PDF uncertainties at the LHC made easy:

Compressed PDF sets, MC to Hessian, specialised PDF sets

Juan Rojo
STFC Rutherford Fellow
Rudolf Peierls Center for Theoretical Physics
University of Oxford

PDF4LHC Meeting
Les Houches, 04/06/2015

Outline

- Compressed Monte Carlo PDFs **CMC-PDFs**
- A Hessian representation of Monte Carlo PDFs **MC2Hessian**
- Specialised Minimal PDFs for specific applications **SM-PDFs**

In collaboration: **S. Carrazza**, **S. Forte**, **K. Zassabov**, **J. I. Latorre** and **G. Watt**

This talks summarises the **completion of an exhaustive program** that aims to deliver **optimal tools for the estimation of PDF uncertainties at Run II**

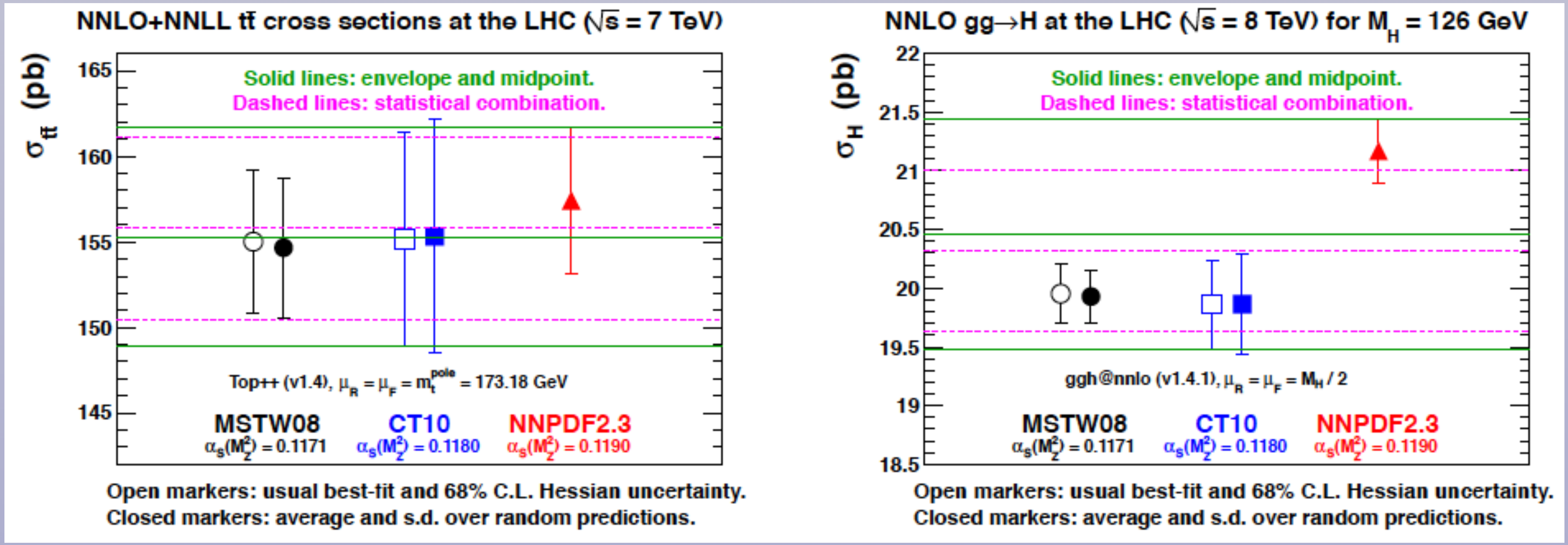
Monte Carlo Combination of PDF sets

Basic strategy

- Select the **PDF sets that enter the combination**. Results here based on NNPDF3.0, CT14 and MMHT14. Assume **equal prior likelihood** of the three sets in the combination.
- Transform the **Hessian PDF sets into their Monte Carlo representation (Watt and Thorne 12)**

$$F(\mathcal{S}_k) = F(\mathcal{S}_0) + \sum_{j=1}^{n} \Bigl[F(\mathcal{S}_j^{\pm}) - F(\mathcal{S}_0) \; |R_{jk}| \qquad (k = 1, \ldots, N_{\rm pdf}$$

