A multivariate muon counter

B. Kégl, K. Louedec, R. Busa-Fekete, M. Urban

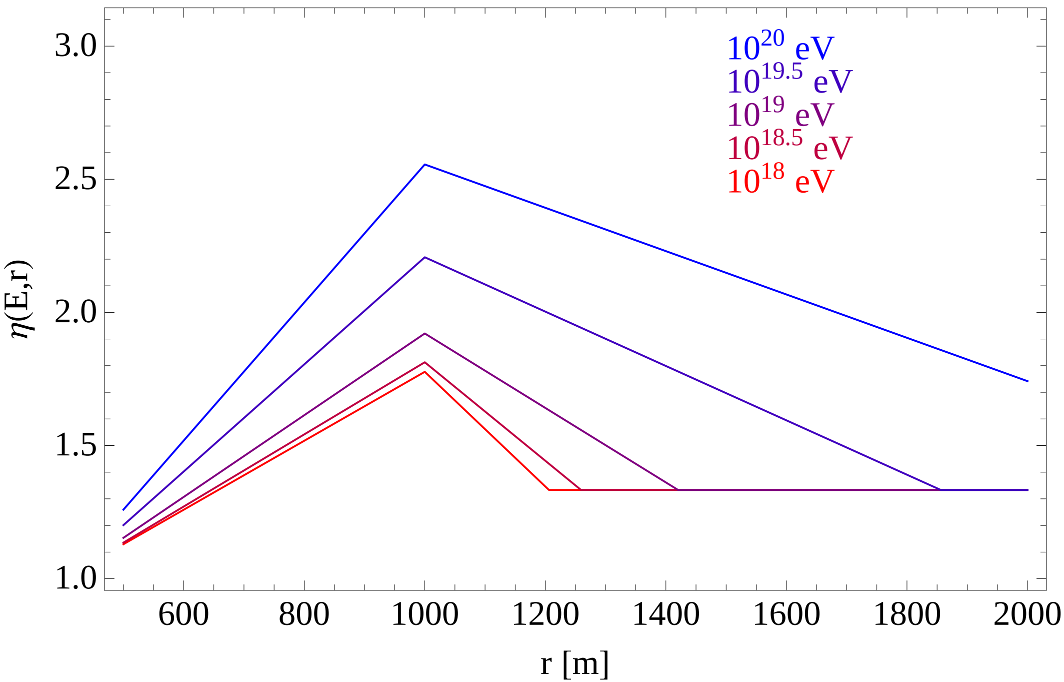
Linear Accelerator Laboratory

CNRS - University of Paris Sud

Auger France meeting, October 21, 2010

The jump method (X. Garrido)

$$\widehat{N}_\mu = \eta(E,r) \sum_{\text{FADC bin } i} \underbrace{(x_{i+1} - x_i)}_{\text{jump}} \mathbb{I}\{x_{i+1} - x_i > 0.5\}$$



The jump method (X. Garrido)

- Main problem: explicit energy dependece
  - If N_μ is undersimulated, then this can generate a bias through $\frac{\partial N_\mu}{\partial E}$ and $\frac{\partial \eta(E,r)}{\partial E}$
- Motivation to revisit the jump method
  - design a “tuning” that does not explicitly depend on E

The idea

- “Forget” the jumps

- Record a large set of statistics on the FADC signal
  - that might be sensitive to muon content

- Train a nonparametric predictor of the number of muons on a large set of simulations

The input parameters

- FADC signal observables
  - total thresholded signal: $\sum_i x_i \mathbb{I}\{x_i > s\}, \ s = 0, 0.1, \ldots, 2.9, 3$
  - thresholded jump signal: $\sum_i (x_{i+1} - x_i) \mathbb{I}\{x_{i+1} - x_i > s\}, \ s = 0, 0.1, \ldots, 2.9, 3$
  - number of thresholded jumps: $\sum_i \mathbb{I}\{x_{i+1} - x_i > s\}, \ s = 0, 0.1, \ldots, 2.9, 3$
  - thresholded jump signal and number of thresholded jumps in the average of the two lowest PMs (getting rid of occasional direct light)
  - risetime $10\%, 20\%, \ldots, 90\%$
  - some PM asymmetry parameters

The input parameters

- Global observables
  - distance from shower axis r, zenith angle θ, but no energy!
  - slightly controversial: it might be possible to implicitly reconstruct the energy using these prameters $\longrightarrow$ bias
- The list of variables is open, ideas are welcome
- We will also prune variables based on what the nonparametric regressors are actually using

Nonparametric regression

- Take a data set $D = \left\{(\mathbf{x}_i, y_i)\right\}$ and "learn" a function

$$f : \mathbf{x} \to y$$

with low expected loss

$$R(f) = \mathbb{E}\left\{L\big(f(\mathbf{x}), y\big)\right\}$$

- The usual loss in regression is the quadratic loss

$$L(y, y') = (y - y')^2$$

Nonparametric regression

- We are shooting for the relative error

$$L(N,\widehat{N}) = \left|\frac{\widehat{N}}{N} - 1\right|$$

  - it overemphasizes small signals, so we use

$$L(N,\widehat{N}) = \left|\frac{\widehat{N}+C}{N+C} - 1\right|$$

  - $C = 1$ in plots, $C = 10$ in training
  - boosting can handle this cost through a calibration technique, but our neural network package needs a quadratic loss, here we use

$$L(N,\widehat{N}) = \big(\log(\widehat{N}+C) - \log(N+C)\big)^2$$

Nonparametric regression

- Neural network
  - standard matlab package
  - feature normalization, PCA
  - one sigmoidal hidden layer, linear output layer
  - 88 hidden neurons
  - resilient back-prop with learning rate $\eta = 0.3$
  - run until convergence

Nonparametric regression

- Boosting (no results yet)
  - in-house implementation (`multiboost.org`)
  - multiclass $\rightarrow$ regression through appropriate initial weighting and calibration
  - will also provide individual error bars

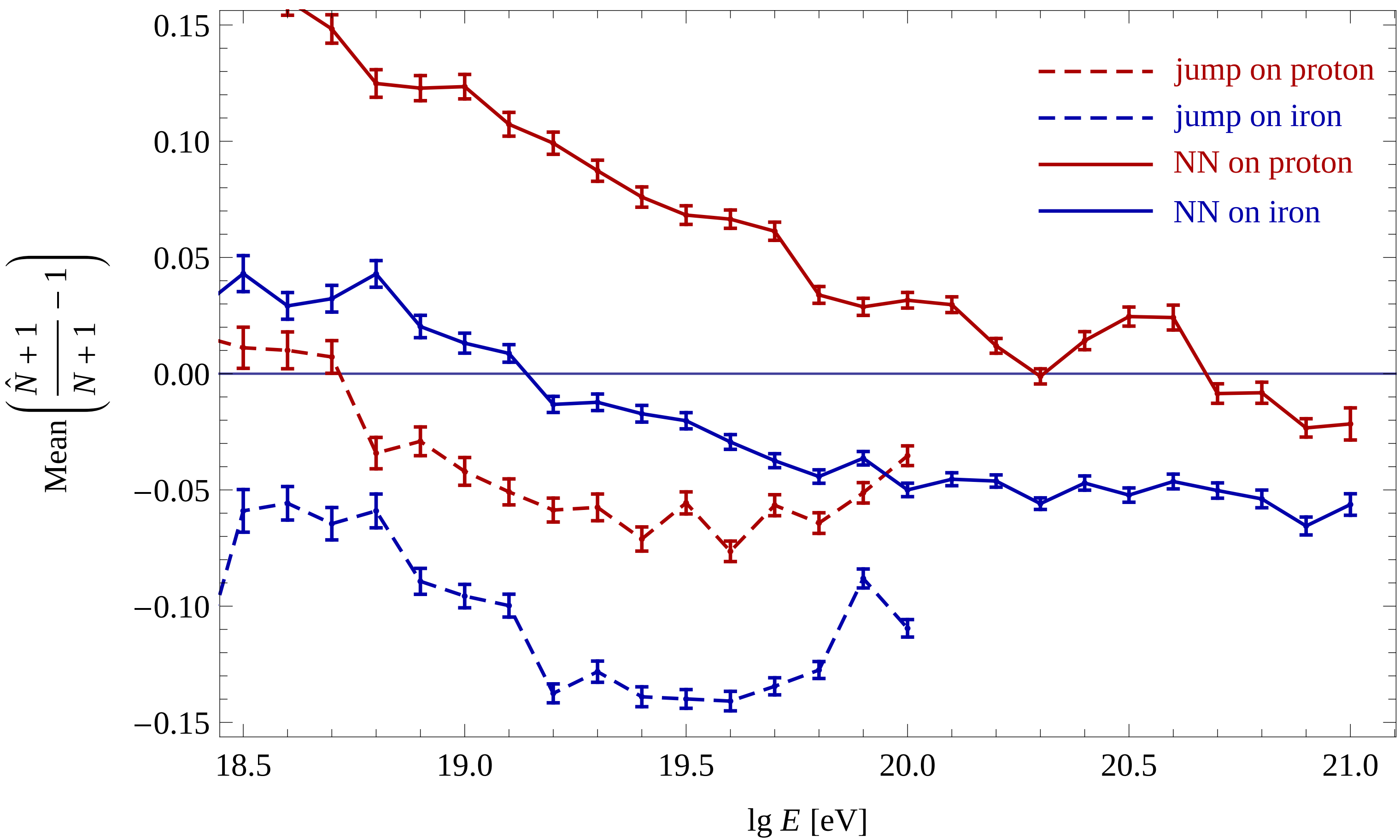
Data

- Official QGSJetII-03 iron and proton simulations at Lyon
  - "quality cut" due to a bug in the old OffLine version
  - cutting stations with $r < 400\,\mathrm{m}$

