Systematics: Upstream correction and MC stat

Shyam Bhuller

University of Bristol

April 25, 2024

Upstream correction systematic

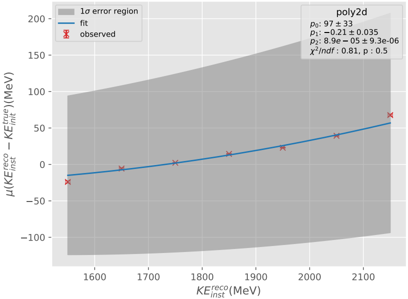
- fit to upstream loss in MC in bins of reco KE inst

$$\Delta E^{MC}_{upstream} = KE^{MC,reco}_{inst} - KE^{MC,true}_{ff} \quad (1)$$

- from the central values, fit a 2nd order polynomial to obtain an energy dependant upstream loss correction which is applied to **both Data and MC**

$$\Delta E_{upstream} \rightarrow \Delta E_{upstream}(KE^{reco}_{inst}, \{p_i\}) \quad (2)$$

- to evaluate systematic, shift p_i by $\pm\epsilon_{p_i}$ uncertainties, re-run the analysis, obtain two measurements of the cross section $xs^{\pm}$



p_0	p_1	p_2
97 ± 33	-0.21 ± 0.03	$(8.9 \pm 0.9) \times 10^{-5}$

Plots

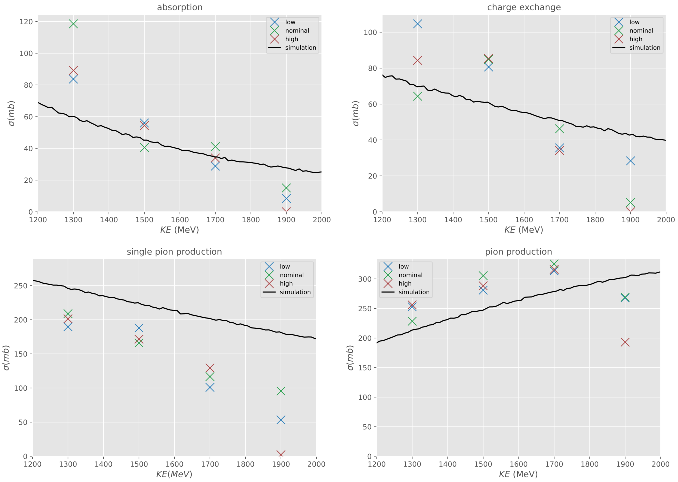
- plots show the nominal central value measurements and the measurements for $p_i \pm \epsilon_{p_i}$
- uncertainty is difference in from the nominal measurement:

$$\epsilon^{\pm} = xs^{nominal} - xs^{\pm}$$

- for a single bin, asymmetric uncertainties are defined as:

$$\epsilon^{low} \text{ if } \epsilon^{\pm} \geq 0, \ \epsilon^{high} \text{ if } \epsilon^{\pm} < 0$$

if both $\epsilon^{\pm} \geq 0$ or $\epsilon^{\pm} < 0$, then take the largest of the two.



Tables

absorption

KE (MeV)	Total	Data stat	Upstream low	Upstream high
1300	0.37	0.22	0.29	-
1500	0.42	0.18	-	0.38
1700	0.36	0.21	0.30	-
1900	1.91	1.63	1.00	-
average	0.77	0.56	0.40	0.10

charge exchange

KE (MeV)	Total	Data stat	Upstream low	Upstream high
1300	0.68	0.27	-	0.63
1500	0.12	0.11	0.05	0.01
1700	0.30	0.14	0.26	-
1900	5.23	2.50	1.00	4.48
average	1.58	0.75	0.33	1.28

single pion production

KE (MeV)	Total	Data stat	Upstream low	Upstream high
1300	0.22	0.20	0.09	-
1500	0.16	0.09	-	0.13
1700	0.20	0.11	0.14	0.11
1900	1.05	0.38	0.98	-
average	0.41	0.20	0.30	0.06

pion production

KE (MeV)	Total	Data stat	Upstream low	Upstream high
1300	0.23	0.20	-	0.12
1500	0.10	0.07	0.08	-
1700	0.06	0.05	0.04	-
1900	0.33	0.16	0.28	-
average	0.18	0.12	0.10	0.03

$$\text{fractional error} = \frac{\epsilon}{xs^{nominal}} \tag{3}$$

- dashed points are when $\epsilon^{\pm} \geq 0$ or $\epsilon^{\pm} < 0$
- uncertainty in the upstream energy loss is largest in the 1900 MeV bin for all measurements
- upstream energy correction is largest for the abs and cex measurement.

