PROCEEDINGS OF SCIENCE

Cosmic ray composition study using machine learning at the IceCube Neutrino Observatory

The IceCube Collaboration*
http://icecube.wisc.edu/collaboration/authors/icrc19_icecube
E-mail: matthias.plum@marquette.edu

The evaluation of mass composition of cosmic rays in the knee region ($\sim$ 3 PeV) is critical to understanding the transition in the origin of cosmic rays from galactic to extragalactic sources. The IceCube Neutrino Observatory at the South Pole is a multi-component detector consisting of the surface IceTop array and the deep in-ice IceCube detector. By applying modern machine-learning techniques to cosmic-ray air showers reconstructed coincidentally in both detector components of IceCube observatory, the energy and the mass of primary cosmic rays in this transition region can be measured. In this contribution, we will discuss the reconstruction performance and composition sensitivity of IceCube observables presently under development.

Corresponding authors: Matthias Plum[†1]
[1] *Department of Physics, Marquette University, Milwaukee, WI, 53201, USA*

36th International Cosmic Ray Conference -ICRC2019-
July 24th - August 1st, 2019
Madison, WI, U.S.A.

*For collaboration list, see PoS(ICRC2019) 1177.
†Speaker.

PoS(ICRC2019)394

1. Importance of Cosmic Ray Composition

High-energy cosmic rays consist mostly of charged particles. Their composition is of key importance to understand their origin and acceleration processes. Additionally, the supposed transition from galactic to extragalactic sources of cosmic rays in the energy range from PeV to EeV should be visible in the evolution of the composition with energy. In air-shower physics, secondary particles are detected and it is presently only possible to measure the composition on a statistical basis.

2. IceCube and IceTop Detectors

The IceCube Neutrino Observatory [1] is a multi-purpose astroparticle detector located at the geographic South Pole. The detector is composed of the deep in-ice IceCube detector and the surface IceTop array. IceCube consists of 86 detector strings with 60 digital optical modules distributed between 1450 m and 2450 m beneath the surface of the ice. The detector strings are arranged in a triangular grid with ∼125 m separation, as shown in Figure 1.

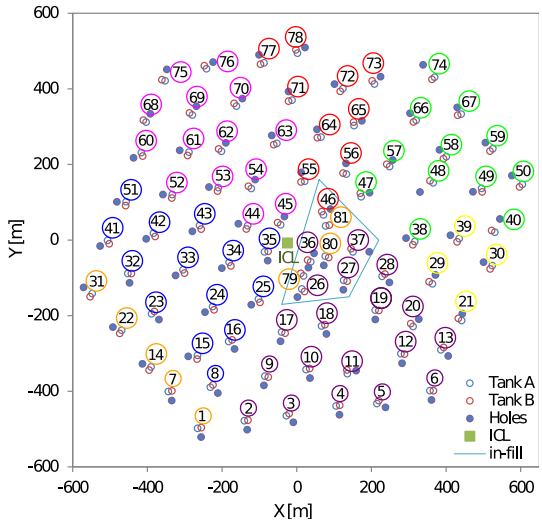

Figure 1: A top view of the IceTop surface array [2]. The color code correspond to deployment period of the IceCube string and the IceTop tanks.

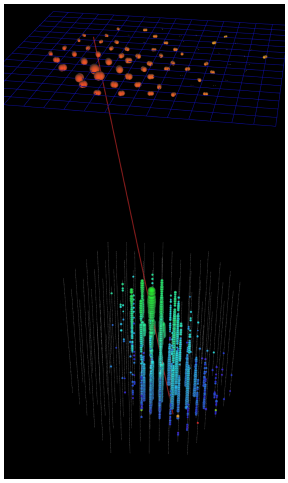

Figure 2: Example coincidence event in IceCube and IceTop. The color represents the timing information from early (red) to late (blue). The size of the circles corresponds to the size of the measured signal in the PMTs. The red line is the reconstructed shower axis.