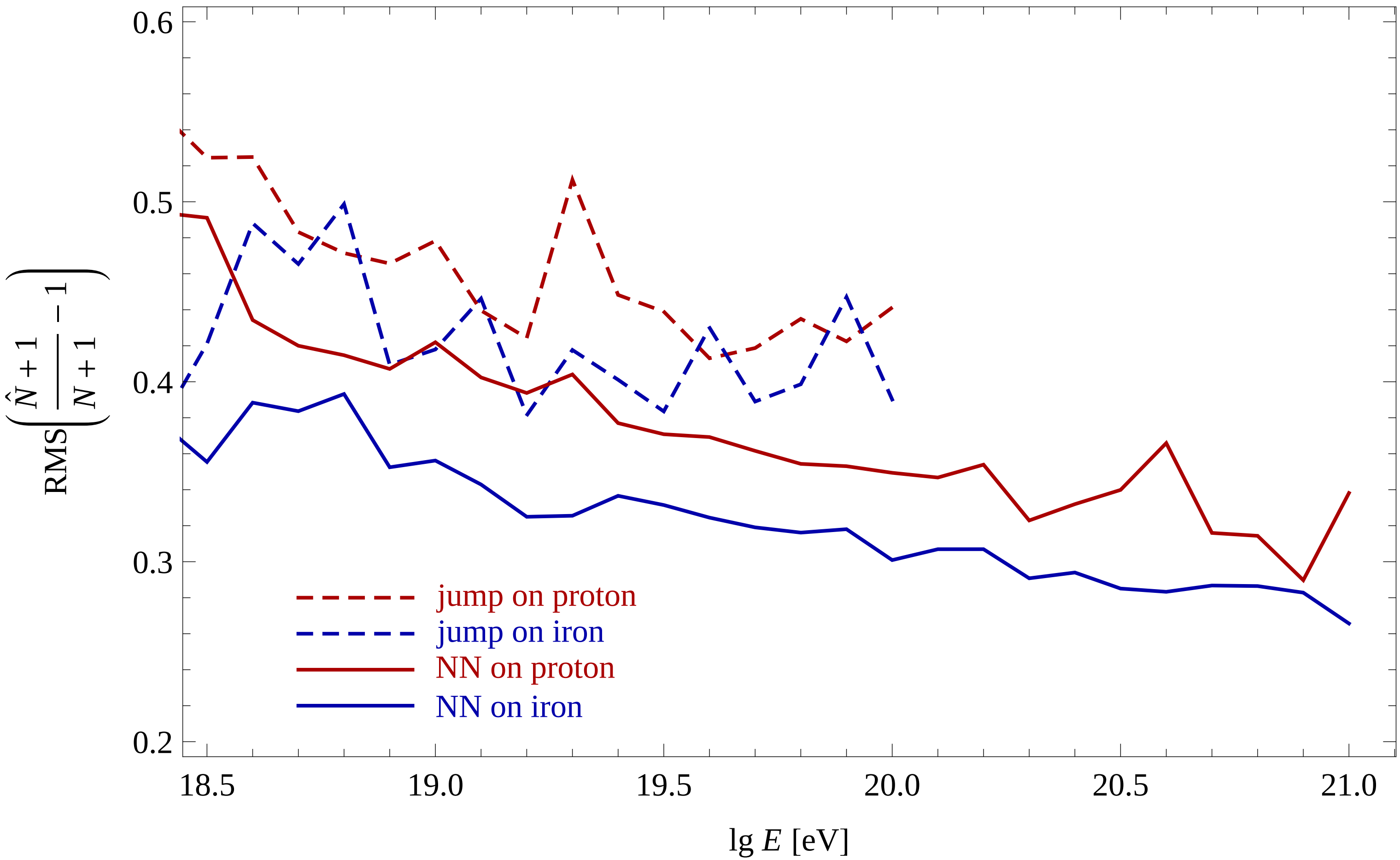
	Before cuts		After cuts	
	nb of showers	nb of tanks	nb of showers	nb of tanks
proton	26372	334466	20328	214401
iron	22374	352738	16766	233794

  - 15+15K tanks are for training, the rest is for testing
  - we could use more training data once the methods are fine tuned, but overfitting is not an issue at this point