MC stat uncertainty

- MC stat uncertainty accounts for limited MC statistics used to define templates for the fit, and response matrices for the unfolding.
- MC stat uncertainty in the fit can be accounted for using nuisance parameters α_{cb}:

$$L = -2 \prod_b \prod_c \left[\log(\mathrm{Pois}(n_{c,b} | \sum_s \mu_s \alpha_{cb} \lambda_{c,b,s})) + \mathrm{Gaus}(\alpha_{c,b} | \gamma_{cb} \delta_{c,b}) \right]$$

  - $\gamma_{c,b} = \sum_s \lambda_{c,b,s} \rightarrow$ total counts in region c, slice b
  - $\delta_{c,b} = \sqrt{\gamma_{c,b}} \rightarrow$ stat uncertainty
  - currently, *pyhf* API only supports using Gaussian constraints for MC stat NPs.
- For unfolding, uncertainty is quantified by a covariance matrix:

$$V = V_{Data} + V_{MC}$$

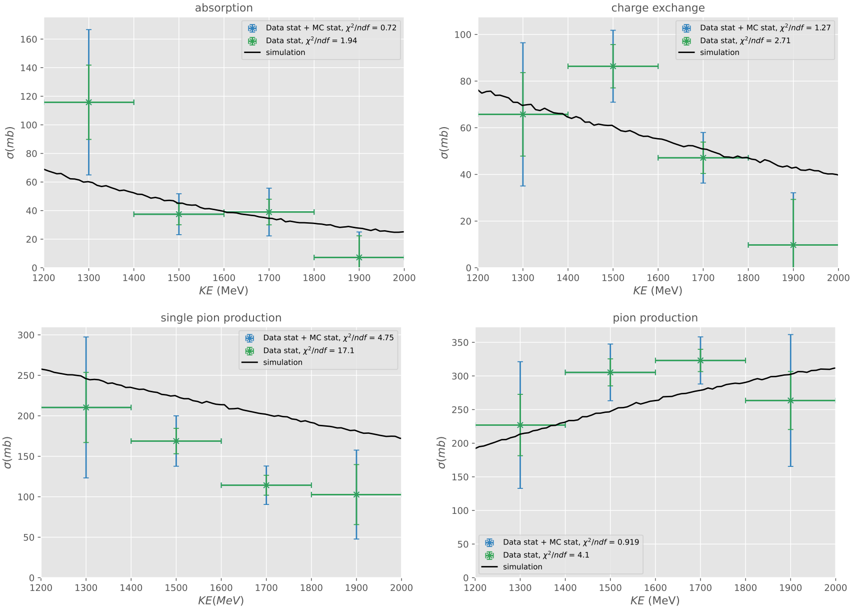
  - $V_{Data} \rightarrow$ covariance of the unfolded distribution (accounts for prior uncertainties, uncertainty for multiple unfolding iterations)
  - $V_{MC} \rightarrow$ covariance of the migration probability i.e. covariance of response matrix. Expressed as the poisson covariance in the *pyunfold* API
  - diagonal component of V is the uncertainty in the unfolded histogram.

How to evaluate the MC stat uncertainty?

