the statistical as well as the systematic uncertainties. Two different theoretical models, Bartol [59] and Honda-Sarcevic [57, 94], are shown for comparison. Note, that the Bartol model does not account for a possible contribution of neutrinos from the decay of charmed mesons. Within uncertainties a good agreement of the unfolded neutrino spectrum with both models is observed. Due to the relatively large errorbars, no statement can be made about a possible contribution of prompt neutrinos.

ding algorithm provided by TRUEE. Especially the use of three energy dependent variables in contrast to the use of only one such variable in reference [22, 80] greatly contributed to the overall quality of the spectrum.

Figure 4.27 shows the final unfolding result weighted by $E^{3.2}$. The spectrum is found to follow a power law proportional to $E^{-3.2}$ up to energies of $E = 10^{4.5}\,\mathrm{GeV}$. Above these energies a steepening of the spectrum is observed.
This steepening of the spectrum may be caused to the knee in the spectrum of the primary cosmic rays. Due to the rather large errorbars, however, no concluding statement can be made on this behaviour of the spectrum.

4.8 Comparison of the Result

Figure 4.28 shows the comparison of the atmospheric ν_μ spectrum for the thesis at hand compared to the flux obtained with IceCube in the 40-string configuration. The Honda-Sarcevic model including a $\pm 15\%$ uncertainty is shown as well. It should be noted, that the IC-40 result was obtained in a forward folding procedure using dEdX as the only input variable [22, 80]. In general, a good agreement between the two individual flux spectra is observed.
In comparison to the IC-40 unfolding, the errorbars were found to be significantly improved, especially in the energy range from $10^{2.5}\,\mathrm{GeV}$ to $10^{4.5}\,\mathrm{GeV}$. A larger errorbar is observed in the very first bin, due to the large sensitivity of this bin to

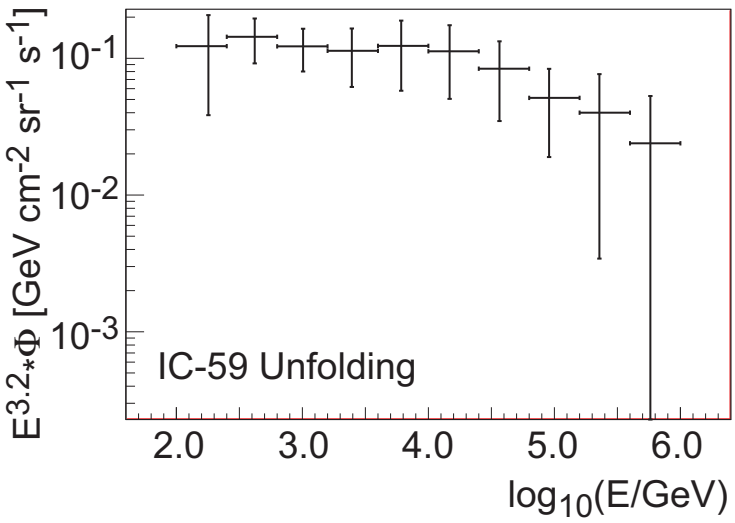


Figure 4.27: Final unfolding result including statistical and systematic uncertainties, weighted by $E^{3.2}$. The spectrum is found to follow a power law of the form proportional to $E^{-3.2}$ up to energies of $E = 10^{4.5}\,\mathrm{GeV}$. Above these energies, a steepening of the spectrum is observed.

changes in the muon interaction cross section. In comparison to the IC-40 unfolding the errorbar estimated for the next to last bin, appears to be roughly 10% larger. In the plot this small deviation appears to be enlarged due to the logarithmic scale. A detailed bin to bin comparison between the two unfoldings cannot be given, due to the different binnings. For the detailed results of the IC-40 unfolding see [22, 80].

Figure 4.29 shows a comparison of the unfolded flux obtained in the analysis at hand to the results obtained in [89]. The Honda-Sarcevic model including a $\pm 15\%$ uncertainty is shown as well. Both results are found to agree well with the theoretical flux within their errorbars. The flux obtained in this analysis, however, appears to be somewhat closer to the theoretical prediction. A possible reason for the larger discrepancies in [89] can be found in the use of LDirC, NDirC and QTot as input parameters for the unfolding. This combination of parameters was found to be particular sensitive to changes in the ice model (see section 4.3). Also note that the x-values in [89] were obtained using a spline fit, whereas the mean of the Monte Carlo distribution for every bin was used in this analysis.

The errorbars obtained with the detailed binwise estimation of uncertainties in this analysis are found to be significantly smaller compared to the ones estimated in [89]. The improvement in the errorbars is particularly prominent for the highest energy bins. One should, however, note that the use of different sets of simulation in the pull mode of Truee was not possible in [89] (this feature has been implemented after the analysis was finished).

A larger errorbar was obtained in the first bin. It is, however, noteworthy that in [89] the impact of changes in the cross section were included in an overall uncertainty. Figure 4.30 shows the comparison of the final unfolding result to

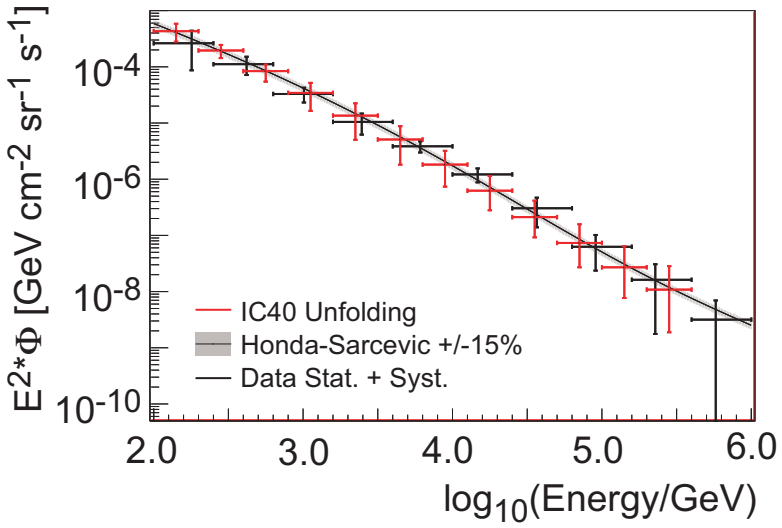


Figure 4.28: Comparison of the unfolded ν_μ spectrum compared to the spectrum obtained with IceCube in the 40 string configuration [22, 80]. The Honda-Sarcevic model including a $\pm 15\%$ uncertainty is shown as well. In general a good agreement between the two measurements and the theoretical prediction is observed. The uncertainties of the measurement at hand were found to improve for most of the bins, especially in energy the range from $10^{2.5}$ GeV to $10^{4.5}$ GeV.

measurements obtained with Frejus [20] and AMANDA [21]. The Frejus results are shown in blue, whereas the AMANDA measurement is depicted in red. In general a good agreement between the measurements is observed and the unfolding result obtained for the analysis at hand is found to nicely extend the previous measurements. Smaller errorbars were obtained compared to the AMANDA results, especially at intermediate energies.

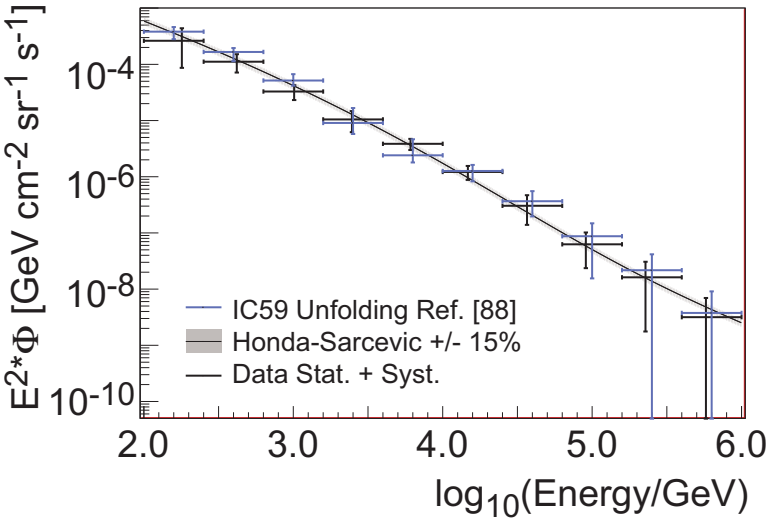


Figure 4.29: Comparison of the unfolding result obtained in the analysis at hand to the results obtained in [89]. The Honda-Sarcevic model including a ±15% uncertainty is shown as well. Although both unfolding results agree with the Honda-Sarcevic flux within uncertainties, the result obtained in this analysis deviates less from the theoretical prediction. Compared to [89] the size of the errorbars was significantly improved. A larger errorbar is obtained for the very first bin.

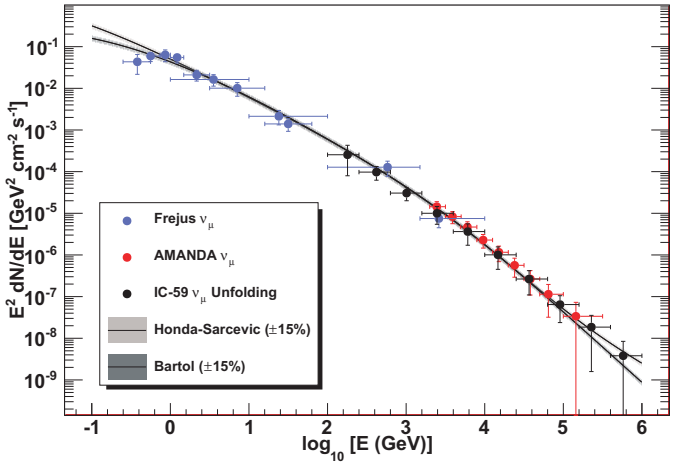


Figure 4.30: Comparison of the final unfolding results (black) to the measurements obtained using Frejus [20] and AMANDA [21]. The Frejus results are shown in blue, whereas the AMANDA measurement is depicted in red. In general, a good agreement between the measurements is observed and the unfolding result obtained for the analysis at hand is found to nicely extend the previous measurements.

Chapter 5

A Data Mining Approach to Unfolding

5.1 Introduction

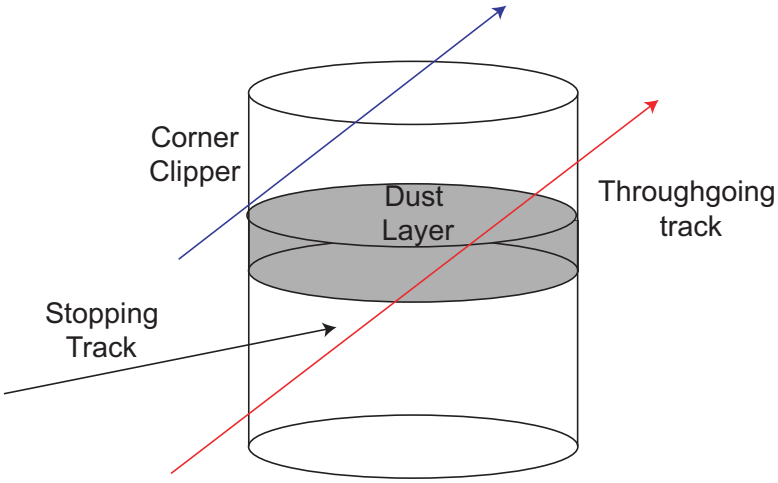


Figure 5.1: Sketch of a large volume neutrino detector utilising a natural medium for detection. Tracks of equal energy can produce significantly different patterns within the detector depending on their geometry (zenith angle, center of gravity, vertex position). The same argument holds for stopping and throughgoing tracks. Inhomogenities in the detection medium, e.g. layers of dust might introduce additional changes to the event patterns.

Despite the fact that unfolding algorithms such as TRUEE [25] and $\mathcal{RUN}$ [90] provide stable and reliable results they suffer from several drawbacks. Firstly, the number of simulated events required for a reliable determination of the response matrix grows exponentially with the number of input variables. This is a common challenge for all density based model building algorithms [95] and limits the number of input variables to three in the most recent version of TRUEE [25, 89].
However, by using three or less input variables, additional and possibly useful information on individual events is discarded. Moreover, the estimation of the response matrix in general implies that the response function of the detector is

identical for all parts of the detector. This implication is of course true for small and homogeneous detectors. For large detectors, especially for those in which natural media are used for detection, this assumption does not hold. Events of the exact same energy may produce different patterns of light depending on their vertex position and zenith angle. A neutrino induced muon of energy 1 TeV stopping inside a large layer of dust, will produce a signature that differs strongly from a 1 TeV muon traversing the detector edge to edge. Therefore, additional information, especially on the geometry of the track, will be helpul in the unfolding process.
Limitations of current unfolding techniques are also observed for small homogeneous detectors, as no information on the individual events is available after unfolding. Thus, the user will in general not be able to determine the contribution of a certain event to a specific part of the energy spectrum. As an additional implication it follows that studies on the behaviour of an energy spectrum with time or angle, require the unfolding of different time slices or zenith bands. This means that studies on the variation of the neutrino energy spectrum over ten different bins in zenith or time, require ten different unfoldings, which need to be tested, optimised and validated individually. Thus, such studies require an additional amount of time and computing resources.
Information on individual events, however, is not only useful for signal events. Studying the contribution of background events to the energy spectrum, might lead to an increase in statistics, as the purity requirements of the sample can be relaxed in case background events do not contribute to regions suffering from small statistics.
Within this chapter a novel data mining based approach towards spectrum unfolding is presented. Section 5.2 will introduce the algorithm itself, whereas section 5.3 is dedicated to a comparison to other existing unfolding approaches. In section 5.4 the performance of the algorithm on a toy Monte Carlo simulation is shown.

