Published for SISSA by Springer

Received: May 21, 2012
Revised: July 18, 2012
Accepted: July 24, 2012
Published: August 8, 2012

Study of Monte Carlo approach to experimental uncertainty propagation with MSTW 2008 PDFs

G. Watt[a] **and R.S. Thorne**[b]

[a] *Theory Group, Physics Department, CERN, CH-1211 Geneva 23, Switzerland*

[b] *Department of Physics and Astronomy, University College London, London, WC1E 6BT, U.K.*

E-mail: Graeme.Watt@cern.ch, thorne@hep.ucl.ac.uk

Abstract: We investigate the Monte Carlo approach to propagation of experimental uncertainties within the context of the established "MSTW 2008" global analysis of parton distribution functions (PDFs) of the proton at next-to-leading order in the strong coupling. We show that the Monte Carlo approach using replicas of the original data gives PDF uncertainties in good agreement with the usual Hessian approach using the standard $\Delta\chi^2 = 1$ criterion, then we explore potential parameterisation bias by increasing the number of free parameters, concluding that any parameterisation bias is likely to be small, with the exception of the valence-quark distributions at low momentum fractions x. We motivate the need for a larger tolerance, $\Delta\chi^2 > 1$, by making fits to restricted data sets and idealised consistent or inconsistent pseudodata. Instead of using data replicas, we alternatively produce PDF sets randomly distributed according to the covariance matrix of fit parameters including appropriate tolerance values, then we demonstrate a simpler method to produce an arbitrary number of random predictions on-the-fly from the existing eigenvector PDF sets. Finally, as a simple example application, we use Bayesian reweighting to study the effect of recent LHC data on the lepton charge asymmetry from W boson decays.

Keywords: QCD Phenomenology, Deep Inelastic Scattering (Phenomenology)

ArXiv ePrint: 1205.4024

Open Access

doi:10.1007/JHEP08(2012)052

Contents

1 Introduction

The parton distribution functions (PDFs) of the proton are best determined from global analysis of a wide variety of deep-inelastic scattering (DIS) and related hard-scattering data taken from both fixed-target experiments and colliders (HERA, the Tevatron, and most recently the LHC). Propagation of the experimental errors on the fitted data points to the uncertainties on the PDFs is a non-trivial task. The traditional Hessian method requires effective error inflation by a *tolerance* parameter to accommodate minor inconsistencies between the fitted data sets. This means that the PDF uncertainties cannot be considered to be statistically rigorous, despite the rôle of PDF uncertainties as an important (and sometimes dominant) source of theoretical uncertainty on predicted quantities, such as the cross sections for Drell-Yan processes or Higgs boson production at the Tevatron and LHC [1, 2]. Moreover, the number of fitted parameters for error propagation in the Hessian method must be kept sufficiently small to avoid large correlations, often requiring several parameters to be held fixed and thereby introducing a potential parameterisation bias. Some insight into these problems may be gained using Monte Carlo techniques [3, 4], recently used in conjunction with a neural-network parameterisation by the NNPDF Collaboration ([5], and references therein), where a large number $N_{\rm rep} \sim \mathcal{O}(10\text{–}1000)$ of fits are performed, each to a sample of replica pseudodata generated by shifting the original data points by random amounts dependent on the data errors. Then the PDF uncertainties can be calculated by simply taking the standard deviation of the resulting $N_{\rm rep}$ PDF sets.

JHEP08(2012)052

In this paper we make a first study of the Monte Carlo approach to experimental error propagation within the context of the established "MSTW 2008" PDF determination [6]. We retain the usual functional-form parameterisation and least-squares χ^2-minimisation (using the Levenberg-Marquardt algorithm) rather than moving to the neural-network parameterisation and genetic-algorithm χ^2-minimisation of the NNPDF approach [5]. We focus on the most widely-used PDF determination at next-to-leading order (NLO) in the strong coupling α_S, although the results would be expected to be similar at leading-order (LO) and at next-to-next-to-leading order (NNLO). Moreover, to avoid complications associated with simultaneously fitting α_S with the PDFs, throughout this paper we keep the value of $\alpha_S(M_Z^2)$ held fixed at the MSTW 2008 NLO best-fit value. First in section 2 we describe the Monte Carlo approach using data replicas and compare results to the usual Hessian method, then in section 3 we explore potential parameterisation bias by increasing the number of free parameters. We then motivate the need for a tolerance parameter by fitting restricted data sets in section 4 and by fitting idealised pseudodata in section 5. In section 6 we explain how to produce PDF sets randomly distributed in the space of parameters rather than in the space of data, which allows the inclusion of a suitable tolerance. As an example application of these random PDFs, in section 7 we demonstrate the use of Bayesian reweighting to study the effect of recent LHC data on the $W \to \ell\nu$ charge asymmetry [7, 8]. Finally, we conclude in section 8.

2 Comparison of Hessian and Monte Carlo uncertainties

2.1 Recap of the Hessian method

The basic procedure for propagating experimental uncertainties in global PDF analyses using the Hessian method is discussed in detail in refs. [6, 9–11]. Here, we briefly review the salient points. We assume that the global goodness-of-fit quantity, χ^2_{global}, is quadratic about the global minimum, which has n best-fit parameters $\{a_1^0, \ldots, a_n^0\}$. In this case we can write

$$\Delta\chi^2_{\text{global}} \equiv \chi^2_{\text{global}} - \chi^2_{\text{min}} = \sum_{i,j=1}^{n} H_{ij}(a_i - a_i^0)(a_j - a_j^0)\,, \tag{2.1}$$

where the Hessian matrix H has components

$$H_{ij} = \frac{1}{2}\frac{\partial^2\chi^2_{\text{global}}}{\partial a_i \partial a_j}\bigg|_{\text{min}}. \tag{2.2}$$

It is convenient to diagonalise the covariance (inverse Hessian) matrix, $C \equiv H^{-1}$, also known as the error matrix, and work in terms of the eigenvectors and eigenvalues. Since the covariance matrix is symmetric it has a set of orthonormal eigenvectors $\vec{v}_k$ defined by

$$\sum_{j=1}^{n} C_{ij} v_{jk} = \lambda_k v_{ik}\,, \tag{2.3}$$

where λ_k is the kth eigenvalue and v_{ik} is the ith component of the kth orthonormal eigenvector ($k = 1, \ldots, n$). The parameter displacements from the global minimum can be

JHEP08(2012)052

expanded in a basis of rescaled eigenvectors $e_{ik} \equiv \sqrt{\lambda_k} v_{ik}$, that is,

$$a_i - a_i^0 = \sum_{k=1}^{n} e_{ik} z_k \,. \tag{2.4}$$

Then it can be shown, using the orthonormality of $\vec{v}_k$, that eq. (2.1) reduces to

$$\chi^2_{\text{global}} = \chi^2_{\text{min}} + \sum_{k=1}^{n} z_k^2 \,, \tag{2.5}$$

that is, $\sum_{k=1}^{n} z_k^2 \leq T^2$ is the interior of a hypersphere of radius T. Pairs of eigenvector PDF sets $S_k^{\pm}$ can then be produced to span this hypersphere, with parameters given by

$$a_i(S_k^{\pm}) = a_i^0 \pm t\, e_{ik} \,. \tag{2.6}$$

In the quadratic approximation, $t = T \equiv (\Delta\chi^2_{\text{global}})^{1/2}$, but particularly for the larger eigenvalues λ_k there are significant deviations from the ideal quadratic behaviour, so in general t is adjusted iteratively to give the desired value of T. Then asymmetric PDF uncertainties on a quantity F, which may be an individual PDF at particular values of x and Q^2, or a derived quantity such as a cross section, can be calculated with the following "master equations":

$$(\Delta F)_+ = \sqrt{\sum_{k=1}^{n} \left\{ \max\left[F(S_k^+) - F(S_0),\, F(S_k^-) - F(S_0),\, 0\right]\right\}^2} \,, \tag{2.7}$$

$$(\Delta F)_- = \sqrt{\sum_{k=1}^{n} \left\{ \max\left[F(S_0) - F(S_k^+),\, F(S_0) - F(S_k^-),\, 0\right]\right\}^2} \,, \tag{2.8}$$

where S_0 is the central PDF set. Symmetric PDF uncertainties can be calculated with

$$\Delta F = \frac{1}{2}\sqrt{\sum_{k=1}^{n} \left[F(S_k^+) - F(S_k^-)\right]^2} \,. \tag{2.9}$$

Ideally, with the standard "parameter-fitting" criterion [12], we would expect the errors to be given by the choice of tolerance $T = 1$ for the 68% (one-sigma) confidence-level (C.L.) limit or $T = 1.64$ for the 90% C.L. limit [13]. This criterion is appropriate if fitting consistent data sets with ideal Gaussian errors to a well-defined theory. However, in practice, there are some inconsistencies between the independent fitted data sets, and unknown experimental and theoretical uncertainties, so the parameter-fitting criterion is not appropriate for global PDF analyses. Historically, the CTEQ [10] and MRST [11] groups defined 90% C.L. uncertainties using $T = \sqrt{100}$ and $T = \sqrt{50}$, respectively. Instead, the "MSTW 2008" analysis [6] introduced a new "dynamic" determination of the tolerance, chosen separately for each eigenvector direction according to a "hypothesis-testing" criterion [12] to maintain an adequate description of each individual data set in the global fit. Therefore, the distance t in eq. (2.6) was replaced by $t_k^{\pm}$, adjusted to give the desired $T_k^{\pm}$, with an average value of $\langle t_k^{\pm} \rangle \approx \langle T_k^{\pm} \rangle \approx 3$ for 68% C.L. uncertainties, and $\langle t_k^{\pm} \rangle \approx \langle T_k^{\pm} \rangle \approx 6$ for 90% C.L. uncertainties; see figure 10 of ref. [6] for the individual $T_k^{\pm}$ values in the MSTW 2008 NLO fit.

JHEP08(2012)052

2.2 Generation of Monte Carlo replica sets

We generate replica data sets with the central values shifted according to

$$D_{m,i} \to \left(D_{m,i} + R_{m,i}^{\text{uncorr.}}\, \sigma_{m,i}^{\text{uncorr.}} + \sum_{k=1}^{N_{\text{corr.}}} R_{m,k}^{\text{corr.}}\, \sigma_{m,k,i}^{\text{corr.}} \right) \cdot \left(1 + R_m^{\mathcal{N}}\, \sigma_m^{\mathcal{N}}\right). \tag{2.10}$$

Here, "m" labels a particular data set, or a combination of data sets, with a common (fitted) normalisation $\mathcal{N}_m$, "i" labels the individual data points in that data set, and "k" labels the individual correlated systematic errors for a particular data set. The individual data points $D_{m,i}$ have uncorrelated (statistical and systematic) errors $\sigma_{m,i}^{\text{uncorr.}}$ and correlated systematic errors $\sigma_{m,k,i}^{\text{corr.}}$. Treating the correlated errors as uncorrelated leads to the alternative form used for most of the data sets in the MSTW 2008 fit:

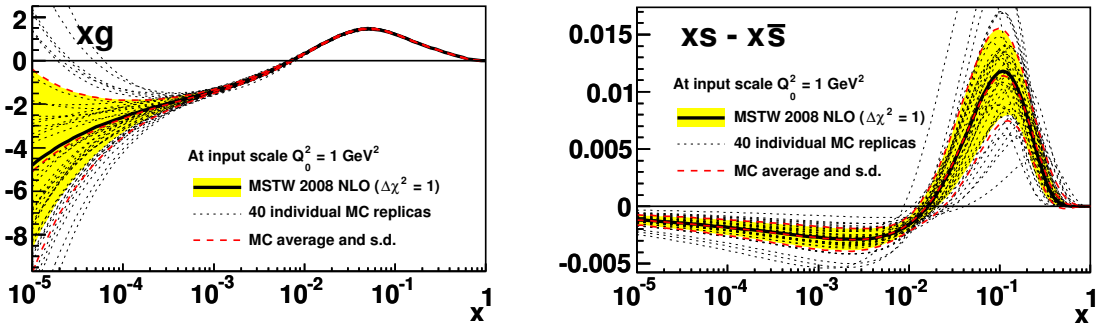
$$D_{m,i} \to \left(D_{m,i} + R_{m,i}^{\text{uncorr.}}\, \sigma_{m,i}^{\text{tot.}} \right) \cdot \left(1 + R_m^{\mathcal{N}}\, \sigma_m^{\mathcal{N}}\right), \tag{2.11}$$

where the total error is simply obtained by adding all errors (except normalisation) in quadrature,

$$\left(\sigma_{m,i}^{\text{tot.}}\right)^2 = \left(\sigma_{m,i}^{\text{uncorr.}}\right)^2 + \sum_{k=1}^{N_{\text{corr.}}} \left(\sigma_{m,k,i}^{\text{corr.}}\right)^2. \tag{2.12}$$

We shift the data points in a way to be as consistent as possible with the χ^2 definition used in the MSTW 2008 fit [6]. The random numbers $R_{m,i}^{\text{uncorr.}}$ or $R_{m,k}^{\text{corr.}}$ are obtained from a Gaussian distribution of mean zero and variance one. A complication arises with the treatment of normalisation uncertainties in the MSTW 2008 analysis, where a *quartic* penalty term was used in the χ^2 definition instead of the usual quadratic penalty term, cf. eqs. (35) and (37) of ref. [6]. This modification was made to discourage large normalisation shifts in the fit. It was partly motivated by claims (see section 6.7.4 on "Normalizations", pg. 170 in ref. [14]) that, for many experiments, quoted normalisation uncertainties represent the limits of a box-shaped distribution rather than the standard deviation of a Gaussian distribution; see further discussion in section 5.2.1 of ref. [6]. The quartic χ^2 penalty term is small if the fitted normalisation $\mathcal{N}_m \in [1-\sigma_m^{\mathcal{N}}, 1+\sigma_m^{\mathcal{N}}]$, then it rises rapidly outside this range, with the effect that the normalisation uncertainty is perhaps closer to being described by a box-shaped distribution than by a Gaussian distribution (which would correspond to a quadratic χ^2 penalty term). Therefore, by default we take $R_m^{\mathcal{N}}$ in eqs. (2.10) and (2.11) to be uniformly distributed in the interval $(-1, 1)$, so that the normalisation $\mathcal{N}_m$ is uniformly distributed in the interval $(1-\sigma_m^{\mathcal{N}}, 1+\sigma_m^{\mathcal{N}})$. However, we have also looked at the effect of obtaining $R_m^{\mathcal{N}}$ from a Gaussian distribution or alternatively simply fixing $R_m^{\mathcal{N}} = 0$, i.e. the case of fixed data set normalisations. As expected, fixing normalisations in the data replicas generally gives slightly smaller PDF uncertainties, while assuming normalisation uncertainties to be Gaussian gives larger PDF uncertainties, particularly for the up-valence distribution. However, it is perhaps inconsistent to assume Gaussian uncertainties in the replica generation with a quartic penalty term in the χ^2: changing to a quadratic penalty term would allow more freedom in the fitted normalisations and so the PDF parameters would move less, likely reducing the PDF uncertainty compared to the

JHEP08(2012)052



(a) (b)

Figure 1. Comparison of Hessian and Monte Carlo results at the input scale of $Q_0^2 = 1\,\mathrm{GeV}^2$ for the (a) gluon distribution and (b) strange asymmetry. Both results allow $n = 20$ free PDF parameters and do not apply a tolerance (i.e. $T = 1$ in the Hessian case). The best-fit (solid curves) and Hessian uncertainty (shaded region) are in good agreement with the average and standard deviation (thick dashed curves) of the $N_{\mathrm{rep}} = 40$ Monte Carlo replica PDF sets (thin dotted curves).

case of a quartic penalty term. The default treatment of uniform $R_m^{\mathcal{N}} \in (-1, 1)$ is probably reasonable and is closer to the treatment of normalisation uncertainties in the χ^2 definition than a Gaussian $R_m^{\mathcal{N}}$. The Hessian error propagation via eigenvector PDF sets includes theoretical uncertainties on the hadronisation corrections for the CDF jet data (treated as a correlated systematic) and the small modification for the nuclear corrections (r_1, r_2, r_3) [6]. It is currently not obvious how to treat these theoretical uncertainties in the replica generation, so the effect on PDF uncertainties will be assumed to be small.

We perform a separate PDF fit to each replica data set, then we can take the average $\langle F \rangle$ and standard deviation ΔF of an observable F calculated with each PDF replica set, $\mathcal{S}_k$ ($k = 1, \ldots, N_{\mathrm{rep}}$), that is,

$$\langle F \rangle = \frac{1}{N_{\mathrm{rep}}} \sum_{k=1}^{N_{\mathrm{rep}}} F(\mathcal{S}_k)\,, \tag{2.13}$$

$$\Delta F = \sqrt{\frac{N_{\mathrm{rep}}}{N_{\mathrm{rep}} - 1} \big(\langle F^2 \rangle - \langle F \rangle^2 \big)}\,. \tag{2.14}$$

The number of replicas N_{rep} is arbitrary, but in all cases we choose to generate $N_{\mathrm{rep}} = 40$ replica PDF sets, where this number is chosen to be equal to the number of eigenvector PDF sets mostly for practical reasons, i.e. to demonstrate that the implementation of the Monte Carlo (MC) method does not necessarily require more computer resources than the Hessian method. It could easily be increased in further studies, but first indications are that $N_{\mathrm{rep}} = 40$ is sufficiently large to avoid significant fluctuations. To allow a fair comparison with the existing Hessian eigenvector PDF sets, we take $n = 20$ free PDF parameters, i.e. 8 PDF parameters are held fixed at their global best-fit values, and we do not apply a tolerance, i.e. we use the Hessian eigenvector PDF sets corresponding to $T = 1$ (see section 6.6 of ref. [6]). In figure 1 we show the input gluon distribution and strange asymmetry for the $N_{\mathrm{rep}} = 40$ MC replica PDF sets (thin dotted curves), and their average

JHEP08(2012)052

and standard deviation (thick dashed curves), and we compare to the best-fit and Hessian uncertainty (solid curves and shaded region). We find good agreement of the Hessian and MC results at all x and Q^2 values, and for all parton flavours, as will be demonstrated more explicitly in the next section.