- run analysis using fit model without NPs and without calculating V_{MC}
- uncertainties in xs will be purely due to Data stat uncertainty.
- run analysis with fit model + V_{MC}, uncertainties in xs are Data + MC stat.

$$\epsilon^2_{\text{MC stat}} = \epsilon^2_{\text{Data stat + MC stat}} - \epsilon^2_{\text{Data stat}} \tag{4}$$

Plots


absorption
Data stat + MC stat, χ^2/ndf = 0.72
Data stat, χ^2/ndf = 1.94
simulation
$\sigma(mb)$
0
20
40
60
80
100
120
140
160
1200
1300
1400
1500
1600
1700
1800
1900
2000
KE (MeV)
charge exchange
Data stat + MC stat, χ^2/ndf = 1.27
Data stat, χ^2/ndf = 2.71
simulation
$\sigma(mb)$
0
20
40
60
80
100
1200
1300
1400
1500
1600
1700
1800
1900
2000
KE (MeV)
single pion production
Data stat + MC stat, χ^2/ndf = 4.75
Data stat, χ^2/ndf = 17.1
simulation
$\sigma(mb)$
0
50
100
150
200
250
300
1200
1300
1400
1500
1600
1700
1800
1900
2000
$KE(MeV)$
pion production
Data stat + MC stat, χ^2/ndf = 0.919
Data stat, χ^2/ndf = 4.1
simulation
$\sigma(mb)$
0
50
100
150
200
250
300
350
1200
1300
1400
1500
1600
1700
1800
1900
2000
KE (MeV)

Tables

absorption

KE (MeV)	Total	Data stat	MC stat
1300	0.43	0.22	0.37
1500	0.34	0.18	0.29
1700	0.38	0.21	0.32
1900	1.92	1.63	1.02
average	0.77	0.56	0.50

charge exchange

KE (MeV)	Total	Data stat	MC stat
1300	0.47	0.27	0.38
1500	0.18	0.11	0.14
1700	0.23	0.14	0.18
1900	2.85	2.50	1.38
average	0.93	0.75	0.52

single pion production

KE (MeV)	Total	Data stat	MC stat
1300	0.41	0.20	0.36
1500	0.18	0.09	0.16
1700	0.20	0.11	0.17
1900	0.55	0.38	0.40
average	0.34	0.20	0.27

pion production

KE (MeV)	Total	Data stat	MC stat
1300	0.41	0.20	0.36
1500	0.14	0.07	0.12
1700	0.11	0.05	0.09
1900	0.37	0.16	0.34
average	0.26	0.12	0.23

- tables show fractional error in the cross section for Data stat and MC stat.

$$\text{fractional error} = \frac{\epsilon}{xs^{nominal}} \tag{5}$$

- for abs and cex, Data stat uncertainty is larger, for spip and pip, MC stat is larger

Summary

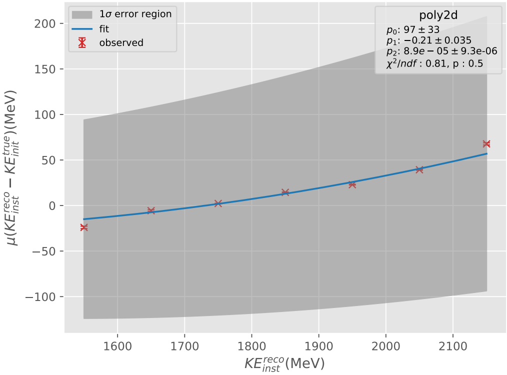
- upstream correction systematic evaluated,
- fit currently has very large uncertainties which propagates to the final measurement try to improve? (see backups)
- MC stat uncertainty has two components, in the fit and unfolding.
- MC stat unfolding uncertainty easy to distinguish from total uncertainty
- MC stat uncertainty in fit is more difficult, as this uncertainty is intrinsically part of of the model.

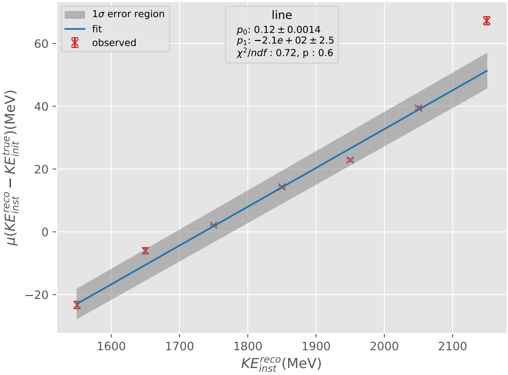
table of systematics and which methods can be used