5.2 The Dortmund Spectrum Estimation Algorithm DSEA

5.2.1 Description of the Algorithm

In general, unfolding aims at reconstructing a distribution $f(x)$ from a measured distribution $g(y)$ (see section 4.1 for details). For most physics cases, however, a detailed reconstruction of the continuous function $f(x)$ is not necessary and the reconstruction of binned a version $\vec{f}(x)$ is fully sufficient.
Several approaches to the reconstruction of $\vec{f}(x)$ exist (see [92] for a review on several unfolding approaches). Most of these approaches state that the inversion of the response matrix is required in order to obtain $\vec{f}(x)$ from $\vec{g}(y)$. From the machine learning point of view this matrix inversion appears to be unnecessary as the individual bins in $\vec{f}(x)$ can be interpreted as individual classes of events. Solving an inverse problem in physics is therefore equivalent to solving a multiclass

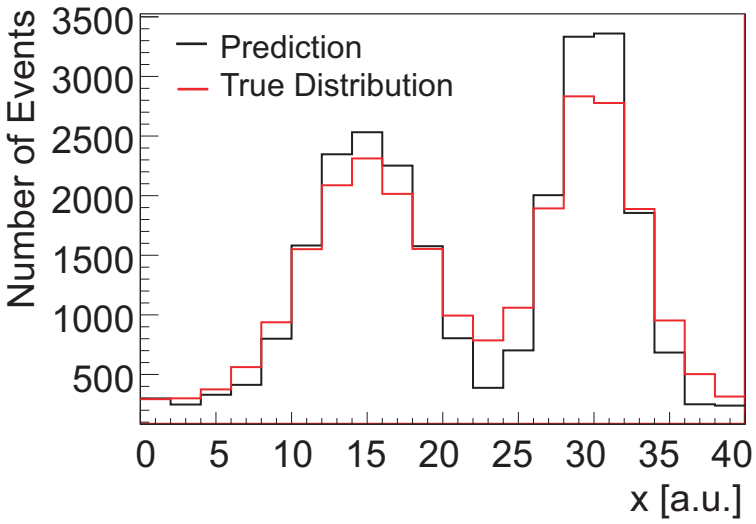
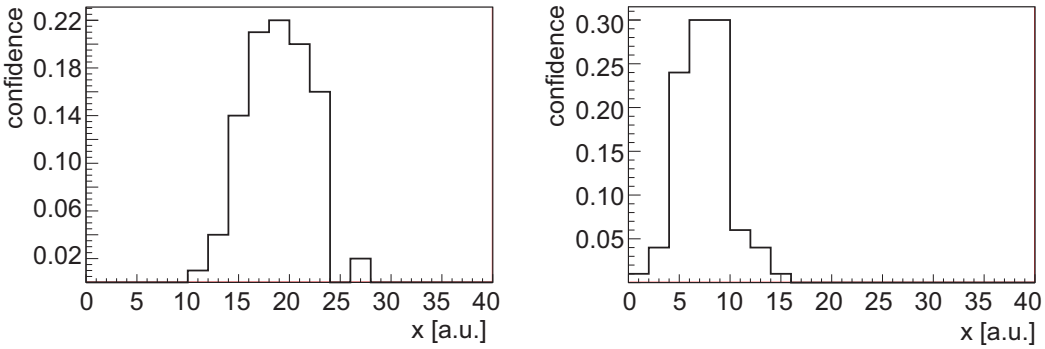


Figure 5.2: Comparison of the predictions returned by a classification algorithm (Random Forest) with the initial spectrum. Despite the fact that characteristic features, e.g. the peak position can be reproduced, the distributions do not agree.

classification problem in machine learning. In fact, reference [87] states that the unfolding of two bins corresponds to a classification problem. Fortunately, several algorithms for solving multinominal classification problems exist.
The result of treating the reconstruction of a spectrum as a multinominal classification task is shown in figure 5.2. The true distribution is shown in red, whereas the reconstructed distribution is depicted in black. Despite the fact, that the major features of the true distribution, the positions of the two peaks, were reconstructed correctly, large discrepancies are observed for several bins. The reconstructed spectrum appears to overestimate the number of events under the peaks, whereas the number of events between the two peaks is underestimated. It should, however, be noted that such an oscillating behaviour of the solution is typical for solutions of inverse problems.
Taking into account the confidence distributions of individual bins, one finds that this oscillating behaviour can be understood. Confidence distributions of two selected events are depicted in figure 5.3. From both distributions one finds, that the maximum of the confidence distribution is not very distinct. The confidences for the neighbouring bins are found to be only marginally different from the maximum value of the distribution. In figure 5.3b the same confidence value is obtained for two adjacent bins. As the maximum of these distributions determines the class into which an event is classified, it can be concluded, that this classification task poses a challenge for the learning algorithm. Furthermore, one should note that confidence values around 0.3 are observed for the class, in which an event was classified.
The small discrepancies in the confidence distributions can be understood, when taking into account that events from neighbouring bins appear to produce similar event patterns, which in turn results in highly similar attribute values.



(a) Confidence distribution of a selected event after application of the forest.

(b) Confidence distribution of a selected event after application of the forest.

Figure 5.3: Confidence distribution of two selected events. Only small differences between the confidence values of neighbouring bins are observed, due to the high similarity of events from neighbouring bins.

The **D**ortmund **S**pectrum **E**stimation **A**lgorithm (DSEA) takes into account the large similarity of adjacent classes by utilising the confidence values c_{ik} of individual events rather than their predictions. These c_{ik} can be interpreted as the probability of an event originating in bin k, given a certain event signature i. Therefore, the confidence distributions obtained from a machine learning algorithm can be interpreted as probability density distributions of individual events. Provided the probability densities of the individual events are estimated reliably, summing up these events will reproduce the correct distribution.
Summarising the previous arguments in one equation one finds that the content of bin k of the final distribution f can be computed via:

$$f_k = \sum_{i=1}^{N_{Events}} c_{ik}. \tag{5.1}$$

It should be noted that the training and testing of a reliable classification algorithm is not part of DSEA, as it only uses the output of these classifiers. This general independence from the learning algorithm, as long as it shows a stable performance, has to be considered one of the greatest advantages of DSEA, as different real life problems might require the use of different learning algorithms.

5.3 Comparison to other Unfolding Approaches

As already outlined in section 4.1 the simplest approach to solving inverse problems is the direct inversion of the response matrix. In addition to the drawbacks already discussed in section 4.1, an overall response matrix for the detector as a whole is computed, which is not desirable for large scale neutrino telescopes (see section 5.1).

The differences between DSEA and the regularised unfolding approach by Blobel [87, 90] are quite numerous. Firstly, in [90] the discretisation is achieved by representing the $f(x)$ as a set of base functions p_j with coefficients a_j. $f(x)$ is then obtained by a maximum likelihood fit, in which the a_j are reconstructed. The conversion into a user defined binning is performed after obtaining $f(x)$.
In DSEA, however, the user defined binning is applied before training and testing of the learning algorithm, which allows the user to pick a binning fully tailored to the problem at hand. For example a more dense binning might be chosen, in regions where a large number of events has been observed. Compared to TRUEE, DSEA therefore offers better control over the distribution of classes that enter the learner.
Furthermore, information on individual events is lost in the case of regularised unfolding and events cannot be unfolded on an event-by-event basis. Moreover, regularised unfolding appears to be a density based method and is therefore inherently limited in the number of input variables, which might cause a loss of relevant information.
It should be pointed out that compared to DSEA at its current stage, regularised unfolding offers the advantage of explicitly using regularisation. The implementation of a regularisation procedure in DSEA was not covered in the thesis at hand, but is one of the key topics in the future development of the algorithm.
As the probability of x given y is obtained on an event by even basis, DSEA can in a sense be seen as a Bayesian unfolding method [96]. This becomes even more understandable, when considering that a Naive Bayes learner, a learner which directly utilises Bayes' theorem for the classification of events, could in principle be used as a machine learning algorithm prior to DSEA. Despite this similarity, however, several differences exist, which will be discussed in the following.
The iterative method presented in [97] utilises Bayes' theorem and also aims at obtaining the conditional probability of x given y. Compared to DSEA, however, the general approach is somewhat different. First of all, the conditional probabilities are estimated in a binwise fashion for all bins in x, rather than event-wise. This binwise computation, has the drawback that the smearing matrix, which is computed from Monte Carlo simulations, is kept constant for every event. The drawbacks of this approach have already been discussed above. The smearing matrix is not explicitly calculated in DSEA, as the pdf of the individual events is obtained using a machine learning algorithm.
Furthermore, no information on individual events is retained in the approach presented in [97]. Although Bayesian Networks are used, a possible utilisation of machine learning algorithms is not discussed.
An approach towards the use of machine learning techniques for solving inverse problems is given in [98], where neural networks are utilised. In contrast to DSEA, where unfolding is handled as a multiclass problem, the inverse problem is converted into a binominal classification task, through a series of variable transformations, which transform $f(t)$ into $g(s)$ [98]. Moreover, the cumulated conditional probability density $G(s|\vec{x})$ is obtained from the neural network output rather than

the conditional probability density $g(s|\vec{x})$. This is due to the two class structure in which the inverse problem was transformed. The neural network is used to determine whether the true value t is greater or smaller than a certain threshold L.
Despite the fact that [98] is able to perform an event-by-event unfolding, without determining the response matrix the method appears to include some unnecessary complexity, which is easily resolved, by treating the inverse problem as a multinominal classification task. It should, however, be noted that the use of neural nets in principal allows for the use of more than three input variables, and therefore for the use of additional information in the unfolding process.
Another machine learning based approach to the solution of inverse problems is presented in [99]. In contrast to Dsea, however, it still aims at determining the transition matrix A, through the use of machine learning techniques. The matrix is obtained by utilising system identification techniques. Validation of the unfolding is carried out using standard validation procedures such as leave-one-out and cross validation. Furthermore, the algorithm presented in [99] aims at reconstructing the spectrum as a whole, rather than at the unfolding of individual events. The explicit use of a learning algorithm is not discussed. One should note that the dimensionality of the algorithm is, at least in principle, not limited.
The largest similarities can be found with the connectionist approach presented in [100], where the regression problem is also converted into a classification problem and solved by the application of a neural net. Furthermore, a fractional binning is discussed, where events are not classified in one bin but in at most two adjacent bins. This is very similar to Dsea. For Dsea, however, an event can be classified into an arbitrary number of bins. Moreover, an explicit ordering of bins is required in [100]. Such an explicit ordering is not required in Dsea. Furthermore, the data is partitioned into equal mass bins, each having approximately the same number of events. The individual bin centers are then calculated after binning. This kind of binning is highly undesirable for the application in astroparticle physics, where power law spectra are observed for a number of phenomena. Furthermore, the algorithm in [100] appears to be originally designed for the purpose of time series predictions. A possible use of an eventwise information extraction is not discussed.
The review of various very different unfolding methods in this section shows that certain elements and ideas of these algorithms can also be found in Dsea. One should, however, note that the explicit use of learning algorithms for the reconstruction of spectra has been discussed only in [98] and [100]. In both references the learning process is built into the spectrum reconstruction process. This is very different for Dsea, as it does not depend on a specific kind of classifier as long as the learner is well trained and stable in performance.
The idea of fractional classification can be traced back to [100]. It has, however been improved by allowing for a partioning of events into more than two adjacent bins.
One of the great advantages of Dsea can be found in the fact that it allows the user to follow the flow of any given event throughout the complete learning and

reconstruction method directly. This firstly removes the black box character of some other unfolding procedures and secondly enables the user to extract physical meaningful information about the spectrum on an event-by-event basis.

5.4 Validation of DSEA on Toy Monte Carlo simulations

5.4.1 Toy Monte Carlo production

Since the use of toy Monte Carlo events did not aim at simulating a specific experiment or a certain type of detector a simple approach using Gaussian smearing was utilised. Ten different observables were computed from a given input spectrum, which was composed of two distinct gaussian peaks of different width. Dependencies of the form:

$$y_i \propto x_i^{\gamma},\ \gamma \in [0.5; 1.5], \tag{5.2}$$

were used to include a certain degree of nonlinearity. The y_i was then smeared using a gaussian distribution:

$$\bar{y}_i = f(y_i, \sigma), \tag{5.3}$$

with σ obtained at random before running the actual Monte Carlo production but fixed throughout the Monte Carlo production. Random offsets were used in order to make the individual observables as distinct as possible.

5.4.2 Training of the Machine Learning algorithm

A Random Forest was chosen as a machine learning algorithm for the validation of the DSEA performance on a toy Monte Carlo simulation. The Random Forest is a quite intuitive choice for the treatment of this problem because of its well known stability and interpretability. Moreover, the Random Forest is able to handle multiclass problems and returns the confidence values for every class, which is required as an input for DSEA. Training and testing of the forest was carried out in a 5-fold cross validation using 2.6×10^5 Monte Carlo events. The number of events used for training was limited to 8×10^4.
The forest was found to perform stably and reliably on the Monte Carlo events. No signs of overtraining were observed.

5.4.3 Performance on Toy Monte Carlo

Figure 5.4 shows the outcome of the unfolding with DSEA compared to the true distribution. In general, a good agreement within statistical uncertainties in observed between the two distributions. Slight deviations are observed in the center and towards the higher end of the distribution.
Figure 5.5 shows the deviation of the unfolded distribution obtained with DSEA from the true distribution in units of the statistical uncertainty σ. For 18 out of 20 bins

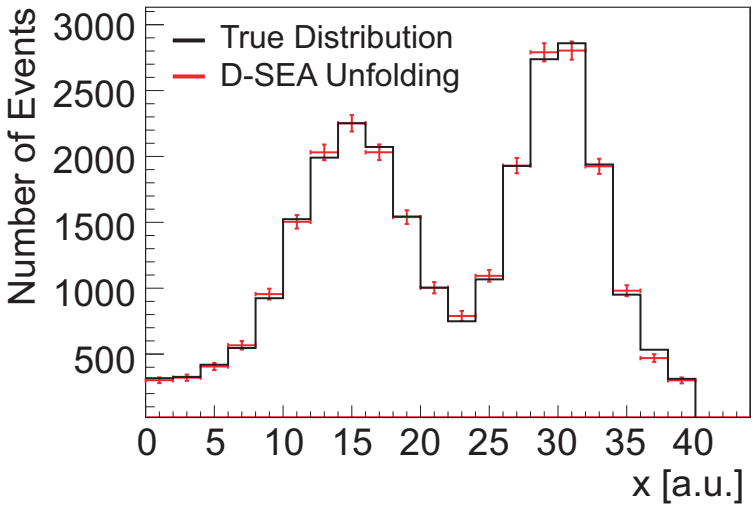


Figure 5.4: Unfolding results obtained with DSEA compared to the true distribution. A good agreement is observed within statistical uncertainties. Slight deviations are observed in the region between the two peaks and towards the higher end of the distribution.

both distribution are found to agree within $1\,\sigma$. One bin is found to show a deviation only slightly above the $1\,\sigma$ limit. One bin is found to show a deviation on the order of $2\,\sigma$. No deviations above $2\,\sigma$ were observed.

5.4.4 Utilising the 2D Binning

The capabilities of DSEA compared to other unfolding approaches can best be demonstrated by analysing 2-dimensional histograms. Due to the fact that the information on the individual events is still available after the unfolding, this information can be utilised to plot other x-dependent or independent variables versus the unfolding variable x. The capabilities of this approach will be outlined in this section.
Figure 5.6 shows the true distribution of x versus the x-dependent variable y. The colour column corresponds to the bin content. Two clear maxima are observed. The reconstructed two dimensional distribution of y versus x is depicted in figure 5.7. Despite the fact that the resolution is quite limited the major features of the distribution, the two peaks, are clearly visible.
The ratio of the true distribution and the reconstructed distribution is shown in figure 5.8. The deviation in units of the statistical uncertainty σ is depicted as the colour column. No deviations above the $3\,\sigma$ limit were observed. Several bins appear to have deviations above $1\,\sigma$ or $2\,\sigma$, respectively. This is, however, expected since the number of bins in this very histogram is $n_{bins} = 400$. Thus, only 272 bins are expected to be below the $1\,\sigma$ limit and 380 bins are expected to be below the $2\,\sigma$ limit.