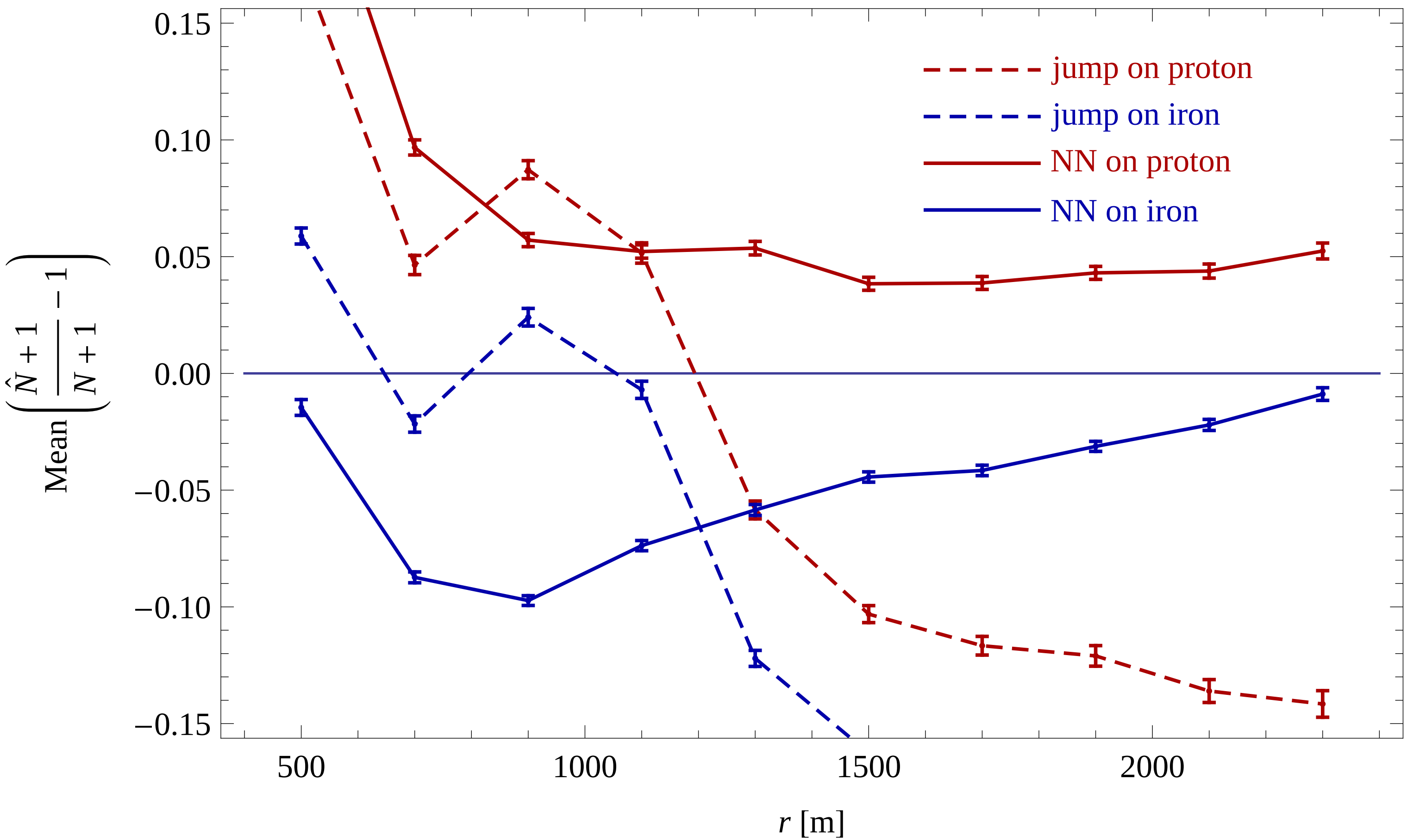
Bias vs. energy



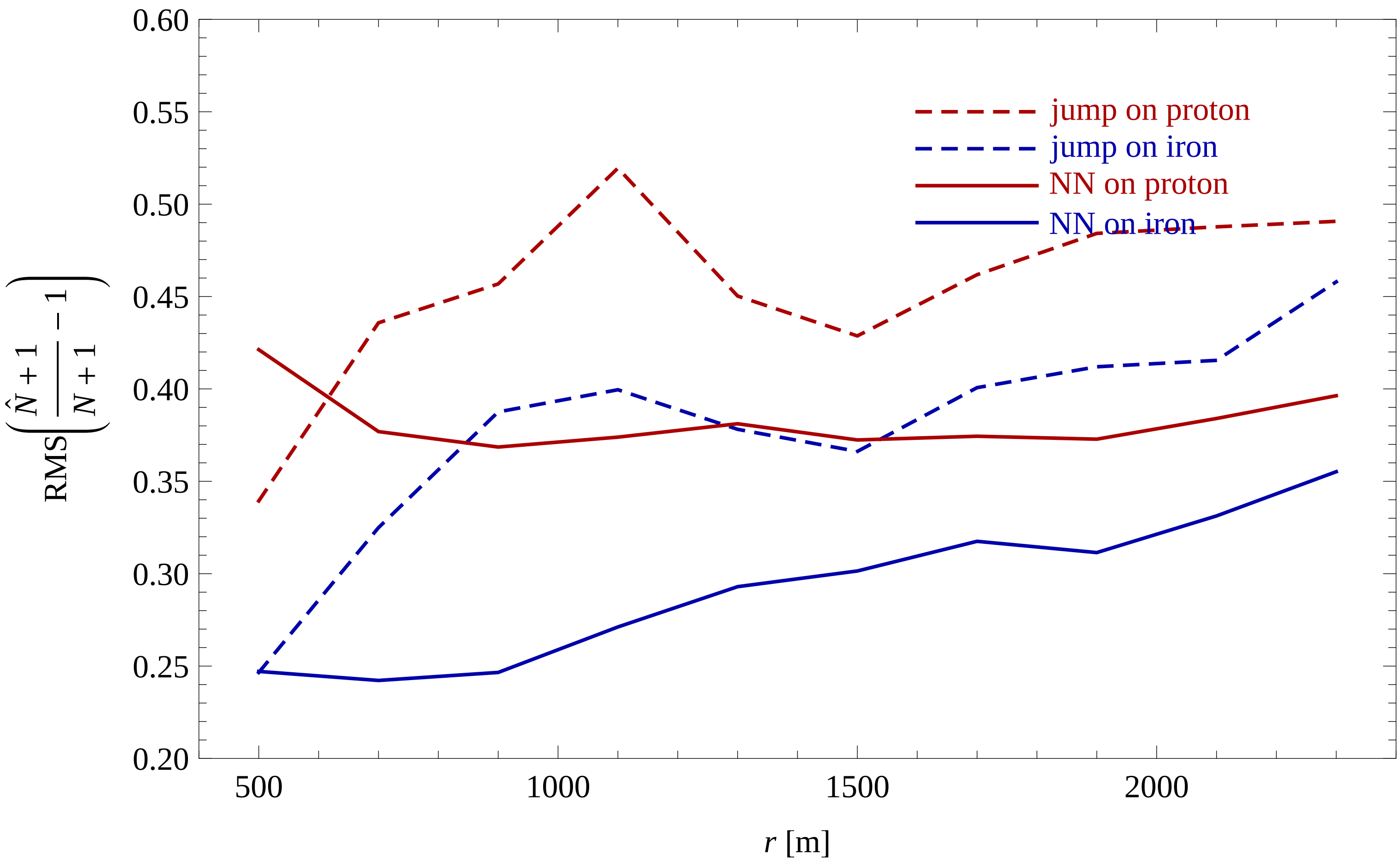
RMS vs. energy



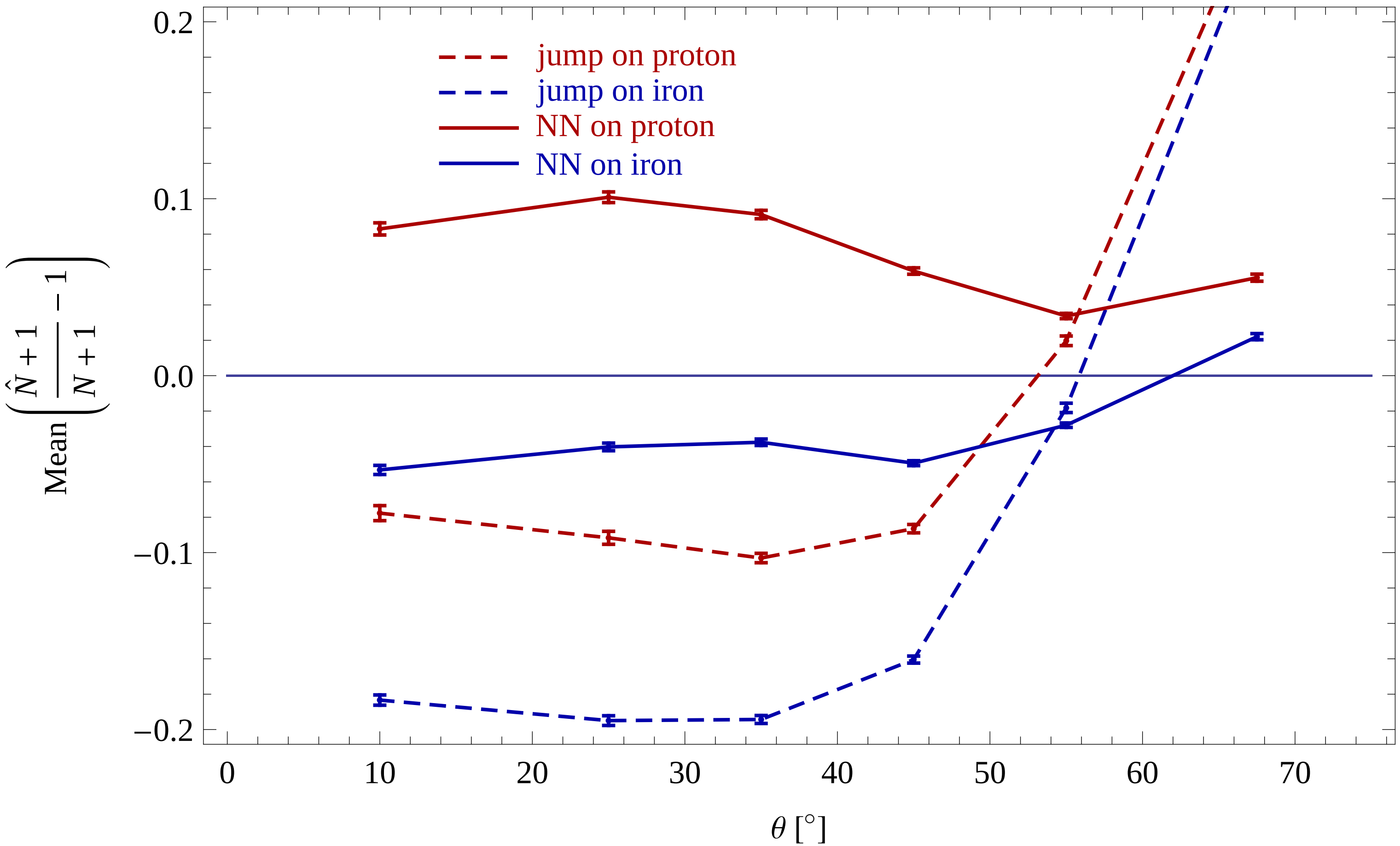
Bias vs. distance from shower axis



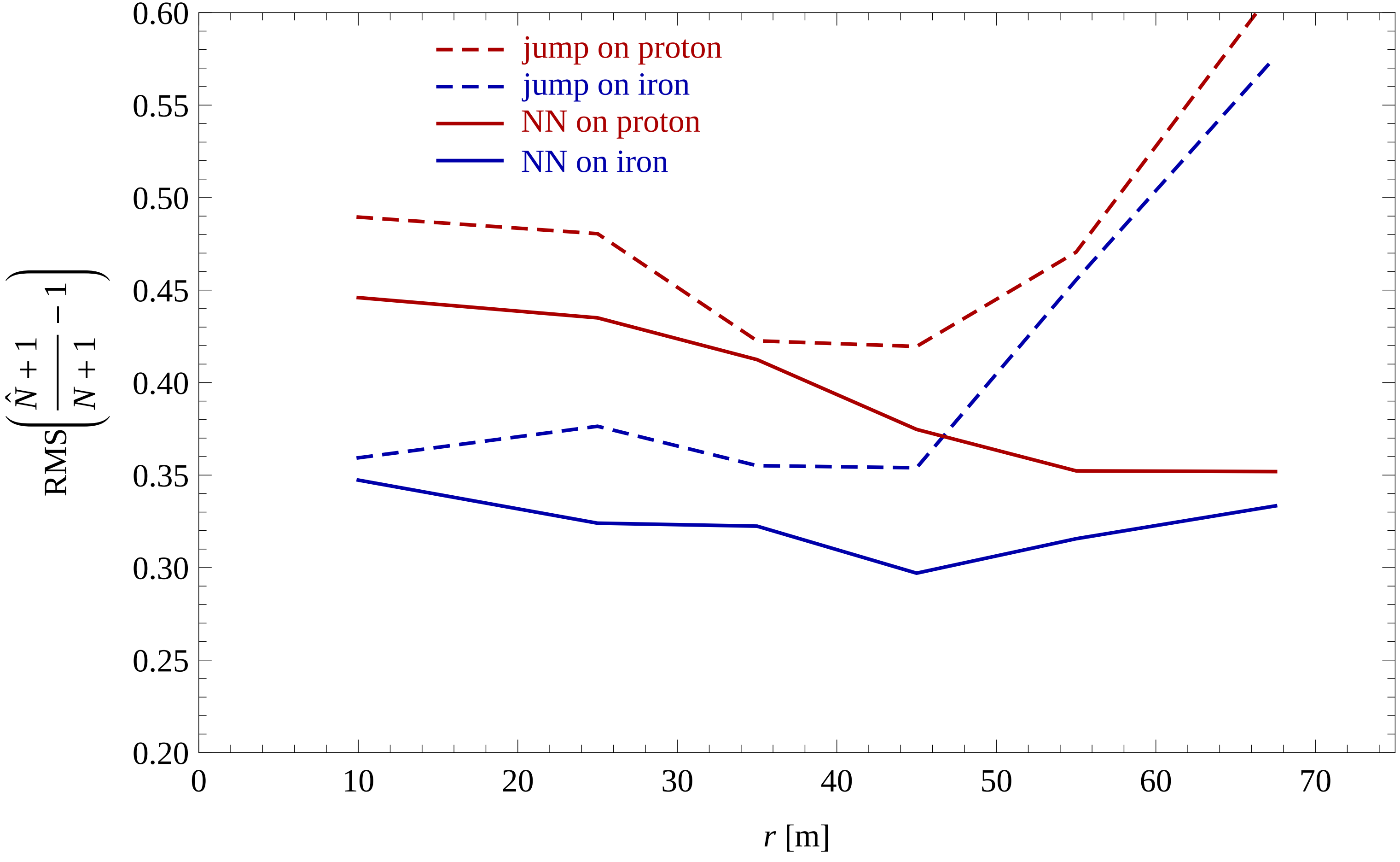