Similar comparisons between Hessian and MC results were performed in a fit only to the H1 data from HERA I on neutral- and charged-current $e^{\pm}p$ cross sections [15], but it is still reassuring that we find a similar good agreement in the context of a more complicated global fit. On the other hand, in section 6.6 of ref. [6] we also performed a fit to a reduced data set consisting of a limited number of inclusive DIS data sets (BCDMS, NMC, H1, ZEUS) with fairly conservative cuts of $Q^2 \geq 9\,\text{GeV}^2$ and $W^2 \geq 15\,\text{GeV}^2$, where eigenvector PDF sets were produced with $n = 16$ free PDF parameters for both a dynamic tolerance and with $T = 1$. We find that there are some differences between the MC results with $n = 16$ free PDF parameters and the Hessian results with $T = 1$. The approximate equivalence between the Hessian and MC methods may break down, therefore, when fitting a limited selection of discrepant data sets that are insufficient to unambiguously constrain all fitted parameters.

3 Investigation of potential parameterisation bias

Recall the MSTW 2008 NLO PDF parameterisation at the input scale $Q_0^2 = 1\,\text{GeV}^2$ [6]:

$$xu_v \equiv xu - x\bar{u} = A_u\, x^{\boldsymbol{\eta_1}}(1-x)^{\boldsymbol{\eta_2}}(1 + \boldsymbol{\epsilon_u}\sqrt{x} + \gamma_u\, x)\,, \tag{3.1}$$

$$xd_v \equiv xd - x\bar{d} = A_d\, x^{\boldsymbol{\eta_3}}(1-x)^{\boldsymbol{\eta_4}}(1 + \boldsymbol{\epsilon_d}\sqrt{x} + \gamma_d\, x)\,, \tag{3.2}$$

$$xS \equiv 2x\bar{u} + 2x\bar{d} + xs + x\bar{s} = \boldsymbol{A_S}\, x^{\delta_S}(1-x)^{\boldsymbol{\eta_S}}(1 + \boldsymbol{\epsilon_S}\sqrt{x} + \gamma_S\, x)\,, \tag{3.3}$$

$$x\Delta \equiv x\bar{d} - x\bar{u} = \boldsymbol{A_\Delta}\, x^{\boldsymbol{\eta_\Delta}}(1-x)^{\eta_S+2}(1 + \boldsymbol{\gamma_\Delta}\, x + \delta_\Delta\, x^2)\,, \tag{3.4}$$

$$xg = A_g\, x^{\boldsymbol{\delta_g}}(1-x)^{\boldsymbol{\eta_g}}(1 + \epsilon_g\sqrt{x} + \gamma_g\, x) + A_{g'}\, x^{\boldsymbol{\delta_{g'}}}(1-x)^{\boldsymbol{\eta_{g'}}}\,, \tag{3.5}$$

$$xs + x\bar{s} = \boldsymbol{A_+}\, x^{\delta_S}\,(1-x)^{\boldsymbol{\eta_+}}(1 + \epsilon_S\sqrt{x} + \gamma_S\, x)\,, \tag{3.6}$$

$$xs - x\bar{s} = \boldsymbol{A_-}\, x^{0.2}(1-x)^{\boldsymbol{\eta_-}}(1 - x/x_0)\,. \tag{3.7}$$

The parameters A_u, A_d, A_g and x_0 were fixed by enforcing number- and momentum-sum rule constraints, while the other parameters were allowed to go free to determine the overall best fit. The 20 highlighted (red) parameters were those allowed to go free when producing the eigenvector PDF sets, where the other 8 (blue) parameters were held fixed, as for the MC results in the previous section. However, this is not in fact necessary in the MC approach where it is only needed to find the best fit for each replica data set, and the Hessian matrix is not used for error propagation. Therefore, we can perform MC replica fits with all 28 free parameters to examine the effect on PDF uncertainties of the greater freedom in parameterisation, and to explore the extent that the Hessian uncertainties are limited by the restricted parameterisation.

Recall [6] that the reason to freeze several parameters before applying the Hessian method was to reduce the large correlations between some parameters, which would lead

JHEP08(2012)052

to severe breaking of the quadratic behaviour of $\Delta\chi^2$, meaning that linear error propagation would not be applicable. (A similar procedure was used in the CTEQ global fits; see, for example, section 5 of ref. [16].) We observed some departure from the ideal quadratic behaviour of $\Delta\chi^2$ even with only 20 parameters; see figures 5 and 6 of ref. [6]. However, even with all 28 parameters free, the Hessian matrix is generally still positive-definite (has positive eigenvalues) and therefore we can still be relatively confident that the best fit is correctly determined. Note that we use the Levenberg-Marquardt algorithm for χ^2-minimisation, which combines the advantages of the inverse-Hessian method and the steepest-descent method, and therefore simply finding the best fit is less reliant on accurate knowledge of the Hessian matrix compared to subsequent error propagation using the Hessian method.

In figure 2 we show percentage uncertainties at the input scale $Q_0^2 = 1\,\text{GeV}^2$, and in figure 3 we show percentage uncertainties after evolving to $Q^2 = (100\,\text{GeV})^2$. We show only the uncertainties since the MC average is very close to the Hessian best-fit, with residual differences likely explained by statistical fluctuations. Again the MC uncertainties with $n = 20$ input PDF parameters are in good agreement with the Hessian uncertainties with $\Delta\chi^2 = 1$, and both are much smaller than the 68% C.L. uncertainties including the dynamic tolerance. We show the effect of moving to $n = 28$ input PDF parameters, which gives significantly larger u_v and d_v uncertainties mainly at low x values (removing some unusual shapes in the x dependence) and slightly larger gluon uncertainties around $x \sim 0.05$ in figure 2(f) and around $x \sim 0.01$ in figure 3(f), but in all cases the MC uncertainties are still much smaller than the Hessian uncertainties at 68% C.L. One can see from the equations above that in going from a total of $20 \to 28$ input PDF parameters, the number of free parameters for both xu_v and xd_v goes from $3 \to 4$, for xS $(\equiv 2x\bar{u} + 2x\bar{d} + xs + x\bar{s})$ goes from $3 \to 5$, and for xg goes from $4 \to 7$. While there is perhaps some degree of parameterisation bias in the valence-quark distributions, the insensitivity of the sea-quark and gluon distributions to the relatively large increase in the number of free parameters suggests that parameterisation bias is likely to be small in those cases. Of course, an exception is the strange-quark and -antiquark distributions which are certainly constrained by the choice of parameterisation outside the limited data region ($0.01 \lesssim x \lesssim 0.2$) of the CCFR/NuTeV dimuon cross sections. For example, the low-x behaviour of s and $\bar{s}$ is assumed to be the same as for $\bar{u}$ and $\bar{d}$, as suggested by arguments based on both Regge theory and perturbative QCD (see discussion in section 6.5.5 of ref. [6]).

The study of potential parameterisation bias presented here is indicative rather than exhaustive. It could be followed up by a more involved study, for example, using Chebyshev polynomials along the lines of refs. [17, 18]. However, switching to an extremely flexible parameterisation brings the danger of fitting the statistical fluctuations of the data unless some method is used to enforce smoothness. We note that the limiting power-law behaviour as $x \to 0$ and $x \to 1$ is well-motivated by Regge theory and counting rules, respectively, and it is difficult to perceive a sensible alternative. More discussion and justification for the MSTW 2008 input parameterisation was given in section 6.5 of ref. [6].

JHEP08(2012)052

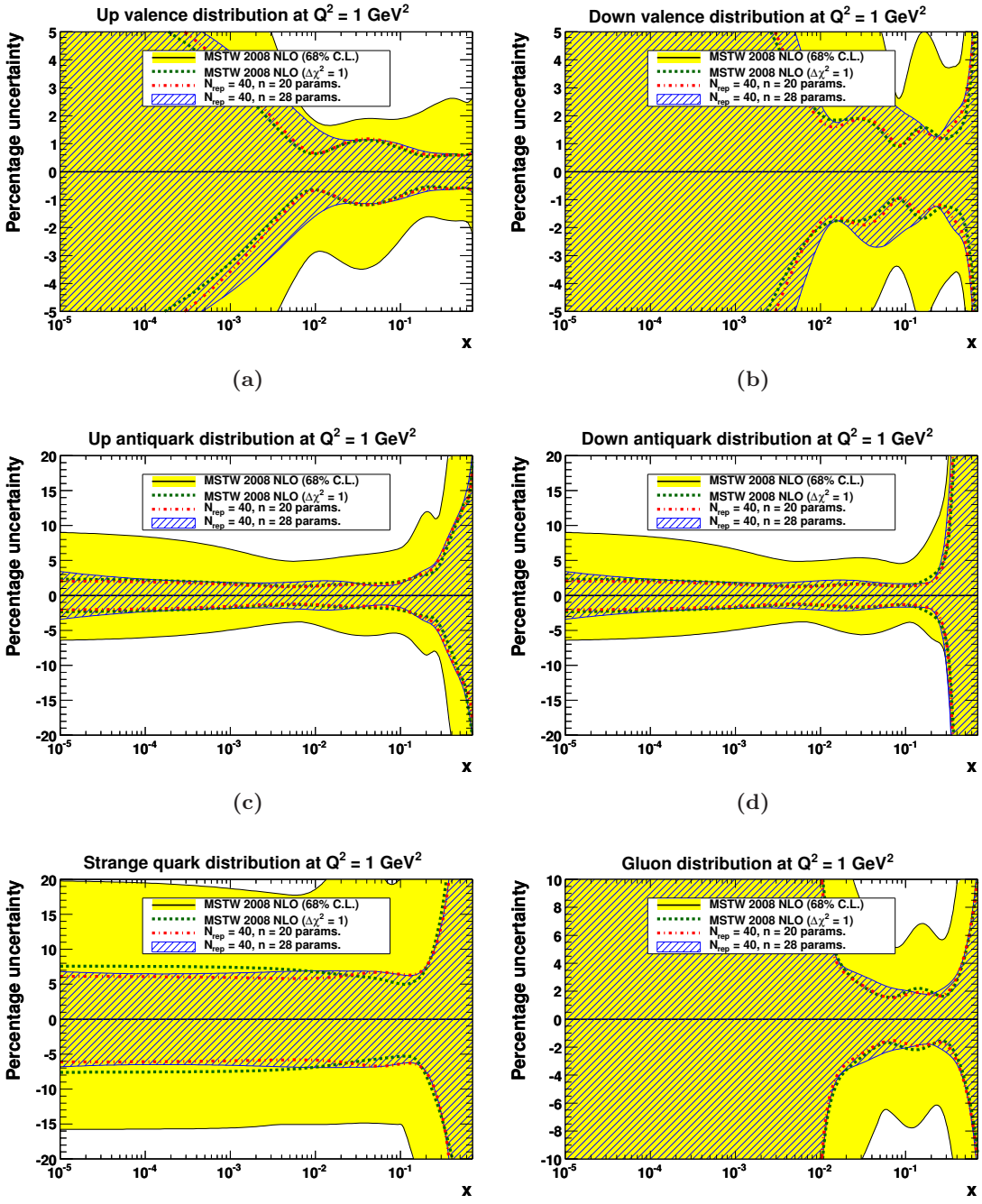


Figure 2. Effect of $n = 20 \to 28$ parameters on percentage PDF uncertainties at $Q^2 = 1\,\mathrm{GeV}^2$.

JHEP08(2012)052

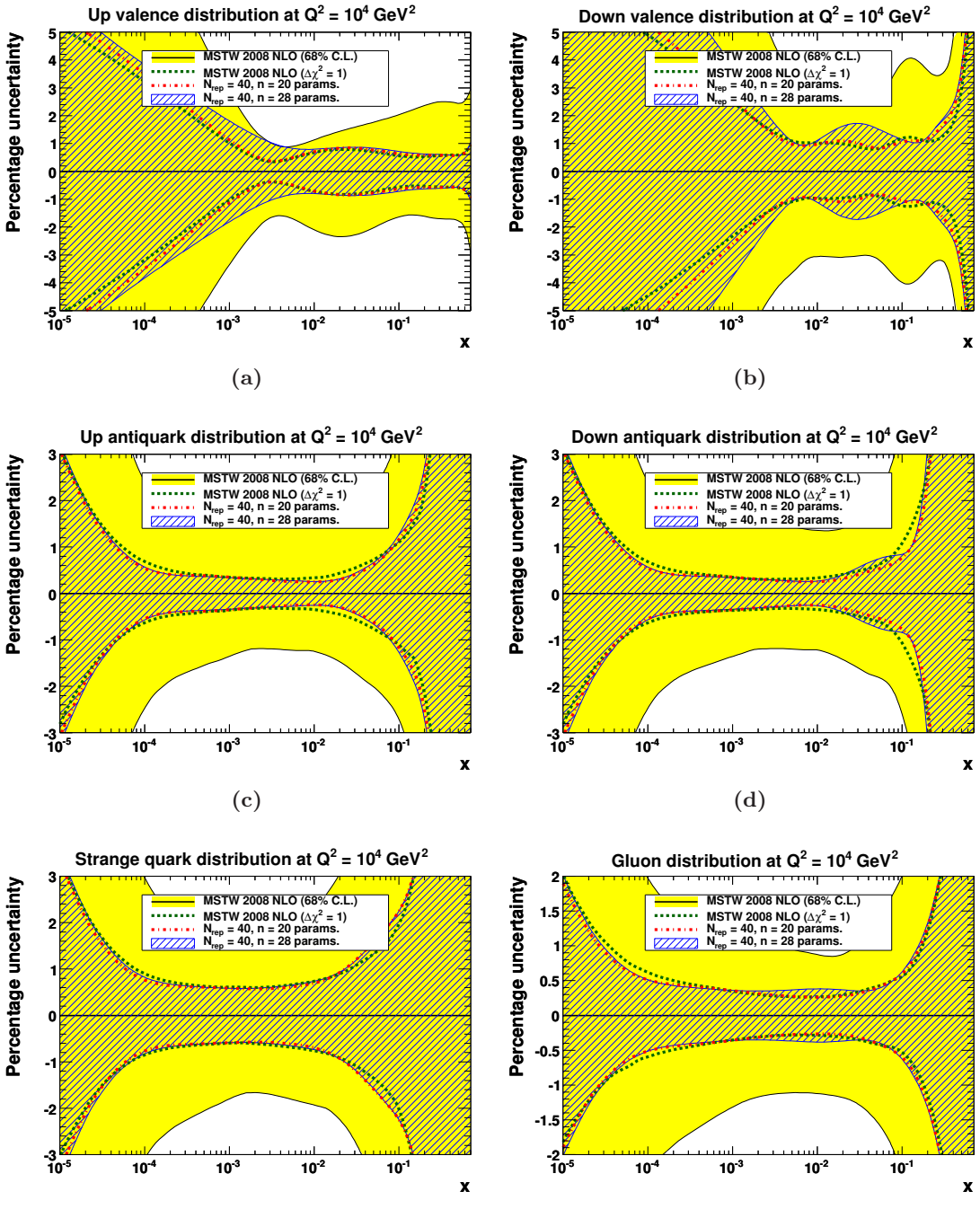


Figure 3. Effect of $n = 20 \to 28$ parameters on percentage PDF uncertainties at $Q^2 = (100\,\mathrm{GeV})^2$.