In total, IceCube covers an effective detector volume of 1 km^3. On top of most IceCube strings, two IceTop [2] tanks are placed, separated by roughly 10 m. Due to the construction of the detector over several years and the snow drift at the South Pole, the tanks are unevenly covered with snow, which is taken into account in the reconstruction and simulation of cosmic-ray events. The deposited charge and the timing information of the cosmic-ray events are used to reconstruct the air-shower geometry and the lateral distribution of the charges at the surface and deep in-ice. This study includes only simulated cosmic-ray events which were reconstructed by the combination of both detector components in coincidence. An example coincidence event is shown in Figure 2. Due to the geometric constraints of the IceCube and IceTop coincidence, only vertical ($\theta \leq 30^\circ$)

events are used. The simulations use the CORSIKA [3] air-shower generator, with FLUKA [4] as the low-energy hadronic interaction model and SIBYLL-2.1 [5] as the high-energy interaction model.

3. Machine-Learning Reconstruction of the Primary Mass

Based on the successful application of machine-learning methods on 3 years of IceCube/IceTop coincidence data to reconstruct the primary energy and to derive a mass composition in [6, 7], a further improvement of the mass composition resolution is studied by testing several new reconstructed detector observables for composition sensitivity.

The *"baseline analysis"* of the simulation (similar to [6, 7]) is performed with the following observables: The logarithmic signal strength in IceTop at a distance of 125 m from the shower axis, S_{125}, and the cosine of the reconstructed zenith angle $\cos(\theta)$ derived from the standard reconstruction; from IceCube, the energy deposit of the high energy muon bundles at a slant depth of 1500 m $\log_{10}(dE/dX_{1500\mathrm{m}})$. $\log_{10}(dE/dX_{1500\mathrm{m}})$ shows a strong primary composition dependence, which is shown in Figure 3 for different primaries. An *"improved analysis"* is additionally performed to evaluate another composition sensitive variable, the shower age parameter β from the 'Double Logarithmic Parabola' fit [2] of the IceTop array, which also exhibits composition dependence over the whole energy range, shown in Figure 4.

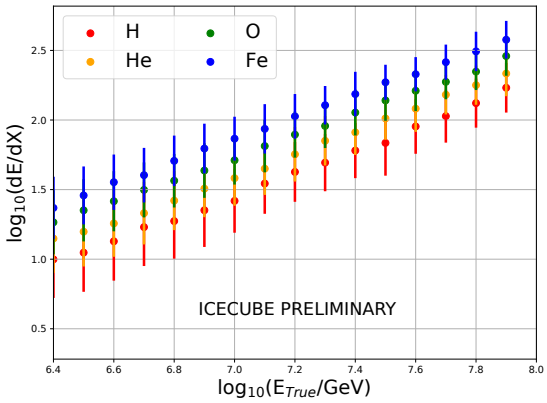

Figure 3: Mean and standard deviation of reconstructed $\log_{10}(dE/dX_{1500\mathrm{m}})$ in IceCube as a function of energy for different primaries.

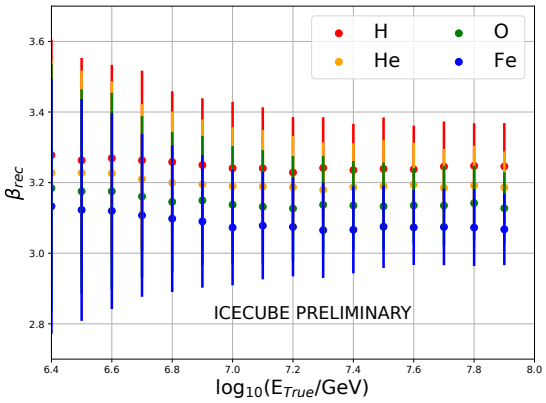

Figure 4: Mean and standard deviation of reconstructed β in IceTop as a function of energy for different primaries.

The primary cosmic-ray elements used are proton, helium, oxygen and iron. Those elements are equally spaced over the $\langle \ln A \rangle$ mass range and are therefore useful to train machine-learning regression algorithms. Due to the time-consuming full Monte Carlo production and detector reconstruction, roughly 2000 high quality full Monte Carlo simulation events per energy bin are used for the training and testing of a random forest tree (RFT) [8] using a 3-fold cross validation technique. Another independent $\sim$ 2000 full Monte Carlo simulation events per energy bin of $\log_{10}(E/\mathrm{GeV})$ are used for verification of the machine-learning output, and for the final mass composition analysis.

