PROCEEDINGS OF SCIENCE

A Markov Chain Monte Carlo determination of Proton PDF uncertainties at NNLO

Peter Risse,[a,*] **Nasim Derakhshanian**,[b] **Tomas Ježo**,[a] **Karol Kovařík**[a] **and Aleksander Kusina**[b]

[a] *Institut für Theoretische Physik, Universität Münster, Wilhelm-Klemm-Straße 9, D-48149 Münster, Germany*

[b] *Institute of Nuclear Physics Polish Academy of Sciences, PL-31342 Krakow, Poland*

E-mail: risse.p@uni-muenster.de

The current scientific standard in PDF uncertainty estimation relies either on repeated fits over artificially generated data to arrive at Monte Carlo samples of best fits or on the Hessian method, which uses a quadratic expansion of the figure of merit, the χ^2-function. Markov Chain Monte Carlo methods allow one to access the uncertainties of PDFs without making use of quadratic approximations in a statistically sound procedure while at the same time preserving the correspondence between the sample and χ^2-value. Rooted in Bayesian statistics the χ^2-function is repeatedly sampled to obtain a set of PDFs that serves as a representation of the statistical distribution of the PDFs in their function space. After removing the dependence between the samples (the so-called autocorrelation) the set can be used to propagate the uncertainties to physical observables. The final result is an independent procedure to obtain PDF uncertainties that can be confronted by the state-of-the-art in order to ultimately arrive at a better understanding of the proton's structure.

31st International Workshop on Deep Inelastic Scattering (DIS2024)
8–12 April 2024
Grenoble, France

*Speaker

PoS(DIS2024)054

1. Introduction

In recent years proton PDF extractions have become more and more precise by including newer and more complete experimental data sets, utilizing more experimental observables, increasing the theoretical accuracy from NNLO to approximate N3LO and including further methodical advancements. Recently, also the estimation of the error PDFs has re-gained interest [1, 2], including the proposal of using advanced statistical tools of uncertainty estimation: Markov Chain Monte Carlo (MCMC). So far this method has only been used in toy models [3, 4] or in extractions using only DIS data from HERA run I and II [5], because the analysis is much more involved computationally. In this talk we present a proton PDF uncertainty estimation from MCMC with a realistic set of data and compare the results with the state-of-the-art for global analyses.

2. Experimental data and theoretical setup

The goal of the analysis is to perform a realistic proton PDF extraction, whilst keeping the computational effort at a reasonable level. As a compromise we only consider a reduced selection of experimental data sets compared to a global PDF analysis, because we exclusively rely on theoretical predictions in the form of fast-convolution grids. This allows for an extremely fast recalculation of theoretical predictions and ultimately makes the statistical investigation with Markov Chains possible. The complete list of data sets is given in table 1 grouped by observable; the kinematic coverage is given in fig. 1. In all cases we use NNLO theoretical accuracy and take correlated uncertainties into account wherever available.

In the following we give a brief description of the considerations behind the selection. Finally we introduce our PDF parametrization.

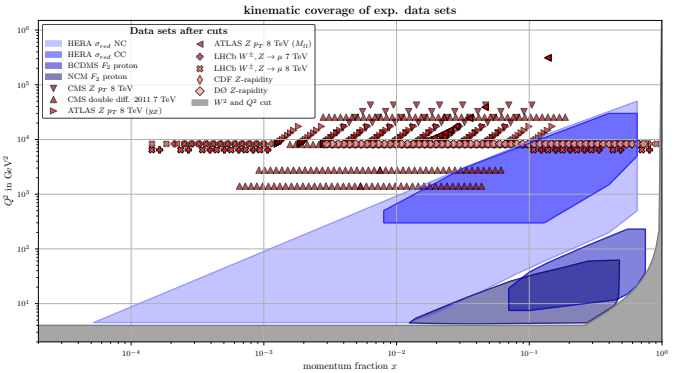

Figure 1: The kinematic coverage of the experimental data sets (see table 1) used for the extraction. DIS data sets are indicated by blue patches, DY data sets by red by individual symbols and cut regions in gray. We used leading order approximations to estimate the (x, Q^2)-point.