JHEP08(2012)052

4 Fits to restricted data sets using data replicas

Although we see little evidence for significant parameterisation bias in the MSTW 2008 *global* fit, this might not be true for some "non-global" fits which tend to be constrained by parameterisation choices in the absence of relevant data. For example, the input parameterisation at $Q_0^2 = 1.9\,\text{GeV}^2$ in the HERAPDF1.0 [19] or HERAPDF1.5 NLO [20] analyses takes the form:

$$\begin{aligned}
xu_v &= A_{u_v}\, x^{B_{q_v}} (1-x)^{C_{u_v}} (1 + E_{u_v}\, x^2)\,, \\
xd_v &= A_{d_v}\, x^{B_{q_v}} (1-x)^{C_{d_v}}\,, \\
x\bar{u} &= A_{\bar{q}}\, x^{B_{\bar{q}}} (1-x)^{C_{\bar{u}}}\,, \\
x\bar{d} &= A_{\bar{q}}\, x^{B_{\bar{q}}} (1-x)^{C_{\bar{d}}}\,, \\
x\bar{s} &= 0.45\, x\bar{d}\,, \\
xs &= x\bar{s}\,, \\
xg &= A_g\, x^{B_g} (1-x)^{C_g}\,.
\end{aligned}$$

There are only 10 parameters used to obtain the central fit and "experimental" uncertainties, although the more recent HERAPDF1.5 NNLO [21] analysis introduces 4 more parameters (2 for g and 1 each for u_v, d_v). The HERAPDF analyses additionally include "model" and "parameterisation" uncertainties that can be much larger than the "experimental" uncertainties. For example, quantities sensitive to the high-x gluon distribution have a very large "model" uncertainty in the HERAPDF1.5 NNLO analysis due to variation of the minimum Q^2 cut [22]. Nevertheless, it is interesting to investigate the potentially more realistic constraint arising only from HERA data with a flexible parameterisation; see also similar studies by the NNPDF Collaboration [23]. This would be difficult to achieve in the Hessian method where several parameters would need to be held fixed to use the covariance matrix for error propagation, but it is straightforward using the MC method. We fit subsets of the global data included in the MSTW 2008 NLO analysis [6], specifically (i) *excluding* all HERA data (neutral-current $e^{\pm}p$ and charged-current $e^{+}p$ cross sections, F_2^{charm}, and inclusive jet production in DIS), (ii) including *only* HERA data, (iii) performing a "collider" fit meaning data from HERA and the Tevatron (inclusive jet production, the $W \to \ell\nu$ charge asymmetry, and the Z rapidity distribution) with no fixed-target data. The HERA-only fit uses the older separate H1 and ZEUS inclusive cross sections compared to the more precise combined HERA I data [19] used in the HERAPDF fits. On the other hand, the public HERAPDF fits [19–21] do not use data on F_2^{charm} or jet production. In all cases we use the MC method with $n = 28$ free parameters wherever possible. However, for the HERA-only and HERA+Tevatron fits, there is no constraint at all on the strange asymmetry since the CCFR/NuTeV dimuon cross sections are missing, so we fix $s - \bar{s}$ at the global best-fit value, leaving $n = 26$ free parameters. The percentage uncertainties on the PDFs at $Q^2 = (100\,\text{GeV})^2$ from the various fits are shown in figure 4. The results reflect what might naïvely be expected. For example, removing HERA data gives a huge increase in the small-x uncertainties for the sea-quarks and gluon, but the valence-quark uncertainties are almost unchanged. With only HERA data, the gluon and antiquarks are

JHEP08(2012)052

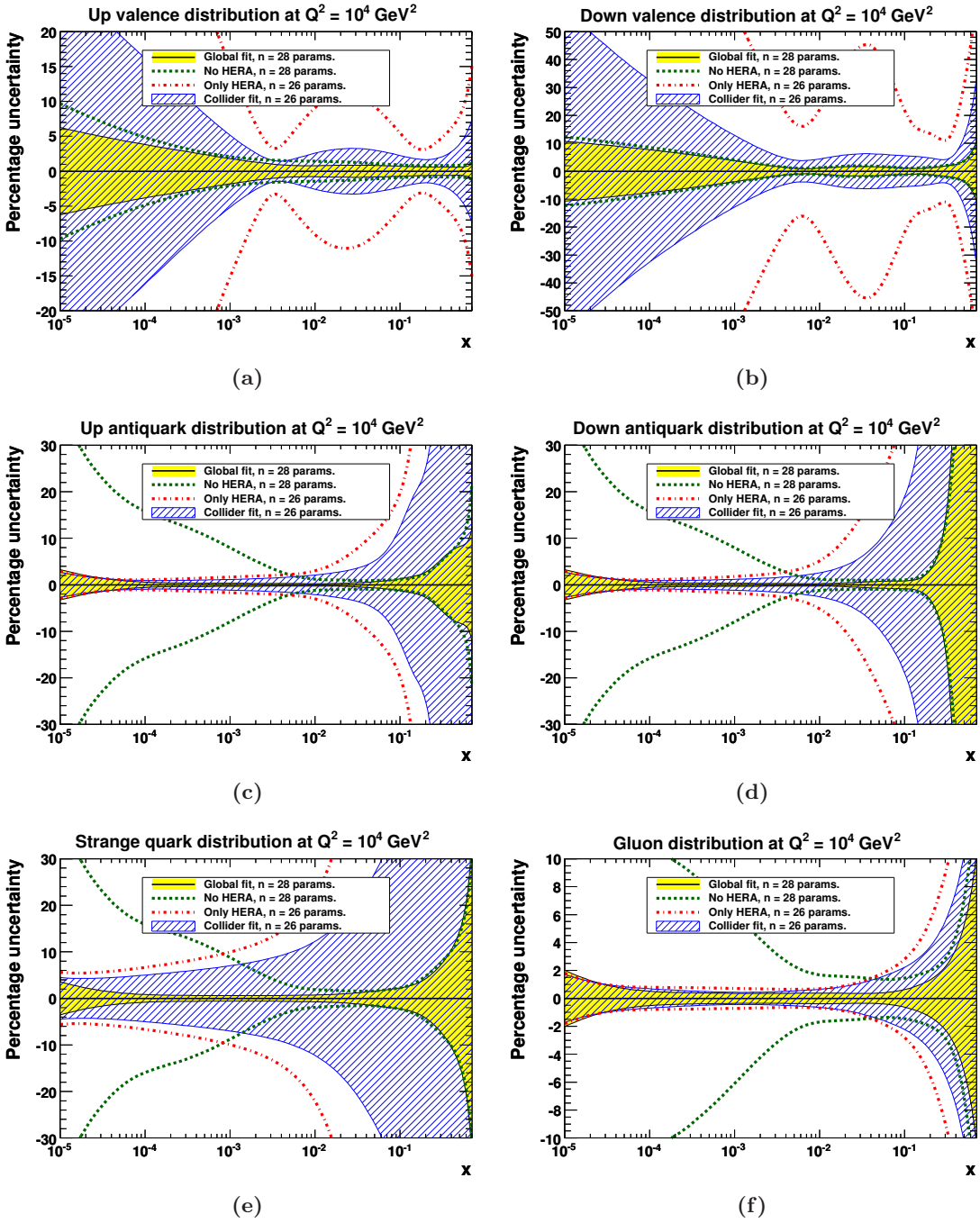


Figure 4. Effect on percentage PDF uncertainties of fitting subsets of MSTW 2008 global data.

still well-constrained at small x, but not at large x, and there are huge uncertainties in the valence- and strange-quark distributions. Adding the Tevatron data helps, but the collider-only uncertainty is still much larger than in the global fit, so really we need data from HERA, the Tevatron *and* the fixed-target experiments to get a meaningful result. The corresponding ratios to the global fit are shown in figure 5. Here, we see that the uncer-

JHEP08(2012)052

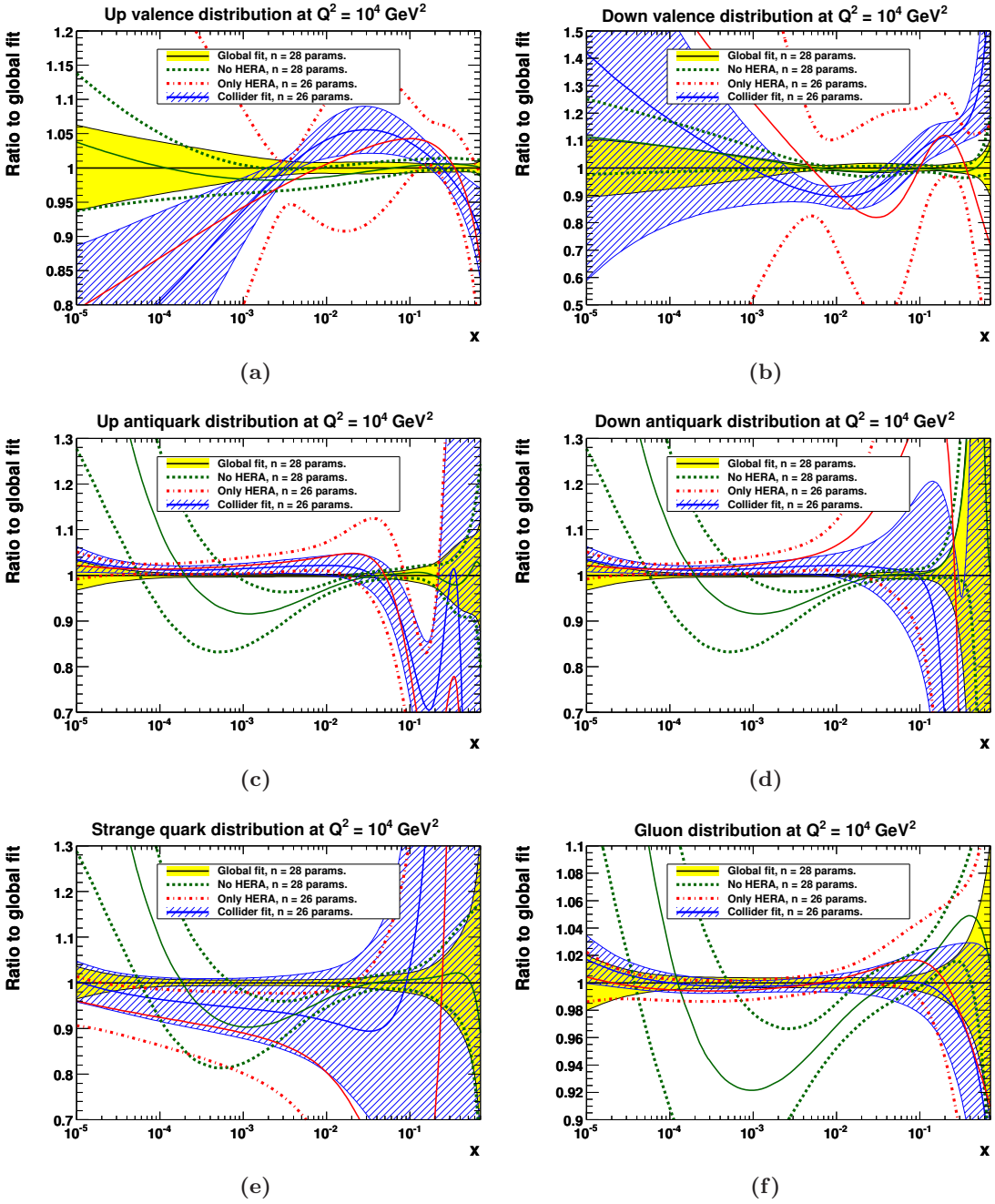


Figure 5. Effect on PDFs of fitting subsets of MSTW 2008 global data.

tainty bands from fits to subsets of the global data do not always overlap with those from the global fit, implying some tension between the different data sets, and suggesting that some kind of error inflation (or *tolerance*) is necessary. A similar exercise was performed in the MSTW 2008 paper [6] to a "reduced" data set, with a slightly more constrained parameterisation, and we find similar results if fitting the same "reduced" data set using the MC method.

JHEP08(2012)052

5 Fits to idealised consistent and inconsistent pseudodata

As a further exercise to examine potential data set inconsistency within the global fit, we generate idealised pseudodata from the best-fit theory predictions, i.e. we replace $D_{m,i}$ by $T_{m,i}$ on the right-hand side of eqs. (2.10) and (2.11), where $T_{m,i}$ are the theory predictions evaluated using the global best-fit parameters. The pseudodata are then simply given by the best-fit theory predictions with appropriate Gaussian noise added, and with uncertainties given by the genuine data uncertainties. We can then introduce deliberate inconsistencies into this idealised pseudodata and investigate the effect on the fitted PDFs. We choose the following deliberate inconsistencies, intended to simulate realistic, if somewhat large, incompatibilities that could potentially be present in the genuine data:

- We introduce a Q^2-dependent offset for the H1 and ZEUS inclusive neutral-current reduced cross sections, such that the pseudodata are multiplied by a factor of $\{1 \pm 0.005 \log[Q^2/(10\,\text{GeV}^2)]\}$, with the "+" sign for H1 and the "−" sign for ZEUS.

- We generate the pseudodata for the CDF and DØ inclusive jet cross sections with a scale choice $\mu_R = \mu_F = p_T/2$, but fit it with $\mu_R = \mu_F = p_T$.

- We normalise the CCFR/NuTeV dimuon cross sections downwards by 10%.

- We normalise the NuTeV/CHORUS xF_3 structure functions upwards by 5%.

- We introduce a rapidity-dependent offset for the CDF Z rapidity distribution, such that the pseudodata are multiplied by a factor of $(1 + 0.03\, y_Z)$.

- We introduce an x-dependent offset for the BCDMS/NMC/SLAC/E665 deuteron structure functions, intended to mimic a possible deuteron correction, such that the F_2^d data are multiplied by a factor of

$$f(x) = \begin{cases} (1+0.005)[1 - 0.003\log^2(x/x_1)] : & x < x_1 \\ (1+0.005)[1 - 0.018\log^2(x/x_1) + 3\cdot 10^{-8}\log^{20}(x/x_1)] : & x \geq x_1 \end{cases},$$

  where $x_1 = \exp(-2.5) \simeq 0.0821$.

- We introduce a Q^2-dependent offset for the BCDMS F_2^p and F_2^d structure functions, such that the pseudodata are multiplied by a factor of $\{1 + 0.01\log[Q^2/(1\,\text{GeV}^2)]\}$.

In figures 6 and 7 we show the effect of fitting the genuine data, then the consistent or inconsistent idealised pseudodata, in each case using MC error propagation with $N_{\text{rep}} = 40$ replica data sets and $n = 20$ input PDF parameters, and we compare to the standard MSTW 2008 NLO fit with dynamic tolerance. Despite the central values of the PDFs from the inconsistent fit shifting by significant amounts, the percentage uncertainties in figure 7 are remarkably almost identical whether fitting either the genuine data, the consistent pseudodata or the inconsistent pseudodata. The MC fit to perfectly consistent pseudodata gives $\chi^2_{\text{global}}/N_{\text{pts.}} = 0.98 \pm 0.03$, which by construction is exactly unity up to the statistical fluctuation, and similarly for the individual data sets included in the global fit; see table 1. On

JHEP08(2012)052

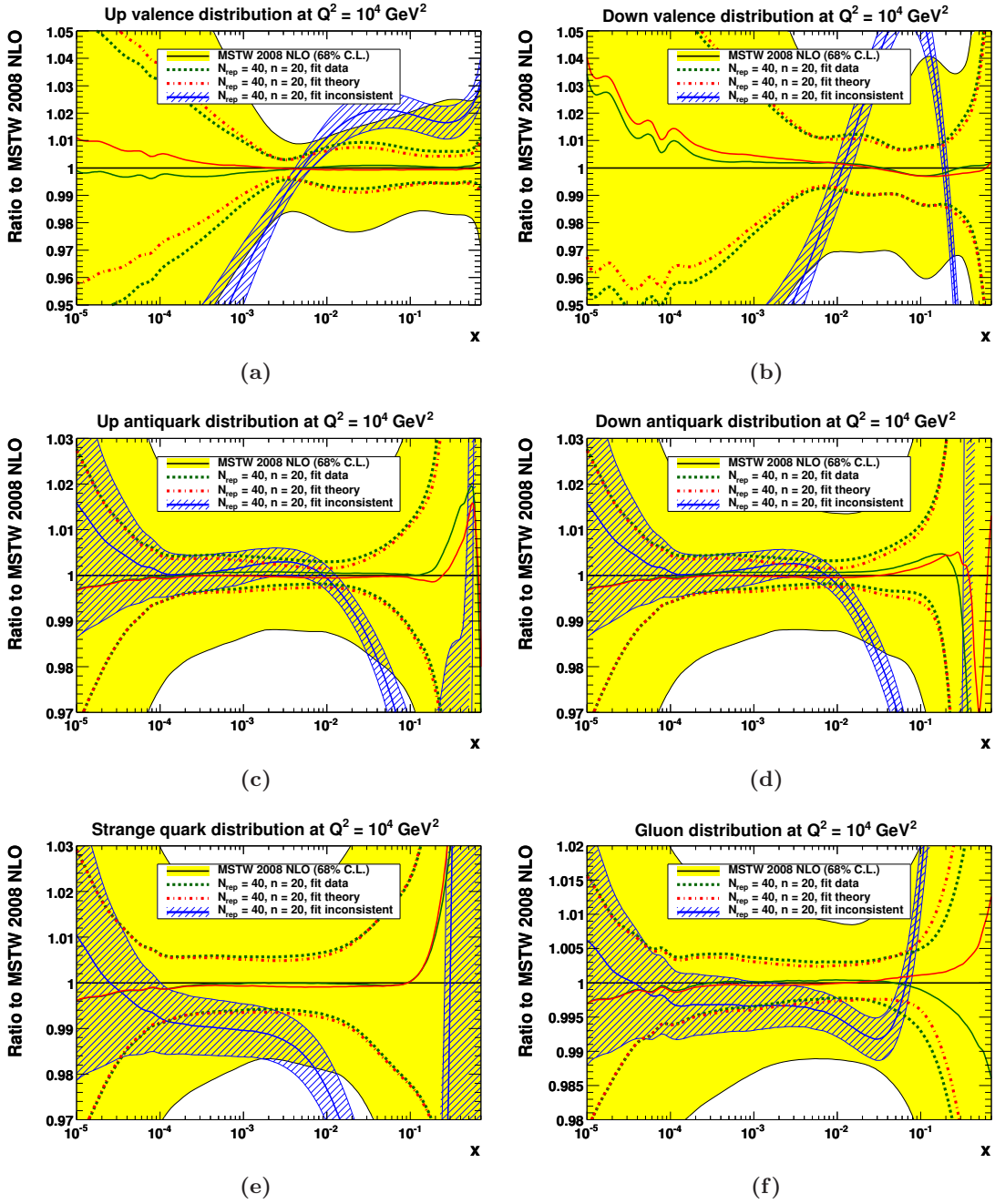


Figure 6. Effect on PDFs of fitting consistent or inconsistent idealised pseudodata.

the other hand, the MC fit to the inconsistent pseudodata gives $\chi^2_{\text{global}}/N_{\text{pts.}} = 1.07 \pm 0.03$, so the fit quality has only deteriorated slightly, despite the central values of some PDFs shifting well outside their original uncertainty band; see figure 6. This result is in contradiction to what seems to be a widely held view that a fit to inconsistent data should lead to a $\chi^2/N_{\text{pts.}} \gg 1$. The values of the $\chi^2/N_{\text{pts.}}$ in table 1 deviate further from unity for a few individual data sets such as BCDMS F_2^d, the NMC F_2^d/F_2^p ratio, NuTeV xF_3 and the CDF Z