The machine-learning output of the verification sample is converted into kernel density estimation (KDE) [9] probability density functions (PDF), which are used as templates for every energy bin and individual elementary groups using the RooFit toolkit [10] as in [6, 7]. An example energy bin is shown in Figure 5. The template shape of the new improved (dashed) templates are distinc-

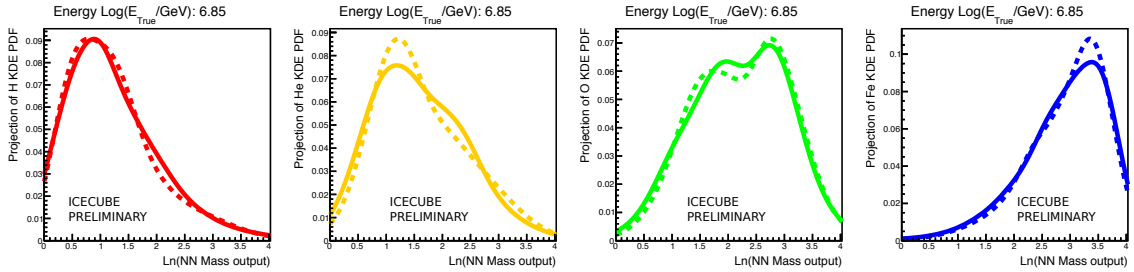

Figure 5: Example bin with the KDE [9] mass template PDFs generated with a RFT regressor of the baseline (solid) and improved (dashed) analysis in the energy bin $\log_{10}(E/\mathrm{GeV})$=6.8-6.9.

tively different from the templates of the baseline (solid) analysis. These template PDF's for each energy bin of the four elementary groups are combined to a joined mass composition response PDF by introducing a weight factor for each primary given by

$$P_{mass}(X) = \sum_{i=H,He,O,Fe} w_i \cdot P_i(X) \text{ with } \sum_{i=H,He,O,Fe} w_i = 1,$$

where X is the mass output of the machine learning. Due to the constrained weight factors w_i, the results for the elementary groups are highly correlated with each other. On a statistical basis, P_{mass} is used to fit the data to determine the contribution of each elementary primary group with an extended likelihood fit.

4. Reconstruction Capabilities of Different Composition Scenarios

The reconstruction and the improvement of the mass resolution of the template method is studied for various scenarios and both cases, with and without β, are compared to each other. The mass composition response PDF for each energy bin can be used to generate fast mock scenario data sets where the number of events and the fractions of each primary group are artificially selected. In this way the reconstruction capabilities of the method are tested for several Monte Carlo scenarios.

The first tested scenario is the maximum mixing scenario of the four elementary groups (Fractions: 25%:25%:25%:25%). A generated Monte Carlo data set with 3000 total events and the corresponding fit is shown in Figure 6. The input PDF is reconstructed within the statistical uncertainties. This procedure is repeated several thousand times and the fitted number of events per elementary group are collected and analyzed. An example energy bin of this Monte Carlo study is shown in Figure 7, where the distribution of the fit parameter of the baseline is slightly broader than for the improved analysis. The average reconstructed fractions and the mass resolutions for each elementary group are measured using a Gaussian fit to these distributions. This study is repeated for all other energy bins with the same parameters. The mass fraction in this scenario is on average accurately reconstructed as shown in Figure 8 for the whole energy range. Both the baseline and

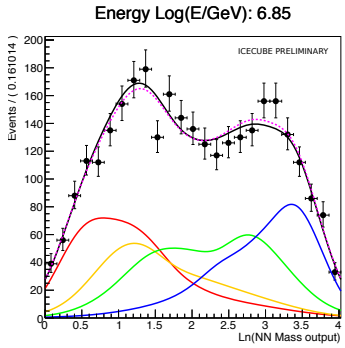

Figure 6: Example fast Monte Carlo data set from the improved analysis from the maximum mixing scenario of the four elementary groups (Fraction: 25%:25%:25%:25%) in the energy bin $\log_{10}(E/\mathrm{GeV})$=6.8-6.9 with in total 3000 MC events. The generator PDF is shown as a solid black line, the fitted PDF is shown as a dashed magenta line. The corresponding weighted elementary group PDF are shown in red for H, yellow for He, green for O and blue for Fe.