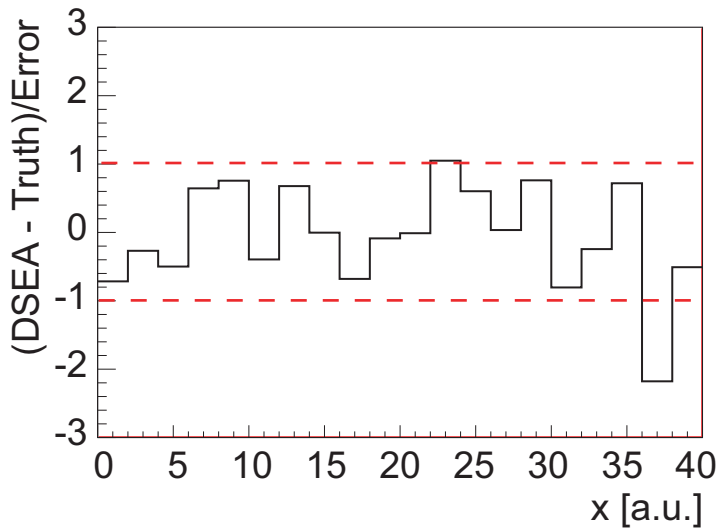


Figure 5.5: Deviation between the distribution unfolded using DSEA and the true distribution in units of the statistical uncertainty σ. Only two bins out of 20 have been found to show deviations larger than $1\,\sigma$.

5.4.5 Comparison to TRUEE

To further investigate the performance of DSEA its results were compared to the ones obtained using TRUEE. The comparison of the reconstructed spectrum obtained with TRUEE (red) and DSEA (black) is depicted in figure 5.9. The shaded area represents the true distribution of events. One finds that the results of both unfolding algorithms produce comparable results. The statistical errorbars obtained with TRUEE, however, appear to be somewhat larger. In general, the results of both algorithms were found to reproduce the true distribution within the estimated statistical uncertainties.
Figures 5.10 and 5.11 show the pull mode results of DSEA and TRUEE, respectively. On average no deviations larger than $1\,\sigma$ are observed for either of the unfolding algorithms. This result indicates, that the slight deviations to the true distribution, observed in section 5.4.3 are due to statistical fluctuations.
Compared to the TRUEE result a slightly oscillating behaviour of the DSEA solution is observed. Such behaviour is characteristic for the solution of inverse problems and indicates that regularisation is explicitly required in DSEA. Note, however, that the observed oscillations are well below the $1\,\sigma$ limit. They can, thus, be tolerated in the application of DSEA.

5.4.6 Dependency on the Input Distribution

Within this chapter the dependency of the obtained spectrum on the input distribution used for training the classifier is examined. The role of the input spectrum of the learner is comparable to that of the prior in Bayesian unfolding. For the case of Bayesian unfolding the unfolding algorithm was found to reproduce the

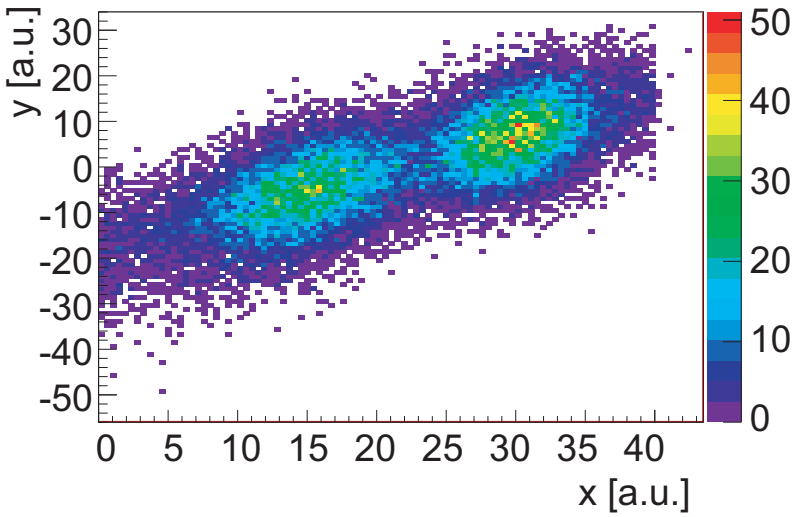


Figure 5.6: True distribution of an x-dependent variable used to illustrate the capabilities of DSEA, when using it for the analysis of two dimensional distributions.

prior only in cases where no additonal information was present in the unfolding attributes [96].
In order to study the behaviour of DSEA with respect to the input spectrum, a uniform distribution was used as input for the learner. A uniform distribution was chosen, as this is the most reasonable distribution to commence with, when no additional information, theoretical or experimental, is available.
The outcome is shown in figure 5.12. The DSEA unfolding result is depicted in red, whereas the true distribution is shown in black. Although the two distributions do not agree, the reconstructed distribution clearly deviates from a uniform distribution.
Furthermore, the algorithm is found to correctly reproduce the position of the two peaks. This result is remarkable as it allows for the detection of features in the true distribution even if those have not been directly simulated. Note, that generally some information will be available on the sought after distribution, either from the theoretical or from the experimental side.
Moreover, the result obtained using a uniform distribution can be used in an iterative procedure. In this case the outcome of the unfolding is used as an input distribution for the generation of Monte Carlo events. These Monte Carlo events can then be used for the training of the machine learning algorithm. This procedure is repeated until a stopping criterion is reached.
The result of such an iterative unfolding is presented in figure 5.13. The outcome of the first iteration is shown in black, whereas the outcome of the seventh iteration is depicted in red. The average deviations are found to significantly decrease between the first and the seventh iteration. Moreover, the oscillatory behaviour of the solution is found to decrease as well. In fact the result obtained after the seventh iteration is comparable to the result obtained using the exact spectrum as input for the training of the learner.

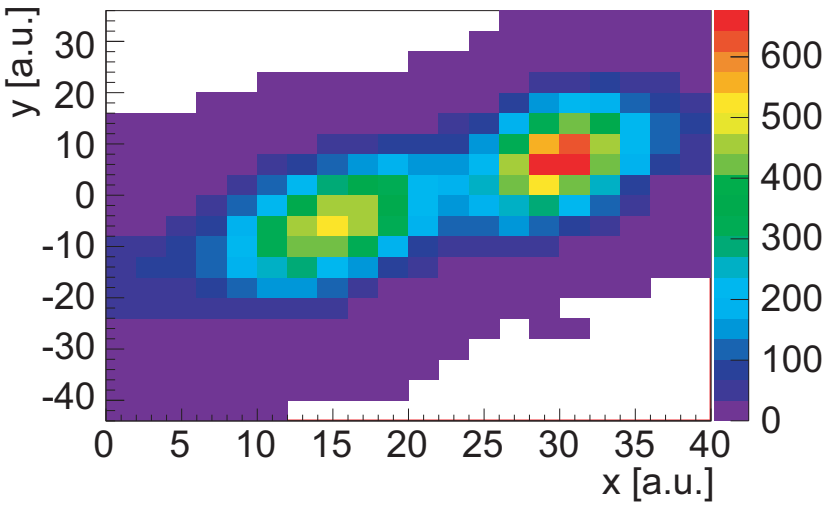


Figure 5.7: Distribution of the x-dependent variable y obtained after unfolding with DSEA. Despite the limited resolution, the major features of the distribution (2 peaks) are restored.

In conclusion, it is found that the output of DSEA does show a certain dependency on the input spectrum. This dependency is, however, found to be small. Furthermore, the algorithm was able to reconstruct the correct peak positions in the true distribution, using a uniform distribution as input. Moreover, the DSEA results can be used in an iterative manner for spectrum reconstruction in case no prior knowledge on the sought distribution exists.

5.5 DSEA as a RAPIDMINER plugin

Using DSEA in an already existing data mining environment such as WEKA, TMVA or RAPIDMINER appears to be the optimal choice since the algorithm relies on the output of already existing machine learning algorithms. Thus, no effort is required for the implementation and debugging of learning algorithms. Morevover, the user is enabled to select from a variety of classification algorithms, which further enables him to select the algorithm best suited for a specific unfolding problem.
Further requirements for utilising DSEA within any data mining environment are high performance, simple installation and good usability. All of these features are offered by RAPIDMINER. Thus, DSEA was implemented as a plugin in RAPIDMINER. The operator list, after successful installation of the plugin, is shown in figure 5.14. The installation of plugins in RAPIDMINER is explained in [75].
An example process utilising all three operators available in the current version of DSEA is depicted in figure 5.15. All three operators require the input of a labelled example set which has been classified using a learner of the user's choice. The actual DSEA operator, which can be run in data and MC mode, realises the binwise summation over the confidence values of the individual attributes. In case

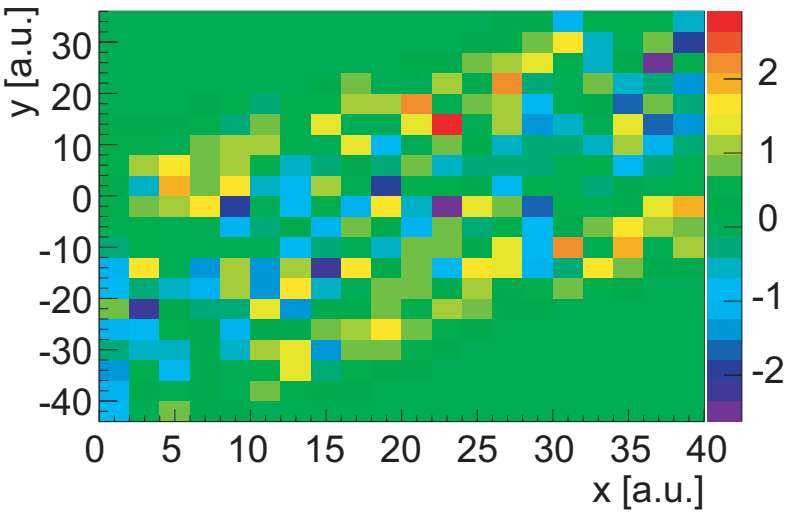


Figure 5.8: Comparison of the true distribution to the distribution unfolded using Dsea. The deviation of both distributions in units of the statistical uncertainty σ is shown as the color column. No deviations $> 3\,\sigma$ are observed. Deviations of the order of $1\,\sigma$ and $2\,\sigma$ are observed for a couple of bins. One should note, however, that such a behaviour is expected, as the total number of bins in this histogram is $n_{bins} = 400$.

the operator is run in MC mode a comparison to the true distribution is performed in addition to the estimation of the spectrum.
Within the D-sea Pull operator a user specified number of test unfoldings is performed on a given fraction of the data. These test unfoldings are then compared to the Monte Carlo truth and the average deviation is computed binwise and in units of the statistical uncertainty.
The Dsea-2D operator can be used to construct a two dimensional histogram using the unfolding variable and an arbitrary additional parameter, e.g. angle or time. The additional parameter needs to be discretised before the application of Dsea-2D. Note, that this operator is not to be confused with a multidimensional unfolding, as no unfolding is carried out for the second parameter.

5.6 Summary and Conclusion

The novel Dsea approach was successfully applied to toy Monte Carlo simulations. The true distribution was fully reproduced by utilising Dsea within statistical uncertainties. All but two bins were found to deviate less than $1\,\sigma$ from the true distribution.
The data mining approach to unfolding was further found to be applicable for the reconstruction of two dimensional distributions. Despite the fact that the resolution of the plot still has some optimisation potential, the major features of the distributions could be reconstructed. No deviations larger than $3\,\sigma$ were observed. The reconstruction of two dimensional distributions is, however, not to be confused with a two dimensional unfolding. Within the toy Monte Carlo no additional

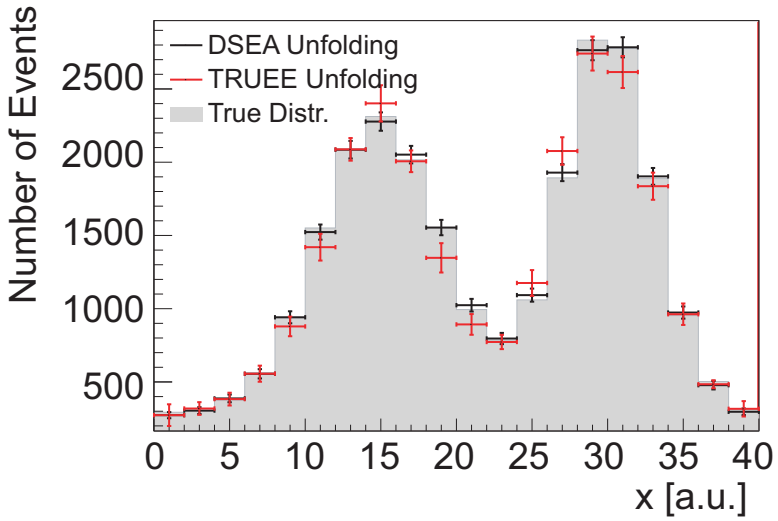


Figure 5.9: Comparison of unfolding results obtained with DSEA (black) and TRUEE (red). Comparable reults are obtained. TRUEE, however, was found to produce slightly larger errorbars.

smearing was introduced to the x-dependent variable y. A true two dimensional unfolding, however, would also require the correction of such an additional smearing.
The results were found to agree with unfolding results obtained with TRUEE.
In summary, a first application of DSEA on toy Monte Carlo simulations was found to be very successful. Furthermore, the great potential of this data mining based approach to unfolding has been illustrated through the reconstruction of a two dimensional distribution. An application of DSEA in astroparticle physics in the near future, will yield deeper insight and a better understanding of astrophysical phenomena. Studies of variations in the atmospheric neutrino spectrum with time and angle, are key candidates for a first application of DSEA, as reference measurements obtained with TRUEE or in a forward folding procedure [22, 80] exist. Moreover, a great improvement of studies on neutrino oscillations with IceCube is expected, since these studies no longer rely on energy estimators, but can obtain the energy versus zenith distribution using DSEA.
Finally, DSEA might be applied in studies on the neutrino mass hierarchy using the Proposed IceCube Next Generation Upgrade (PINGU) [101, 102].

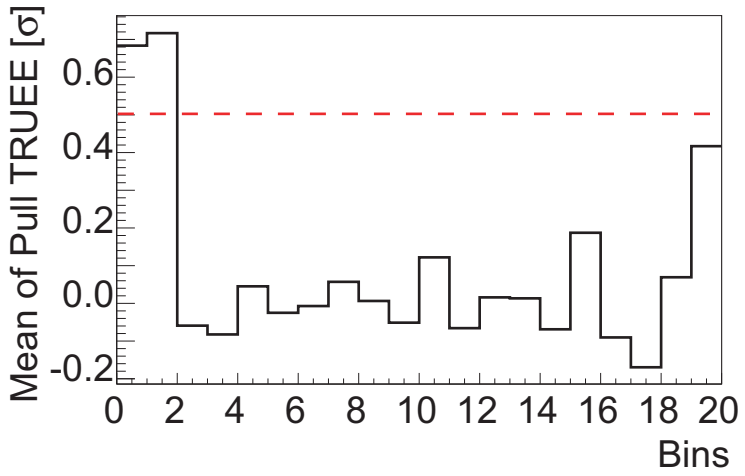


Figure 5.10: Result of the pull mode implemented in TRUEE. 500 pulls were performed unfolding 10% of the data and comparing it to the MC truth. On average the correct distribution is obtained for every bin within the statistical error.