RMS vs. distance from shower axis



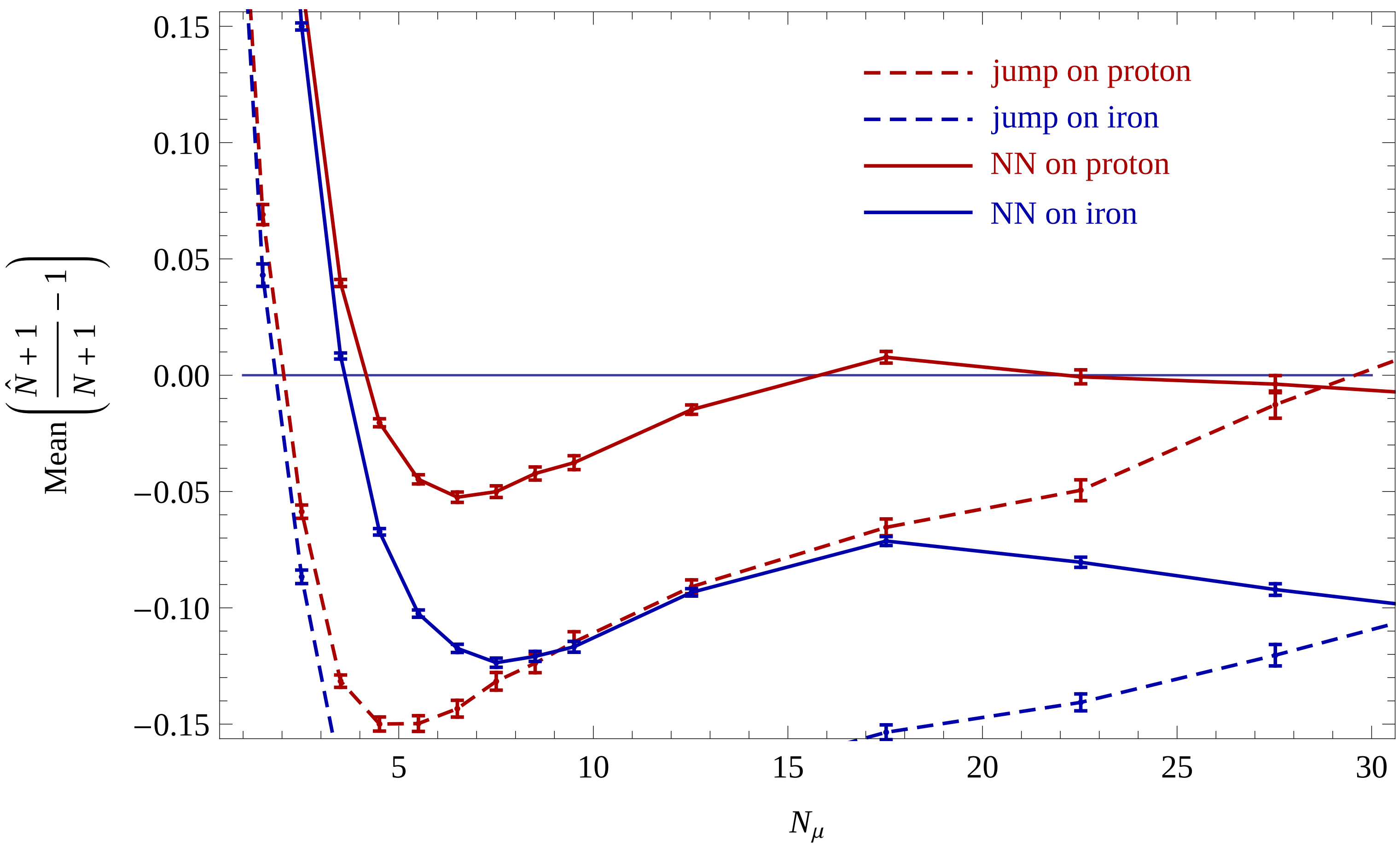
Bias vs. zenith angle



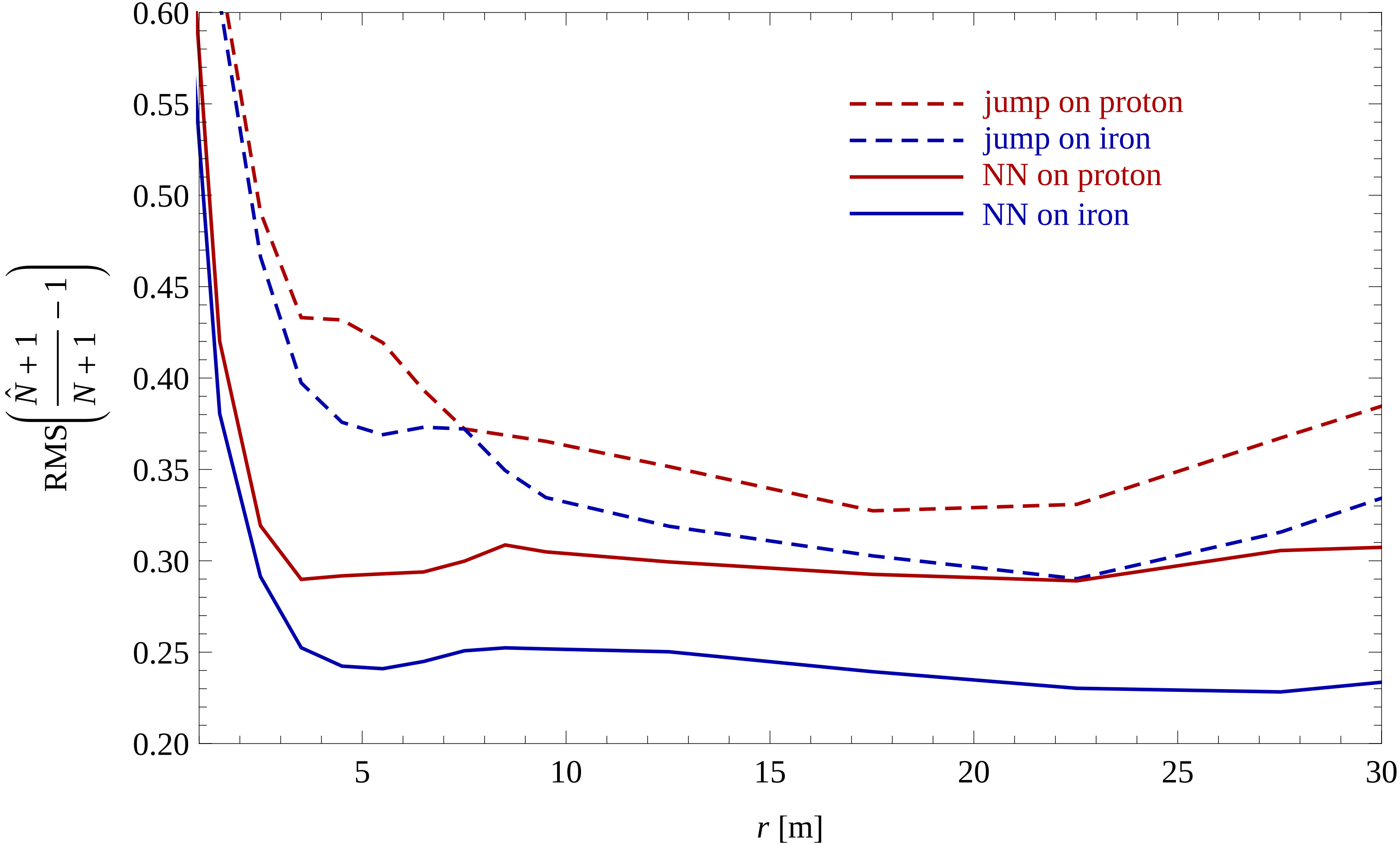
RMS vs. zenith angle



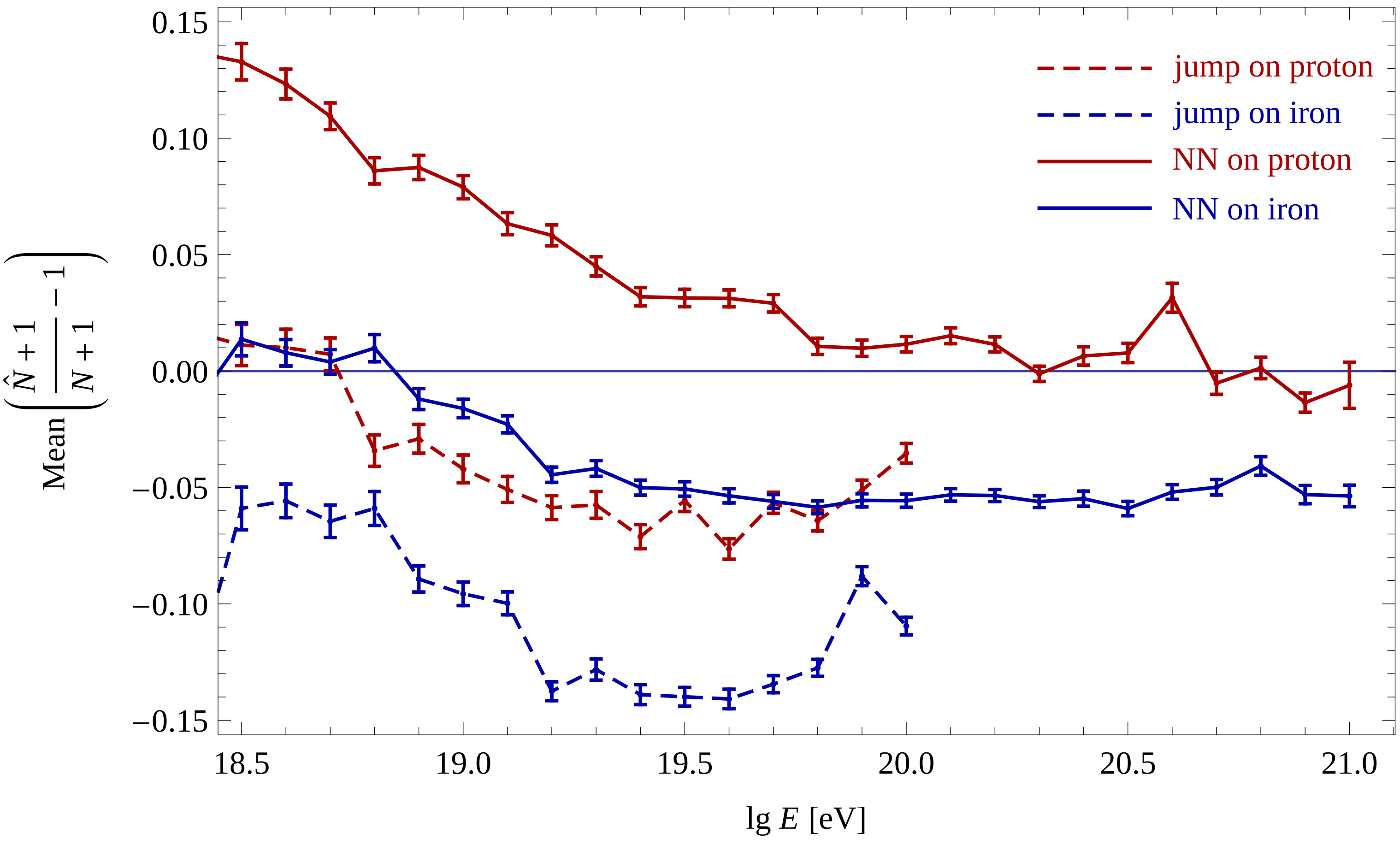