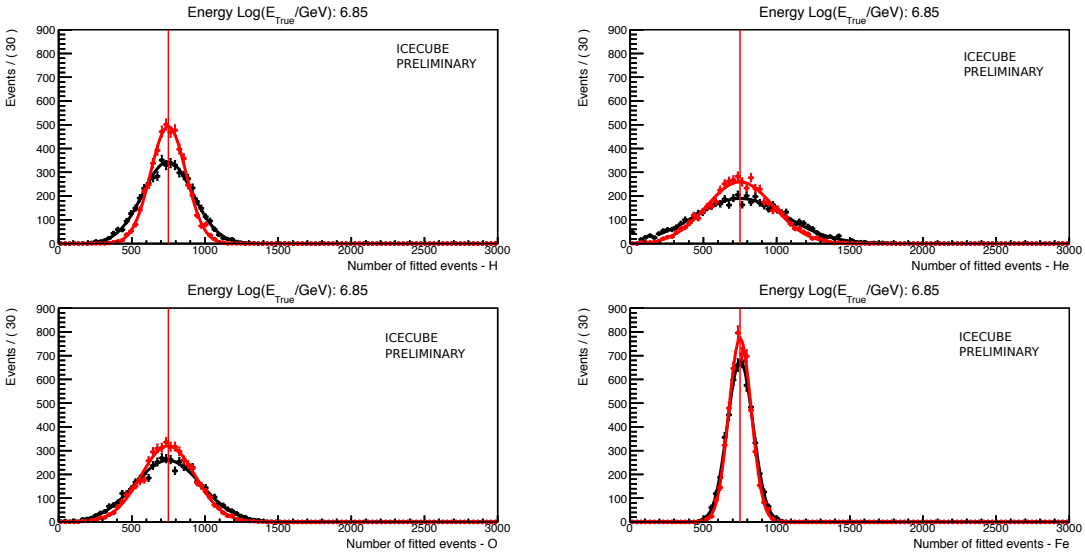

Figure 7: Example bin of the reconstructed number of events from the Monte Carlo reconstruction study for the maximum mixing scenario of the four elementary groups (Fractions: 25%:25%:25%:25%) in the energy bin $\log_{10}(E/\mathrm{GeV})$=6.8-6.9 with in total 3000 MC events per bin. The baseline analysis results are shown in black, the improved ones are shown in red. A Gaussian fit was applied to each distribution to measure the average reconstructed fraction and the resolution. The vertical line shows the Monte Carlo truth in this scenario.

the improved analysis reconstructed the true composition with high accuracy inside the statistical uncertainties. The comparison of the mass resolution from the baseline and the improved analysis is shown in Figure 9, which shows that the additional information of a new composition sensitive variable improves the mass resolution over the whole energy range. The fractions of the intermediate element groups of helium and oxygen are intrinsically uncertain due to the large overlap with their neighboring distributions and are showing a slightly larger improvement inside the statistical