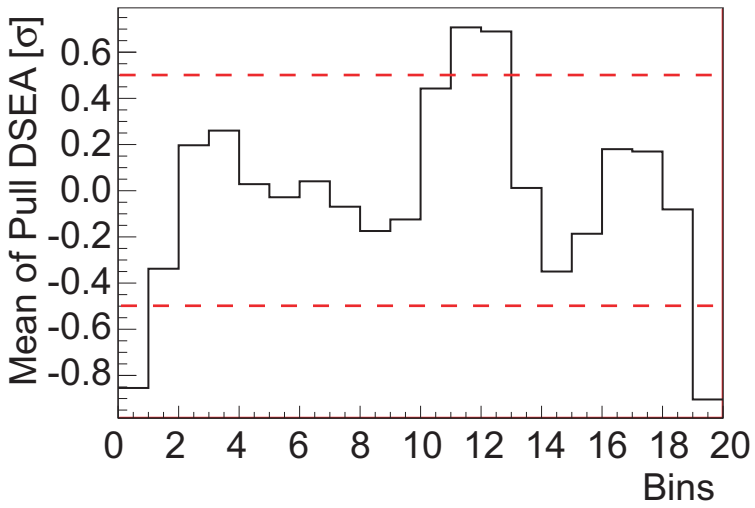


Figure 5.11: Pull results obtained with DSEA in a bootstrapping procedure. 10% of all MC events were treated as pseudodata. After unfolding the result was compared to the MC truth. A slightly oscillating behaviour of the solution is observed.

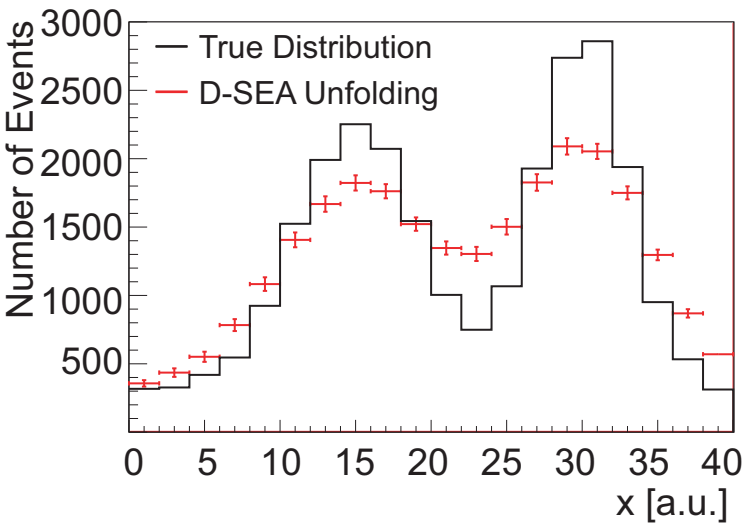


Figure 5.12: Dsea output, depicted in red, using a uniform distribution for the training of the learner. The true distribution, in black, is shown for comparison. Although the distributions do not agree within the estimated statistical uncertainties, the positions of the two peaks have been correctly reconstructed using Dsea.

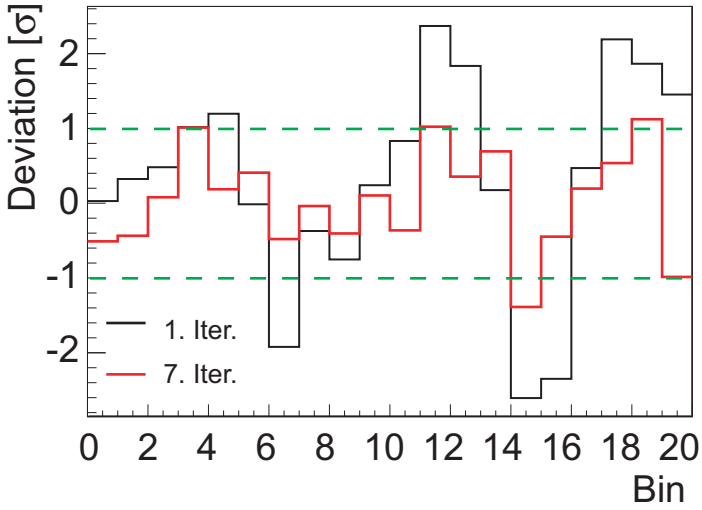


Figure 5.13: Pull mode results for the first (black) and the seventh (red) iteration. The unfolding result of iteration i was used to generate Monte Carlo events used in the training of iteration $i+1$. A uniform distribution was used for the training of the first iteration.

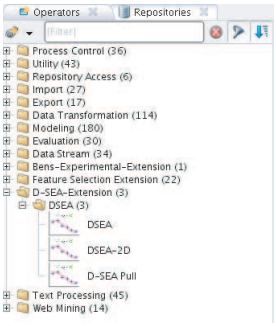


Figure 5.14: View of the list of available operators in RapidMiner, after a successful installation of the Dsea-plugin. In the current version three operators are available.

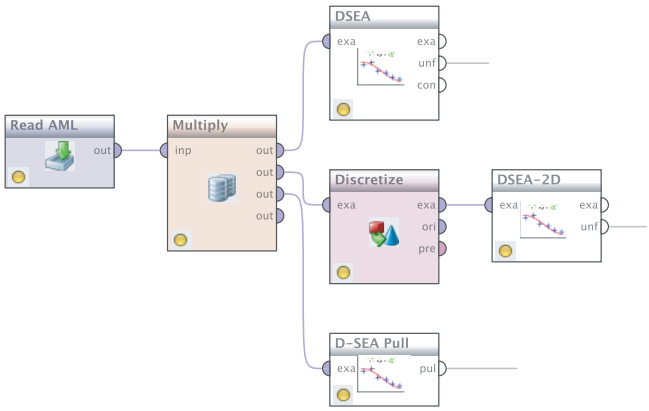


Figure 5.15: RapidMiner example process depicting the use of the individual operators implemented in the Dsea-plugin for RapidMiner.

Chapter 6

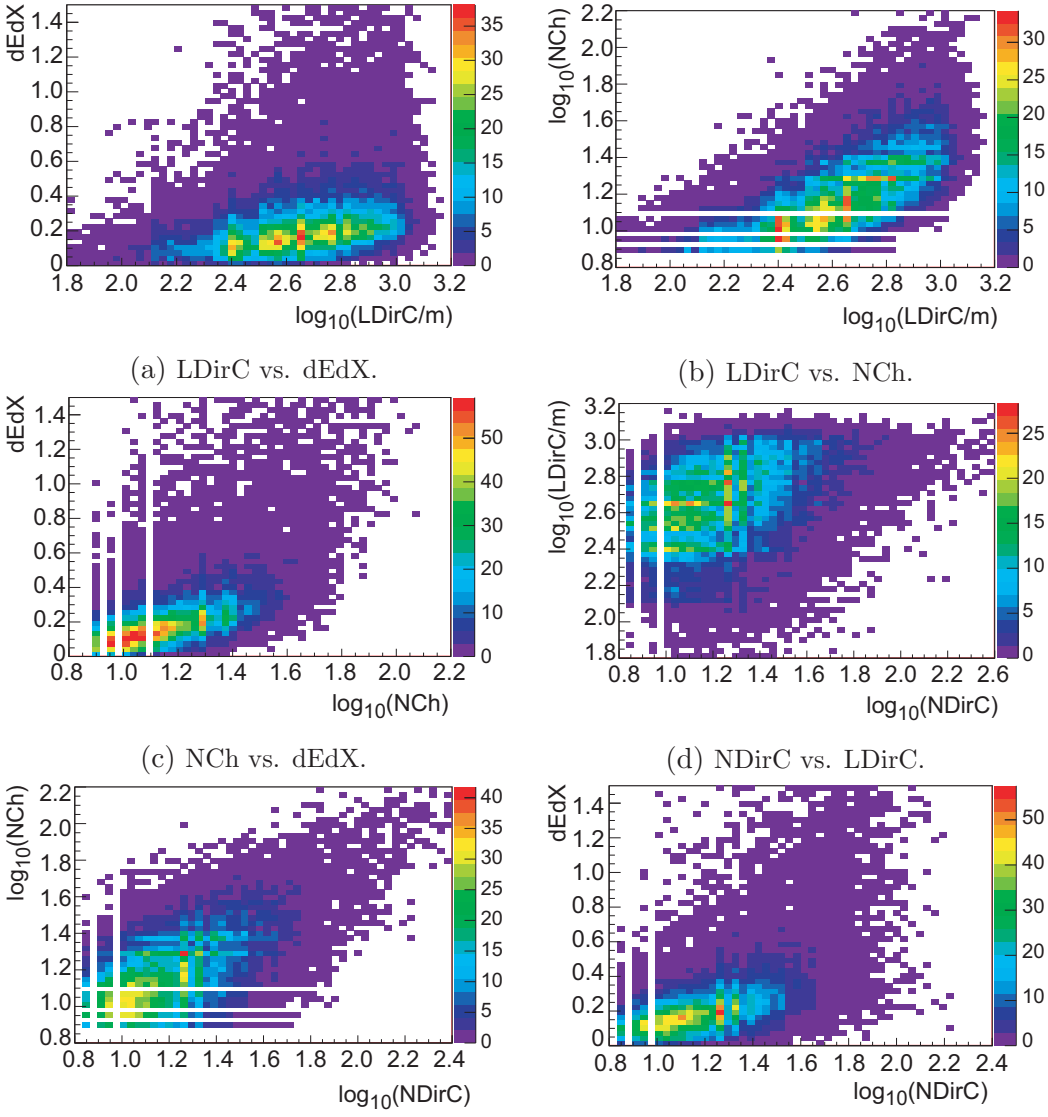
Summary and Outlook

A measurement of the atmospheric neutrino spectrum up to an energy of 1 PeV using IceCube in the 59 string configuration was presented in this thesis. This corresponds to the highest energy ever accessed in a measurement of atmospheric neutrinos.
The analysis consisted of two basic parts, the first one being an event selection, aiming at separating neutrino induced events from atmospheric muons. This was achieved by utilising state of the art machine learning techniques, consisting of a Random Forest, preceeded by a dedicated feature selection. An improvement from 49.3 neutrino events per day, obtained in an IC-40 analysis by using Boosted Decision Trees, to 80 neutrino events per day was achieved. Correcting for the increased detector volume one finds that this corresponds to an increase of approximately 8%, which can be attributed to the utilisation of a Random Forest preceeded by a dedicated MRMR Feature Selection.
A more direct comparison was obtained to the IC-59 point source sample, which als utilised boosting but had a higher contamination of atmopheric muons. When this contamination is taken into account, an increase in event rate of approximately 50% was achieved.
Furthermore, it was observed that the increase in event rate was primarily achieved at low and intermediate energies. The IC-59 point source event selection was found to outperform the Random Forest event selection at the highest energies. Neutrinos from point sources follow a harder spectrum and will, therefore, be dominant over the background of atmospheric neutrinos at higher energies. Accordingly, a point source event selection is optimised for high energy neutrinos.
The unfolding of the obtained neutrino sample by utilising the unfolding software Truee was presented in the second part of the thesis. Within this unfolding the atmospheric ν_μ spectrum was probed between 100 GeV and 1 PeV. Compared to a measurement obtained with IC-40 this corresponds to an increase of more than half an order of magnitude in energy. Furthermore, the errorbars, especially at intermediate energies, were significantly reduced compared to previous measurements [22].
Furthermore, systematic uncertainties were derived in a statistically reliable way, by using the pull mode of Truee on different sets of Monte Carlo simulations.

Several sources of uncertainties including the ice model were investigated in detail. The error estimation using different DOM efficiencies was found to fail, due to unknown reasons. A detailed discussion on this matter was given, concluding that an estimation of errors due to uncertainties in the DOM efficiency is not necessarily required since it might be mimicked by other systematic effects, e.g. increased/decreased cross sections.
As different sources of systematic uncertainties, which result in similar effects inside the detector, cannot be disentangled on real data, cross checks were carried out dividing the detector into different parts and unfolding each part of the detector separately. The spectra obtained for these individual parts were found to agree with the full detector spectrum within the estimated uncertainties, after the ice model uncertainties of bins 5 and 6 were scaled by a factor of two. The scaling became necessary, since the ice model uncertainties derived for these bins were particularly small.
A novel approach towards unfolding was given in the third part of the thesis. This new approach is based on treating unfolding as a mutinominal classification problem and utilising the output of an arbitrary but well trained learner. The algorithm corresponds to an event-by-event unfolding in which all information on the individual events is retained in the process. In comparison to the popular regularised unfolding this approach offers the possibility of using a larger number of input variables. This leads to an increase in the available information, which in turn leads to a more accurate estimation of the sought distribution.
In addition to theoretical considerations the algorithm was applied in toy Monte Carlo studies. From an unfolding carried out using TRUEE it was found that both algorithms yield comparable results. It was observed that the algorithm can be used in an iterative procedure in case no prior information is available on the sought spectrum.

Appendix A

Correlation Between Unofolding Variables



(a) LDirC vs. dEdX.

(b) LDirC vs. NCh.

(c) NCh vs. dEdX.

(d) NDirC vs. LDirC.

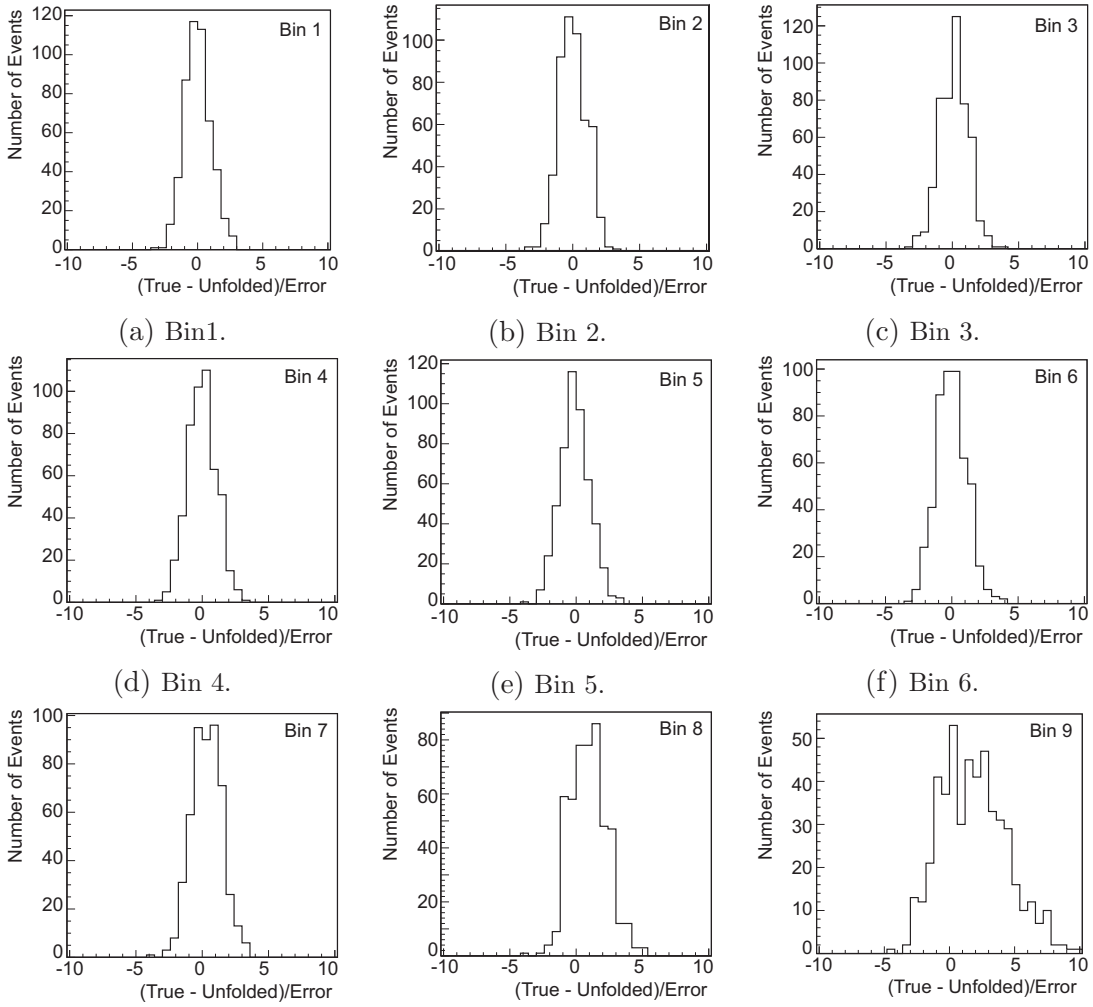
(e) NDirC vs. NCh.