JHEP08(2012)052

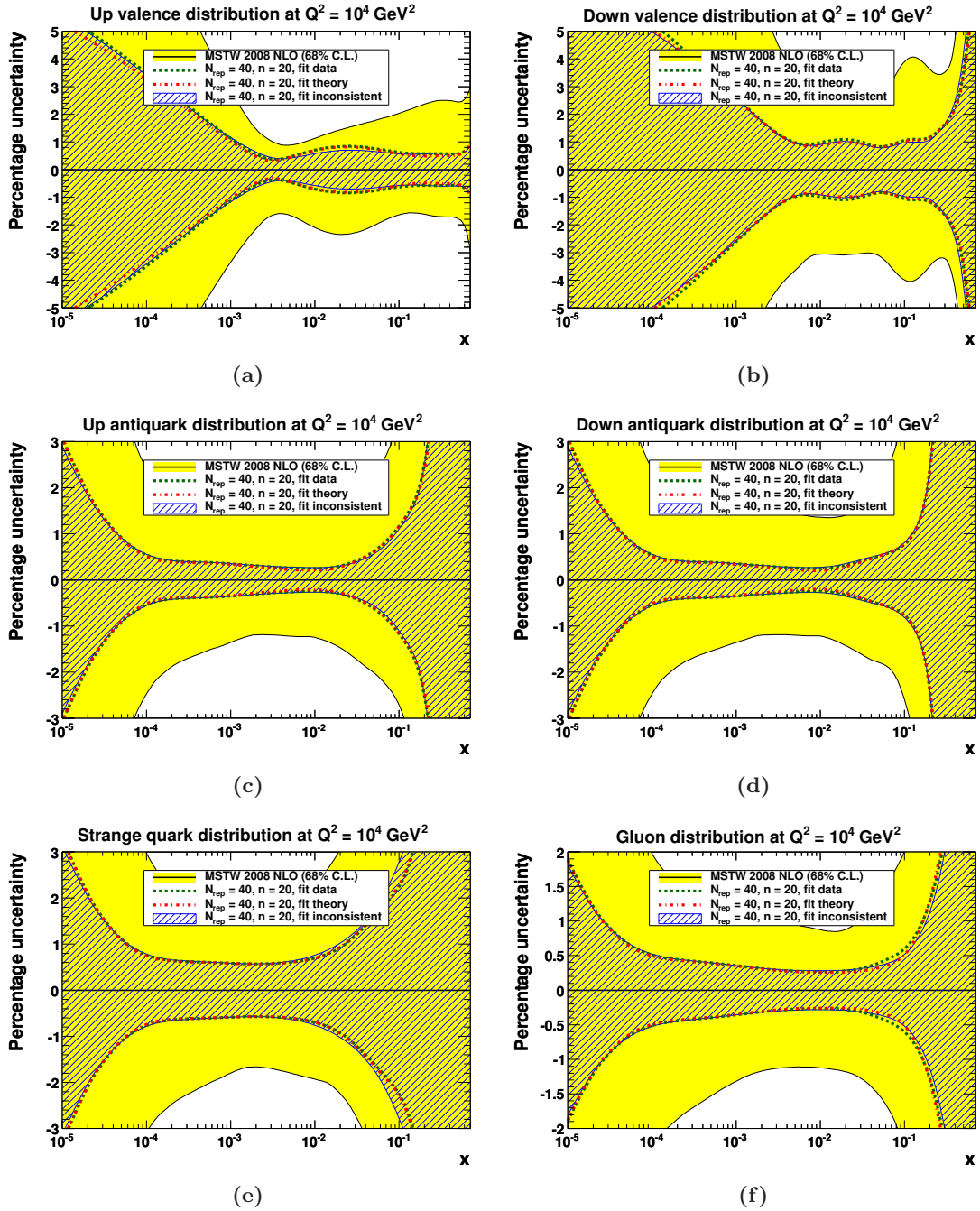


Figure 7. Effect on percentage PDF uncertainties of fitting consistent or inconsistent pseudodata.

rapidity distribution, but not by such large amounts that the inconsistent fit would not be judged to be an "acceptable" fit. Despite the fairly significant Q^2-dependent offset of the H1 and ZEUS inclusive cross sections, amounting to almost 4% at $Q^2 = 500\,\mathrm{GeV}^2$, there is only a slight increase in the χ^2 values in going from the consistent to the inconsistent fit. Simi-

JHEP08(2012)052

Data set	MSTW08	Fit consistent pseudodata	Fit inconsistent pseudodata
BCDMS μp F_2	1.12	0.96 ± 0.13	1.10 ± 0.15
BCDMS μd F_2	1.26	0.99 ± 0.13	1.44 ± 0.17
NMC μp F_2	0.98	0.96 ± 0.12	0.97 ± 0.12
NMC μd F_2	0.83	1.00 ± 0.12	1.05 ± 0.13
NMC $\mu n/\mu p$	0.88	1.02 ± 0.12	1.25 ± 0.13
E665 μp F_2	1.08	0.99 ± 0.18	0.99 ± 0.18
E665 μd F_2	1.01	1.00 ± 0.18	1.02 ± 0.18
SLAC ep F_2	0.80	0.97 ± 0.22	0.98 ± 0.23
SLAC ed F_2	0.78	0.98 ± 0.16	1.03 ± 0.18
NMC/BCDMS/SLAC F_L	1.22	1.04 ± 0.27	1.04 ± 0.27
E866/NuSea pp DY	1.24	0.92 ± 0.10	0.98 ± 0.10
E866/NuSea pd/pp DY	0.93	0.86 ± 0.35	0.96 ± 0.35
NuTeV νN F_2	0.92	0.93 ± 0.19	1.07 ± 0.19
CHORUS νN F_2	0.62	1.01 ± 0.24	1.08 ± 0.27
NuTeV νN xF_3	0.89	0.99 ± 0.19	1.42 ± 0.22
CHORUS νN xF_3	0.93	0.89 ± 0.21	1.14 ± 0.25
CCFR $\nu N \to \mu\mu X$	0.77	0.98 ± 0.14	1.03 ± 0.14
NuTeV $\nu N \to \mu\mu X$	0.46	0.96 ± 0.16	1.00 ± 0.17
H1 MB 99 e^+p NC	1.15	0.87 ± 0.44	0.92 ± 0.44
H1 MB 97 e^+p NC	0.66	0.99 ± 0.20	1.01 ± 0.20
H1 low Q^2 96–97 e^+p NC	0.56	1.00 ± 0.15	1.03 ± 0.15
H1 high Q^2 98–99 e^-p NC	0.97	0.98 ± 0.12	1.00 ± 0.12
H1 high Q^2 99–00 e^+p NC	0.89	1.02 ± 0.10	1.05 ± 0.10
ZEUS SVX 95 e^+p NC	1.16	0.94 ± 0.25	0.94 ± 0.25
ZEUS 96–97 e^+p NC	0.60	1.01 ± 0.11	1.04 ± 0.11
ZEUS 98–99 e^-p NC	0.59	0.98 ± 0.14	1.00 ± 0.14
ZEUS 99–00 e^+p NC	0.70	1.02 ± 0.16	1.05 ± 0.16
H1 99–00 e^+p CC	1.04	1.00 ± 0.23	1.03 ± 0.24
ZEUS 99–00 e^+p CC	1.27	0.95 ± 0.20	1.02 ± 0.21
H1/ZEUS ep F_2^{charm}	1.29	1.00 ± 0.12	1.00 ± 0.12
H1 99–00 e^+p incl. jets	0.78	1.00 ± 0.30	1.03 ± 0.30
ZEUS 96–97 e^+p incl. jets	0.99	1.07 ± 0.26	1.07 ± 0.25
ZEUS 98–00 $e^\pm p$ incl. jets	0.56	0.95 ± 0.25	0.98 ± 0.26
DØ II $p\bar{p}$ incl. jets	1.04	0.96 ± 0.14	1.03 ± 0.15
CDF II $p\bar{p}$ incl. jets	0.73	1.01 ± 0.22	1.08 ± 0.23
CDF II $W \to \ell\nu$ asym.	1.32	1.00 ± 0.30	1.03 ± 0.33
DØ II $W \to \ell\nu$ asym.	2.51	0.94 ± 0.40	1.08 ± 0.47
DØ II Z rap.	0.68	1.05 ± 0.29	1.07 ± 0.30
CDF II Z rap.	1.70	1.05 ± 0.29	1.62 ± 0.43
All data sets	**0.93**	$\mathbf{0.98 \pm 0.03}$	$\mathbf{1.07 \pm 0.03}$

Table 1. Values of $\chi^2/N_{\text{pts.}}$ for the data sets in various NLO global fits. The "MSTW08" column shows the best-fit numbers [6]. The pseudodata numbers in the other two columns are the average and standard deviation of the $\chi^2/N_{\text{pts.}}$ over $N_{\text{rep}} = 40$ replica fits. See ref. [6] for data references.

larly, by looking at the MSTW08 fit to the genuine data in table 1, there are only a few individual data sets with values of $\chi^2/N_{\text{pts.}}$ significantly above unity, perhaps giving the false impression that there is not a large degree of incompatibility between individual data sets.

JHEP08(2012)052

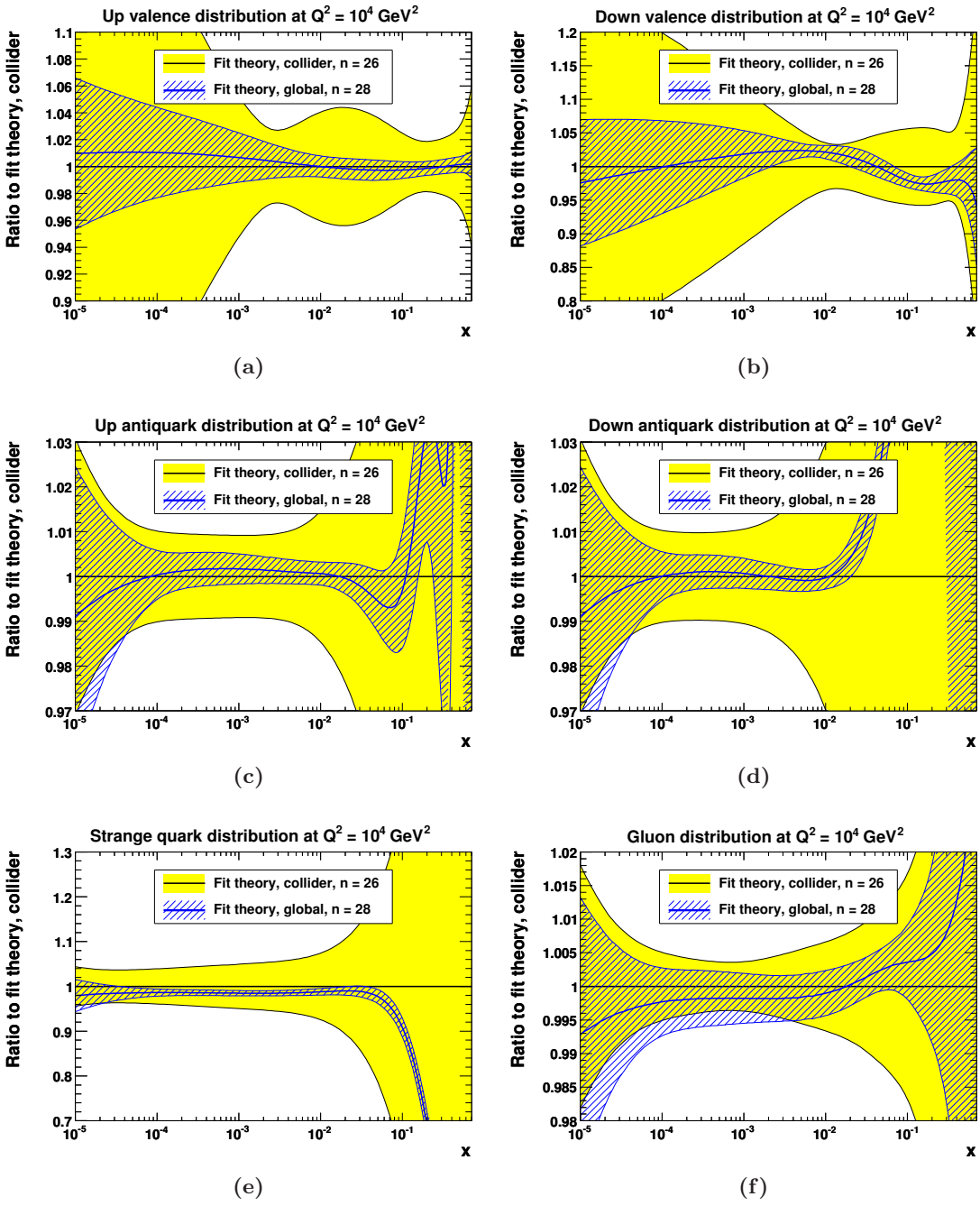


Figure 8. Effect on PDFs of fitting consistent idealised pseudodata, either collider-only or global.

In figures 8 and 9 we show the result of another study using the same consistent or inconsistent idealised pseudodata. First we show the PDFs obtained from fitting only the collider (HERA+Tevatron) subset of the pseudodata, then we show the effect of adding the remaining fixed-target pseudodata. In the "theory" case in figure 8, the fixed-target pseudodata are perfectly consistent with the collider pseudodata (by construction), so the

JHEP08(2012)052

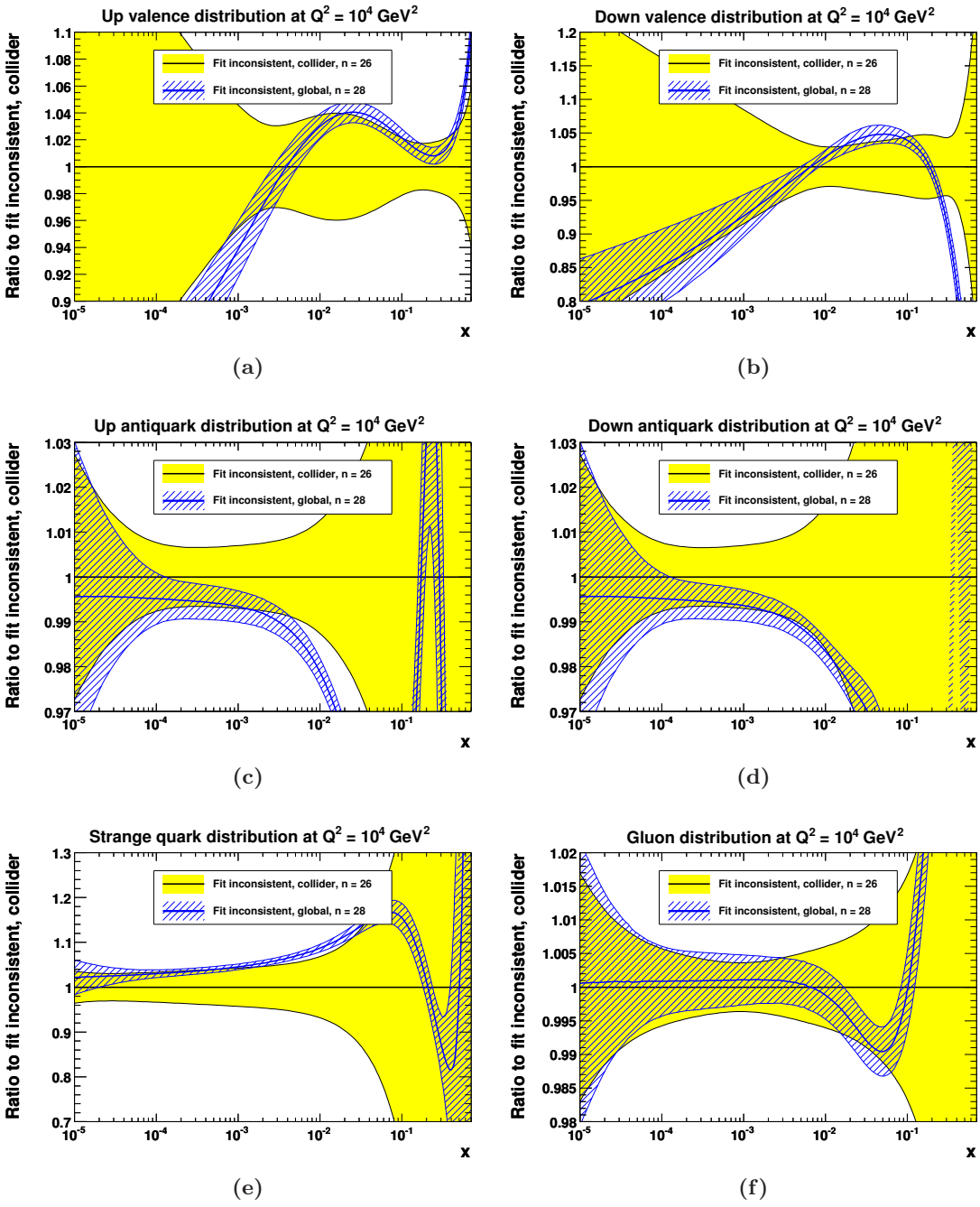


Figure 9. Effect on PDFs of fitting inconsistent idealised pseudodata, either collider-only or global.

global fit gives PDFs consistent with the collider fit, but with much smaller uncertainties. This is not true for the "inconsistent" case in figure 9, where the global fit gives PDFs often lying outside the uncertainty band for the collider fit. The latter situation arises when fitting the genuine data in figure 5, implying that the real collider data are inconsistent with the real fixed-target data. Note that the peculiar behaviour at large x in figures 8(c,d)

JHEP08(2012)052

and 9(c,d) is due to the antiquark distributions going negative in the collider fit at high x where there is no data constraint.