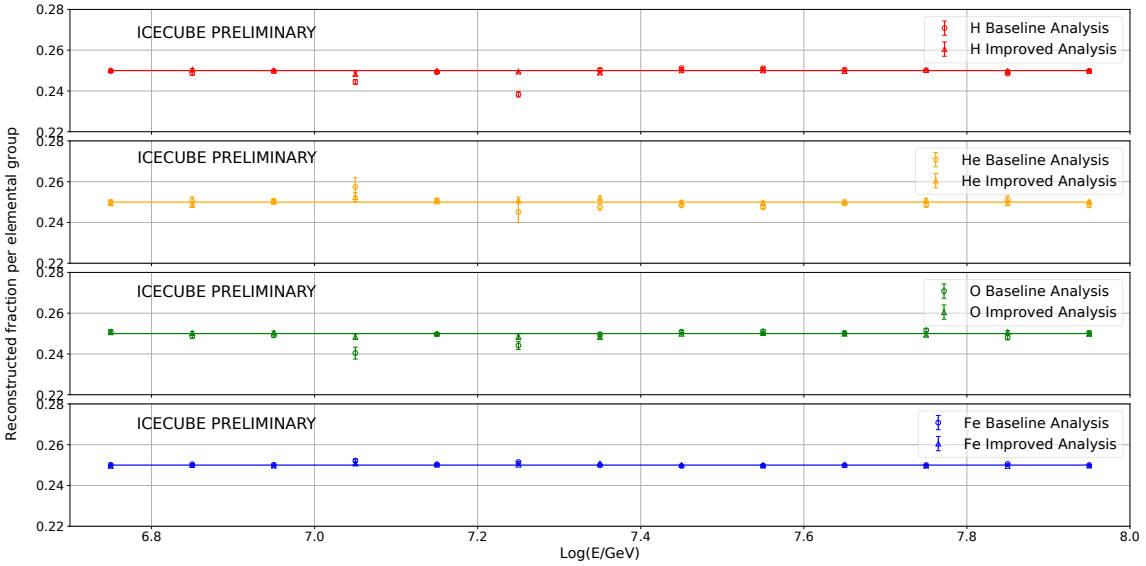

Figure 8: Comparison of average reconstructed mass fractions by the baseline and improved analysis with each bin containing 3000 events. Both analyses reconstruct on average the Monte Carlo truth, represented by the solid line. Note the zoom on the y-axis around 25% to emphasize the very minor change in results.

uncertainties of the fit method than the proton an iron groups.

A realistic cosmic-ray scenario assuming an H4a [11] composition is also tested in this energy range. The reconstructed average composition is shown in Figure 10. The plot shows that both analyses are capable of reconstructing the primary mass composition in a realistic source scenario.

5. Summary & Outlook

The mass composition template analysis of the IceCube/IceTop coincident events is verified with Monte Carlo scenarios. Full Monte Carlo simulation was used to train and test a random forest tree regressor and elementary group probability density templates for every energy bin were created. The templates were used to generate fast Monte Carlo data sets for both a maximum mixing and an H4a[11] input scenario. The mass resolution is measured by generating and fitting several thousand fast Monte Carlo data sets and analyzing the fit results. A comparison of the mass resolution between the baseline analysis used in [6, 7] and the future improved analysis using the shower age parameter β is presented and shows a slight improvement over the whole energy range.

In the future, an improved coincidence reconstruction [12] will add several new composition sensitivity variables like the air shower curvature and the muon density at the surface. The joint measurements of the proposed scintillator array [13] and IceTop will add information sensitive to composition by the inclusion of the shower parameter as shown in [14]. Also, the proposed IceAct [15] array will provide additional air shower information about the electromagnetic shower component like the center-of-gravity, which will further improve cosmic-ray measurements of the composition and also produces opportunities to investigate and constrain hadronic interaction models.

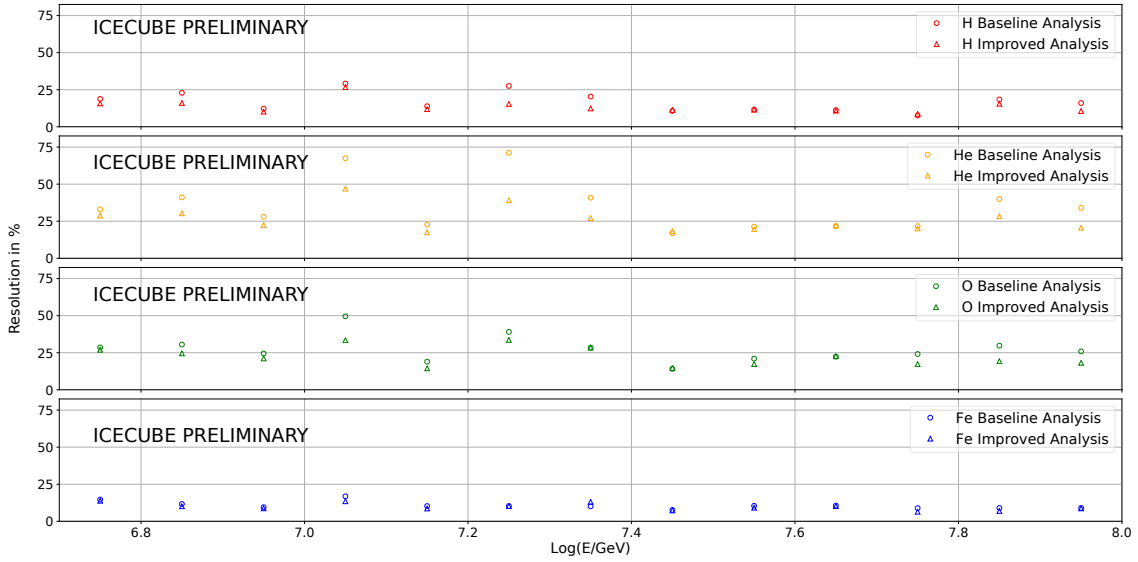

Figure 9: Comparison of mass composition resolution derived from a Monte Carlo study with 3000 MC events per bin. The improved analysis shows a slight improvement over the baseline analysis.