Systematic	NPs	Repeat Data analysis	Toy MC method	Propagate	Input
MC stat uncertainty	✓(implemented)	-	✓	-	stat err
theory uncertainty	-	-	✓(implemented)	-	±20%
background subtraction	-	-	-	✓(implemented)	±20%
upstream energy loss	-	✓(implemented)	-	-	$\pm 1\sigma$ fit err
beam momenta mis-modelling	-	✓(implemented)	?	-	$\pm 1\sigma$ fit err
shower energy correction	-	✓(implemented)	-	-	$\pm 1\sigma$ fit err
track length resolution	-	✓(implemented)	?	-	2.6%
beam momentum resolution	-	-	?	✓	2.5%
space charge correction	-	✓(run with SCE off)	-	-	-

Backup

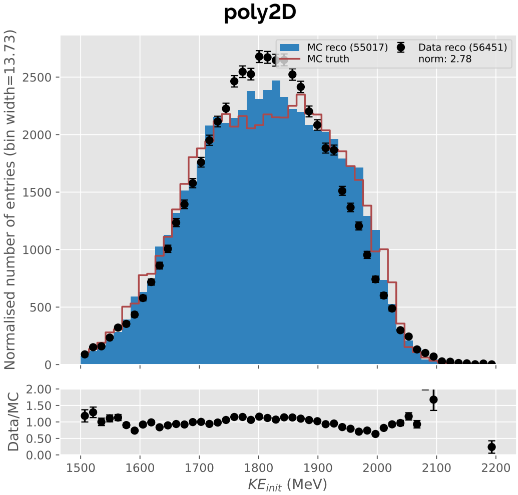
Alternate Upstream correction

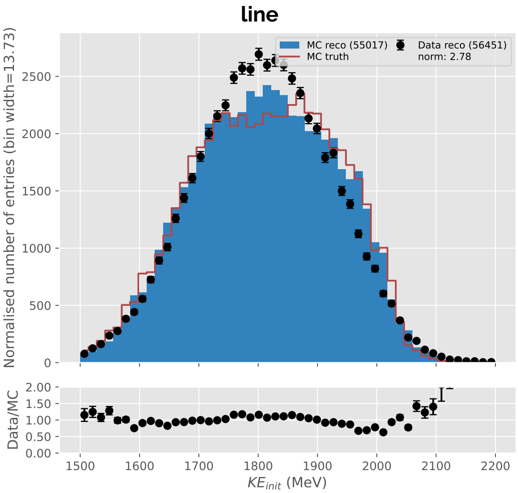




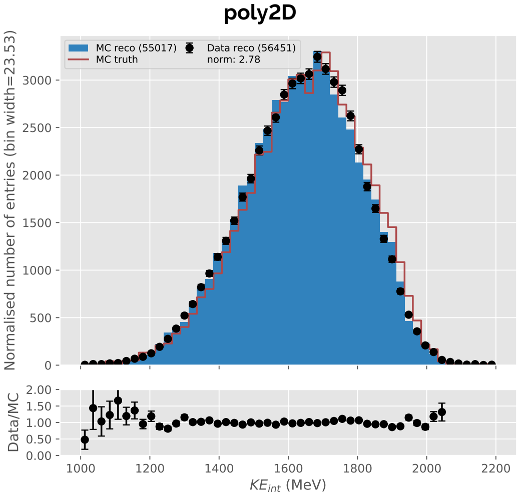
- Tried alternate fit function for $DeltaE_{upstream}$,
- from left $p_2 << 1$, 2nd order polynomial term is very small, so tried a linear fit instead
- fit is also reasonable, and 1σ error band is much smaller.
- both have good χ^2/ndf

Alternate Upstream correction

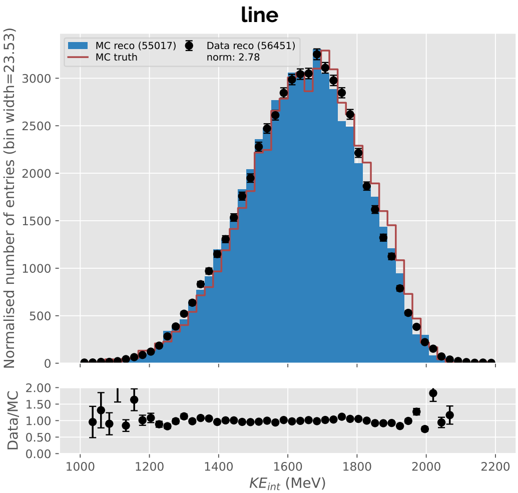




Alternate Upstream correction



poly2D



line

- calculated KE distributions are fairly similar, so switch to line fit.