The conclusion of these studies is that defining experimental uncertainties via $\Delta\chi^2_{\rm global} = 1$ is overly optimistic for *global* PDF analysis and that the more conservative "dynamic" tolerance [6] based on a "hypothesis-testing" criterion [12] is much more appropriate.[1] As a final example of a situation where we believe it would make sense to introduce a tolerance to account for a potential discrepancy between data sets, consider the recent ATLAS determination [25] of the ratio of the strange-to-down sea-quark distributions, $r_s(x,Q^2) \equiv 0.5(s+\bar{s})/\bar{d}$, from a fit to inclusive $W^\pm$ and Z differential cross sections at the LHC, combined with inclusive DIS data from HERA. This ratio took the surprising values of

$$r_s(x=0.023,\, Q_0^2 = 1.9\,\text{GeV}^2) = 1.00^{+0.25}_{-0.28} \quad \text{and} \quad r_s(x=0.013,\, Q^2 = M_Z^2) = 1.00^{+0.09}_{-0.10}\,,$$

where the r_s uncertainty is dominated by the experimental PDF uncertainty, determined using $\Delta\chi^2 = 1$, of ± 0.20 and ± 0.07, respectively. These values being consistent with unity indicate no strange suppression, contrary to previous determinations from CCFR/NuTeV dimuon cross sections ($\nu N \to \mu\mu X$), where the strange-quark distributions are suppressed to about half of the $\bar{d}$ and $\bar{u}$ distributions at the lower Q^2 value. Even the HERA DIS data included in the ATLAS analysis [25] shows some tension with the result of no strange suppression; the χ^2 for the HERA DIS data increases by 2.9 units in going from fixed $r_s(x,Q_0^2) = 0.5$ to free r_s with two extra parameters. The MSTW 2008 NNLO analysis [6], which included the CCFR/NuTeV dimuon cross sections, found central values and 68% C.L. PDF uncertainties (including the "dynamic" tolerance) of

$$r_s(x=0.023,\, Q^2 = 1.9\,\text{GeV}^2) = 0.48 \pm 0.04 \quad \text{and} \quad r_s(x=0.023,\, Q^2 = M_Z^2) = 0.79 \pm 0.02\,.$$

Rescaling the experimental PDF uncertainty of the ATLAS determination [25] by a tolerance of ≈ 3, corresponding to $\Delta\chi^2 \approx 9$, would be enough to bring it into agreement with the MSTW08 result. The conclusion that the uncertainty on r_s in the ATLAS determination [25] has been underestimated was also reached by the NNPDF Collaboration [26].

6 Random PDFs generated in space of fit parameters

Given that we have now established that we need an appropriate tolerance, the question arises of how to include this into the MC method. We can introduce a tolerance in the generation of the data replicas simply by rescaling all experimental errors in eqs. (2.10) and (2.11) by $\langle t \rangle \approx \langle T \rangle \approx 3$, corresponding to the average tolerance for 68% C.L. uncertainties. We find that this simple approach, using $n = 20$ input PDF parameters, reproduces the Hessian uncertainties with a dynamic tolerance surprisingly well for most parton flavours and kinematic regions. However, it is not possible to implement exactly the "dynamic" tolerance (different for each eigenvector direction) in the MC method, since no reference is being made to the eigenvectors of the covariance matrix.

[1] A similar conclusion has been reached using very different arguments in ref. [24].

JHEP08(2012)052

Instead of sampling the probability density by working in the space of data, we can produce random PDFs directly in the space of fit parameters.[2] In fact, this was done in the original work of Giele and Keller [3] using the covariance matrix of parameters from Alekhin's fit [27]. A convenient way to generate the random PDFs is to use the eigenvectors of the covariance matrix. Recall from eq. (2.4) that the parameter displacements from the global minimum can be expanded in a basis of the rescaled eigenvectors $e_{ik} \equiv \sqrt{\lambda_k}\, v_{ik}$, that is,

$$a_i - a_i^0 = \sum_{j=1}^{n} e_{ij}\, z_j\,, \tag{6.1}$$

with $n = 20$ the number of input PDF parameters. Usually the $\pm k$th eigenvector PDF set is defined by taking $z_j = (\pm t_j^{\pm})\delta_{jk}$ in eq. (6.1), that is, the usual eigenvector PDF sets are generated with input parameters:

$$a_i(S_k^{\pm}) = a_i^0 \pm t_k^{\pm}\, e_{ik} \qquad (k = 1, \ldots, n)\,, \tag{6.2}$$

with $t_k^{\pm}$ adjusted to give the desired $T_k^{\pm} = (\Delta\chi^2_{\text{global}})^{1/2}$. However, we can instead randomly sample the parameter space such that the kth random PDF set is generated with input parameters obtained by taking $z_j = (\pm t_j^{\pm})|R_{jk}|$ in eq. (6.1), that is,

$$a_i(\mathcal{S}_k) = a_i^0 + \sum_{j=1}^{n} e_{ij} \left(\pm t_j^{\pm}\right) |R_{jk}| \qquad (k = 1, \ldots, N_{\text{pdf}})\,, \tag{6.3}$$

where R_{jk} is a Gaussian-distributed random number of mean zero and variance one, and either $+t_j^+$ or $-t_j^-$ is chosen depending on the sign of R_{jk}. There are therefore $n = 20$ random numbers R_{jk} $(j = 1, \ldots, n)$ associated with the kth random PDF set $(k = 1, \ldots, N_{\text{pdf}})$. The number of random PDF sets N_{pdf} is arbitrary, but again we choose $N_{\text{pdf}} = 40$ mostly for practical reasons. Each random PDF set has equal probability defined by the covariance matrix of fit parameters, and therefore statistical quantities such as the mean and standard deviation can easily be calculated using formulae such as eqs. (2.13) and (2.14) with the obvious replacement $N_{\text{rep}} \to N_{\text{pdf}}$. A comparison of the average and standard deviation of $N_{\text{pdf}} = 40$ PDFs constructed with eq. (6.3) to the best-fit and Hessian uncertainty is made in figure 10. There is generally good agreement, with some shift of the average compared to the best-fit that can be attributed mostly to asymmetric tolerance values $(t_j^+ \neq t_j^-)$. We have verified this explanation by repeating the same studies without a tolerance $(T_j^{\pm} = 1)$. Alternative ad hoc treatments of the asymmetric tolerance values are possible. For example, if $t_j^+ > t_j^-$ proportionally more random PDF sets could be produced for a "$-$" sign than for a "$+$" sign in eq. (6.3) so that the average would be closer to the best-fit, or one could simply symmetrise with the replacement $t_j^{\pm} \to (t_j^+ + t_j^-)/2$ in eq. (6.3). However, since the degree of asymmetry is generally small, we will not explore these possibilities in practice. As some measure of the amount of statistical fluctuation, we produce another $N_{\text{pdf}} = 40$ PDFs constructed with eq. (6.3) using different random numbers R_{jk}. The results are shown in figures 11 and 12 and we conclude that $N_{\text{pdf}} = 40$

[2] We thank H. Prosper for making this suggestion.

JHEP08(2012)052

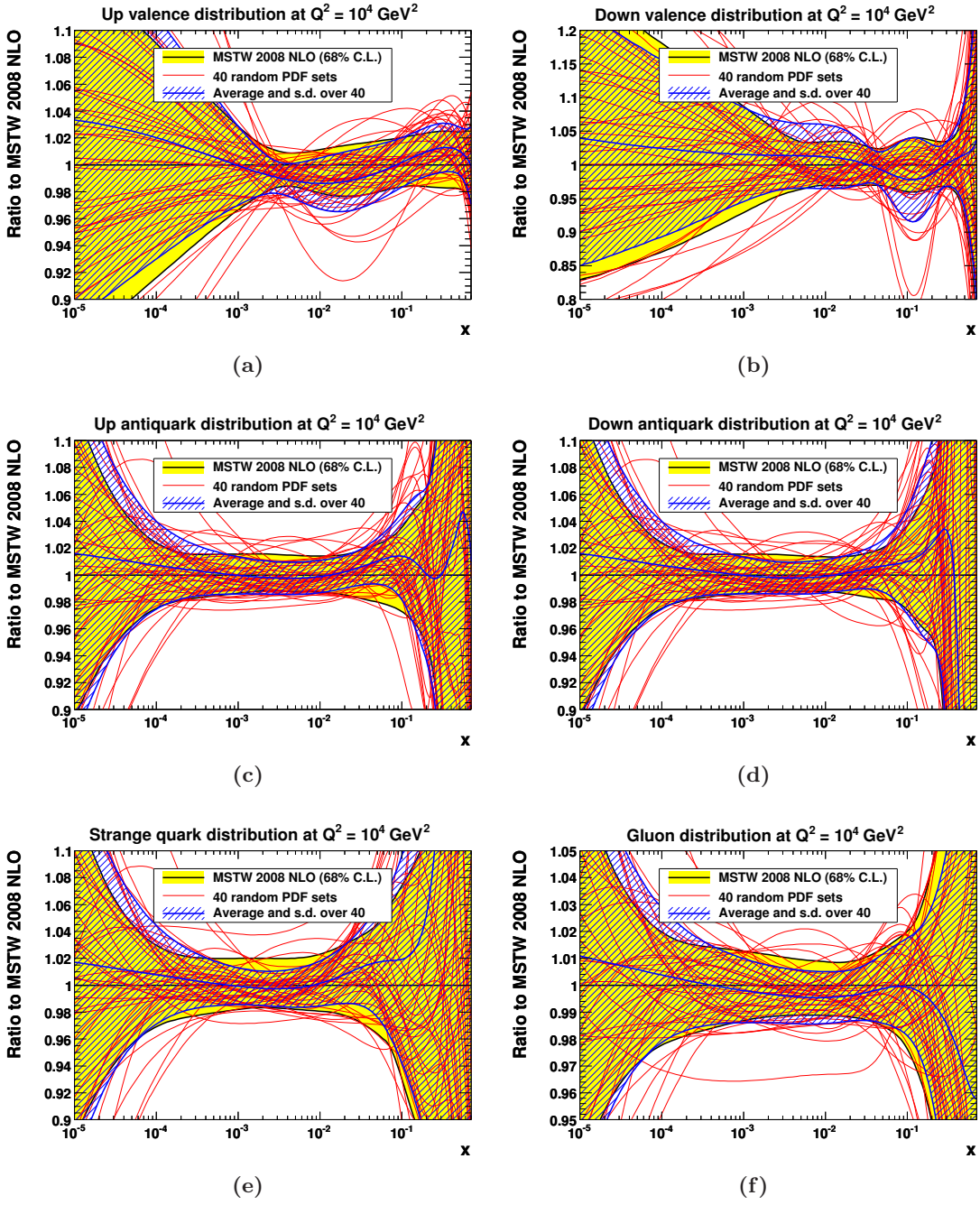


Figure 10. $N_{\rm pdf} = 40$ random sets generated with eq. (6.3) as a ratio to the best-fit PDF set.

is enough to avoid significant fluctuations, although there is some moderate variation due to the limited statistics (for example, in d_v at $x \sim 0.1$).

In principle, there is some amount of non-linearity in going from the input PDF parameters a_i to the input PDFs $f(x, Q_0^2)$, then to the evolved PDFs $f(x, Q^2)$ and to physical observables F calculated using these evolved PDFs (for example, hadronic cross sections with a quadratic PDF dependence). However, we find that, in practice, the apparent de-

JHEP08(2012)052

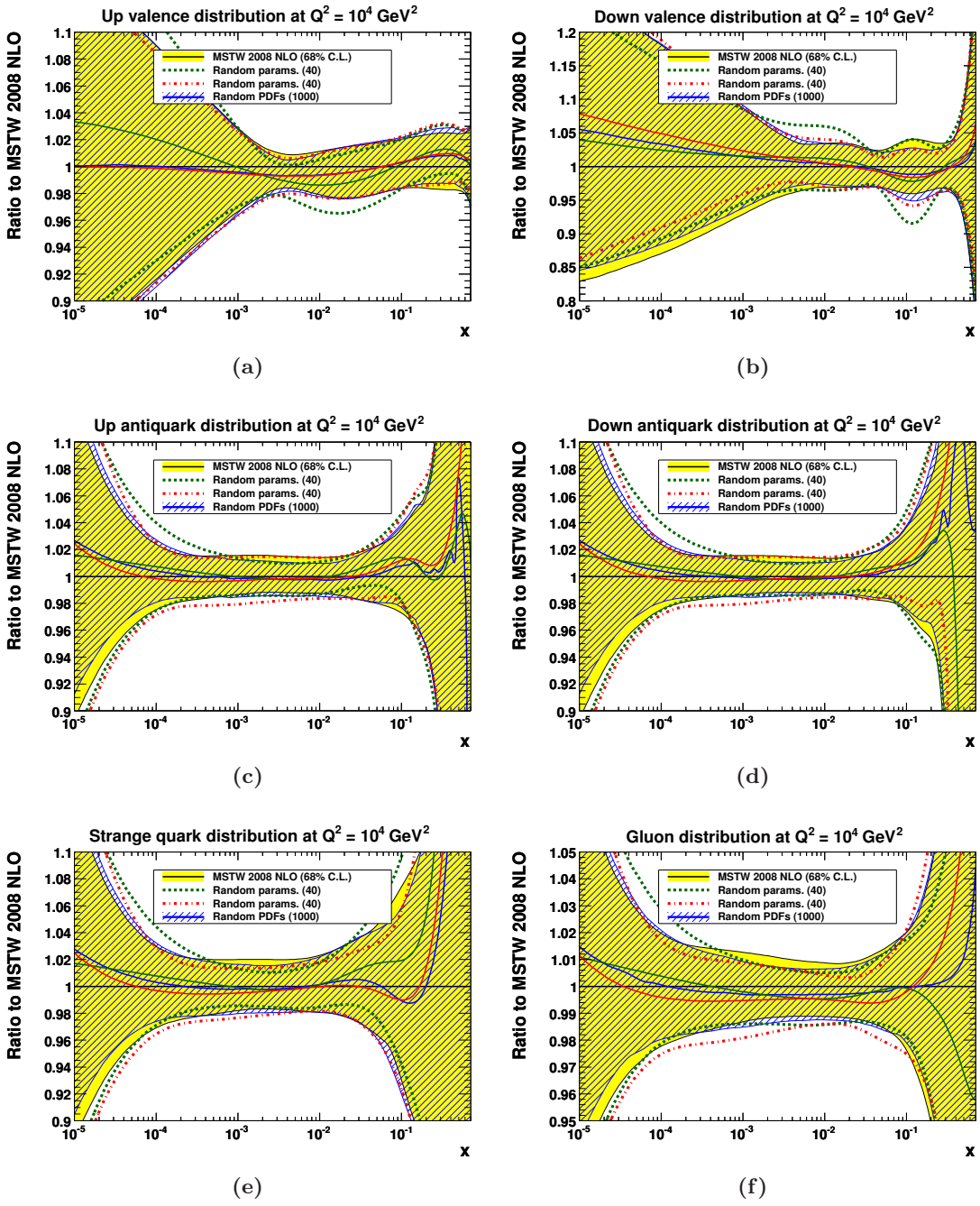
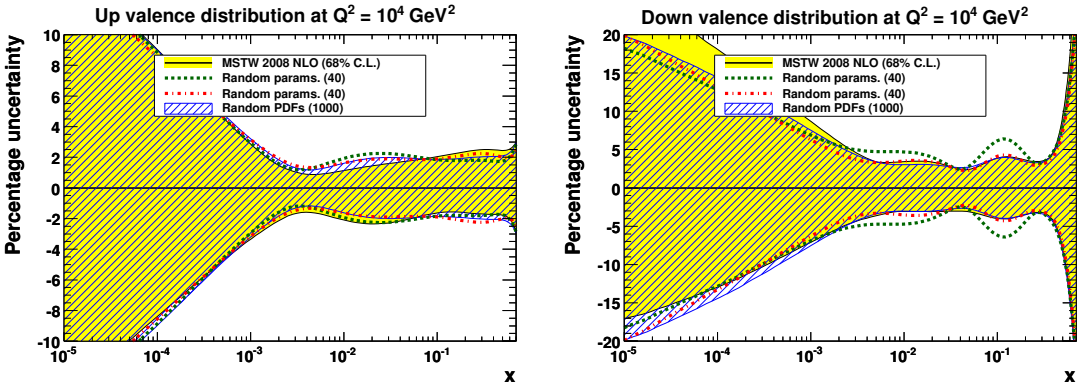


Figure 11. Comparison of best-fit and Hessian uncertainty to the average and standard deviation of two sets of $N_{\rm pdf} = 40$ PDFs generated with different random parameters given by eq. (6.3) and one set of $N_{\rm pdf} = 1000$ random PDFs generated with eq. (6.4).

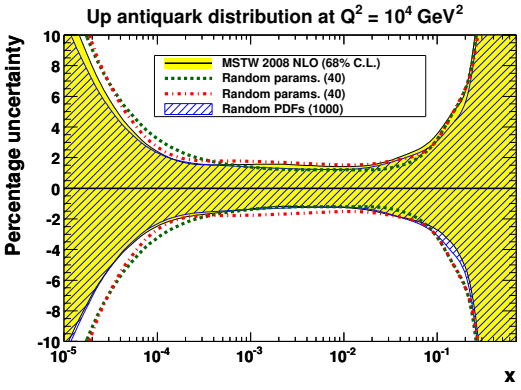
gree of non-linearity is small, an assumption that is inherent in the Hessian method for propagating uncertainties. Making this assumption of linearity, an alternative and simpler way to generate random PDFs is to work with the existing eigenvector PDF sets directly at the level of the quantity of interest F such as the evolved PDF or the hadronic cross

JHEP08(2012)052

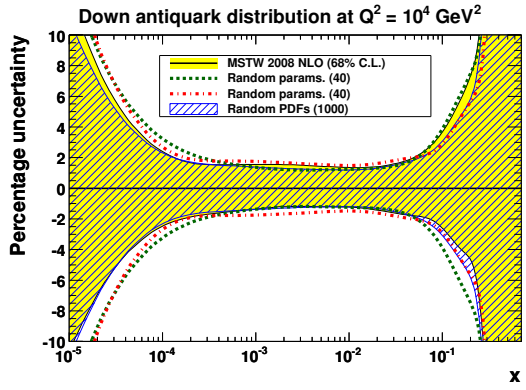


(a)

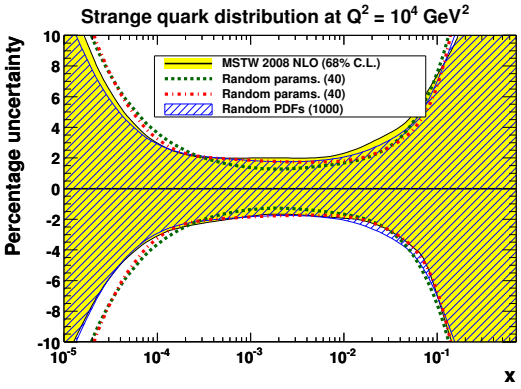
(b)



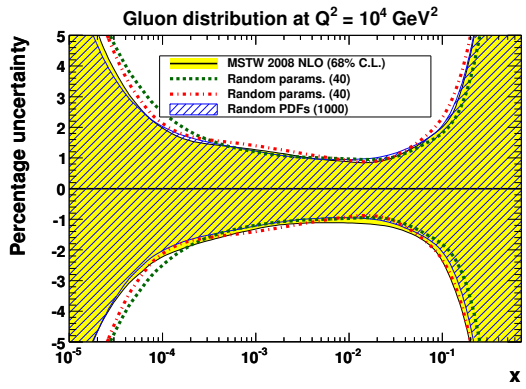
(c)



(d)



(e)



(f)

Figure 12. Similar to figure 11 but percentage uncertainties rather than the ratio to the best-fit.

section. Then we can build random values of F according to[3]

$$F(\mathcal{S}_k) = F(S_0) + \sum_{j=1}^{n} \left[F(S_j^{\pm}) - F(S_0)\right] |R_{jk}| \qquad (k = 1, \ldots, N_{\rm pdf}) \,, \tag{6.4}$$

[3]Cf. the studies of F. De Lorenzi: see eq. (3.1) of ref. [28] or eq. (6.1) of ref. [29].