Bias vs. number of muons



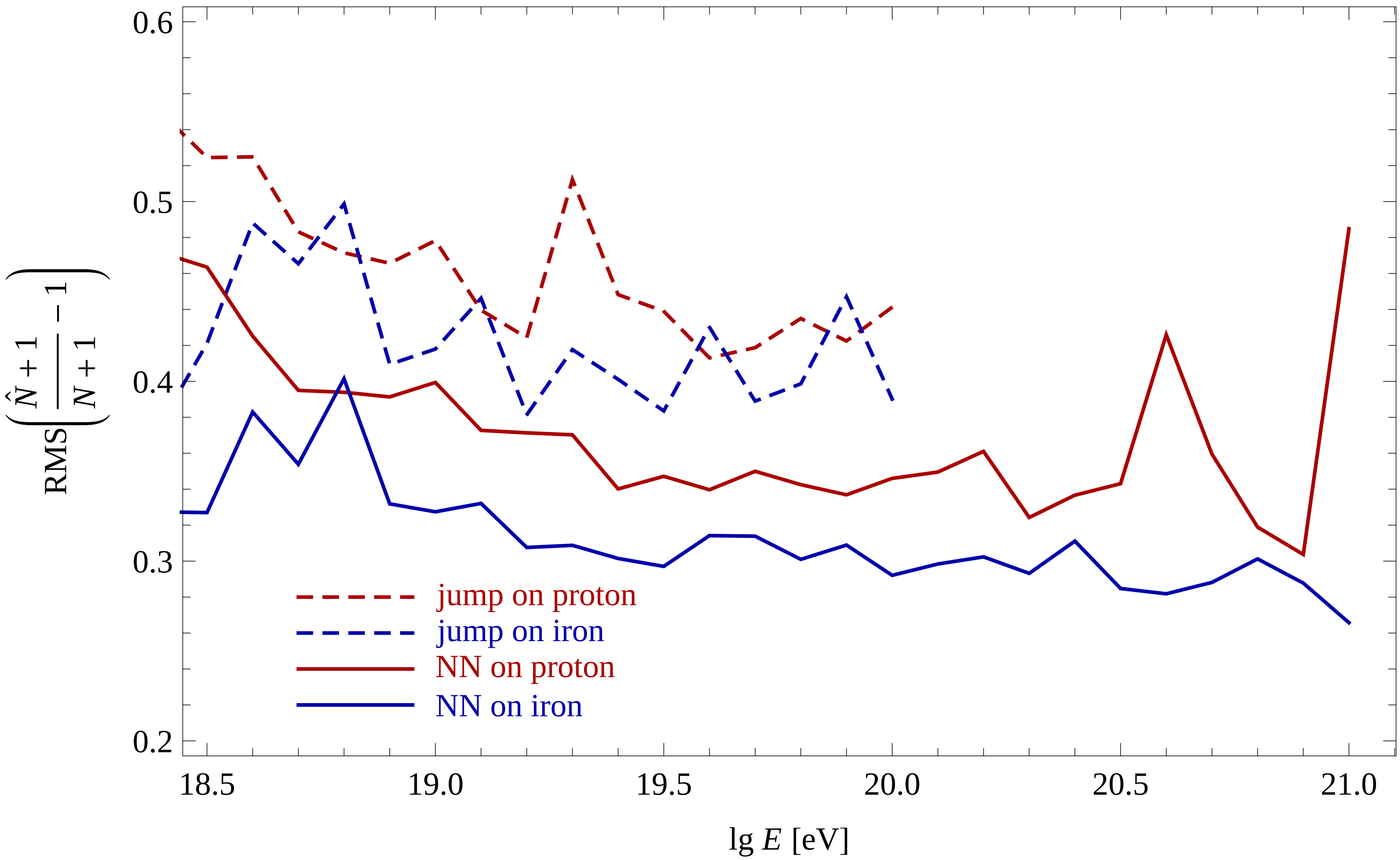
RMS vs. number of muons



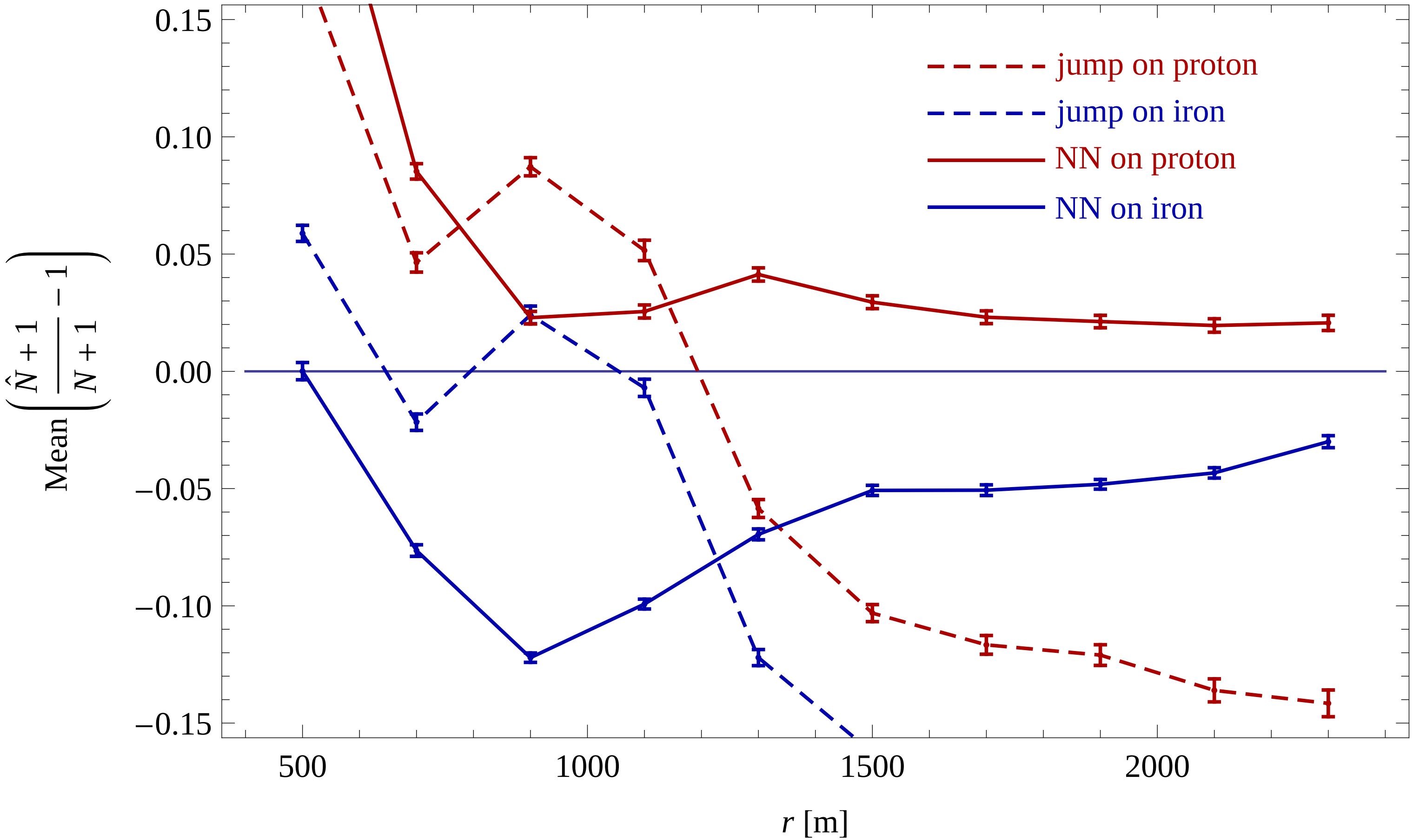
Bias vs. energy



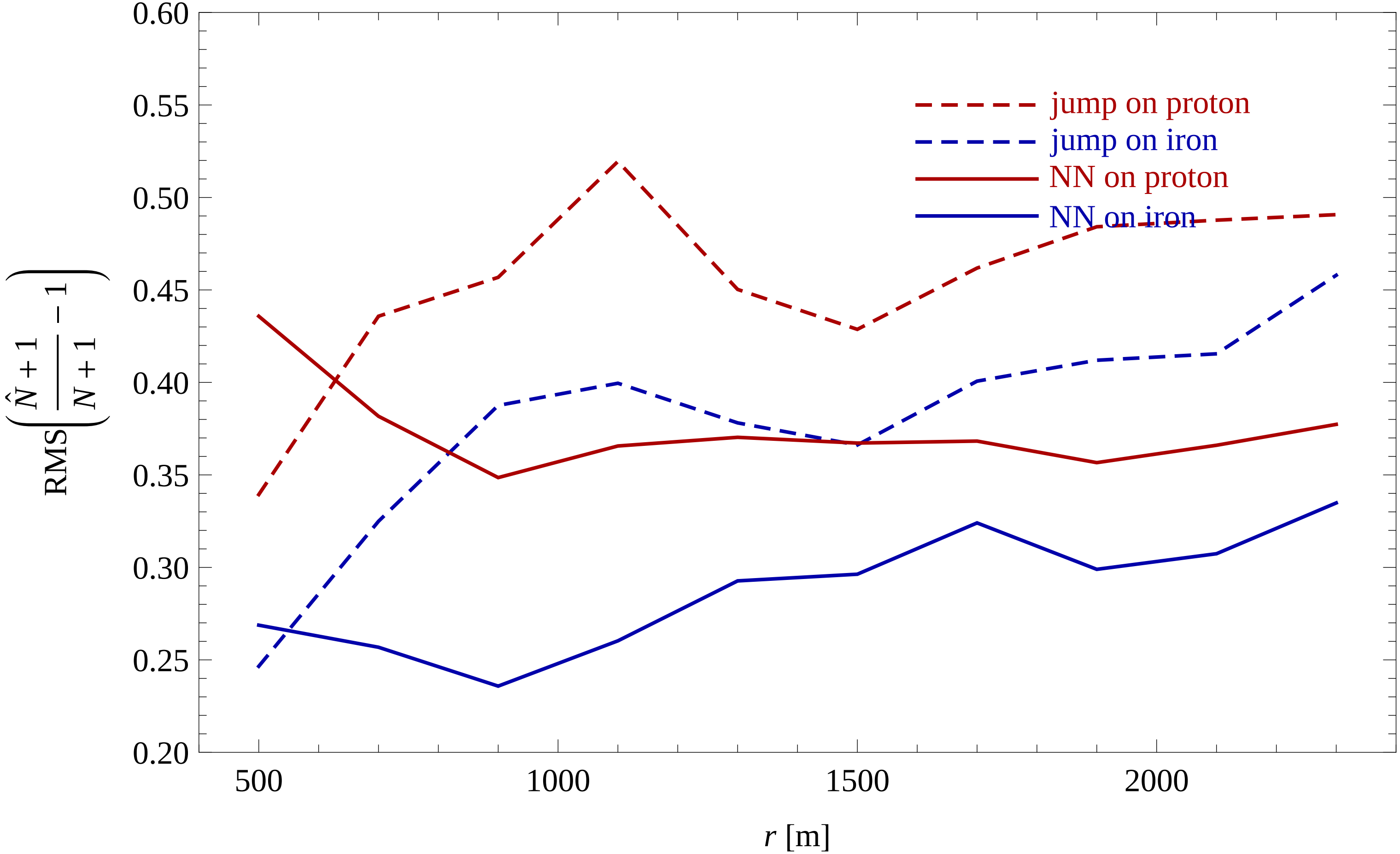
RMS vs. energy



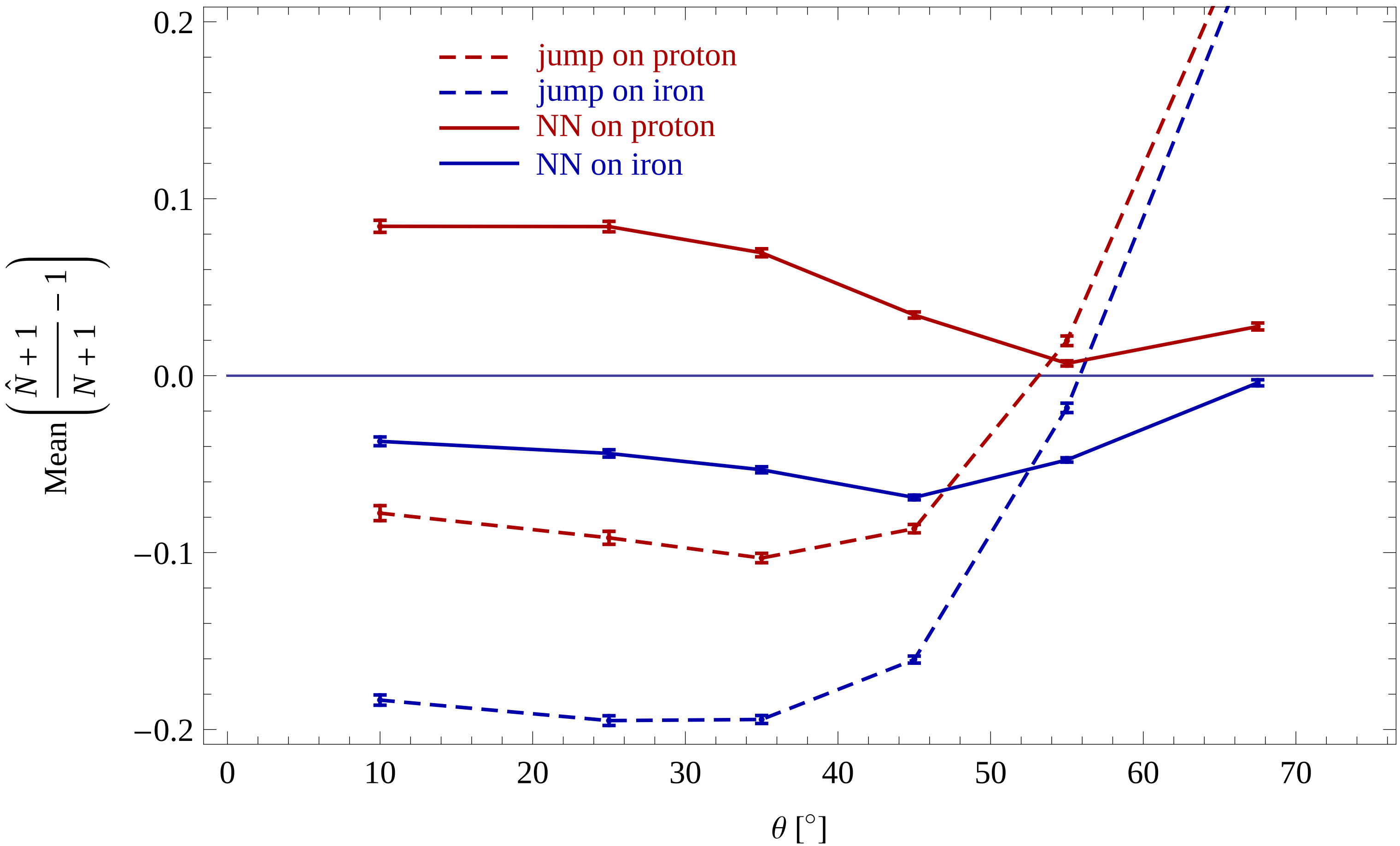