Nuisance parameters

- fit model can facilitate nuisance parameters to quantify different systematic effects.
- if a systematic can be expressed as a fractional error on the number of events, it can be incorporated into the model.
- if we allow more than 1 KE bin in the fit model, systematics which can be expressed as a fractional error in the interacting KE
- **benefit:**
  1. implementation is simple, adjust model, re-run fit and subsequent steps
  2. adding nuisance parameters can help the model mitigate the effects on the fit results
- **disadvantages:**
  1. higher number of NPs results in fit being underconstrained, resulting in more unstable fitting.
  2. must rerun pull study and normalisation cross checks
- Current model has takes 4 observations, and has 8 free parameters (4 POIs, 4 NPs), so model is already underconstrained.
- MC stat uncertainty is currently being incorporated using NPs

Repeat data analysis

- analysis performs various corrections by using values extracted from fits
- systematic effect on a specific correction can be determined by changing values used to calculate the correction
- example:
  - upstream energy correction is determined from fit values p_i.
  - fit values have some uncertainty determined by the fit: $p_i \pm \epsilon_i$
  - determine uncertainty in upstream energy correction by re-running analysis for $p_i + \epsilon_i$, and $p_i - \epsilon_i$, obtain σ^{high} and σ^{low}
  - systematic is $(\sigma^{high} - \sigma^{low})/2$
- **benefits:**
  1. no additional changes required to analysis or fit model
  2. does not require rerunning toy studies if the fit results don't impact the region identification
- **disadvantages:**
  1. evaluation of systematics is very simple, does not account for correlations between other effects
  2. magnitude of systematic may to be compatible to the measurement i.e. can't be expressed as a fractional error

MC method

- systematic is evaluated by running multiple pseudo experiments using the toy estimating the effect some systematic has on the measured cross sections. Uncertainties are then expressed as fractional errors and applied to the Data MC measurement.
- **benefits:**
  1. Data MC analysis does not need to be re-run at all
  2. multiple systematic effects can be varied simultaneously (handles correlations between effects)
- **disadvantages:**
  1. depending on the number of pseudo experiments, method may be time consuming
  2. not all systematics can be incorporated e.g. upstream energy correction, beam momentum resolution, selection is not incorporated into the toy.

Systematics

1. **MC stat uncertainty** (uncertainty due to limited MC statistics)
2. **Theory uncertainty** (uncertainty in Geant 4 cross sections)
3. **Upstream energy loss uncertainty** (used to calculate initial and interacting KE)
4. **Beam momenta mismodelled in MC**(reweigh to Data)
5. **Shower energy correction** (neutral pion id)
6. **Beam momentum resolution**(uncertainty in measured momenta from beam instrumentation)
7. **Space charge correction** (used at multiple places in the selection and track length calculation)
8. **Track length resolution** (used to calculate interacting KE)

MC stat uncertainty

- MC stat uncertainty is built into region fit, and unfolding
- reigon fit incorporates this as a nuisance parameter
- unfolding propagates the statistical uncertainty in the response matrix
- to calculate the MC stat error, run analysis without nuisance parameters and without response matrix error propagation

Theory uncertainty

- cross section model uncertainty is $\pm 20\%$, fit tries to determine normalisation in Data.
- using toys, vary true cross section by $\pm 20\%$, keep template fixed and re-run analysis.
- uncertainty in normalisation systematic is $\epsilon = \sigma^{meas} - \sigma^{true}$.
- repeat experiment multiple times, obtain average $\bar{\epsilon}$
- convert $\bar{\epsilon}$ to fractional error, apply to data measurements