Deep inelastic scattering The DIS data come as measurements of the neutral current F_2 structure function (BCDMS, NMC) and as the reduced cross section (combined data set of H1 and ZEUS) for neutral and charged current. The theoretical predictions are obtained in the aSACOT-χ mass-scheme [6] that was recently implemented in the numerical library `APFEL++` [7, 8]. In this library the numerical predictions are pre-calculated in tables that only have to be interpolated by a PDF set at runtime. This yields a massive speed advantage compared to naive implementations. We employ kinematic cuts of

$$W^2 \geq 12.25\,\text{GeV}^2 \quad \text{and} \quad Q^2 \geq 4\,\text{GeV}^2\,. \tag{1}$$

The kinematic coverage of the data is given by the blue patches in fig. 1.

Drell-Yan In this setup we use the NLO `APPLgrid`-tables published in Ref. [9] by the NNPDF collaboration and translate these to FastKernel-tables [10] in order to increase the evaluation speed by $O(100)$ over the `APPLgrid`-tables. NNLO accuracy is achieved by using K-factors, which have been published by NNPDF in Refs. [11, 12]. In order to stay consistent, we align our cuts with the NNPDF analysis, see Ref. [12, table 2.4]. The data sets are represented by the red symbols in fig. 1.

Data Set	Ref.	Points	χ^2/data	Data Set	Ref.	Points	χ^2/data
DIS				**DY**			
HERA σ_{red} neutral current	[13]	1039	1.26	CDF Z-rapidity	[14]	28	1.10
HERA σ_{red} charged current	[13]	81	1.08	DØ Z-rapidity	[15]	28	0.60
BCDMS F_2 proton	[16]	339	1.09	ATLAS Z p_T 8 TeV (M_{ll})	[17]	44	1.06
NCM F_2 proton	[18]	201	1.54	ATLAS Z p_T 8 TeV (y_Z)	[17]	48	0.65
				CMS Z p_T 8 TeV	[19]	28	0.46
				CMS double diff. 2011 7 TeV	[20]	88	1.02
				LHCb $W^{\pm}, Z \to \mu$ 7 TeV	[21]	29	1.07
				LHCb $W^{\pm}, Z \to \mu$ 8 TeV	[22]	31	1.18
DIS total		1660	1.25	DY total		324	0.91
Total		1984	**1.20** (per dof)				

Table 1: The experimental data sets considered alongside with the χ^2/DATA value of the *best-fit* sample. In total the analysis included 1984 data points (after kinematic cuts), which the MCMC samples are able to describe up to a χ^2-value of 1.2 per degree of freedom.

PDF parametrization Motivated by the proton PDF extractions in the CJ family [23, 24], we use the parametric form

$$x f(x, Q_0) = p_0 x^{p_1} (1-x)^{p_2} \left(1 + p_3 \sqrt{x} + p_4 x\right) \tag{2}$$

for the flavor combinations u_v, d_v, $\bar{d}+\bar{u}$, g and $s+\bar{s}$, where $p_{0\ldots4}$ are the fit-able parameters. The parametrization scale Q_0 is placed at the charm threshold $m_c = 1.3\,\text{GeV}$. In u_v, d_v and $\bar{d}+\bar{u}$ we fix p_0 through sum rules and set p_3 to $\{0, -3.503, 0\}$ respectively. For the gluon distribution we open all five parameters and for the strange-combination only p_0, whilst keeping $p_{1\ldots4}$ fixed at $\{-0.20775, 0, 0, 14.606\}$. Lastly we add $p_5 x^{p_6} x u_v(x, Q_0)$ to the down-valence distribution, where we set $p_5 = 0.0036$ and $p_6 = 2$. In total we fit 15 parameters, whose vector we denote in the following by $\mathtt{p}$.

3. Markov Chain Monte Carlo setup

Rooted in Bayesian statistics, the probability density of finding the 15 PDF parameters $\mathtt{p}$ given the experimental data $\{D_i\}$ can be written as

$$\pi(\mathtt{p}|D) = \frac{1}{\mathcal{N}} p(\mathtt{p}) \exp\left(-\frac{1}{2}\chi^2(D, T(\mathtt{p}))\right) . \tag{3}$$

Here χ^2 is the usual χ^2-function (e.g. [1]) taking correlated and normalization uncertainties into account, $T(\mathtt{p})$ are the theoretical predictions calculated from the PDF parameters, $\mathcal{N}$ is an irrelevant normalization constant and $p(\mathtt{p})$ is the prior distribution for the PDF parameters, which we define later.