Bias vs. distance from shower axis



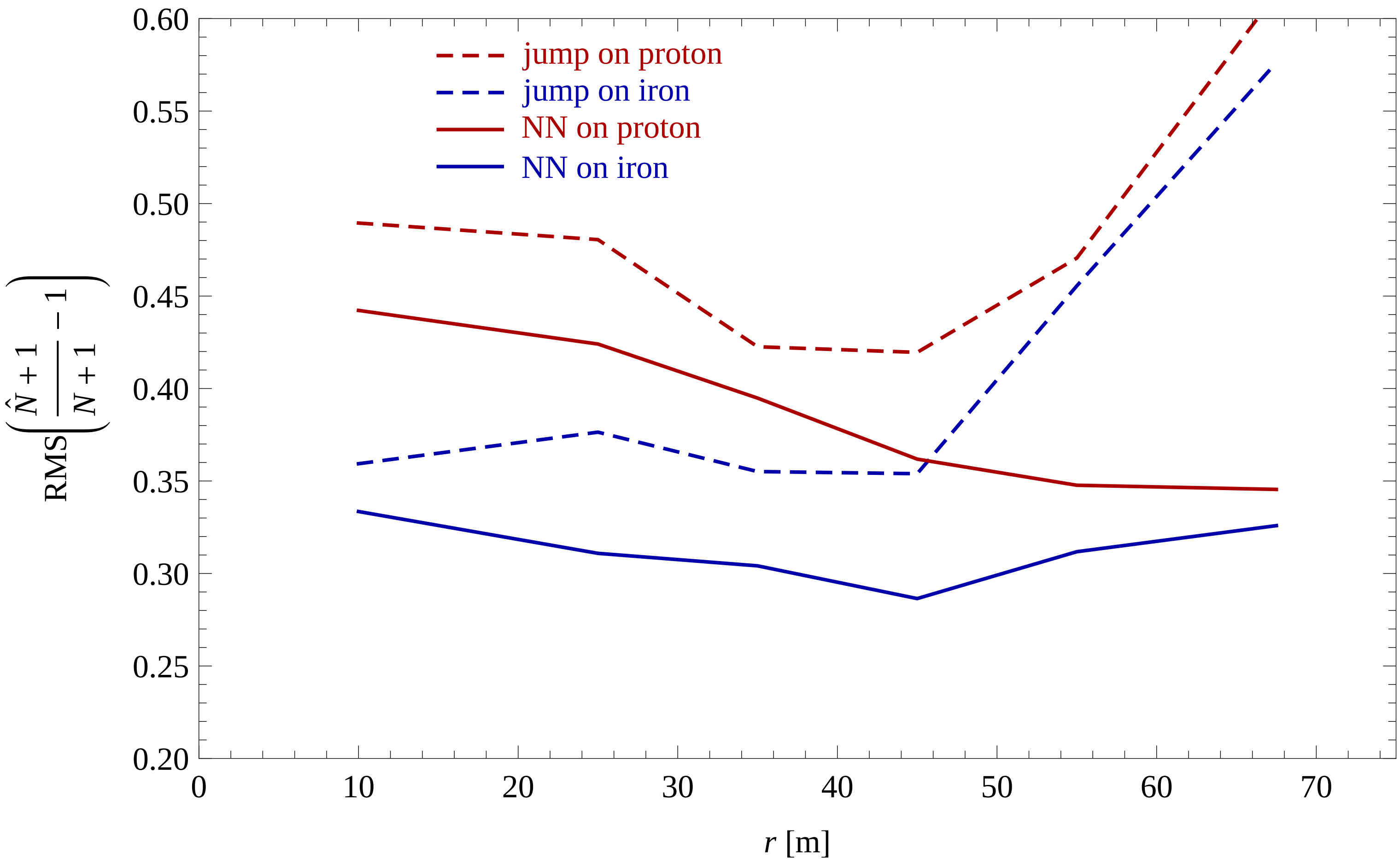
RMS vs. distance from shower axis



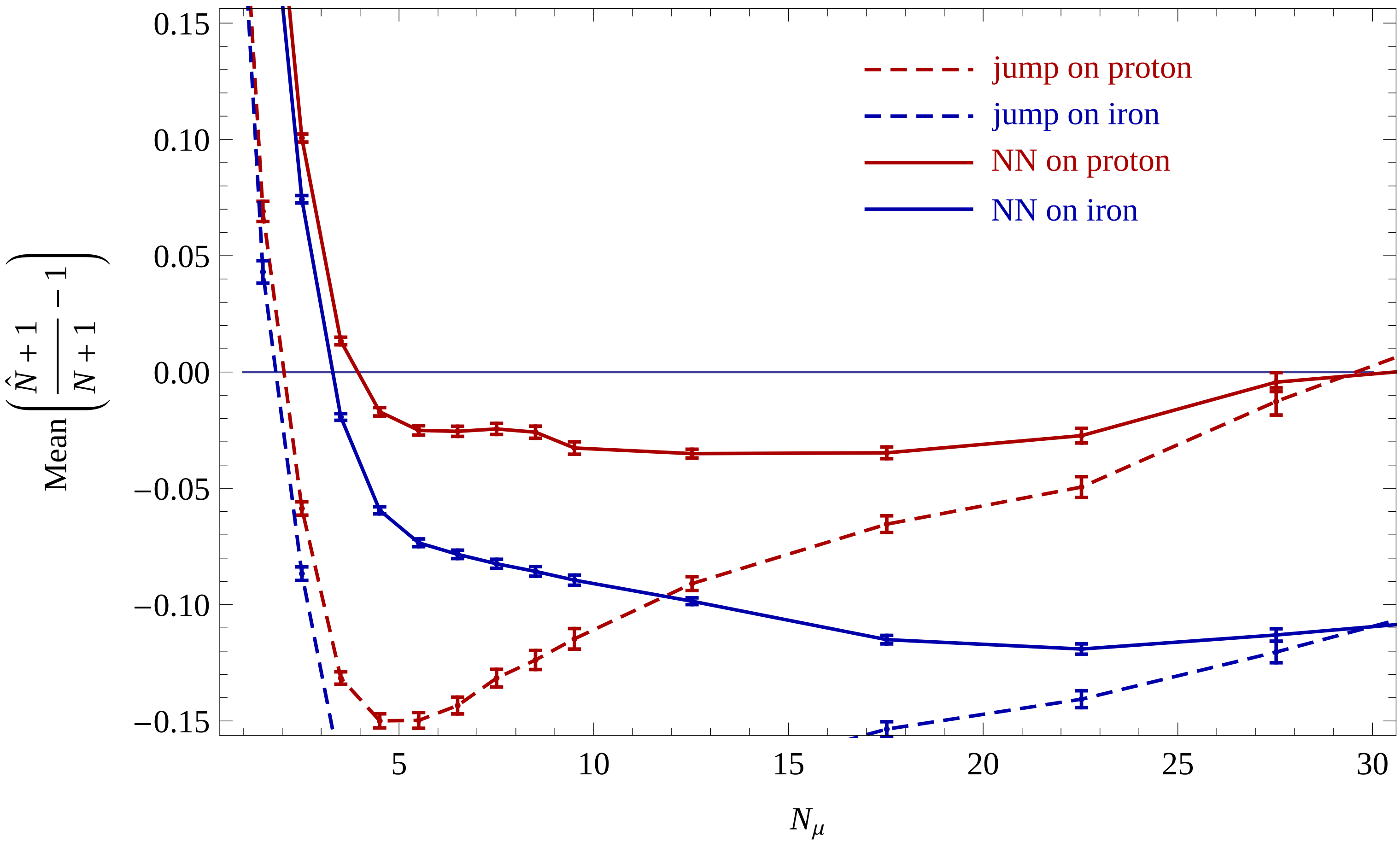
Bias vs. zenith angle



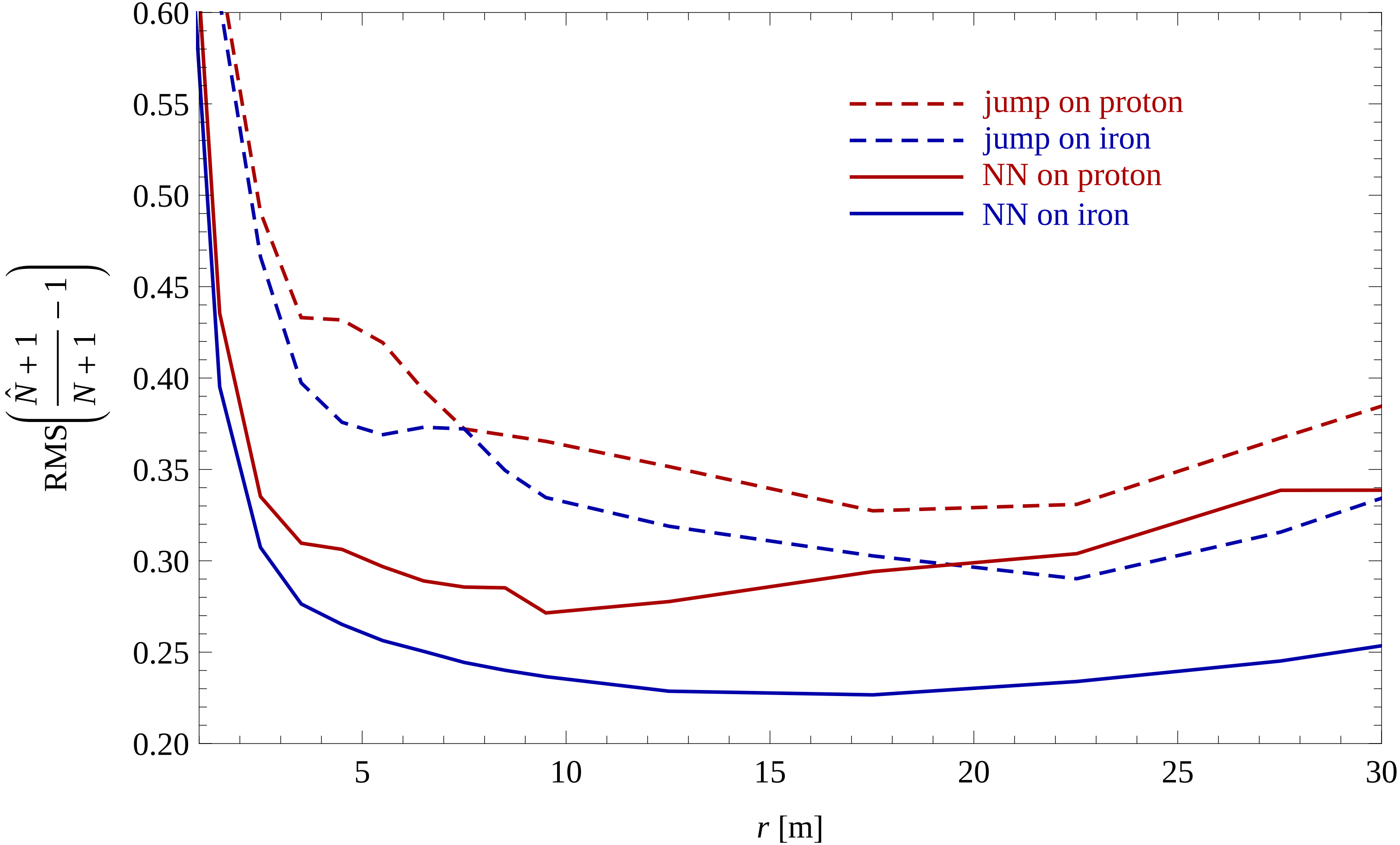
RMS vs. zenith angle



Bias vs. number of muons



RMS vs. number of muons



Conclusions

- A new multivariate muon-counter
  - energy-dependent bias is similar to that of the jump method
  - the estimator no longer depends explicitly on the energy
  - better RMS than that of the jump method
  - valid also for energies larger than $10^{20}\,\mathrm{eV}$
  - these are our very first results, expect improvement
  - heavy simulation-dependence