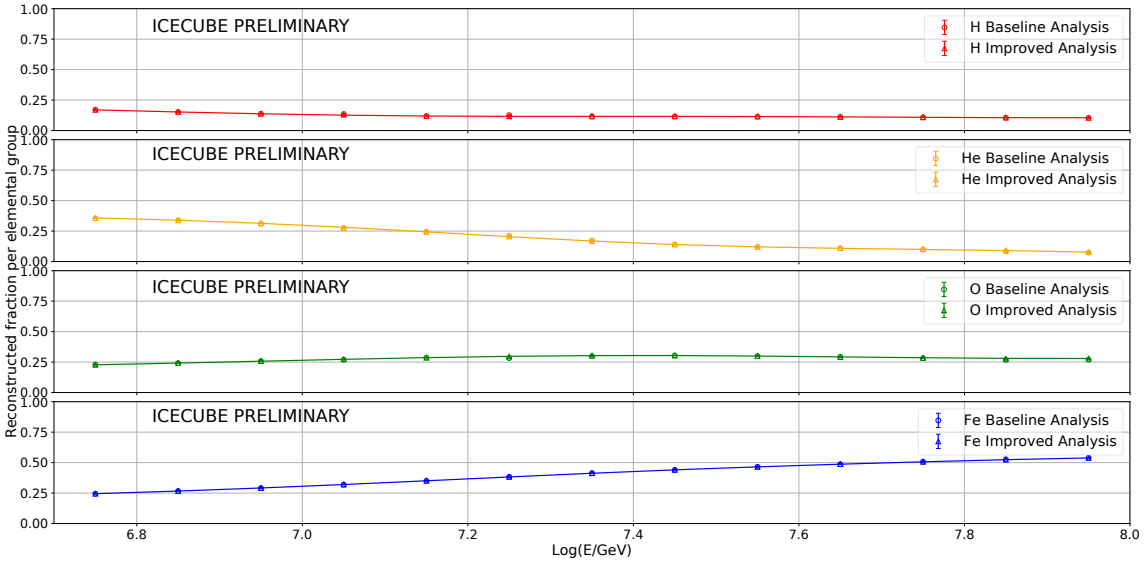

Figure 10: Comparison of mass reconstruction based on the H4a [11] model of the baseline and the improved analysis with 3000 MC events per bin.

References

[1] **IceCube** Collaboration, M. G. Aartsen et al., *JINST* **12** (2017) P03012.

[2] **IceCube** Collaboration, R. Abbasi et al., *Nucl. Instr. and Meth. A* **700** (2013) 188–220.

[3] D. Heck et al., *CORSIKA: A Monte Carlo code to simulate extensive air showers, Report FZKA 6019, Forschungszentrum Karlsruhe*, 1998.

[4] G. Battistoni et al., *AIP Conference Proceedings* **896** (2007) 31–49.

[5] E. Ahn, R. Engel, T. Gaisser, P. Lipari, and T. Stanev, *Physical Review D* **80** (2009) 94003.

[6] **IceCube** Collaboration, K. Andeen and M. Plum, PoS(ICRC2019)172 (these proceedings).

[7] **IceCube** Collaboration, arXiv:1906.04317.

[8] F. Pedregosa et al., *Journal of Machine Learning Research* **12** (2011) 2825–2830.

[9] K. Cranmer, *Computer Physics Communications* **136** (2001) 198 – 207.

[10] W. Verkerke and D. Kirkby, *arXiv e-prints* (2003) physics/0306116.

[11] T. K. Gaisser, *Astroparticle Physics* **35** (2012) 801.

[12] **IceCube** Collaboration, B. Xinhua and E. Dvorak, PoS(ICRC2019)244 (these proceedings).

[13] **IceCube** Collaboration, M. Kauer, PoS(ICRC2019)309 (these proceedings).

[14] **IceCube** Collaboration, A. Leszczyńska and M. Plum, PoS(ICRC2019)332 (these proceedings).

[15] **IceCube** Collaboration, M. Schaufel and K. Andeen, PoS(ICRC2019)179 (these proceedings).