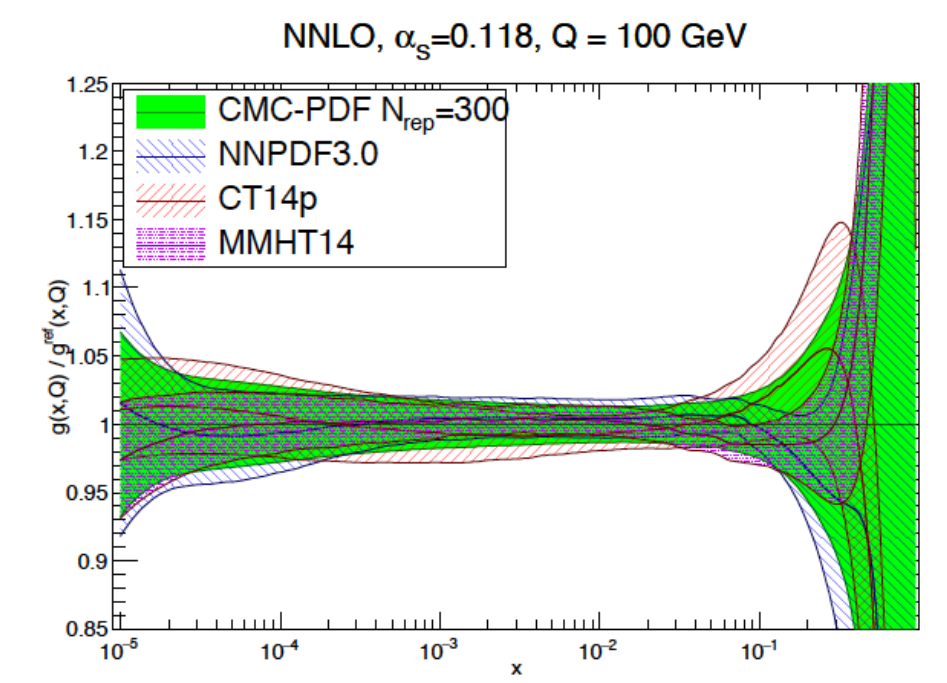
- Now combine the **same number of replicas** from each of the three sets. First proposed by **Forte 12**
- The resulting Monte Carlo ensemble has a **robust statistical interpretation,** and in many cases leads to similar results, with somewhat smaller uncertainties, compared to the **original PDF4LHC envelope**.