(f) NDirC vs. dEdX.

Figure A.1: Mutual correlation of the individual variables considered for unfolding.

Appendix B

Pull Distributions for Individual Bins

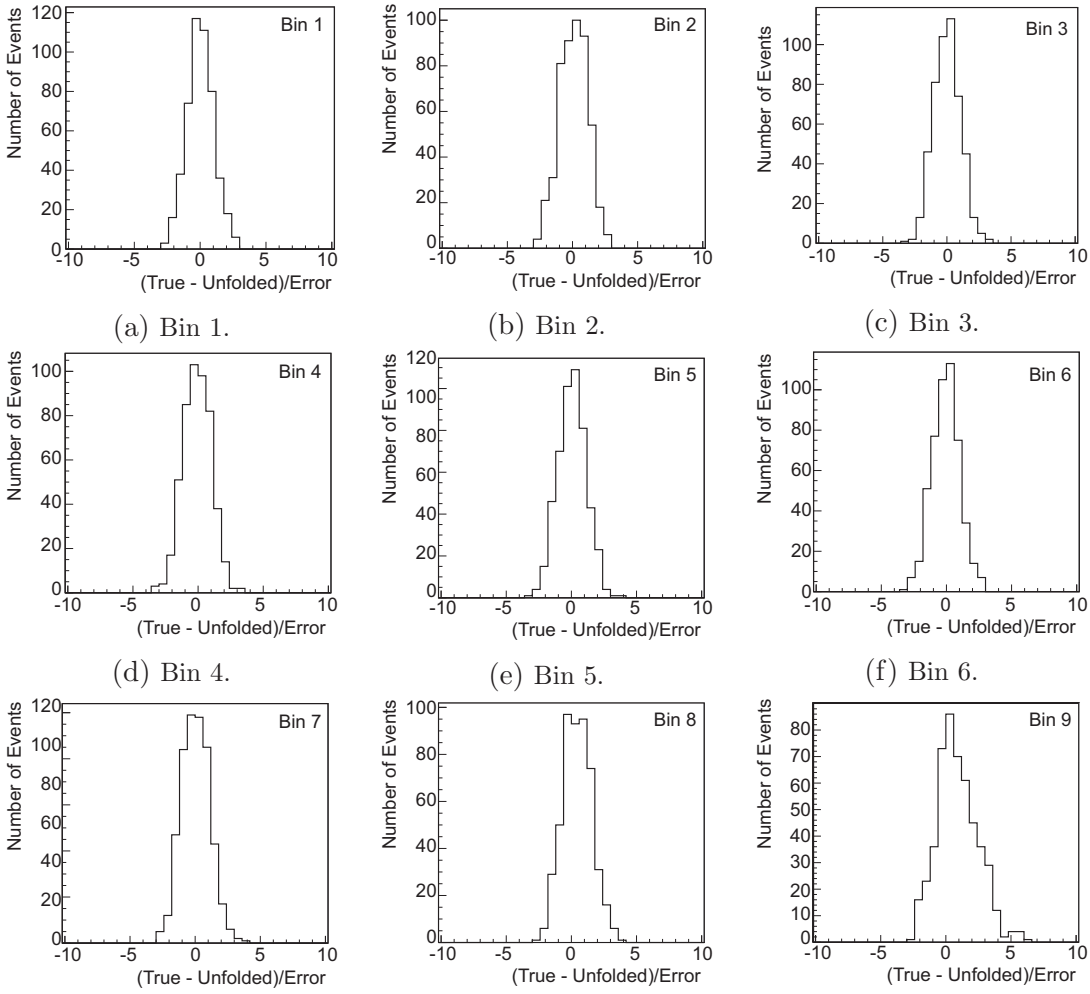


(a) Bin1. (b) Bin 2. (c) Bin 3.

(d) Bin 4. (e) Bin 5. (f) Bin 6.

(g) Bin 7. (h) Bin 8. (i) Bin 9.

Figure B.1: Pull mode distributions for bins 1 to 9 using LDirC, NDirC and NCh as input variables. The number of degrees of freedom was set to $ndf = 5$.

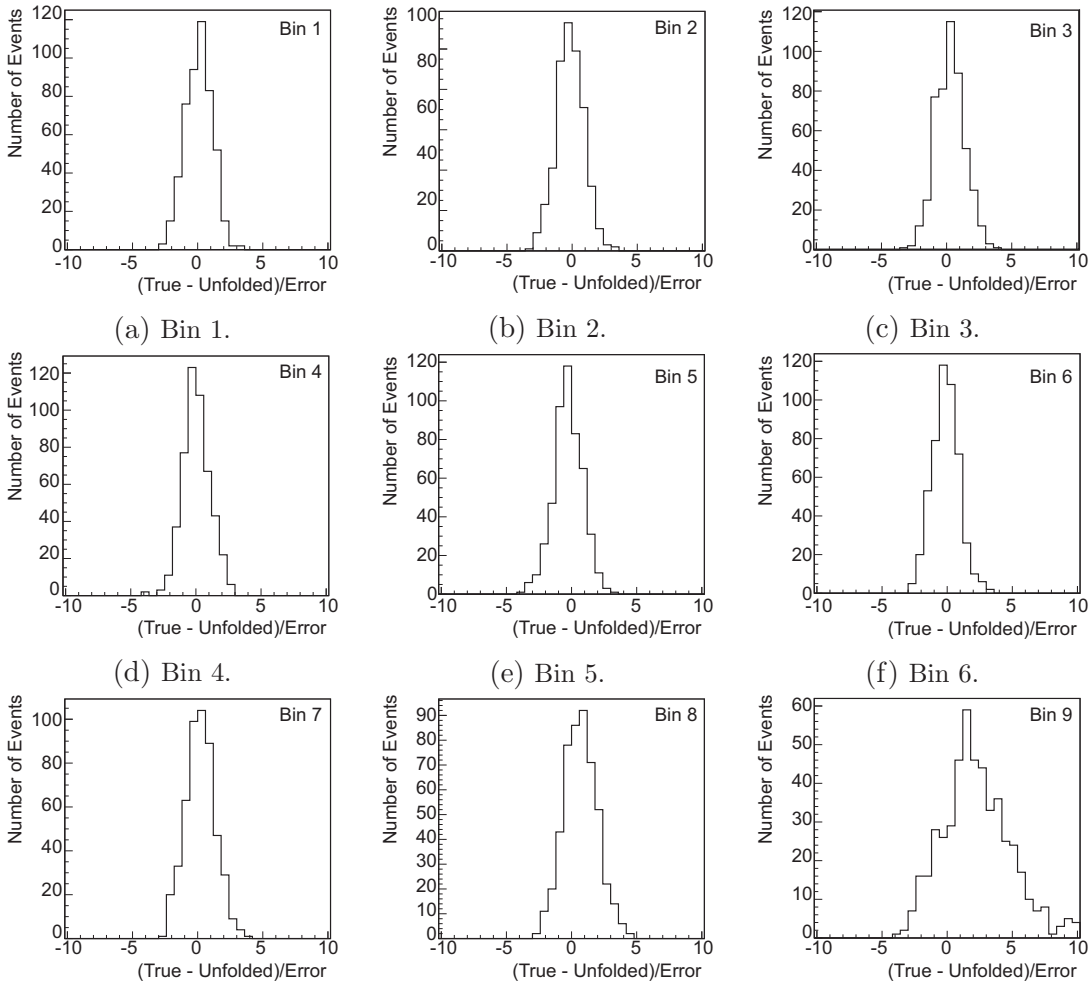


(a) Bin 1. (b) Bin 2. (c) Bin 3.

(d) Bin 4. (e) Bin 5. (f) Bin 6.

(g) Bin 7. (h) Bin 8. (i) Bin 9.

Figure B.2: Pull mode distributions for bins 1 to 9 using LDirC, NDirC and NCh as input variables. The number of degrees of freedom was set to $ndf = 6$.

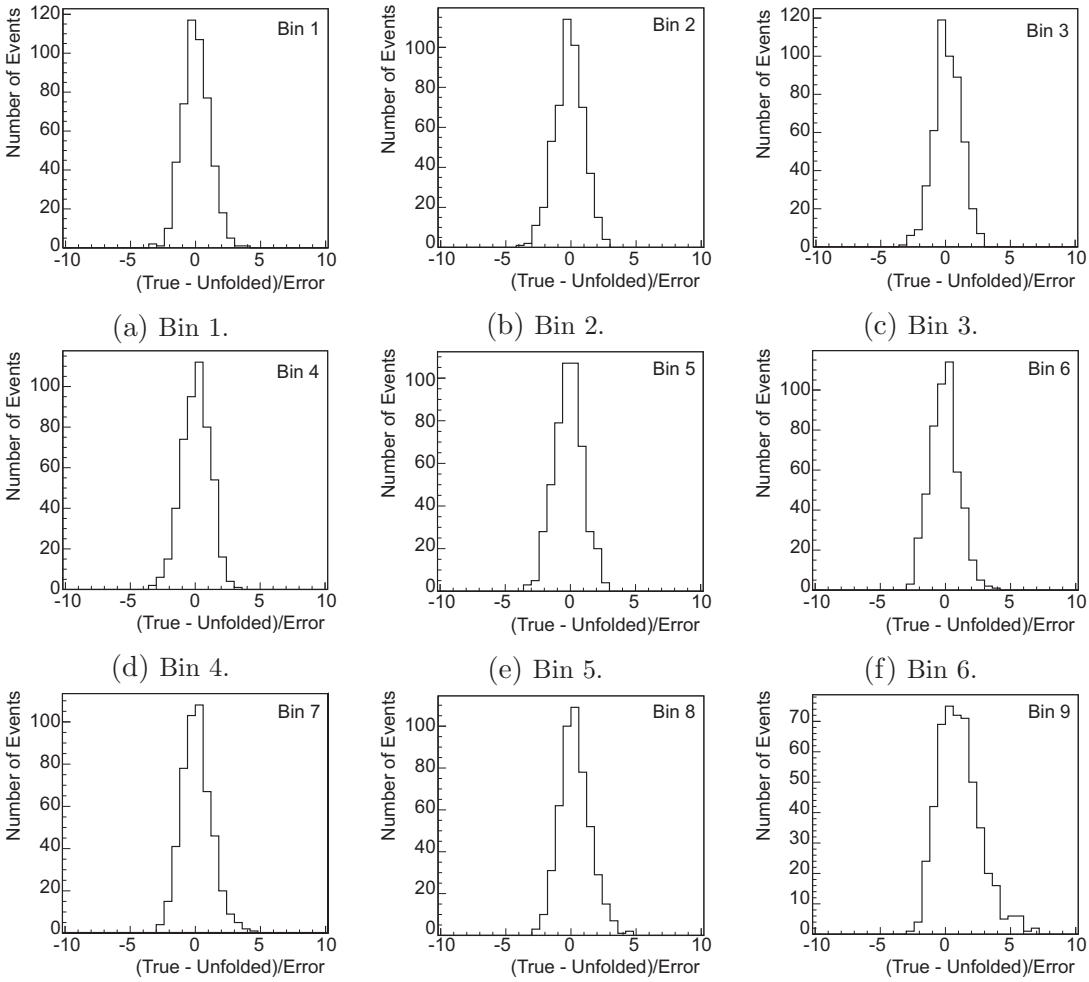


(a) Bin 1. (b) Bin 2. (c) Bin 3.

(d) Bin 4. (e) Bin 5. (f) Bin 6.

(g) Bin 7. (h) Bin 8. (i) Bin 9.

Figure B.3: Pull mode distributions for bins 1 to 9 using LDirC and NDirC as unfolding variables. The number of degrees of freedom was set to $ndf = 5$.

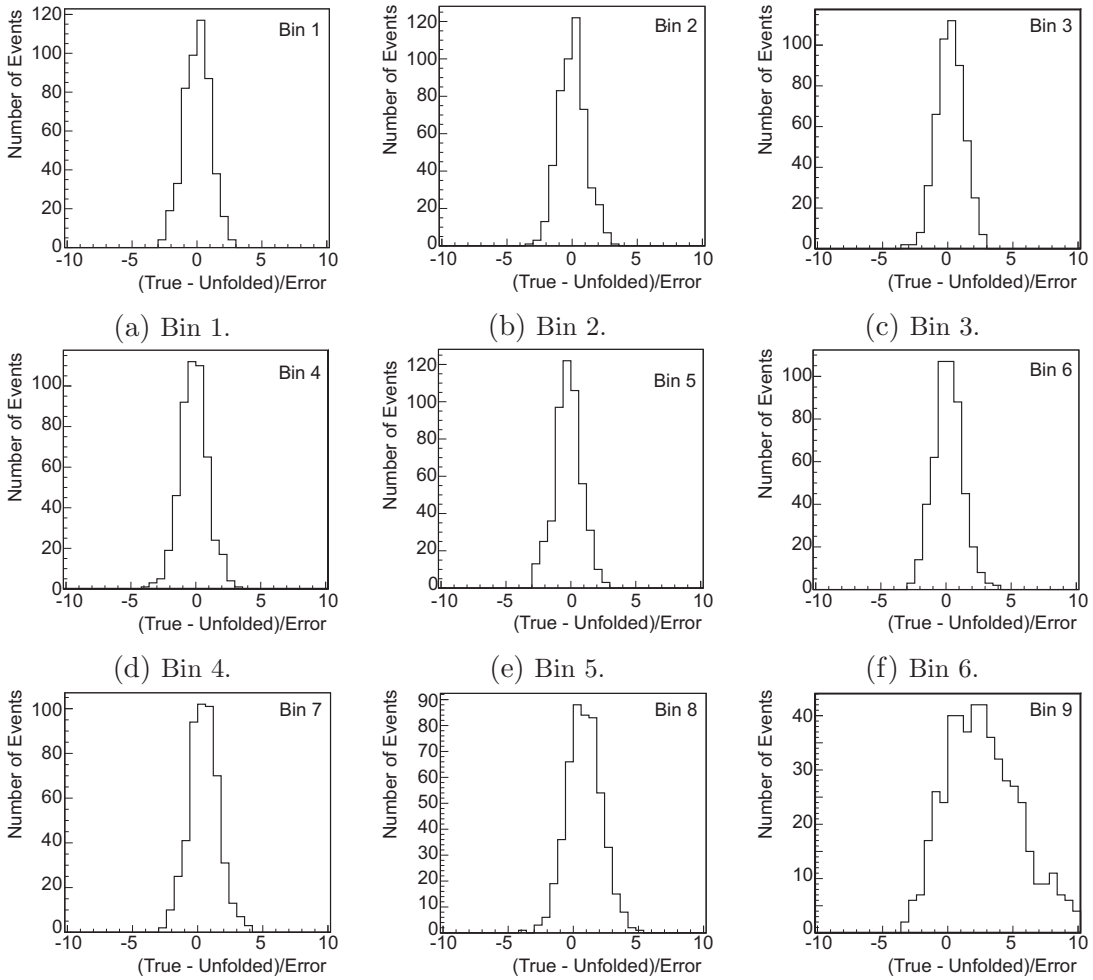


(a) Bin 1. (b) Bin 2. (c) Bin 3.