The goal of the MCMC analysis is to find a set of parameter samples $\mathtt{p}_t$ (with $t = 1 \ldots N$), which approximates the probability distribution $\pi(\mathtt{p}|D)$, i.e. the samples approximate expectation values of any observable $O(\mathtt{p})$:

$$\langle O(\mathtt{p})\rangle_\pi = \int \mathrm{d}\mathtt{p}\, O(\mathtt{p})\pi(\mathtt{p}) \approx \frac{1}{N}\sum_t^N O(\mathtt{p}_t) = \frac{1}{N}\sum_t^N O_t . \tag{4}$$

The MCMC samples are generated in a procedural algorithm, where every new sample $\mathtt{p}_{t+1}$ is first proposed from the current sample by the adaptive Metropolis-Hastings algorithm [25] and then accepted/rejected by the Metropolis-Hastings [26, 27] acceptance probability $a(\mathtt{p}_t, \mathtt{p}_{t+1})$. Here, $a(\mathtt{p}_t, \mathtt{p}_{t+1})$ includes information from eq. (3) such that the density of the samples follows the density given by the experimental measurements. However, each newly proposed sample requires a re-evaluation of the χ^2-function making the procedure computationally expensive. If $\mathtt{p}_{t+1}$ is accepted, it gets appended to the list, otherwise the current sample is repeated. Since $\mathtt{p}_{t+1}$ was proposed based on $\mathtt{p}_t$, there exists a dependence between the samples, which is called autocorrelation (see e.g. Ref. [28]) and can be intuitively understood as a reduced gain of

information compared to the information gained from an independently generated (i.e. uncorrelated) sample. The starting point can be chosen (or generated) freely. In the beginning the chain will drift towards the region of highest probability. This is so-called thermalization time has to be removed to avoid bias.

In the following we briefly discuss the priors for the parameters, the global settings for generating the chain and the purification, where we remove the thermalization region and autocorrelation to ultimately arrive at a reduced set of samples that can be used in the uncertainty estimation.

Priors We set the priors of the parameters to a constant value (which is absorbed by the normalization constant), except for the prior for p_4 of the down-valence distribution. This parameter value is unconstrained from above by the experimental data due to a flaw in the parametrization, where the PDF becomes effectively independent of p_4 if it becomes too large. Instead of fixing its value, we use a uniform prior with the bounds $\left[-10^3, 10^4\right]$[1], which sets the acceptance probability to zero, if it is proposed outside of the limits. The prior is constant if p_4 is proposed inside the limits and thus absorbed by the normalization keeping the correspondence between the sample and the χ^2-value intact.

Generating the samples We generate 36 independent chains, each consisting of 479,000 samples yielding 17,244,000 in total after 14 days of computing time (on 36 cores in parallel). The starting point for each chain was found by first running a minimization algorithm to find the region of highest probability and then individually perturbing them in a random fashion from the minimum to keep the chains independent. The proposal algorithm was reset at the 20,000th and 40,000th step to boost convergence.

Purification The thermalization is roughly finished after the 120,000th iteration. To be conservative, we choose to remove the data before the 140,000 sample. As the chain exhibits strong autocorrelation, we thin the chain, i.e. instead of using every sample we only consider every η-th sample. This does not only reduce the chain and therefore the computational costs of eq. (4) greatly, but also simplifies the interpretation of each sample individually. It is to be noted that thinning reduces the statistics [29, 30] (i.e. our results are less precise after thinning), but we still end up with a sufficiently large sample.

The autocorrelation is estimated by the Γ-method [31] and after applying a thinning factor of $\eta = 3000$ (for each chain individually) we arrive at the estimate that the next independent sample is on average found after $2\tau_{int} = 1.14 \pm 0.07$ steps, very close to the optimal value of one. Applying a larger thinning factor does not yield improved results. The final number of samples is $N = 4068$, which we consider as approximately uncorrelated and free from thermalization bias.

4. Final PDF uncertainty estimation

The uncertainty estimation on observables or the PDFs themselves based on the samples can be carried out in several ways. Usually the task is to find a central value O_* along with the confidence interval $\left[O_-, O_+\right]$ corresponding to some probability p, often $p = 90\%$. One of the simplest symmetric estimations is based on the moments of the observable: $\langle O\rangle \pm z_p\sqrt{\langle O^2 - \langle O\rangle^2\rangle}$, where z_p is the quantile for p. Following Ref. [5], an asymmetric estimation can be defined by setting the central value to the best fit sample and then performing a quantile estimation on the upper and lower value of O.