JHEP08(2012)052

where S_j^+ or S_j^- is chosen depending on the sign of R_{jk}. Note that for the case $F = a_i$ in eq. (6.4), then $a_i(S_0) \equiv a_i^0$ and inserting $a_i(S_j^\pm)$ from eq. (6.2) then we recover eq. (6.3). This construction of a random $F(\mathcal{S}_k)$ using eq. (6.4) can be done "on the fly" for an almost arbitrarily large value of $N_{\rm pdf}$, after the initial computation of $F(S_0)$ and $F(S_j^\pm)$ $(j = 1, \ldots, n)$ requiring only $2n + 1$ (= 41 for the MSTW 2008 PDFs) evaluations of F. We choose $N_{\rm pdf} = 1000$ for the results shown in figures 11 and 12, although the results are similar with a much smaller value. Here we take "F" in eq. (6.4) to be the evolved PDF at $Q = 100\,\text{GeV}$ for the particular parton flavour shown in each plot, then we construct $N_{\rm pdf} = 1000$ values of $F(\mathcal{S}_k)$ and take the average and standard deviation, finding good agreement with the best-fit and Hessian uncertainty. Again, the slight shift of the average compared to the best-fit can be attributed mostly to asymmetric tolerance values, which we confirm by repeating the same exercise starting from eigenvector PDF sets generated with $\Delta\chi^2_{\rm global} = 1$. As already mentioned, ad hoc modifications to the procedure could be adopted to better account for asymmetric tolerance values, but we choose not to explore these possibilities in this work given the relatively small size of the effect. For example, a symmetrised version of eq. (6.4) could be obtained using

$$F(\mathcal{S}_k) = F(S_0) + \frac{1}{2}\sum_{j=1}^{n} \left|F(S_j^+) - F(S_j^-)\right| R_{jk} \qquad (k = 1, \ldots, N_{\rm pdf})\,, \tag{6.5}$$

analogous to the symmetric formula for PDF uncertainties given in eq. (2.9).

We note that an unsuccessful attempt to generate random PDFs directly in the space of fit parameters was made in section 6.5 of ref. [30]. This attempt was flawed in that all random PDF sets were constructed with the unnecessary constraint of a fixed $\Delta\chi^2 = 100$, with the n parameters distributed on the surface of an n-dimensional hypersphere using the eigenvectors as basis vectors, leading to an envelope of the random PDF sets covering a much smaller range than the usual Hessian uncertainty. By contrast, if we generate random PDF sets according to eq. (6.3), then the $\Delta\chi^2$, or equivalently $t_j^\pm$, is only used to define the distance along a particular eigenvector direction. At a general point in parameter space, given by stepping along all eigenvector directions by a random amount, the $\Delta\chi^2$ is irrelevant and it can be very large. It is not necessary or desirable that each random PDF set should have $\Delta\chi^2$ below a certain value. A fixed $\Delta\chi^2$ will only be recovered in the specific (and very unlikely) case that $|R_{jk}| = \delta_{jk}$, then eq. (6.3) reduces to eq. (6.2).

Another argument that a Monte Carlo approach in the space of fit parameters involves exploring a space too wide to be sampled efficiently with a small number of random PDFs was made in section 3.2.1 of ref. [31]. There it was argued that if the probability distribution for each parameter is given as a histogram with three bins, say the one-sigma region around the central value and the two outer regions, then naïvely one might expect the need to randomly sample $3^n \gtrsim 3 \times 10^9$ PDF sets for $n = 20$ free parameters. However, the n parameters are certainly not independent, and the complete correlation information is provided by the covariance matrix obtained from the global fit. Working in the basis of eigenvectors then provides an optimally efficient way to sample the parameter space randomly along each eigenvector direction.

JHEP08(2012)052

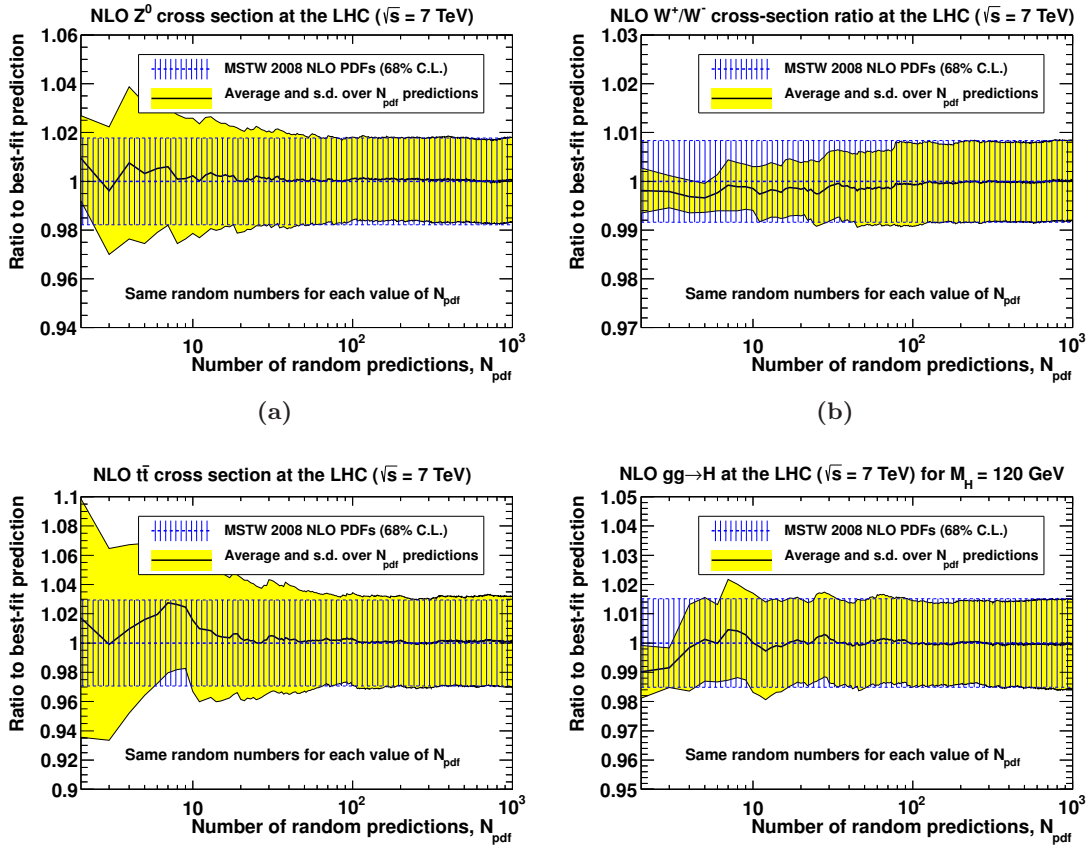


Figure 13. Convergence of average and standard deviation of $N_{\rm pdf}$ random predictions as a function of $N_{\rm pdf}$, each time adding one more random prediction to the $N_{\rm pdf}-1$ previous random predictions, normalised to the best-fit prediction and compared to the Hessian uncertainty.

Nevertheless, it is instructive to perform a numerical exercise in order to explicitly demonstrate roughly how many random predictions are necessary to adequately sample the parameter space. We consider the 7 TeV LHC total cross sections for four typical processes corresponding to inclusive production of (a) Z^0 bosons, (b) W^+ relative to W^- bosons, (c) top-pairs and (d) Standard Model Higgs bosons with $M_H = 120\,\text{GeV}$ from gluon-gluon fusion. These four processes are chosen to sample a variety of parton flavours and momentum fractions x. We use the existing NLO calculations from ref. [1] with the MSTW 2008 NLO best-fit and Hessian eigenvector PDF sets at 68% C.L. For each of the four processes, we generate the minimal $N_{\rm pdf} = 2$ random predictions computed using eq. (6.5) for $F = \{\sigma_{Z^0}, \sigma_{W^+}/\sigma_{W^-}, \sigma_{t\bar{t}}, \sigma_H\}$ and calculate the average and standard deviation. Then the number of random predictions, $N_{\rm pdf}$, is incremented by one, and the average and standard deviation recomputed, until $N_{\rm pdf} = 1000$. The results are shown in figure 13 normalised to the best-fit prediction and compared with the symmetric Hessian uncertainty of eq. (2.9). We use the symmetrised formulae of eqs. (2.9) and (6.5) to allow a direct comparison between the best-fit prediction and the average over the random predictions,

JHEP08(2012)052

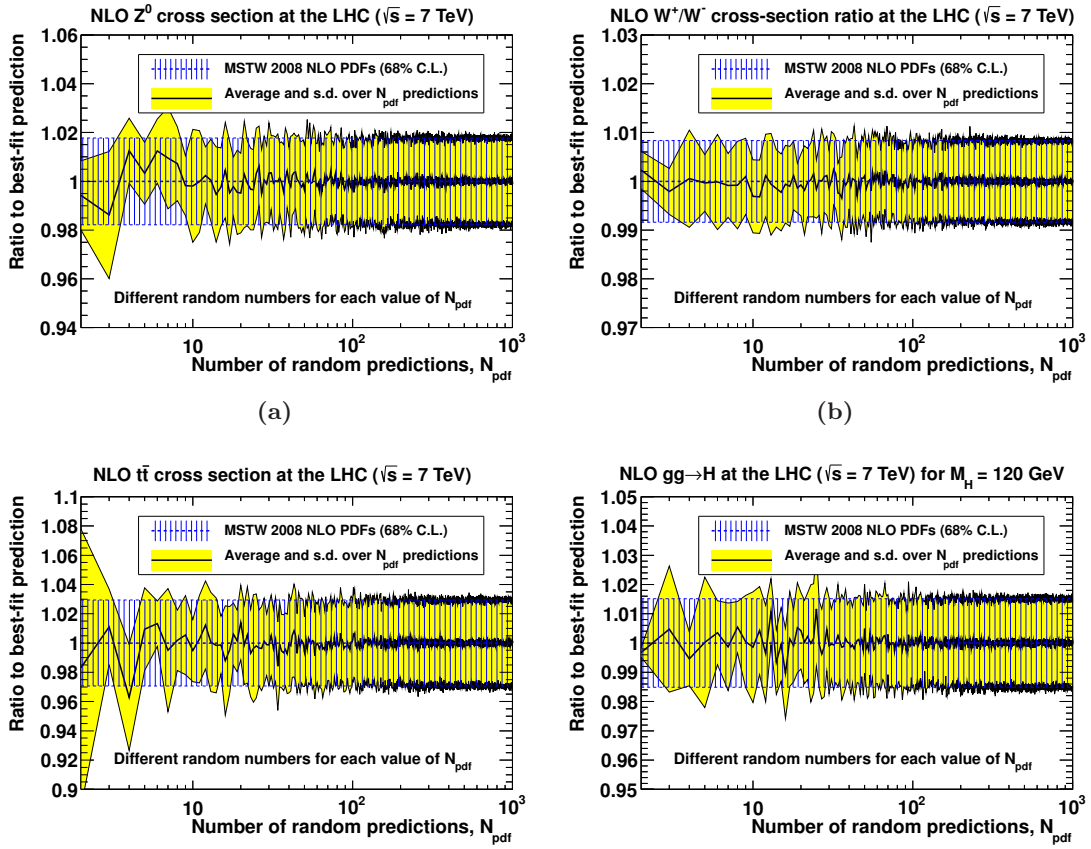


Figure 14. Convergence of average and standard deviation of $N_{\rm pdf}$ random predictions as a function of $N_{\rm pdf}$, each time generating $N_{\rm pdf}$ *independent* random predictions with different random numbers, normalised to the best-fit prediction and compared to the Hessian uncertainty.

without the complications arising from asymmetric tolerance values discussed elsewhere. We show a similar set of plots in figure 14 where each value of $N_{\rm pdf}$ now corresponds to the average and standard deviation over $N_{\rm pdf}$ *independent* random predictions. The results for adjacent $N_{\rm pdf}$ values therefore indicate the size of the statistical fluctuations, which decrease going to larger $N_{\rm pdf}$ values, but are still not completely negligible even for $N_{\rm pdf} \sim 1000$. However, although there is little computational overhead in taking $N_{\rm pdf}$ to be very large when the random predictions are generated "on the fly", one would not expect to see noticeable differences when $N_{\rm rep}$ is much larger than around 1000. In fact, the statistical fluctuations are very small compared to the PDF uncertainty for $N_{\rm pdf} \gtrsim 100$ and even $N_{\rm pdf} = 40$ may be sufficiently accurate for many practical purposes.

7 Reweighting to describe the LHC $W \to \ell\nu$ charge asymmetry data

Updating a PDF set with new data using a Bayesian reweighting method based on statistical inference was originally proposed by Giele and Keller [3] and later developed further

JHEP08(2012)052

by the NNPDF Collaboration [32, 33]. Suppose we have a set of $N_{\rm pdf}$ random PDFs $\{\mathcal{S}_k\}$ with equal probability. It is irrelevant whether they are generated in the space of data (section 2.2) or in the space of parameters (section 6). We can then apply the Bayesian reweighting technique exactly as for the NNPDF sets. The key formulae are summarised below, but we refer to refs. [32, 33] for the derivation and more details of the method. We first compute the χ^2_k for the new data set (comprising $N_{\rm pts.}$ data points) using each $\mathcal{S}_k$, then we can calculate the mean value of any PDF-dependent quantity $F(\mathcal{S}_k)$ as:

$$\langle F\rangle_{\rm old} = \frac{1}{N_{\rm pdf}}\sum_{k=1}^{N_{\rm pdf}} F(\mathcal{S}_k)\,, \qquad \langle F\rangle_{\rm new} = \frac{1}{N_{\rm pdf}}\sum_{k=1}^{N_{\rm pdf}} w_k(\chi^2_k)\,F(\mathcal{S}_k)\,, \tag{7.1}$$

where the *weights* are given by

$$w_k(\chi^2_k) = \frac{W_k(\chi^2_k)}{\frac{1}{N_{\rm pdf}}\sum_{j=1}^{N_{\rm pdf}} W_j(\chi^2_j)}\,, \qquad W_k(\chi^2_k) \equiv (\chi^2_k)^{\frac{1}{2}(N_{\rm pts.}-1)}\exp\left(-\frac{1}{2}\chi^2_k\right), \tag{7.2}$$

with the denominator of $w_k(\chi^2_k)$ ensuring the normalisation condition:

$$\sum_{k=1}^{N_{\rm pdf}} w_k(\chi^2_k) = N_{\rm pdf}\,. \tag{7.3}$$

Note that the expression for the weights in eq. (7.2) differs from the original formula in ref. [3] due to subtle arguments explained in ref. [32]. The standard deviation ΔF after reweighting can be calculated using eq. (2.14) with the trivial replacement $N_{\rm rep} \to N_{\rm pdf}$ and using the weighted averages $\langle F^2\rangle_{\rm new}$ and $\langle F\rangle_{\rm new}$. The effective number of random PDF sets left after reweighting, referred to as the "Shannon entropy" [32], is given by

$$N_{\rm eff} = \exp\left(\frac{1}{N_{\rm pdf}}\sum_{k=1}^{N_{\rm pdf}} w_k \ln\left(\frac{N_{\rm pdf}}{w_k}\right)\right). \tag{7.4}$$

As a simple application of this reweighting technique, we will consider the 7 TeV LHC data from the 2010 running period on the $W \to \ell\nu$ charge asymmetry from CMS [7] and ATLAS [8]. The $W \to \ell\nu$ charge asymmetry is defined differentially as a function of the pseudorapidity η_ℓ of the charged-lepton from the W-boson decay, i.e.

$$A_\ell(\eta_\ell) = \frac{{\rm d}\sigma(\ell^+)/{\rm d}\eta_\ell - {\rm d}\sigma(\ell^-)/{\rm d}\eta_\ell}{{\rm d}\sigma(\ell^+)/{\rm d}\eta_\ell + {\rm d}\sigma(\ell^-)/{\rm d}\eta_\ell}\,. \tag{7.5}$$

We will consider the CMS data [7] with charged-lepton transverse momentum cut of $p_T^\ell > 25\,{\rm GeV}$ in both the electron ($\ell = e$) and muon ($\ell = \mu$) channels. The ATLAS data [8] combine the electron and muon channels with cuts of $p_T^\ell > 20\,{\rm GeV}$, missing transverse energy $\not{E}_T^\nu > 25\,{\rm GeV}$ and transverse mass $M_T = \sqrt{2p_T^\ell\,\not{E}_T^\nu\,(1-\cos\Delta\phi_{\ell\nu})} > 40\,{\rm GeV}$, where $\Delta\phi_{\ell\nu}$ is the azimuthal separation between the directions of the charged-lepton and neutrino. The pseudorapidity distributions, ${\rm d}\sigma(\ell^\pm)/{\rm d}\eta_\ell$, calculated from the public DYNNLO code [34] using the MSTW 2008 NLO best-fit PDFs with $\mu_R = \mu_F = M_W$, are shown in figure 15(a,b)