(d) Bin 4. (e) Bin 5. (f) Bin 6.

(g) Bin 7. (h) Bin 8. (i) Bin 9.

Figure B.4: Pull mode distributions for bins 1 to 9 using LDirC and NDirC as input variables. The number of degrees of freedom was set to $ndf = 6$.

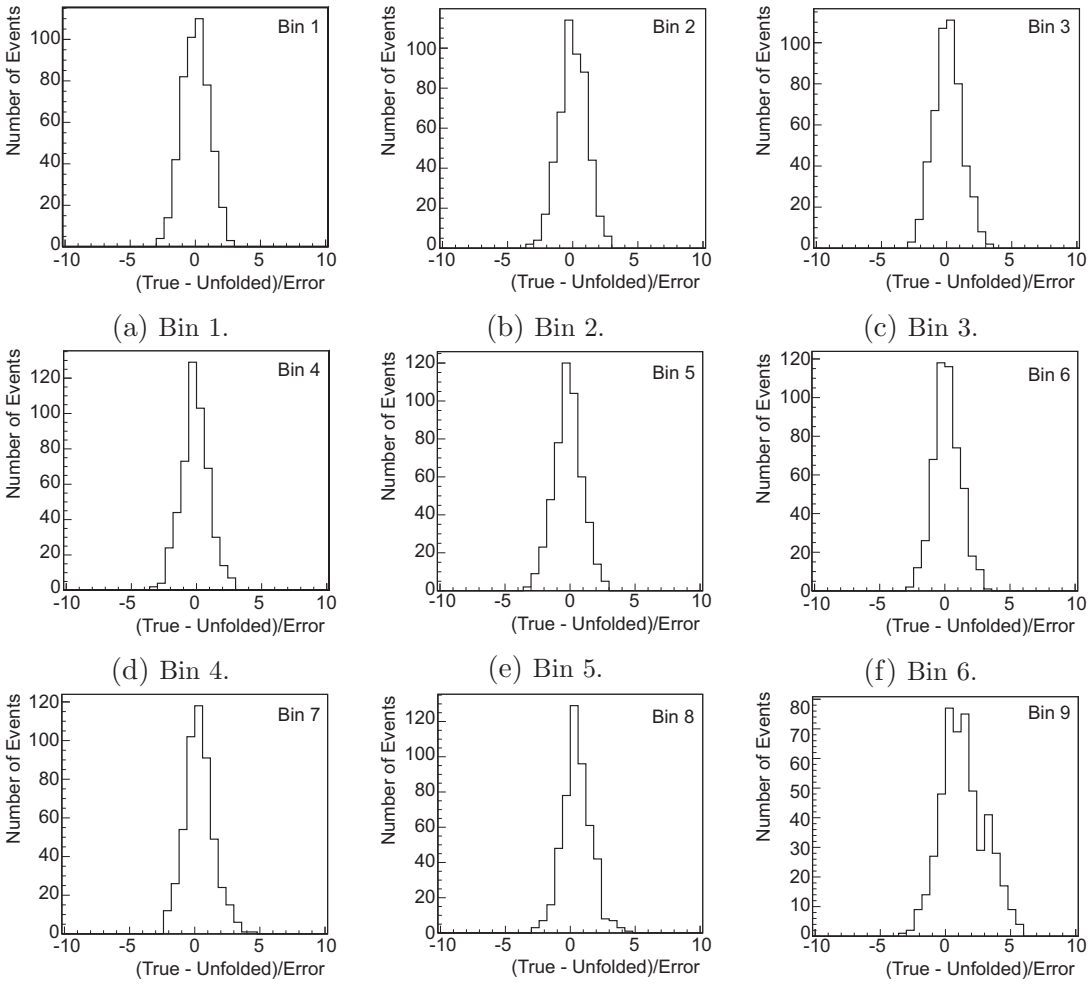


(a) Bin 1. (b) Bin 2. (c) Bin 3.

(d) Bin 4. (e) Bin 5. (f) Bin 6.

(g) Bin 7. (h) Bin 8. (i) Bin 9.

Figure B.5: Pull mode distributions for bins 1 to 9, using LDirC, NDirC and dEdX as input variables. The number of degrees of freedom was set to $ndf = 5$.



(a) Bin 1. (b) Bin 2. (c) Bin 3.

(d) Bin 4. (e) Bin 5. (f) Bin 6.

(g) Bin 7. (h) Bin 8. (i) Bin 9.

Figure B.6: Pull mode distributions for bins 1 to 9 using LDirC, NDir and dEdX as input for the unfolding. The number of degrees of freedom was set to $ndf = 6$.

Appendix C

Summary of Systematic Uncertainties

Bin	mean (ndf=5) [σ]	RMS (ndf=5)	mean (ndf=6) [σ]	RMS (ndf=6)
1	0.03	1.01	0.03	1.04
2	0.02	1.06	0.01	1.06
3	0.19	1.05	0.12	1.08
4	0.15	1.07	0.11	1.05
5	0.22	1.03	0.16	1.06
6	0.16	1.08	0.17	1.03
7	0.45	1.13	0.33	1.10
8	0.85	1.33	0.52	1.10
9	2.85	3.02	1.36	1.69
10	5.26	8.51	2.74	4.89

Table C.1: Binwise summary of the pull mode results obtained using LDirC, NDirC and dEdX as input variables for the unfolding.

Bin	mean (ndf=5) [σ]	RMS (ndf=5)	mean (ndf=6) [σ]	RMS (ndf=6)
1	0.01	1.00	0.01	1.02
2	0.01	1.04	0.10	1.06
3	0.13	1.08	0.02	1.03
4	0.03	1.09	0.08	1.07
5	0.09	1.12	0.06	1.08
6	0.03	1.16	0.05	1.07
7	0.34	1.14	0.04	1.08
8	0.93	1.36	0.44	1.10
9	1.90	2.60	0.89	1.57
10	5.56	9.55	3.15	5.65

Table C.2: Binwise summary of the pull mode results obtained using LDirC, NDirC and NCh as input variables for the unfolding.

Bin	mean (ndf=5) [σ]	RMS (ndf=5)	mean (ndf=6) [σ]	RMS (ndf=6)
1	0.08	1.03	0.05	1.04
2	0.14	1.06	0.13	1.11
3	0.28	1.11	0.17	1.04
4	0.03	1.05	0.05	1.08
5	0.28	1.10	0.16	1.10
6	0.12	1.03	0.07	1.08
7	0.20	1.13	0.11	1.15
8	0.69	1.28	0.33	1.17
9	2.36	2.76	1.04	1.61
10	4.91	7.92	2.52	4.49

Table C.3: Binwise summary of the pull mode results obtained using LDirC and NDirC as input variables for the unfolding.

Bin	dEdX	NCh	NDirC
1	4020.77 ± 128.72	3830.19 ± 95.51	3998.24 ± 133.16
2	8348.71 ± 233.56	8425.01 ± 193.26	8475.10 ± 244.96
3	6925.63 ± 234.96	6888.01 ± 220.44	7009.54 ± 245.48
4	4528.70 ± 215.74	4370.43 ± 190.19	4241.30 ± 229.37
5	2438.19 ± 127.58	2475.81 ± 110.52	2387.26 ± 131.32
6	796.17 ± 72.28	904.09 ± 60.64	898.68 ± 76.10
7	139.62 ± 28.18	198.03 ± 25.88	195.18 ± 29.75
8	17.40 ± 6.32	31.37 ± 5.99	30.60 ± 6.57
9	1.56 ± 0.82	4.00 ± 0.94	3.24 ± 0.84
10	0.15 ± 0.10	0.33 ± 0.09	0.33 ± 0.10

Table C.4: Binwise summary of the unfolding results obtained on real data using $ndf = 5$. The results for LDirC, NDirC and dEdX are shown in the column labelled dEdX, whereas the results obtained using LDirC, NDirC and NCh are depicted in the column labelled NCh. The column labelled NDirC represents the results obtained using LDirC and NDirC only.

Bin	dEdX	NCh	NDirC
1	4057.08 ± 164.07	3944.88 ± 122.98	3991.56 ± 164.78
2	8412.93 ± 333.88	8214.53 ± 267.62	8521.99 ± 347.02
3	6920.64 ± 403.39	7133.18 ± 324.14	7282.83 ± 437.22
4	4167.92 ± 331.04	4178.14 ± 272.44	3765.77 ± 356.48
5	2748.21 ± 215.21	2558.75 ± 169.58	2535.63 ± 231.80
6	862.69 ± 89.00	888.74 ± 74.03	956.59 ± 95.97
7	90.51 ± 43.50	191.50 ± 36.86	182.33 ± 48.72
8	-7.90 ± 14.17	32.67 ± 10.15	25.31 ± 12.75
9	-2.20 ± 2.49	4.38 ± 1.74	2.52 ± 1.78
10	-0.22 ± 0.24	0.37 ± 0.18	0.25 ± 0.21

Table C.5: Unfolding results obtained on real data using $ndf = 6$. The results for LDirC, NDirC and dEdX are shown in the column labelled dEdX, whereas the results obtained using LDirC, NDirC and NCh are depicted in the column labelled NCh. The column labelled NDirC represents the results obtained using LDirC and NDirC only.

Bin	dEdX [σ]	dEdX [%]	NDirC [σ]	NDirC [%]	NCh [σ]	NCh [%]	QTot [%]
1	19.04	35.7	18.96	35.6	13.97	26.1	33.3
2	0.86	2.0	0.57	1.3	11.48	26.8	23.7
3	17.95	48.6	17.34	47.0	10.25	27.8	16.1
4	1.43	5.9	1.39	5.4	8.69	34.0	37.8
5	10.08	50.4	10.11	50.3	3.20	15.9	65.2
6	3.66	32.3	3.86	34.1	2.09	18.4	24.2
7	1.37	23.1	1.12	18.9	2.18	36.9	122
8	3.23	69.0	3.06	65.4	2.39	51.0	153
9	5.68	110.0	5.29	102.5	3.41	66.6	90
10	8.87	83.1	8.91	85.1	7.58	72.4	1

Table C.6: Binwise summary of the ice model dependency of the various sets of input parameters. The derived errorbars are shown in units of the statistical uncertainty σ as well as in percent. The results obtained using LDirC, NDirC and dEdX are summarised in the column labelled dEdX, whereas the uncertainties derived using LDirC, NDirC and NCh are depicted in the column labelled NCh. The errorbars estimated using LDirC and NDirC only are shown in the column labelled NDirC. The uncertainties derived for the unfolding settings used in [89] are shown in the column labelled QTot for comparison. The number of knots was set to $nknots = 17$, wheras the number of degrees of freedom was set to $ndf = 5$.

Bin	A + 25% [σ]	A + 25% [%]	A -25% [σ]	A - 25% [%]
1	0.09	0.2	0.17	0.4
2	0.04	0.1	0.11	0.2
3	0.13	0.4	0.05	0.1
4	0.07	0.3	0.13	0.5
5	0.18	0.9	0.02	0.0
6	0.18	1.6	0.23	2.1
7	0.13	2.3	0.03	0.5
8	0.71	16.7	0.52	12.4
9	1.68	38.4	1.67	37.3
10	5.11	62.8	4.85	59.6

Table C.7: Binwise summary of the uncertainties derived for a 25% increased and decreased normalisation of the atmospheric ν_μ flux.

Bin	$E^{-(\gamma-0.03)}$ [σ]	$E^{-(\gamma-0.03)}$ [%]	$E^{-(\gamma+0.03)}$ [σ]	$E^{-(\gamma+0.03)}$ [%]
1	0.18	0.3	0.46	1.1
2	0.10	0.2	0.43	0.9
3	0.04	0.1	0.16	0.4
4	0.04	0.2	0.34	1.2
5	0.01	0.1	0.19	0.9
6	0.07	0.7	0.40	3.3
7	0.07	1.5	0.08	1.3
8	0.73	19.8	0.94	19.3
9	1.11	30.6	2.00	38.2
10	3.83	65.9	4.68	57.5

Table C.8: Binwise summary of the errorbars derived due to uncertainties in the spectral index γ, which was varied by ± 0.03. The errorbars are depicted in units of the statistical uncertainty σ as well as in percent.

Bin	σ_{pp}+5% [σ]	σ_{pp}+5% [%]	σ_{pp}-5% [σ]	σ_{pp}-5% [%]
1	21.16	61.7	19.45	56.5
2	10.44	23.7	15.71	34.8
3	3.81	9.4	3.29	8.3
4	6.39	21.9	9.38	32.2
5	3.36	15.2	7.92	34.8
6	2.50	19.5	0.89	6.9
7	2.49	37.5	2.76	38.8
8	0.93	16.5	3.15	64.1
9	0.33	5.4	1.99	38.8
10	2.74	33.6	2.95	28.1

Table C.9: Binwise summary of the errorbars derived due to changes in the pair production cross section σ_{pp}. The obtained uncertainties are depicted in units of the statistical error σ as well as in percent.

Bin	$log(E/\text{GeV})$	$E^2 \times \Phi$ [GeV cm^{-2} sr^{-1} s^{-1}]	Rel. Error [%]
0	2.25	2.42×10^{-4}	67.0
1	2.62	1.03×10^{-4}	35.9
2	3.01	3.02×10^{-5}	29.5
3	3.39	9.63×10^{-6}	40.7
4	3.78	3.55×10^{-6}	35.6
5	4.17	1.12×10^{-6}	42.3
6	4.56	2.80×10^{-7}	54.3
7	4.96	5.77×10^{-8}	62.4
8	5.36	1.49×10^{-8}	89.1
9	5.76	2.91×10^{-9}	119.6

Table C.10: Binwise summary of the acceptance corrected unfolding result, which corresponds to the differential flux of atmospheric neutrinos, scaled by E^2. The relative uncertainties, containing systematic as well as statistical errors is depicted as well.