Background subtraction uncertainty

- background subtraction uses background shapes from MC, thus, also propagate ±20% theory uncertainty through the background subtraction
- For a region c' the background samples are when $c' \neq s'$.
- fit predicts the estimated **counts** of each process in each region $\nu_{c',s'} \pm \Delta\nu_{c',s'}$
- subtract background from N_{int} to get $N_{int,c'}$ **in each energy slice** i.e. we require **shapes** of $N_{int,s'}$
- shape of N_{int} for each process is determined from MC $S_{b,s'} = \frac{\sum_c N^{MC}_{c,b,s'}}{\sum_{c,b} N^{MC}_{c,b,s'}}$
- background subtracted interacting counts in each region is $N_{c',b} \pm \Delta N_{c',b}$ (includes Data stat + MC stat uncertainty):

$$N_{c',b} = N^{Data}_{c',b} - \sum_{s'} \nu_{c',s'} = N^{Data}_{c',b} - \sum_{s'} \nu_{c',s'} S_{b,s'} \tag{6}$$

$$(\Delta N_{c',b})^2 = N^{Data}_{c',b} + \sum_{s'} \left[(S_{b,s'})^2 (\Delta\nu_{c',s'})^2 + \frac{(\nu_{c',s'})^2}{\sum_{c,b} N^{MC}_{c,b,s'}} S_{b,s'}(1 + S_{b,s'}) + f^2 S^2_{b,s'} \right] \tag{7}$$

- $f = 0.2$
- not really a systematic, need to propagate

Re-weighing systematic

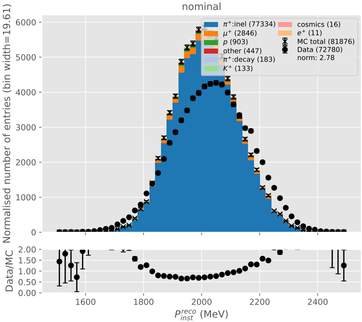
- Data MC discrepancy in beam profile treated by re-weighing MC
- fit to ratio in a sideband sample is done to derive weights for each event

$$r_i = \frac{N_i^{Data}}{N_i^{MC}} \frac{\sum_j N_j^{MC}}{\sum_j N_j^{Data}} \qquad (8)$$

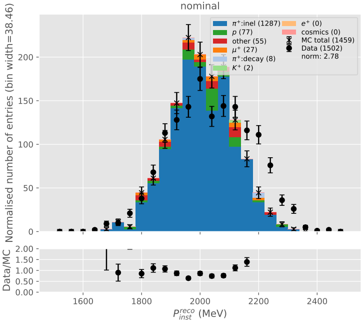
$$w = \frac{1}{r(P_{inst}^{reco}, \{p_i\})}, \qquad (9)$$

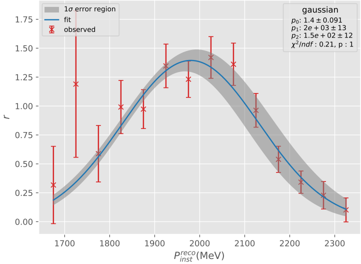
- sideband is beam particle selection, except $nPFO == 0$
- **to evaluate systematic, shift uncertainties in fitted parameters by ± 1 standard deviation, re-run analysis with PDSP MC + Data**

Selected beam pions



Sideband





Shower energy correction

- shower energy is corrected improve pi^0 mass reconstruction
- mass used in π^0 selection
- correction is:

$$C(E_{shower}) = p_0 \ln(E_{shower} - p_1) + p_2 \tag{10}$$

- parameters $p_i, i \in \{0, 1, 2\}$ obtained from fit, and have uncertainties
- vary p_i by $\pm 1\sigma$, re-run analysis with PDSP Data/MC

Beam momentum resolution

- individual beam momenta measured by the beam spectrometer has a 2.5% uncertainty.
- for 2GeV beam, this is 50MeV, thus energy slice bins should be no less than 100MeV (currently 200MeV)
- one option is to propagate uncertainty through KE calculation

  (how does this manifest in errors in a histogram?)
- others (Kang, Francesca) vary beam momenta in data by $\pm 1\sigma$, re-run analysis
- could run psuedo experiments using toy, randomly assign a beam momenta offset to toy data

track length resolution

- using ProtoDUNE MC fit distribution to:

$$\frac{l^{reco} - l^{true}}{l^{reco}} \tag{11}$$

- from fit, resolution is rms of distribution
- similar approach to beam momenta resolution can be used