JHEP08(2012)052

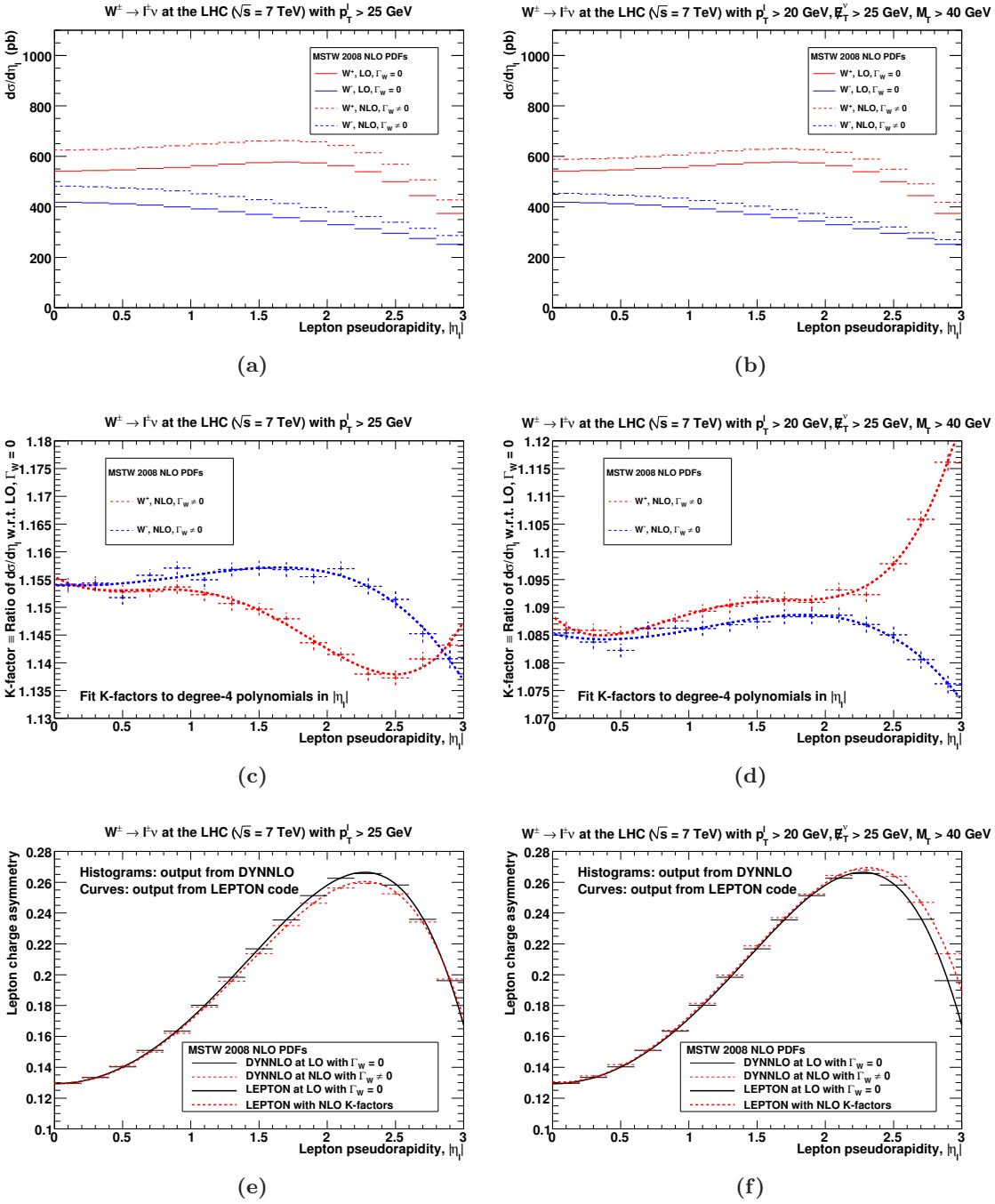


Figure 15. (a,b) $\mathrm{d}\sigma(\ell^{\pm})/\mathrm{d}\eta_{\ell}$ distributions, (c,d) K-factors, (e,f) lepton charge asymmetry, for kinematic cuts corresponding to the (a,c,e) CMS data [7] and (b,d,f) ATLAS data [8].

for (a) CMS cuts and (b) ATLAS cuts. For LO kinematics ($p_T^W = 0$) with zero W width ($\Gamma_W = 0$), then $p_T^{\ell} = \not{E}_T^{\nu}$ and $M_T = 2p_T^{\ell}$, and the predictions are identical for the CMS and ATLAS cuts, but not after accounting for NLO and finite W width effects. In figure 15(c,d) we define a K-factor by taking the ratio of the DYNNLO histograms, then we fit to quartic

JHEP08(2012)052

polynomials in $|\eta_\ell|$ to provide a convenient parameterisation and to smooth statistical fluctuations from the VEGAS multidimensional integration. A fast calculation of the $W \to \ell\nu$ charge asymmetry can then be obtained using a simple LO calculation with zero W width (denoted "LEPTON"), including the parameterised K-factors for $\mathrm{d}\sigma(\ell^\pm)/\mathrm{d}\eta_\ell$, making the assumption that the K-factors are independent of the PDF choice. In figure 15(e,f) we compare the LEPTON calculation, without and with the inclusion of K-factors, with the DYNNLO histograms, finding good agreement (by construction). It can be seen that the NLO corrections and finite-width effects are very small over most of the $|\eta_\ell|$ range. The effect on the $W \to \ell\nu$ charge asymmetry of neglecting the PDF dependence of the K-factors should then be completely negligible. We have also computed the NNLO corrections using the DYNNLO code and find them to be much smaller than the NLO corrections, but we will consider only NLO QCD in making comparisons to data, as done elsewhere in this paper.

We will focus on demonstrating the reweighting technique rather than aiming to make an exhaustive study of the impact of the LHC data. With this aim in mind, we will not consider in this work the 2010 CMS data with $p_T^\ell > 30\,\mathrm{GeV}$ [7], the preliminary CMS measurements using 2011 data with $p_T^\mu > 25\,\mathrm{GeV}$ [35] or $p_T^e > 35\,\mathrm{GeV}$ [36], or the recent LHCb measurements using 2010 data with $p_T^\mu > \{20, 25, 30\}\,\mathrm{GeV}$ [37]. The ATLAS Collaboration [8] provide the differential cross sections, $\mathrm{d}\sigma(\ell^\pm)/\mathrm{d}\eta_\ell$, separately for $W^+ \to \ell^+\nu$ and $W^- \to \ell^-\bar{\nu}$ with the complete information on correlated systematic uncertainties, which is potentially more useful for PDF fits than simply the asymmetry $A_\ell(\eta_\ell)$. A future study could perhaps investigate the use of reweighting with the ATLAS $W^\pm$ (and Z/γ^*) differential cross sections rather than the asymmetry $A_\ell(\eta_\ell)$. In this study, we simply calculate the χ^2_k values with all experimental uncertainties added in quadrature.

In figure 16(a,b) we compare the (a) CMS and (b) ATLAS data on the $W \to \ell\nu$ charge asymmetry to predictions using the MSTW 2008 NLO PDFs, firstly with the usual best-fit and Hessian uncertainty. We then generate $N_{\mathrm{pdf}} = 1000$ random predictions for the asymmetry by taking $F = A_\ell(\eta_\ell)$ in eq. (6.4), and take the average and standard deviation, giving results slightly different from the best-fit and Hessian uncertainty (mainly due to the asymmetric tolerance values). The χ^2 values of the average $A_\ell(\eta_\ell)$, displayed in the plot legends, are then slightly larger than the χ^2 of the best-fit predictions. Next we compute weights for each of the N_{pdf} predictions according to eq. (7.2), then finally we plot the weighted average and standard deviation in figure 16(a,b). The χ^2 of the weighted average $A_\ell(\eta_\ell)$ improves significantly compared to the unweighted average. The effective number of random predictions N_{eff} after reweighting, computed according to eq. (7.4), is about half the original number for CMS and almost a quarter the original number for ATLAS. The most significant change in the predictions after reweighting is for $\eta_\ell \approx 0$ where $A_\ell(\eta_\ell)$ depends on the combination $u_v - d_v$ at momentum fractions x slightly above $x \sim M_W/\sqrt{s} \sim 0.01$. In figure 16(c,d) we plot this combination for $Q^2 = (100\,\mathrm{GeV})^2$ for the same three sets of predictions shown in figure 16(a,b). We compare the best-fit and Hessian uncertainty with the unweighted/weighted average and standard deviation of $N_{\mathrm{pdf}} = 1000$ random PDFs constructed by taking $F = x(u_v - d_v)(x, Q^2)$ in eq. (6.4), with the *same* random numbers

JHEP08(2012)052

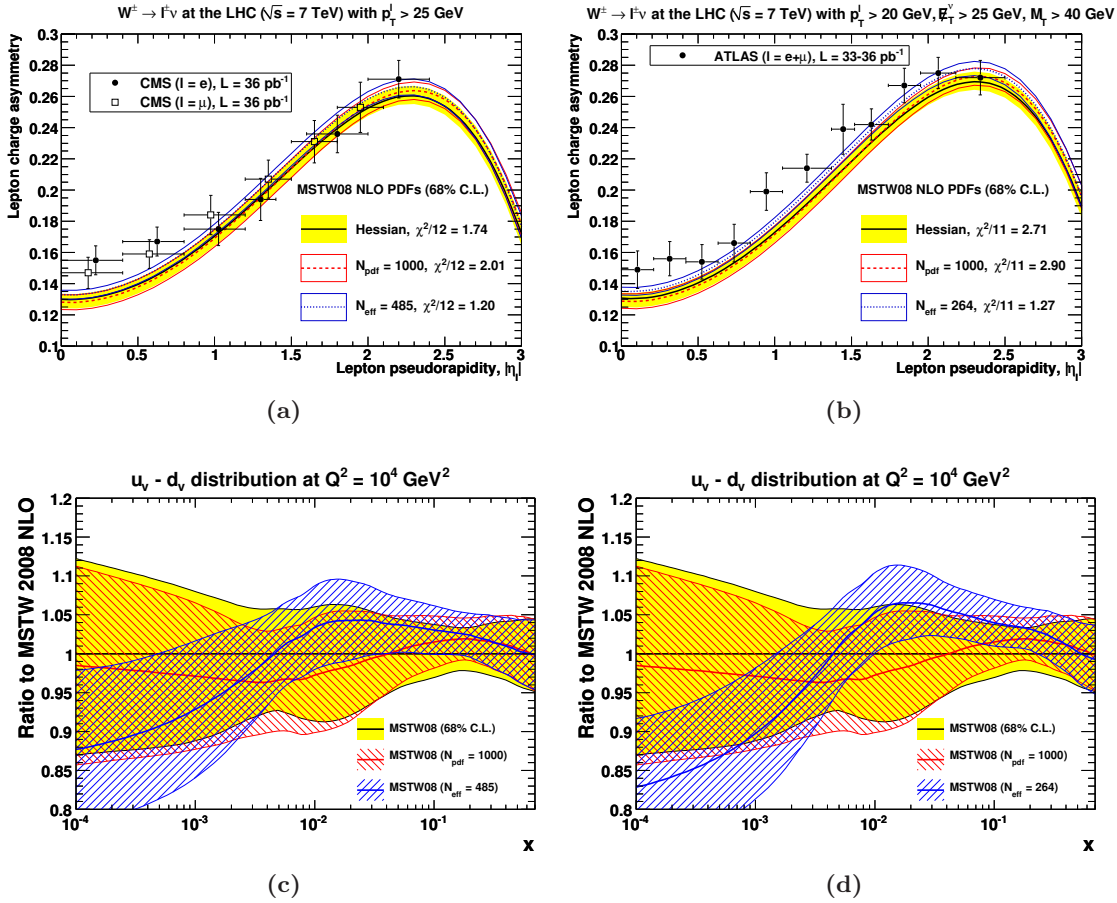


Figure 16. Lepton charge asymmetry $A_\ell(\eta_\ell)$ predictions compared to (a) CMS [7] and (b) ATLAS [8] data, then change in $u_v - d_v$ after reweighting using (c) CMS and (d) ATLAS data.

R_{jk} and weights w_k used in figure 16(a,b). As expected from figure 16(a,b), the shift in $u_v - d_v$ is largest at $x \sim 0.01$, and the average value after reweighting using the ATLAS data even lies outside the original uncertainty band. There is also a distinct reduction in the size of the uncertainty band after reweighting.

The procedure just described is not completely unambiguous. Alternative prescriptions could be formulated which are equivalent in a linear approximation, but which might differ due to some degree of non-linearity. For example, rather than starting by generating random predictions for the asymmetry by taking $F = A_\ell(\eta_\ell)$ in eq. (6.4), we could instead generate $N_{\rm pdf} = 1000$ random PDF sets by taking $F = xf(x, M_W^2)$ in eq. (6.4), where $f = \{g, d, u, s, c, b, \bar{d}, \bar{u}, \bar{s}, \bar{c}, \bar{b}\}$, then calculate $A_\ell(\eta_\ell)$ for each of these $N_{\rm pdf}$ random PDF sets, before calculating weights according to eq. (7.2) as before. This alternative method will give slightly different results since $A_\ell(\eta_\ell)$ depends on $xf(x, M_W^2)$ in a non-linear manner. In figure 17(a,b) we compare the distribution of weights w_k computed using the two different methods, using the same random numbers R_{jk} to allow a direct comparison of individual weights with the same label k. The distribution of weights is very similar using the two

JHEP08(2012)052

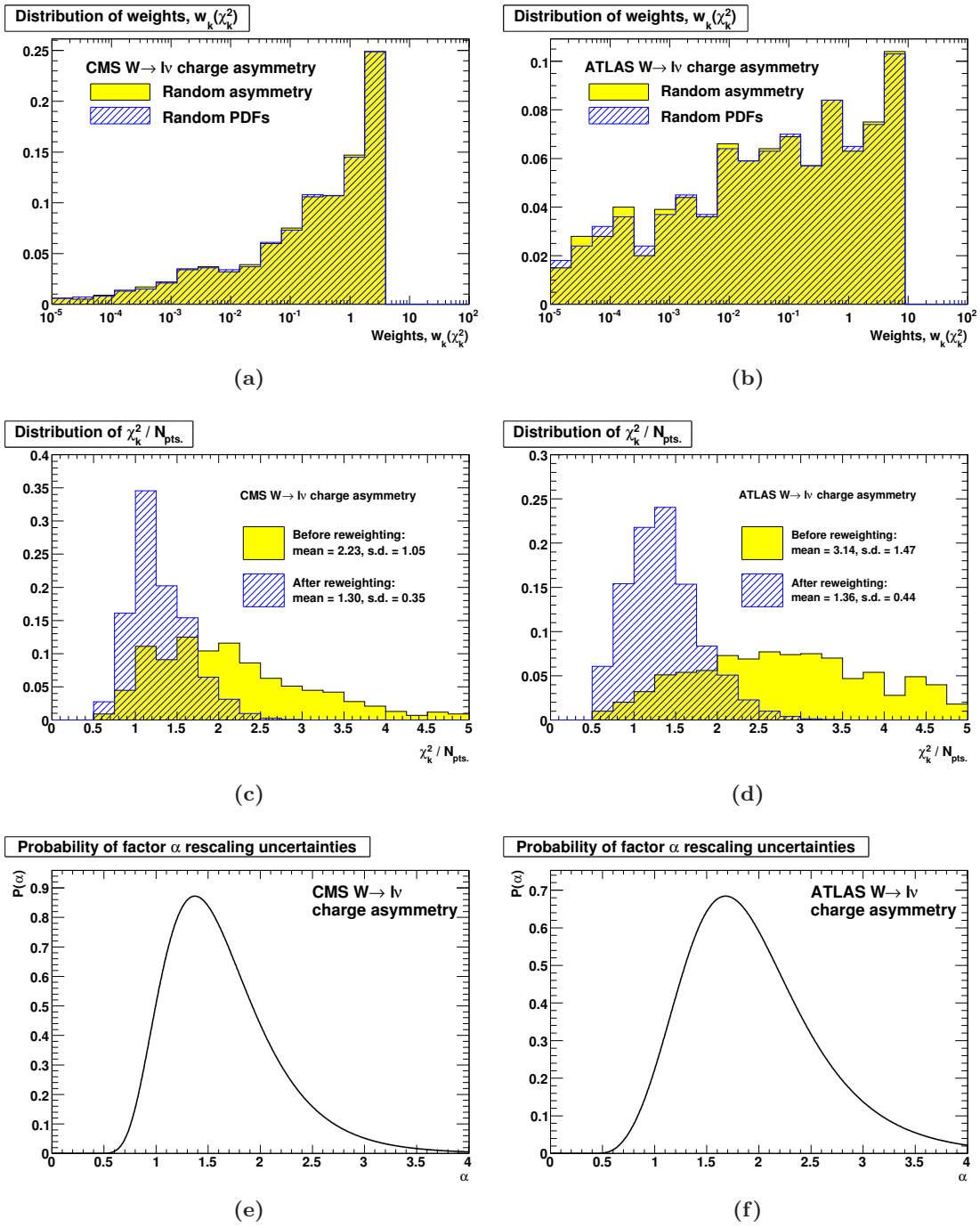


Figure 17. Distributions of (a,b) w_k, (c,d) $\chi^2_k/N_{\text{pts.}}$, (e,f) $\mathcal{P}(\alpha)$, for (a,c,e) CMS and (b,d,f) ATLAS.

methods. The individual weights typically agree to within a few percent and differ by only a few tens of percent in the worst cases. The values of N_{eff} agree to the nearest integer and the values of $\chi^2/N_{\text{pts.}}$ agree to two decimal places. The plots of figure 16 produced using the alternative method are indistinguishable. We conclude that the degree of non-linearity

JHEP08(2012)052

is small and both techniques may be useful in practice. For example, it might be useful to first generate $N_{\rm pdf} = 1000$ random PDF sets as grid files by taking $F = xf(x,Q^2)$ in eq. (6.4), then these grid files can be processed in exactly the same way as the NNPDF grid files. On the other hand, that method would require substantial disk storage and would require the observable $A_\ell(\eta_\ell)$ to be evaluated $N_{\rm pdf}$ times, which is potentially time-consuming. With the first method described above, it is unnecessary to store intermediate grid files, and only $2n+1$ (= 41 for the MSTW 2008 PDFs) evaluations of $A_\ell(\eta_\ell)$ are needed for the best-fit and $2n$ eigenvector PDF sets, exactly as for the usual computation of Hessian uncertainties. The first method will therefore be used for subsequent results.