Appendix D

D-SEA

Inverse problems are described by the Fredholm integral of first kind [87]:

$$g(y) = \int_a^b A(E, y) f(E) dE \tag{D.1}$$

Since one is interested in a binned solution of $f(E)$, equation (D.1) can be rewritten as a matrix equation:

$$\Rightarrow \vec{g}(x) = A\vec{f}(E) \tag{D.2}$$

$$\vec{g} = (g_1, \ldots, g_n) \tag{D.3}$$

$$\vec{f} = (f_1, \ldots, f_m) \tag{D.4}$$

$$g_i = \sum_{j=1}^{m} A_{ij} f_j \tag{D.5}$$

The normalisations of $\vec{g}$ and $\vec{f}$ yield:

$$\sum_{i=1}^{n} g_i = \sum_{i=1}^{n} \sum_{j=1}^{m} A_{ij} f_j = N_{Events} \tag{D.6}$$

$$\sum_{j=1}^{m} f_j = N_{Events} \tag{D.7}$$

The output of any machine learning algorithm returning confidence values can be interpreted as a conditional probability. According to Bayes theorem this can be written as:

$$P(E_j|\vec{x}) = \frac{p(\vec{x}|E_j)P(E_j)}{p(\vec{x})}. \tag{D.8}$$

Comment: Introducing equation (D.8) at this stage is not required for the proof itself. It is, however, useful to keep in mind, that the confidence values returned by a Naive Bayes Classifier, are computed using this very equation.
One should also keep in mind that a confidence c_{jk} depicts the probability of the

kth event to have an energy corresponding to bin j. Summation over all j is of course normalised:

$$\sum_{j=1}^{m} c_{jk} = 1 \Rightarrow \sum_{k=1}^{N_{Events}} \sum_{j=1}^{m} c_{jk} = N_{Events} \tag{D.9}$$

Identification of (D.9) and (D.6) yields:

$$\sum_{i=1}^{n} \sum_{j=1}^{m} A_{ij} f_j = \sum_{k=1}^{N_{Events}} \sum_{j=1}^{m} c_{jk} \tag{D.10}$$

Rewriting the sums:

$$\sum_{j=1}^{m} \sum_{i=1}^{n} A_{ij} f_j = \sum_{j=1}^{m} \sum_{k=1}^{N_{Events}} c_{jk} \tag{D.11}$$

$$\Leftrightarrow \sum_{j=1}^{m} \sum_{i=1}^{n} A_{ij} f_j - \sum_{j=1}^{m} \sum_{i=k}^{N_{Events}} c_{jk} = 0 \tag{D.12}$$

$$\Leftrightarrow \sum_{j=1}^{m} \left(\sum_{i=1}^{n} A_{ij} f_j - \sum_{i=k}^{N_{Events}} c_{jk} \right) = 0 \Leftrightarrow \sum_{j=1}^{m} \Delta_j = 0 \tag{D.13}$$

Analysing the individual Δ_j:

$$\Delta_j = \sum_{i=1}^{n} A_{ij} f_j - \sum_{i=k}^{N_{Events}} c_{jk} \Leftrightarrow \sum_{i=k}^{N_{Events}} c_{jk} = \sum_{i=1}^{n} A_{ij} f_j - \Delta_j \tag{D.14}$$

Since the summation in $\sum_{i=1}^{n} A_{ij} f_j$ is only in i, f_j can be treated as a constant factor:

$$\sum_{i=k}^{N_{Events}} c_{jk} = f_j \sum_{i=1}^{n} A_{ij} - \Delta_j \tag{D.15}$$

Examining the case $f_j \sum_{i=1}^{n} A_{ij} \gg |\Delta_j|$ yields:

$$\sum_{i=k}^{N_{Events}} c_{jk} \approx f_j \sum_{i=1}^{n} A_{ij} \Leftrightarrow f_j \approx \frac{\sum_{i=k}^{N_{Events}} c_{jk}}{\sum_{i=1}^{n} A_{ij}} \tag{D.16}$$

Comment: at this stage a special case is studied. This might seem random for the moment, but will deliver a criterion in order to determine, whether the classifier is well trained. The same result can be obtained starting out from the assumption of a well trained classifier, which will be defined in equation (D.27).
For better readability $\sum_{i=1}^{n} A_{ij} = \gamma_j$. This yields:

$$f_j = \frac{1}{\gamma_j} \sum_{i=k}^{N_{Events}} c_{jk} \tag{D.17}$$

This result can be interpreted as follows: The content of bin j can be estimated as the sum over all confidence values for this very bin, multiplied by a normalisation factor. This normalisation factor can be determined by studying the matrix elements A_{ij}:

$$A_{ij} = P_{ij} B_j \tag{D.18}$$

P_{ij} is the probability of an event of type j, producing a signature of type i. B_j is the detector acceptance for events of type j. Furthermore:

$$\sum_{i=1}^{n} P_{ij} = 1 \tag{D.19}$$

$$B_j \leq 1 \tag{D.20}$$

$$\Rightarrow \gamma_j = \sum_{i=1}^{n} A_{ij} = B_j \sum_{i=1}^{n} P_{ij} = B_j \tag{D.21}$$

This yields: f_j equals the sum of all confidences multiplied by an acceptance correction factor.
In order to derive a quality criterion for the machine learning method $f_j \sum_{i=1}^{n} A_{ij} \gg |\Delta_j|$ is examined:

$$\sum_{k=1}^{N_{Events}} c_{jk} = f_j \gamma_j - \Delta_j \tag{D.22}$$

$$\Leftrightarrow \Delta_j = f_j \gamma_j \sum_{k=1}^{N_{Events}} c_{jk} \tag{D.23}$$

$$\tag{D.24}$$

$f_j \sum_{i=1}^{n} \gamma_j = f_{j,MC}$ is known from Monte Carlo simulations:

$$|\Delta_j| = f_{j,MC} - \sum_{k=1}^{N_{Events}} c_{jk} \tag{D.25}$$

$$\Leftrightarrow \frac{|\Delta_j|}{f_{j,MC}} = \frac{f_{j,MC} - \sum_{k=1}^{N_{Events}} c_{jk}}{f_{j,MC}} \tag{D.26}$$

$f_{j,MC} \gg |\Delta_j|$ yields:

$$\Rightarrow 1 \gg \frac{f_{j,MC} - \sum_{k=1}^{N_{Events}} c_{jk}}{f_{j,MC}} \tag{D.27}$$

Thus, for every bin the relative difference between the reconstructed distribution and the true distribution needs to be small.

LIST OF FIGURES

LIST OF TABLES

BIBLIOGRAPHY

[1] C. L. Cowan, Jr., F. Reines, F. B. Harrison, H. W. Kruse, and A. D. McGuire. Detection of the Free Neutrino: a Confirmation. *Science*, 124:103–104, July 1956. doi: 10.1126/science.124.3212.103.

[2] F. Reines and C. L. Cowan. The Neutrino. *Nature*, 178:446–449, September 1956. doi: 10.1038/178446a0.

[3] R. Davis. A Review of the Homestake Solar Neutrino Experiment. *Progress in Particle and Nuclear Physics*, 32:13–32, 1994.

[4] D. Vignaud. The GALLEX solar neutrino experiment. *Nuclear Physics B-Proceedings Supplements*, 60(3):20–29, 1998.

[5] J.N. Abdurashitov, V.N. Gavrin, S.V. Girin, et al. The Russian-American Gallium Experiment (SAGE) Cr Neutrino Source Measurement. *Physical Review Letters*, 77(23):4708–4711, 1996.

[6] Y. Suzuki. Kamiokande solar neutrino results. *Nuclear Physics B-Proceedings Supplements*, 38(1):54–59, 1995.

[7] Y. Fukuda, T. Hayakawa, E. Ichihara, et al. Evidence for Oscillation of Atmospheric Neutrinos. *Physical Review Letters*, 81:1562–1567, August 1998. doi: 10.1103/PhysRevLett.81.1562.

[8] S. Fukuda, Y. Fukuda, M. Ishitsuka, et al. Constraints on Neutrino Oscillations Using 1258 Days of Super-Kamiokande Solar Neutrino Data. *Physical Review Letters*, 86:5656–5660, June 2001. doi: 10.1103/PhysRevLett.86.5656.

[9] M.B. Smy. Neutrino Experiments with Super-Kamiokande. In *The Launching of La Belle Epoque of High Energy Physics and Cosmology*, volume 1, pages 265–272, 2004.

[10] J. Bonn, B. Bornschein, L. Bornschein, et al. Results from the Mainz Neutrino Mass Experiment. *Progress in Particle and Nuclear Physics*, 48(1): 133–139, 2002.

[11] V.M. Lobashev. Study of the Tritium Beta-spectrum in Experiment Troitsk ν-mass. *Progress in Particle and Nuclear Physics*, 48(1):123–131, 2002.

[12] Ch. Weinheimer. Katrin, a Next Generation Tritium β Decay Experiment in Search for the Absolute Neutrino Mass Scale. *Progress in Particle and Nuclear Physics*, 48(1):141–150, 2002.

[13] L. Wolfenstein. Neutrino mass in particle physics and astrophysics. *Contemporary Physics*, 37(3):175–182, 1996.

[14] R.M. Bionta, G. Blewitt, C.B. Bratton, et al. Observation of a Neutrino Burst in Coincidence with Supernova 1987A in the Large Magellanic Cloud. *Physical Review Letters*, 58(14):1494–1496, 1987.

[15] K. Hirata, T. Kajita, M. Koshiba, et al. Observation of a Neutrino Burst from the Supernova SN1987A. *Physical Review Letters*, 58(14):1490–1493, 1987.

[16] T. K. Gaisser. *Cosmic Rays and Particle Physics*. Cambridge University Press, 1991.

[17] T. Stanev. *High Energy Cosmic Rays (Springer Praxis Books / Astronomy and Planetary Sciences)*. Springer, 2004. ISBN 3540406530.

[18] K. Zuber. *Neutrino Physics (Series in High Energy Physics, Cosmology and Gravitation)*. Taylor & Francis, 2003. ISBN 0750307501.

[19] N. Schmitz. *Neutrinophysik (Teubner Studienbücher Physik) (German Edition)*. Vieweg+Teubner Verlag, 1997. ISBN 3519032368.

[20] K. Daum, W. Rhode, P. Bareyre, et al. Determination of the atmospheric neutrino spectra with the Fréjus detector. *Zeitschrift fur Physik C Particles and Fields*, 66:417–428, September 1995. doi: 10.1007/BF01556368.

[21] R. Abbasi, Y. Abdou, T. Abu-Zayyad, et al. The energy spectrum of atmospheric neutrinos between 2 and 200 TeV with the AMANDA-II detector. *Astroparticle Physics*, 34:48–58, August 2010. doi: 10.1016/j.astropartphys.2010.05.001.

[22] R. Abbasi, Y. Abdou, T. Abu-Zayyad, et al. Measurement of the atmospheric neutrino energy spectrum from 100 GeV to 400 TeV with IceCube. *Physical Review D*, 83(1):012001, January 2011. doi: 10.1103/PhysRevD.83.012001.

[23] A. Fedynitch, J. Becker Tjus, and P. Desiati. Influence of hadronic interaction models and the cosmic ray spectrum on the high energy atmospheric muon and neutrino flux. *Physical Review D*, 86(11):114024, 2012.

[24] J. G. Learned and K. Mannheim. High-Energy Neutrino Astrophysics. *Annual Review of Nuclear and Particle Science*, 50:679–749, 2000. doi: 10.1146/annurev.nucl.50.1.679.

[25] S. Klepser D. Mazin. V. Blobel W. Rhode N. Milke, M. Doert. Solving inverse problems with the unfolding program truee: Examples in astroparticle physics. *Nuclear Instruments and Methods in Physics Research A*, 697, 2013.

[26] C.H.Q. Ding and H. Peng. Minimum Redundancy Feature Selection from Microarray Gene Expression Data. *J. of Bioinformatics and Computational Biology*, 3(2), 2005.

[27] L. Breiman. Random Forests. *Machine Learning*, 45, 2001.

[28] W. Wagner. *Design and Realisation of a new AMANDA Data Acquisition System with Transient Waveform Recorders*. PhD thesis, Technische Universität Dortmund, Dortmund, Germany, 2004.

[29] V.F. Hess. *Beobachtungen der durchdringenden Strahlung bei sieben Freiballonfahrten*. Kaiserlich-Königlichen Hof-und Staatsdruckerei, in Kommission bei Alfred Hölder, 1912.

[30] A. A. Watson and Pierre Auger Collaboration. Recent results from the Pierre Auger Observatory - Including comparisons with data from AGASA and HiRes. *Nuclear Instruments and Methods in Physics Research A*, 588: 221–226, April 2008. doi: 10.1016/j.nima.2008.01.083.

[31] J. Behringer et al. (Particle Data Group). *Phys. Rev. D*, 86, 2012.

[32] M. Boezio, V. Bonvicini, E. Mocchiutti, et al. The Space Experiment PAMELA. *Nuclear Physics B Proceedings Supplements*, 134:39–46, September 2004. doi: 10.1016/j.nuclphysbps.2004.08.006.

[33] R. Battiston and AMS-02 Collaboration. The Anti Matter Spectrometer (AMS-02): a Particle Physics Detector In Space. *Nuclear Physics B Proceedings Supplements*, 166:19–29, April 2007. doi: 10.1016/j.nuclphysbps.2006.12.091.

[34] W. B. Atwood, A. A. Abdo, M. Ackermann, et al. The Large Area Telescope on the Fermi Gamma-Ray Space Telescope Mission. *The Astrophysical Journal*, 697:1071–1102, June 2009. doi: 10.1088/0004-637X/697/2/1071.

[35] C. Amsler. *Kern- und Teilchenphysik (Uni-Taschenbücher M)*. ISBN 3825228851.

[36] J. Aleksić et al. Performance of the MAGIC stereo system obtained with Crab Nebula data. *Astroparticle Physics*, 35:435–448, February 2012. doi: 10.1016/j.astropartphys.2011.11.007.

[37] J. A. Hinton. The status of the HESS project. *New Astronomy Reviews*, 48: 331–337, April 2004. doi: 10.1016/j.newar.2003.12.004.

[38] H. Anderhub et al. FACT - The first Cherenkov telescope using a G-APD camera for TeV gamma-ray astronomy. *Nuclear Instruments and Methods in Physics Research A*, 639:58–61, May 2011. doi: 10.1016/j.nima.2010.10.081.

[39] M. S. Longair. *High Energy Astrophysics.* Cambridge University Press, 1981. ISBN 0521235138.

[40] A. A. Penzias and R. W. Wilson. A Measurement of Excess Antenna Temperature at 4080 Mc/s. *The Astrophysical Journal*, 142:419–421, July 1965. doi: 10.1086/148307.

[41] E. Komatsu, K. M. Smith, J. Dunkley, et al. Seven-year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Cosmological Interpretation. *The Astrophysical Journal Supplement Series*, 192:18, February 2011. doi: 10.1088/0067-0049/192/2/18.

[42] B. Falkenburg and W. Rhode, editors. *From Ultra Rays to Astroparticles: A Historical Introduction to Astroparticle Physics.* Springer, 2012. ISBN 9400754213.