Here we follow the definition of Ref. [32], which we call the "Cumulative χ^2"-method: The central value is set to the best fit sample (i.e. the sample with the minimal χ^2-value). Then we perform a quantile estimation of the distribution of the χ^2-values as depicted in fig. 2. The lower/upper bound on the observable is defined as the minimal/maximal value found within the quantile. With our samples we find $\chi^2_{\text{max}} = 22$ for the 90% quantile, which is in agreement with a χ^2-distribution with 15 degrees of freedom. Intuitively the confidence interval of this method can be understood as the "maximum reach" an observable can have, whilst still being in the 90%-quantile of describing the experimental data.

[1] Only the upper limit is relevant, as the parameter is constraint by the data from below.

Finally, we compare the MCMC uncertainty estimation with the Hessian method [33]. Thus we employ a standard minimization algorithm and calculate the asymmetric error PDFs. For this purpose we need to set a tolerance for the error PDFs, which the Hessian method does not provide. Therefore we resort to the MCMC analysis and use the quantile estimation of the distribution of the χ^2-values and set $\Delta\chi^2_{\text{Hessian}} = \chi^2_{\text{max}}$. In fig. 3 we show the error PDFs for both methods for the u_v, d_v (left) and $\bar{u} + \bar{d}, g$ (right) distributions. From the ratio plots (lower panels) we can see that the error estimations agree on the right figures, whilst the cumulative method gives larger uncertainties on the left. This agrees with the marginal distributions of the parameters: The parameters corresponding to $\bar{u} + \bar{d}, g$ follow Gaussian distributions closely and can therefore be captured by the Hessian method. This no longer holds for the parameters of u_v, d_v. Here the marginal distributions differ from Gaussian significantly and are therefore not captured well by the Hessian method. Correspondingly the uncertainties do not agree, with the Hessian ones being markedly smaller, in some regions more than a factor of two. Even though the tolerance is in principle a free parameter of the Hessian method, increasing its value such that the uncertainty bands from the two different methods agree for valence distributions, would lead to overestimated uncertainty in the anti-quark and gluon distributions. In any case: Without the MCMC analysis this issue would have been very hard to catch as the marginal parameter distributions are not available in the Hessian method. Instead, one- or two-dimensional parameter scans are performed, which keep the remaining parameter values fixed and are therefore difficult to interpret correctly.

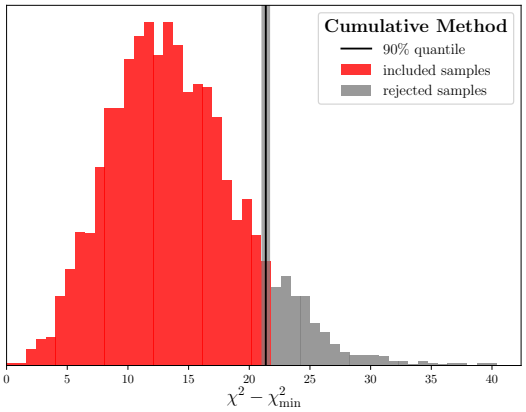

Figure 2: 90%-quantile estimation of the distribution of the χ^2-values.

We conclude that although a Markov Chain Monte Carlo analysis is computationally intensive, it yields the benefits of a more sophisticated statistical analysis tool: We obtain an independent procedure to estimate PDF uncertainties without approximations, which agrees with the Hessian method in regions, where its approximations hold and brings insights in regions, where the approximations break. Furthermore, this analysis can be used to estimate the tolerance.