The χ^2 distribution of the new data set after reweighting can easily be histogrammed:

$$\mathcal{P}(\chi^2_a < \chi^2 < \chi^2_b) = \frac{1}{N_{\rm pdf}} \sum_{k=1}^{N_{\rm pdf}} w_k(\chi^2_k)\, \Theta(\chi^2_k - \chi^2_a)\Theta(\chi^2_b - \chi^2_k)\,, \tag{7.6}$$

where the χ^2 distribution before reweighting is trivially obtained by setting all weights w_k equal to unity. Both these distributions are shown in figure 17(c,d). The plot legends indicate the mean χ^2 and the standard deviation. The reweighting procedure shifts the χ^2 distribution so that larger weights are given to the random predictions with $\chi^2_k/N_{\rm pts.} \sim 1$.

If we rescale the data uncertainties by a factor α, then the probability density for the rescaling parameter α is given by [32]

$$\mathcal{P}(\alpha) \propto \frac{1}{\alpha} \sum_{k=1}^{N_{\rm pdf}} W_k\left(\frac{\chi^2_k}{\alpha^2}\right), \tag{7.7}$$

that is, the sum of the unnormalised weights given by eq. (7.2) with the replacement $\chi^2_k \to \chi^2_k/\alpha^2$. The overall normalisation of eq. (7.7) can be determined from the condition that the integral of $\mathcal{P}(\alpha)$ over α gives unity. The probability distribution $\mathcal{P}(\alpha)$ is shown in figure 17(e,f). These plots suggest that the LHC data on $A_\ell(\eta_\ell)$ are somewhat inconsistent with the data in the MSTW 2008 NLO fit and that the uncertainties on the LHC $A_\ell(\eta_\ell)$ data should be rescaled by a factor 1.37 for CMS and 1.68 for ATLAS to achieve consistency, where these are the most probable values of α. Conversely, a most probable value of α much less than 1 would suggest that the experimental uncertainties are overestimated to some extent. In that case, it might be desirable to repeat the reweighting procedure with the replacement $\chi^2_k \to \chi^2_k/\alpha^2$ in eq. (7.2), where α is the most probable value.

It is clear (see, for example, the discussion in ref. [1]) that there is some considerable tension between the LHC $W \to \ell\nu$ charge asymmetry data and some of the data already included in the MSTW 2008 fit, such as the Tevatron $W \to \ell\nu$ asymmetry, the NMC F_2^d/F_2^p ratio, and the E866/NuSea Drell-Yan σ^{pd}/σ^{pp} ratio. Other tensions have been observed with the more recent and precise Tevatron data on the $W \to \ell\nu$ charge asymmetry, and partially resolved by more flexible nuclear corrections for deuteron structure functions [38] and extended parameterisation choices for the functional form of the input PDFs. Indeed, we note that the LHC asymmetry $A_\ell(\eta_\ell)$ depends on valence-quark parameterisations near $x \sim 0.01$, and the studies in section 3 suggested that this is the single place where the MSTW 2008 parameterisation is likely to be inadequate. Further attempts to resolve these

JHEP08(2012)052

tensions will be necessary for any future update of the MSTW 2008 fit. Therefore, the reweighting technique is instructive, but does not indicate the ultimate impact of including the new data in a global PDF fit after closer scrutiny of potential sources of tension. Nevertheless, we hope that the new method presented in this section of generating random predictions on-the-fly from the existing eigenvector PDF sets, followed by subsequent Bayesian reweighting, will be useful for a wide range of potential studies by third parties from both the experimental and theoretical communities.

8 Conclusions

We have made a first study of the Monte Carlo approach to experimental uncertainty propagation in the context of the MSTW 2008 NLO PDF fit [6], either using data replicas or alternatively working directly in parameter space. The main findings of this study are as follows:

- The Hessian method and the Monte Carlo method using data replicas are approximately equivalent in a global fit when using the same parameterisation and (lack of) tolerance, i.e. $\Delta\chi^2 = 1$. Similar findings have previously been observed in a fit only to H1 data [15].

- The Monte Carlo approach using data replicas is better suited to exploring parameterisation bias due to the potentially restrictive input functional form. Increasing the number of parameters from 20 $\rightarrow$ 28 has only a small effect on PDF uncertainties, with the exception of the valence-quark distributions at low x values where there is a moderate increase in PDF uncertainties. This gives some confidence that, in general, PDF uncertainties in the MSTW 2008 fit are not significantly underestimated due to parameterisation bias, with the possible exception of the strange-quark and -antiquark distributions where the imposed parameterisation constraint is more severe due to the lack of available data constraints.

- The previous findings raise the question why the MSTW/CTEQ uncertainties (*with* a tolerance) are similar to the NNPDF uncertainties (*without* a tolerance) [1], if the tolerance in the former is not compensating for the more restricted functional-form parameterisation rather than the more flexible neural-network parameterisation. One possibility is that the procedural uncertainties for NNPDF associated with splitting data into training and validation sets mimic the effect of a tolerance for MSTW/CTEQ (see discussion in section 3.2 of ref. [39]). Further investigation would be needed by the NNPDF Collaboration to clarify this possible explanation.

- The Monte Carlo approach using data replicas is also better suited when making fits to limited data sets without the need to restrict the input parameterisation. We compared the global-fit PDFs to those extracted using a similar flexible parameterisation from more limited data sets either excluding HERA data, including only HERA data, or including only collider (HERA and Tevatron) data. The fits to limited data sets gave much larger PDF uncertainties for some parton combinations, implying that

JHEP08(2012)052

we need data from HERA, the Tevatron *and* the fixed-target experiments to get a meaningful result. The PDF uncertainty bands from the fits to the limited data sets are not close to overlapping in many cases, implying that some kind of *tolerance* is needed to accommodate inconsistencies between the various data subsets.

- As a further exercise to examine the effect of data set inconsistency, we generated idealised pseudodata from the best-fit theory predictions, then we introduced deliberate inconsistencies. The fractional PDF uncertainties were very similar whether fitting the real data, the consistent pseudodata or the inconsistent pseudodata. On the other hand, the central values obtained when fitting the inconsistent pseudodata were incompatible accounting for the uncertainty bands, even though the $\chi^2_{\rm global}$ only increased by around 10% and the $\chi^2/N_{\rm pts.}$ for individual data sets did not deviate far above unity. Given that a good fit should have $\chi^2/N_{\rm pts.}$ approximately in the range $1 \pm \sqrt{2/N_{\rm pts.}}$ [12], giving 1.0 ± 0.1 for $N_{\rm pts.} \sim 200$, it is far from obvious to spot genuine inconsistencies in the real data of the size we introduced into the idealised pseudodata. It is definitely not the case that the PDF uncertainties will automatically expand to accommodate any inconsistencies. Again, this suggests the need for a tolerance to accommodate potential data set inconsistencies in the real data.

- Having established the need for an appropriate tolerance, we pointed out that it could be introduced by rescaling all experimental uncertainties by a common factor (say, 3) in the generation of data replicas. However, the introduction of a "dynamic" tolerance for each eigenvector direction is not possible, since no use is made of the covariance matrix of fit parameters in the Monte Carlo error propagation.

- Instead, we proposed sampling the covariance matrix of fit parameters by stepping along each eigenvector direction by a random amount, including the appropriate tolerance values. This method of generating random PDF sets is close to the usual generation of eigenvector PDF sets in the Hessian method where one steps along each eigenvector direction in turn by a fixed amount.

- In fact, assuming linearity between the input PDF parameters and derived quantities such as evolved PDFs or hadronic cross sections, an assumption that is inherent in the Hessian method, then it is more convenient to generate random predictions on-the-fly from the existing eigenvector PDF sets.

- As a simple example application to demonstrate the benefits of having randomly-distributed theory predictions, we used Bayesian reweighting to investigate the effect on the PDFs of recent LHC data on the $W \to \ell\nu$ charge asymmetry. Similar studies can now easily be performed by third parties using any PDF determination where eigenvector PDF sets are provided. The reweighting technique is therefore no longer limited only to using the PDF sets provided by the NNPDF Collaboration.

We conclude that the Monte Carlo method using data replicas is, on balance, not superior to the Hessian method in a global fit when using a conventional functional-form parameterisation of the input PDFs. In particular, one of the key benefits of the Monte Carlo approach,

JHEP08(2012)052

namely the use of Bayesian reweighting, can even be accomplished more efficiently using the existing eigenvector PDF sets. Therefore, any future update of the "MSTW 2008" analysis will continue to use the Hessian method with a "dynamic" tolerance.

Acknowledgments

We thank J. Andersen, R. Cousins, G. Cowan, S. Forte, S. Lauritzen, L. Lyons, A.D. Martin, R. McNulty, H. Prosper, J. Pumplin, J. Rojo, G. Salam and W.J. Stirling for useful discussions. The work of R.S.T. is supported partly by the London Centre for Terauniverse Studies (LCTS), using funding from the European Research Council via the Advanced Investigator Grant 267352.

References

[1] G. Watt, *Parton distribution function dependence of benchmark standard model total cross sections at the 7 TeV LHC*, *JHEP* **09** (2011) 069 [arXiv:1106.5788] [INSPIRE].

[2] R.S. Thorne and G. Watt, *PDF dependence of Higgs cross sections at the Tevatron and LHC: response to recent criticism*, *JHEP* **08** (2011) 100 [arXiv:1106.5789] [INSPIRE].

[3] W.T. Giele and S. Keller, *Implications of hadron collider observables on parton distribution function uncertainties*, *Phys. Rev.* **D 58** (1998) 094023 [hep-ph/9803393] [INSPIRE].

[4] W.T. Giele, S.A. Keller and D.A. Kosower, *Parton distribution function uncertainties*, hep-ph/0104052 [INSPIRE].

[5] NNPDF collaboration, R.D. Ball et al., *Parton distributions: determining probabilities in a space of functions*, arXiv:1110.1863 [INSPIRE].

[6] A.D. Martin, W.J. Stirling, R.S. Thorne and G. Watt, *Parton distributions for the LHC*, *Eur. Phys. J.* **C 63** (2009) 189 [arXiv:0901.0002] [INSPIRE].

[7] CMS collaboration, S. Chatrchyan et al., *Measurement of the lepton charge asymmetry in inclusive W production in pp collisions at $\sqrt{s} = 7$ TeV*, *JHEP* **04** (2011) 050 [arXiv:1103.3470] [INSPIRE].

[8] ATLAS collaboration, G. Aad et al., *Measurement of the inclusive $W^{\pm}$ and Z/γ^* cross sections in the electron and muon decay channels in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector*, *Phys. Rev.* **D 85** (2012) 072004 [arXiv:1109.5141] [INSPIRE].

[9] J. Pumplin et al., *Uncertainties of predictions from parton distribution functions. 2. The Hessian method*, *Phys. Rev.* **D 65** (2001) 014013 [hep-ph/0101032] [INSPIRE].

[10] J. Pumplin et al., *New generation of parton distributions with uncertainties from global QCD analysis*, *JHEP* **07** (2002) 012 [hep-ph/0201195] [INSPIRE].

[11] A.D. Martin, R.G. Roberts, W.J. Stirling and R.S. Thorne, *Uncertainties of predictions from parton distributions. 1: Experimental errors*, *Eur. Phys. J.* **C 28** (2003) 455 [hep-ph/0211080] [INSPIRE].

JHEP08(2012)052

[12] J.C. Collins and J. Pumplin, *Tests of goodness of fit to multiple data sets*, hep-ph/0105207 [INSPIRE].

[13] Particle Data Group collaboration, K. Nakamura et. al., *Review of particle physics*, *J. Phys.* **G 37** (2010) 075021 [INSPIRE].

[14] R. Devenish and A. Cooper-Sarkar, *Deep inelastic scattering*, Oxford University Press, Oxford U.K. (2004) [INSPIRE].

[15] J. Feltesse, A. Glazov and V. Radescu, *Experimental error propagation*, in *Parton distributions*, M. Dittmar et al. eds., arXiv:0901.2504 [INSPIRE].

[16] W.K. Tung et al., *Heavy quark mass effects in deep inelastic scattering and global QCD analysis*, *JHEP* **02** (2007) 053 [hep-ph/0611254] [INSPIRE].

[17] J. Pumplin, *Parametrization dependence and $\Delta\chi^2$ in parton distribution fitting*, *Phys. Rev.* **D 82** (2010) 114020 [arXiv:0909.5176] [INSPIRE].

[18] A. Glazov, S. Moch and V. Radescu, *Parton distribution uncertainties using smoothness prior*, *Phys. Lett.* **B 695** (2011) 238 [arXiv:1009.6170] [INSPIRE].

[19] H1 and ZEUS collaborations, F. Aaron et al., *Combined measurement and QCD analysis of the inclusive $e^{\pm}p$ scattering cross sections at HERA*, *JHEP* **01** (2010) 109 [arXiv:0911.0884] [INSPIRE].

[20] H1 and ZEUS collaborations, *PDF fits including HERA-II high Q^2 data*, H1prelim-10-142, ZEUS-prel-10-018 (2010).

[21] H1 and ZEUS collaborations, *HERAPDF1.5 NNLO*, H1prelim-11-042, ZEUS-prel-11-002 (2011).

[22] G. Watt, *MSTW PDFs and impact of PDFs on cross sections at Tevatron and LHC*, *Nucl. Phys. Proc. Suppl.* **222-224** (2012) 61 [arXiv:1201.1295] [INSPIRE].

[23] NNPDF collaboration, R.D. Ball et al., *Unbiased global determination of parton distributions and their uncertainties at NNLO and at LO*, *Nucl. Phys.* **B 855** (2012) 153 [arXiv:1107.2652] [INSPIRE].

[24] J. Pumplin, *Experimental consistency in parton distribution fitting*, *Phys. Rev.* **D 81** (2010) 074010 [arXiv:0909.0268] [INSPIRE].

[25] ATLAS collaboration, G. Aad et al., *Determination of the strange quark density of the proton from ATLAS measurements of the $W \to \ell\nu$ and $Z \to \ell\ell$ cross sections*, *Phys. Rev. Lett.* **109** (2012) 012001 [arXiv:1203.4051] [INSPIRE].

[26] N. Hartland, *LHC data and the proton strangeness*, arXiv:1205.3508 [INSPIRE].

[27] S. Alekhin, *Extraction of parton distributions and α_s from DIS data within the Bayesian treatment of systematic errors*, *Eur. Phys. J.* **C 10** (1999) 395 [hep-ph/9611213] [INSPIRE].

[28] F. De Lorenzi, *Parton distribution function sensitivity studies using electroweak processes at LHCb*, arXiv:1011.4260 [INSPIRE].

[29] F. De Lorenzi, *Parton distribution function studies and a measurement of Drell-Yan produced muon pairs at LHCb*, Ph.D. Thesis, University College Dublin, Dublin Ireland (2011), CERN-THESIS-2011-237.

[30] J. Pumplin et al., *Collider inclusive jet data and the gluon distribution*, *Phys. Rev.* **D 80** (2009) 014019 [arXiv:0904.2424] [INSPIRE].

JHEP08(2012)052

[31] S. Forte, *Parton distributions at the dawn of the LHC*, *Acta Phys. Polon.* **B 41** (2010) 2859 [arXiv:1011.5247] [INSPIRE].

[32] NNPDF collaboration, R.D. Ball et al., *Reweighting NNPDFs: the W lepton asymmetry*, *Nucl. Phys.* **B 849** (2011) 112 [*Erratum ibid.* **B 854** (2012) 926] [arXiv:1012.0836] [INSPIRE].

[33] NNPDF collaboration, R.D. Ball et al., *Reweighting and unweighting of parton distributions and the LHC W lepton asymmetry data*, *Nucl. Phys.* **B 855** (2012) 608 [arXiv:1108.1758] [INSPIRE].

[34] S. Catani, L. Cieri, G. Ferrera, D. de Florian and M. Grazzini, *Vector boson production at hadron colliders: a fully exclusive QCD calculation at NNLO*, *Phys. Rev. Lett.* **103** (2009) 082001 [arXiv:0903.2120] [INSPIRE].

[35] CMS collaboration, *Measurement of the muon charge asymmetry in inclusive W production in pp collisions at* $\sqrt{s} = 7$ *TeV*, CMS PAS EWK-11-005 (2011).

[36] CMS collaboration, *Measurement of the electron charge asymmetry in inclusive W production in pp collisions at* $\sqrt{s} = 7$ *TeV*, CMS PAS SMP-12-001 (2012).

[37] LHCb collaboration, R. Aaij et al., *Inclusive W and Z production in the forward region at* $\sqrt{s} = 7$ *TeV*, *JHEP* **06** (2012) 058 [arXiv:1204.1620] [INSPIRE].

[38] R.S. Thorne, A.D. Martin, W.J. Stirling and G. Watt, *The effects of combined HERA and recent Tevatron* $W \to \ell\nu$ *charge asymmetry data on the MSTW PDFs*, PoS(DIS 2010)052 [arXiv:1006.2753] [INSPIRE].

[39] A. De Roeck and R.S. Thorne, *Structure functions*, *Prog. Part. Nucl. Phys.* **66** (2011) 727 [arXiv:1103.0555] [INSPIRE].

JHEP08(2012)052