[43] I. A. Belolaptikov, L. B. Bezrukov, B. A. Borisovets, et al. The Baikal underwater neutrino telescope: Design, performance, and first results. *Astroparticle Physics*, 7:263–282, August 1997. doi: 10.1016/S0927-6505(97)00022-4.

[44] M. Ageron, J. A. Aguilar, I. Al Samarai, et al. ANTARES: The first undersea neutrino telescope. *Nuclear Instruments and Methods in Physics Research A*, 656:11–38, November 2011. doi: 10.1016/j.nima.2011.06.103.

[45] E. Andres, P. Askebjer, S. W. Barwick, et al. The AMANDA neutrino telescope: principle of operation and first results. *Astroparticle Physics*, 13: 1–20, March 2000. doi: 10.1016/S0927-6505(99)00092-4.

[46] IceCube Collaboration, A. Achterberg, M. Ackermann, et al. First year performance of the IceCube neutrino telescope. *Astroparticle Physics*, 26: 155–173, October 2006. doi: 10.1016/j.astropartphys.2006.06.007.

[47] T. DeYoung. Neutrino Astronomy with IceCube. *Modern Physics Letters A*, 24:1543–1557, 2009. doi: 10.1142/S0217732309031417.

[48] E. Resconi and the IceCube Collaboration. Status and prospects of the IceCube neutrino telescope. *Nuclear Instruments and Methods in Physics Research A*, 602:7–13, April 2009. doi: 10.1016/j.nima.2008.12.013.

[49] C. Wiebusch for the IceCube Collaboration. Physics Capabilities of the IceCube DeepCore Detector. *ArXiv e-prints*, July 2009.

[50] A. Tamburro. Measurements of Cosmic Rays with IceTop/IceCube: Status and Results. *Modern Physics Letters A*, 27:1230038, December 2012. doi: 10.1142/S0217732312300388.

[51] IceCube Collaboration, R. Abbasi, Y. Abdou, et al. IceTop: The surface component of IceCube. *Nuclear Instruments and Methods in Physics Research A*, 700:188–220, February 2013. doi: 10.1016/j.nima.2012.10.067.

[52] J. Ahrens, X. Bai, R. Bay, et al. Muon track reconstruction and data selection techniques in AMANDA. *Nuclear Instruments and Methods in Physics Research A*, 524:169–194, May 2004. doi: 10.1016/j.nima.2004.01.065.

[53] D. J. Boersma, L. Gladstone, and A. Karle for the IceCube Collaboration. Moon Shadow Observation by IceCube. *ArXiv e-prints*, February 2010.

[54] L. Gladstone for the IceCube Collaboration. The Shadow of the Moon in IceCube. *ArXiv e-prints*, November 2011.

[55] Icecube Collaboration, R. Abbasi, M. Ackermann, et al. The IceCube data acquisition system: Signal capture, digitization, and timestamping. *Nuclear Instruments and Methods in Physics Research A*, 601:294–316, April 2009. doi: 10.1016/j.nima.2009.01.001.

[56] Y. Totsuka. Results from Super-Kamiokande. *Nuclear Physics A*, 663:218–218, January 2000. doi: 10.1016/S0375-9474(99)00593-X.

[57] M. Honda, T. Kajita, K. Kasahara, S. Midorikawa, and T. Sanuki. Calculation of atmospheric neutrino flux using the interaction model calibrated with atmospheric muon data. *Physical Review D*, 75(4):043006, feb 2007. doi: 10.1103/PhysRevD.75.043006.

[58] M. Honda, T. Kajita, K. Kasahara, and S. Midorikawa. Calculation of the Flux of Atmospheric Neutrinos. *Physical Review D*, 52:4985–5005, nov 1995. doi: 10.1103/PhysRevD.52.4985.

[59] G. D. Barr, T. K. Gaisser, P. Lipari, S. Robbins, and T. Stanev. Three-dimensional calculation of atmospheric neutrinos. *Physical Review D*, 70(2): 023006, July 2004. doi: 10.1103/PhysRevD.70.023006.

[60] D. Chirkin and W. Rhode. Propagating leptons through matter with Muon Monte Carlo (MMC). *ArXiv High Energy Physics - Phenomenology e-prints*, July 2004.

[61] J.H. Koehne, K. Frantzen, M. Schmitz, et al. PROPOSAL: A tool for propagation of charged leptons. *Computer Physics Communications*, 2013.

[62] Ch. Wiebusch. *The Detection of Faint Light in Deep Underwater Neutrino Telescopes*. PhD thesis, DESY Zeuthen, 1995.

[63] H. Johansson. *Searching for an Ultra-High-Energy Diffuse Flux of Extraterrestrial Neutrinos with IceCube 40*. PhD thesis, Stockholm University, Stockholm, Sweden, 2011.

[64] Y. Freund and R. E. Schapire. Experiments with a New Boosting Algorithm. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, pages 148–156. MORGAN KAUFMANN PUBLISHERS, INC., 1996.

[65] J. H. Friedman. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Statist*, 29(5):1189–1232, 2001.

[66] R. Duda. *Pattern Classification*. Wiley, New York, 2001. ISBN 0471056693.

[67] Th. Cover and P. Hart. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27, 1967.

[68] J. M Keller, M. R. Gray, and J. A Givens. A Fuzzy K-Nearest Neighbor Algorithm. *Systems, Man and Cybernetics, IEEE Transactions on*, (4):580–585, 1985.

[69] I. Guyon and A. Elisseeff. An Introduction to Variable and Feature Selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[70] B. Schowe. Feature Selection for high-dimensional data in RapidMiner. In S. Fischer and I. Mierswa, editors, *Proceedings of the 2nd RapidMiner Community Meeting And Conference (RCOMM 2011)*, Aachen, 2011. Shaker Verlag.

[71] J.-N. Capdevielle, P. Grieder, J. Knapp, et al. The Karlsruhe extensive air shower simulation code CORSIKA. *The Karlsruhe extensive air shower simulation code CORSIKA., by Capdevielle, JN; Grieder, P.; Knapp, J.; Gabriel, P.; Gils, HJ; Heck, D.; Mayer, HJ; Oehlschläger, J.; Rebel, H.; Schatz, G.; Thouw, T.. Kernforschungszentrum Karlsruhe GmbH (Germany). Inst. für Kernphysik, Nov 1992, 60 p.,*, 1, 1992.

[72] Alex Olivas et al. Icecube simulation documentation. http://wiki.icecube.wisc.edu/index.php/Simulation_Documentation_Wiki.

[73] S. Fischer, R. Klinkenberg, I. Mierswa, and O. Ritthoff. YALE: Yet Another Learning Environment – Tutorial. Technical Report CI-136/02, Collaborative Research Center 531, University of Dortmund, Dortmund, Germany, June 2002. URL `http://www-ai.cs.uni-dortmund.de/DOKUMENTE/fischer_etal_2002a.ps.gz`. ISSN 1433-3325. http://yale.sf.net/ .

[74] I. Mierswa. Rapidminer website. http://rapid-i.com/.

[75] K. Morik T. Ruhe and W. Rhode. *Application of RapidMiner in Neutrino Astronomy.* to be published.

[76] M. Wornowizki. Variablenselektion über Filterverfahren auf Daten des Neutrinodetektors IceCube, 2011.

[77] I. Witten and F. Eibe. *Data Mining: Praktische Werkzeuge und Techniken für das maschinelle Lernen.* Hanser, München Wien, 2001. ISBN 3446215336.

[78] L.I. Kuncheva. A Stability Index for Feature Selection. In *Proceedings of the 25th IASTED International Multi-Conference*, 2007.

[79] M. Ackermann, J. Ahrens, X. Bai, et al. Optical properties of deep glacial ice at the South Pole. *Journal of Geophysical Research (Atmospheres)*, 111: D13203, July 2006. doi: 10.1029/2005JD006687.

[80] W. Huelsnitz. *Search For Quantum Gravity With IceCube And High Energy Atmospheric Neutrinos.* PhD thesis, University of Maryland, College Park, MD, USA, 2010.

[81] R. K. Bock, A. Chilingarian, M. Gaug, et al. Methods for multidimensional event classification: a case study using images from a Cherenkov gamma-ray telescope. *Nuclear Instruments and Methods in Physics Research A*, 516: 511–528, January 2004. doi: 10.1016/j.nima.2003.08.157.

[82] G. Holmes B. Pfahringer P. Reutemann I. H. Witten M. Hall, F. Eibe. The WEKA Data Mining Software: An Update.

[83] F. Brosowski. Random Forest Stabilitätstests von IceCube-22 Monte Carlo Simulationen.

[84] D. Chirkin. Study of South Pole ice transparency with IceCube flashers. Technical report.

[85] IceCube Collaboration, M.G. Aartsen, R. Abbasi, et al. Measurement of South Pole ice transparency with the IceCube LED calibration system. *arXiv preprint arXiv:1301.5361*, 2013.

[86] J. A. Aguilar for the IceCube Collaboration. Time-independent searches for astrophysical neutrino sources with the combined data of 40 and 59 strings of IceCube. In *Proceedings of the 32nd International Cosmic Ray Conference.*

[87] V. Blobel and E. Lohrmann. *Statistische und numerische Methoden der Datenanalyse.* Teubner Verlag, 1998.

[88] J. Albert et al. Unfolding of differential energy spectra in the MAGIC experiment. *Nuclear Instruments and Methods in Physics Research A*, 583: 494–506, December 2007. doi: 10.1016/j.nima.2007.09.048.

[89] N. Milke. *Unfolding of the atmospheric neutrino spectrum with the new program TRUEE and IceCube*. PhD thesis, Technische Universität Dortmund, Dortmund, Germany, 2012.

[90] V. Blobel. The RUN manual. *Regularized unfolding for high-energy physics experiments. Technical Note TN361, OPAL*, 1996.

[91] M. Doert. *The talkative AGN next door*. PhD thesis, Technische Universität Dortmund, Dortmund, Germany, 2013.

[92] G. Cowan. A Survey of Unfolding Methods for Particle Physics. In *Prepared for Conference on Advanced Statistical Techniques in Particle Physics, Durham, England*, pages 18–22, 2002.

[93] A. Groß. *Search for High Energy Neutrinos from Generic AGN classes with AMANDA-II*. PhD thesis, Universität Dortmund, Dortmund Germany, 2006.

[94] R. Enberg, M. H. Reno, and I. Sarcevic. Prompt neutrino fluxes from atmospheric charm. *Physical Review D*, 78(4):043005, aug 2008. doi: 10.1103/PhysRevD.78.043005.

[95] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction: with 200 full-color illustrations*. New York: Springer-Verlag, 2001.

[96] V. Dose. Die Bayes'sche Variante. *Physik Journal*, 8/9, 2004.

[97] G. D'Agostini. Improved iterative Bayesian unfolding. *arXiv preprint arXiv:1010.0632*, 2010.

[98] M. Feindt. A neural Bayesian Estimator for Conditional Probability Densities. *arXiv preprint physics/0402093*, 2004.

[99] N. Gagunashvili. Machine learning approach to inverse problem and unfolding procedures. *ArXiv e-prints*, April 2010.

[100] A. S. Weigend and A. N. Srivastava. Predicting conditional probability distributions: A connectionist approach. *International Journal of Neural Systems*, 6, 1995.

[101] IceCube/PINGU Collaboration, K. Clark, and D. F. Cowen. IceCube/DeepCore and IceCube/PINGU: Prospects for Few-GeV Scale ν Physics in the Ice. *Nuclear Physics B Proceedings Supplements*, 233:223–228, December 2012. doi: 10.1016/j.nuclphysbps.2012.12.081.

[102] M. Ribordy and A. Y. Smirnov. Improving the neutrino mass hierarchy identification with inelasticity measurement in PINGU and ORCA. *ArXiv e-prints*, March 2013.

AUTHOR'S PUBLICATIONS

Publications which are associated with this thesis and which are signed by the author as (one of) the main author(s):

i. Aartsen et al., Measurement of the atmospheric $\nu_\mu$flux with IceCube in the 59-string configuration, *in preparation*.

ii. T. Ruhe et al., A Data Mining Approach to Unfolding, *in preparation*.

iii. T. Ruhe, K. Morik and W. Rhode, Application of RapidMiner in Neutrino Astronomy in M. Hofmann and R. Klinkenberg (editors), *Use Cases for RapidMiner*, *accepted for publication*.

iv. T. Ruhe et al., D-SEA: A data mining approach to unfolding, *Proceedings of the 33rd International Cosmic Ray Conference*, *to be published*.

v. Aartsen et al., Measurement of the atmospheric $\nu_\mu$spectrum with IceCube 59, *Proceedings of the 33rd International Cosmic Ray Conference*, *to be published*.

vi. T. Ruhe for the IceCube collaboration and K. Morik, Data Mining IceCube, *Proceedings of the XXI ADASS*, (2012).

vii. T. Ruhe for the IceCube collaboration, K. Morik and B. Schowe, Data Mining on Ice, *Astrostatistics and Data Mining, Springer Series in Astrostatistics*, Volume 2 (2012).

I WOULD LIKE TO THANK

· · · Prof. Dr. Dr. Wolfgang Rhode for giving me the opportunity to work on such an interesting topic and for offering his help whenever needed. I am further grateful for being given the opportunity to participate in various conferences, meetings and schools.

· · · Prof. Dr. Katharina Morik for her help and the numerous invaluable discussions on data mining and machine learning.

· · · Prof. Dr. Claus Gössling for his willingness and his efforts involved in being the second assessor of this thesis.

· · · The entire Astroparticle Physics Group at TU Dortmund in the current and former cast, while special thanks are dedicated to Fabian Clevermann for being great officemate, to Dominik Neise for being willing to discuss Everything again and again until solved and to Jan-Hendrik Köhne for his numerous programming advices and for jumping in on a discussion on soccer whenever physics became too frustrating.

· · · The entire IceCube Diffuse/Atmospheric Working Group for their help and encouragement provided over the years. Special thanks are dedicated to the reviewers Lars Mohrmann, Kotoyo Hoshina, Tyce DeYoung and Henrike Wissing, for their efforts in reviewing the analysis and for their patience and insightful questions, which greatly contributed to the overall quality of the analysis.

· · · The inner core of Dsea, namely Martin Schmitz, Tobias Vogt and Max Wornowitzki for valuable discussions and for their efforts in transforming Dsea from an idea into a scientific project.

· · · Everyone who helped in the proofreading of this thesis: Florian Scheriau, Martin Schmitz, Sabrina Einecke, Jan-Hendrik Köhne, Fabian Clevermann.

· · · Caroline, who proofread the thesis as a native speaker and who provided comments of immense value on language and grammar.

· · · the entire Disc Golf Crew for taking my mind off physics over putts, drives and a handful of beers.

· · · my family who offered me the possibility of studying physics and who always supported me during my time as a PhD student.

· · · Paul for numerous visits to soccer matches. Zusammen stehen, zusammen feiern: Bochum und die Bayern!

· · · and finally Birte, who endured my absence during meetings, schools and conferences. Who never complained about endless IceCube phone conferences or the final stages of thesis insomnia and who miraculously prevented me from going totally bananas. Thank you!