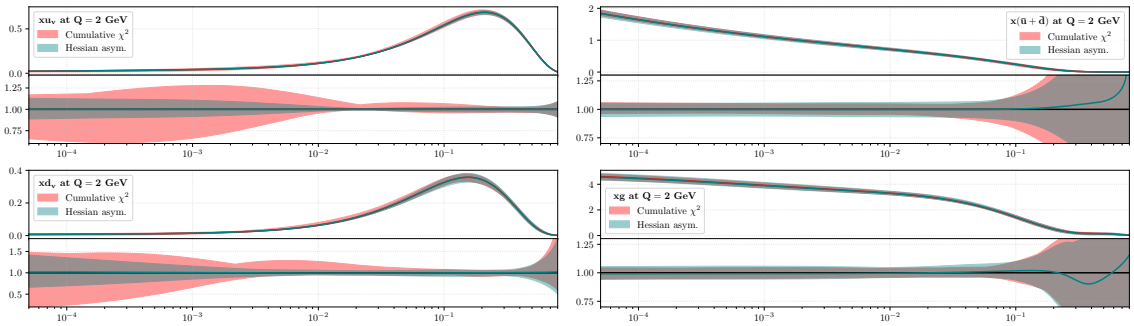

Figure 3: Comparison of PDF error estimations for the $u_v, d_v, \bar{d} + \bar{u}$ and gluon distribution using the Cumulative χ^2 (red) and Hessian (blue) method. The upper panels show the absolute PDFs and the bottom shows the ratio to the central value of the Cumulative χ^2 method.

Acknowledgements

Calculations (or parts of them) for this publication were performed on the HPC cluster PALMA II of the University of Münster, subsidised by the DFG (INST 211/667-1). P.R., T.J. and K.K. acknowledge support of the DFG through the Research Training Group GRK 2149. A.K. and N.D. are grateful for the support of Narodowe Centrum Nauki under grant no. 2019/34/E/ST2/00186.

References

[1] K. Kovařík, P.M. Nadolsky and D.E. Soper, *Hadronic structure in high-energy collisions*, *Rev. Mod. Phys.* **92** (2020) 045003 [1905.06957].

[2] N.T. Hunt-Smith, A. Accardi, W. Melnitchouk, N. Sato, A.W. Thomas and M.J. White, *Determination of uncertainties in parton densities*, *Phys. Rev. D* **106** (2022) 036003 [2206.10782].

[3] N.T. Hunt-Smith, W. Melnitchouk, F. Ringer, N. Sato, A.W. Thomas and M.J. White, *Accelerating Markov Chain Monte Carlo sampling with diffusion models*, *Comput. Phys. Commun.* **296** (2024) 109059 [2309.01454].

[4] F. Capel, R. Aggarwal, M. Botje, A. Caldwell, O. Schulz and A. Verbytskyi, *PartonDensity.jl: a novel parton density determination code*, 2401.17729.

[5] Y.G. Gbedo and M. Mangin-Brinet, *Markov chain Monte Carlo techniques applied to parton distribution functions determination: Proof of concept*, *Phys. Rev. D* **96** (2017) 014015 [1701.07678].

[6] T. Stavreva, F.I. Olness, I. Schienbein, T. Jezo, A. Kusina, K. Kovarik et al., *Heavy Quark Production in the ACOT Scheme at NNLO and N3LO*, *Phys. Rev. D* **85** (2012) 114014 [1203.0282].

[7] P. Risse, V. Bertone, T. Ježo, M. Klasen, K. Kovařík, F.I. Olness et al., *Fast evaluation of heavy-quark contributions to DIS in APFEL++*, in *30th International Workshop on Deep-Inelastic Scattering and Related Subjects*, 7, 2023 [2307.08269].

[8] P. Risse, V. Bertone, T. Jezo, K. Kovarik, F.I. Kusina, A. Olness and I. Schienbein, "Heavy Quark mass effects in charged current Deep-Inelastic Scattering at NNLO in the ACOT scheme." (in preparation).

[9] NNPDF collaboration, "applgrids." https://github.com/NNPDF/applgrids, 2017.

[10] V. Bertone, S. Carrazza and N.P. Hartland, *APFELgrid: a high performance tool for parton density determinations*, *Comput. Phys. Commun.* **212** (2017) 205 [1605.02070].

[11] NNPDF collaboration, "nnpdf." https://github.com/NNPDF/nnpdf, 2017.

[12] NNPDF collaboration, *Parton distributions from high-precision collider data*, *Eur. Phys. J. C* **77** (2017) 663 [1706.00428].

[13] H1, ZEUS collaboration, *Combination of measurements of inclusive deep inelastic $e^{\pm}p$ scattering cross sections and QCD analysis of HERA data*, *Eur. Phys. J. C* **75** (2015) 580 [1506.06042].

[14] CDF collaboration, *Measurement of $d\sigma/dy$ of Drell-Yan e^+e^- pairs in the Z Mass Region from $p\bar{p}$ Collisions at $\sqrt{s}$ = 1.96 TeV*, *Phys. Lett. B* **692** (2010) 232 [0908.3914].

[15] D0 collaboration, *Measurement of the Shape of the Boson Rapidity Distribution for $p\bar{p} \to Z/\gamma^* \to e^+e^- + X$ Events Produced at $\sqrt{s}$ of 1.96-TeV*, *Phys. Rev. D* **76** (2007) 012003 [hep-ex/0702025].

[16] BCDMS collaboration, *A High Statistics Measurement of the Proton Structure Functions F(2) (x, Q**2) and R from Deep Inelastic Muon Scattering at High Q**2*, *Phys. Lett. B* **223** (1989) 485.

[17] ATLAS collaboration, *Measurement of the transverse momentum and ϕ^*_η distributions of Drell–Yan lepton pairs in proton–proton collisions at $\sqrt{s}$ = 8 TeV with the ATLAS detector*, *Eur. Phys. J. C* **76** (2016) 291 [1512.02192].

[18] New Muon collaboration, *Measurement of the proton and deuteron structure functions, F2(p) and F2(d), and of the ratio sigma-L / sigma-T*, *Nucl. Phys. B* **483** (1997) 3 [hep-ph/9610231].

[19] CMS collaboration, *Measurement of the Z boson differential cross section in transverse momentum and rapidity in proton–proton collisions at 8 TeV*, *Phys. Lett. B* **749** (2015) 187 [1504.03511].

[20] CMS collaboration, *Measurement of the Differential and Double-Differential Drell-Yan Cross Sections in Proton-Proton Collisions at $\sqrt{s}$ = 7 TeV*, *JHEP* **12** (2013) 030 [1310.7291].

[21] LHCb collaboration, *Measurement of the forward Z boson production cross-section in pp collisions at $\sqrt{s}$ = 7 TeV*, *JHEP* **08** (2015) 039 [1505.07024].

[22] LHCb collaboration, *Measurement of forward W and Z boson production in pp collisions at $\sqrt{s}$ = 8 TeV*, *JHEP* **01** (2016) 155 [1511.08039].

[23] A. Accardi, L.T. Brady, W. Melnitchouk, J.F. Owens and N. Sato, *Constraints on large-x parton distributions from new weak boson production and deep-inelastic scattering data*, *Phys. Rev. D* **93** (2016) 114017 [1602.03154].

[24] A. Accardi, X. Jing, J.F. Owens and S. Park, *Light quark and antiquark constraints from new electroweak data*, *Phys. Rev. D* **107** (2023) 113005 [2303.11509].

[25] H. Haario, E. Saksman and J. Tamminen, *An adaptive metropolis algorithm*, *Bernoulli* **7** (2001) 223.

[26] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller and E. Teller, *Equation of State Calculations by Fast Computing Machines*, *The Journal of Chemical Physics* **21** (2004) 1087.

[27] W.K. Hastings, *Monte carlo sampling methods using markov chains and their applications*, *Biometrika* **57** (1970) 97.

[28] S. Brooks, A. Gelman, G.L. Jones and X.-L. Meng, *Handbook of Markov Chain Monte Carlo*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods, CRC Press (2011).

[29] C.J. Geyer, *Practical Markov Chain Monte Carlo*, *Statistical Science* **7** (1992) 473 .

[30] W.A. Link and M.J. Eaton, *On thinning of chains in mcmc*, *Methods in ecology and evolution* **3** (2012) 112.

[31] U. Wolff, *Monte Carlo errors with less errors*, *Computer Physics Communications* **156** (2004) 143.

[32] A. Putze, L. Derome, D. Maurin, L. Perotto and R. Taillet, *A Markov Chain Monte Carlo for Galactic Cosmic Ray physics: I. Method and results for the Leaky Box Model*, *Astron. Astrophys.* **497** (2009) 991 [0808.2437].

[33] J. Pumplin, D. Stump, R. Brock, D. Casey, J. Huston, J. Kalk et al., *Uncertainties of predictions from parton distribution functions. 2. The Hessian method*, *Phys. Rev. D* **65** (2001) 014013 [hep-ph/0101032].

PoS(DIS2024)054