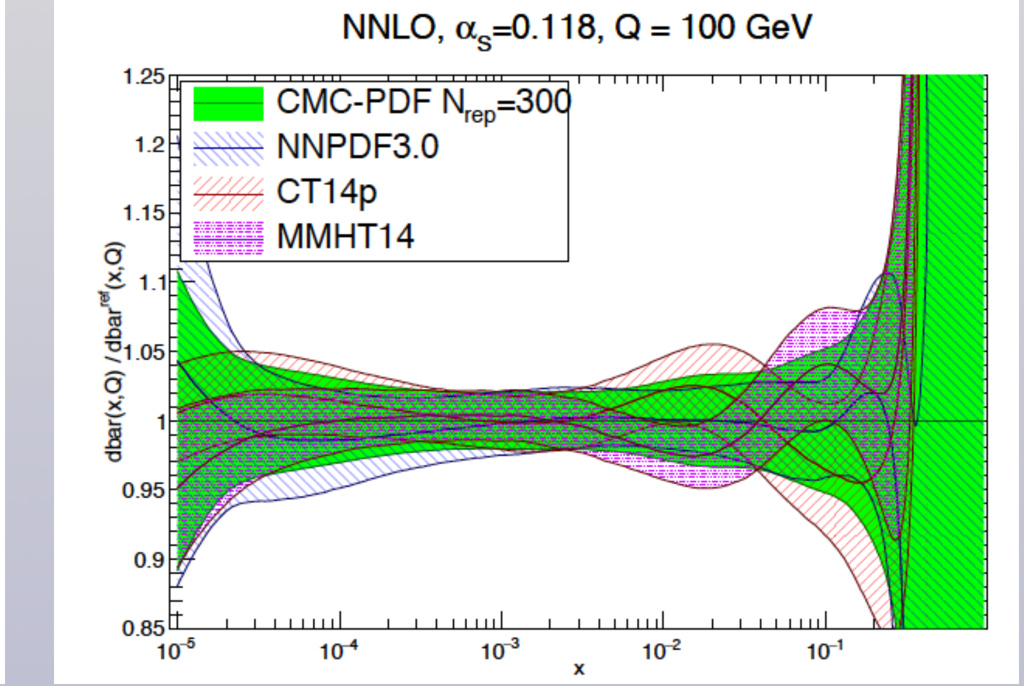


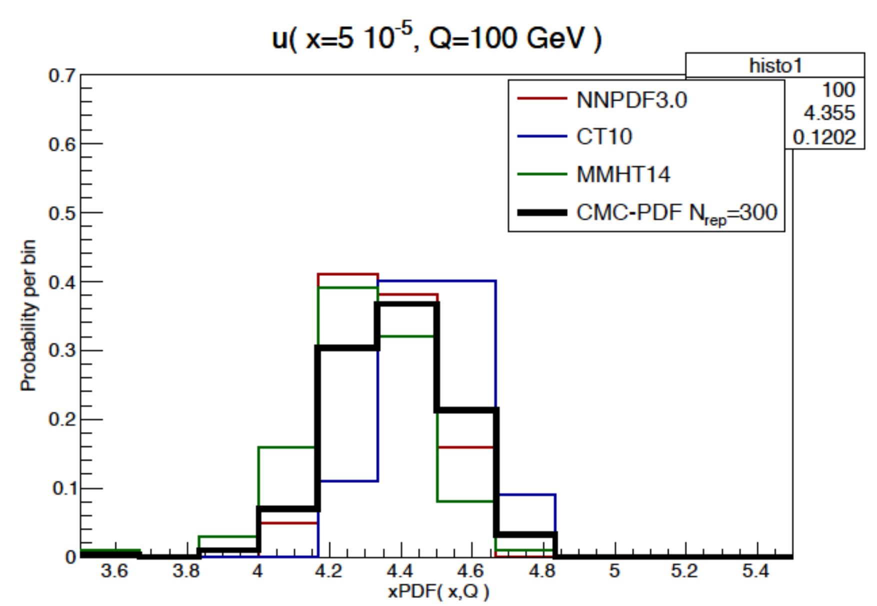
Forte and Watt 13

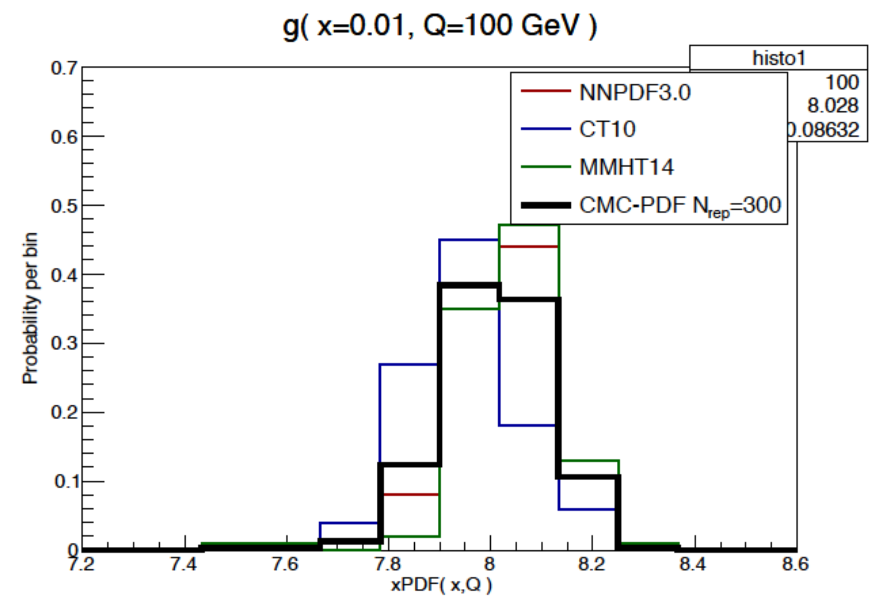
The combined Monte Carlo PDF set

- **NNPDF3.0, MMHT14 and CT14p, all at NNLO, with alphas=0.118**





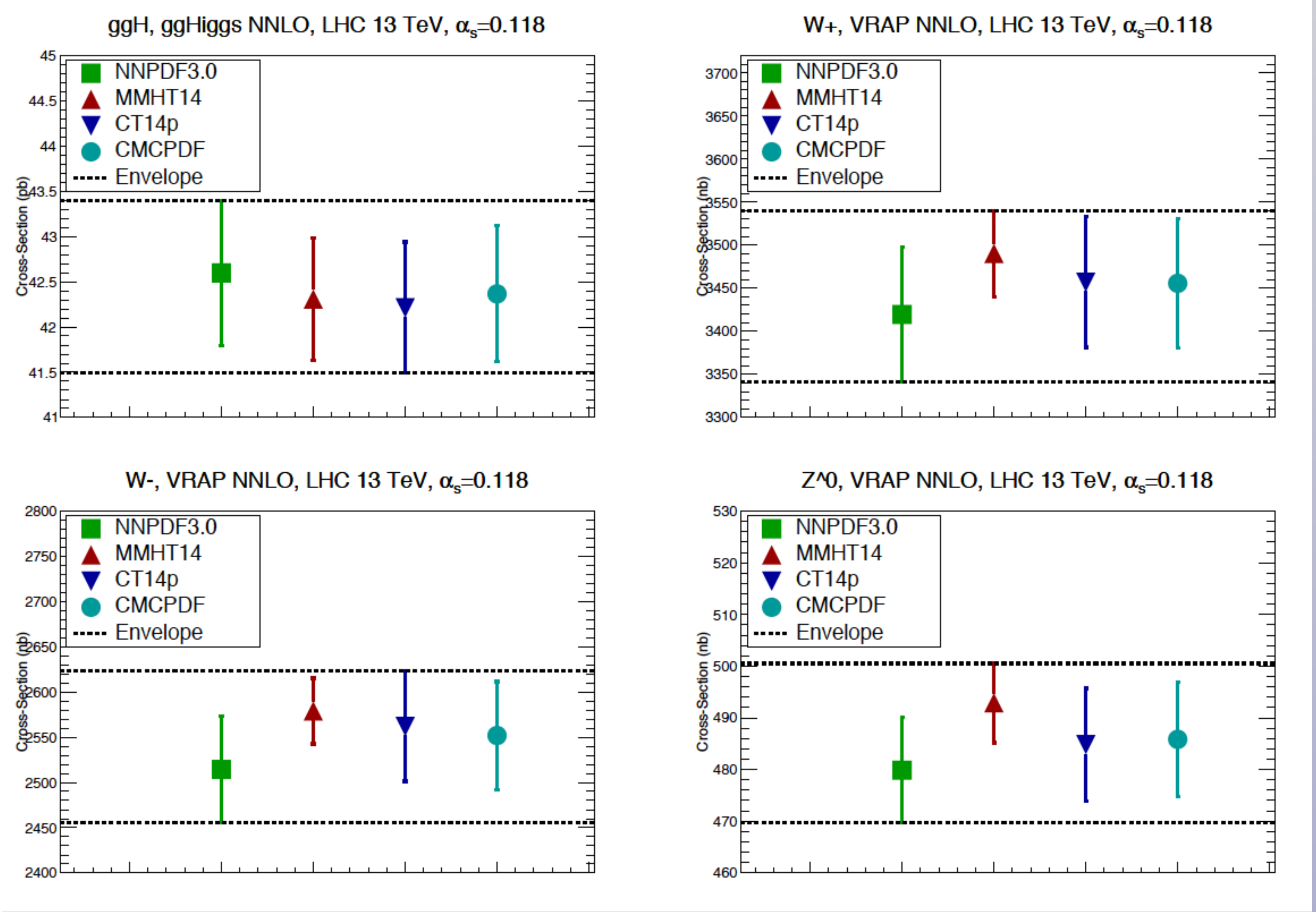




Combined MC set vs PDF4LHC envelope

- As already noted in the **Forte-Watt** study, the **MC combination** leads to somewhat smaller uncertainties than the **PDF4LHC envelope** (same here and in the Meta-PDFs)

- This can be understood because now **each PDF set receives the same weight**, while the PDF4LHC envelope effectively gives more weight to the **outliers**



- Main drawback of this combination method: **too many MC replicas!**

- Between 300 and 1000.

- Need to **reduce this number** …

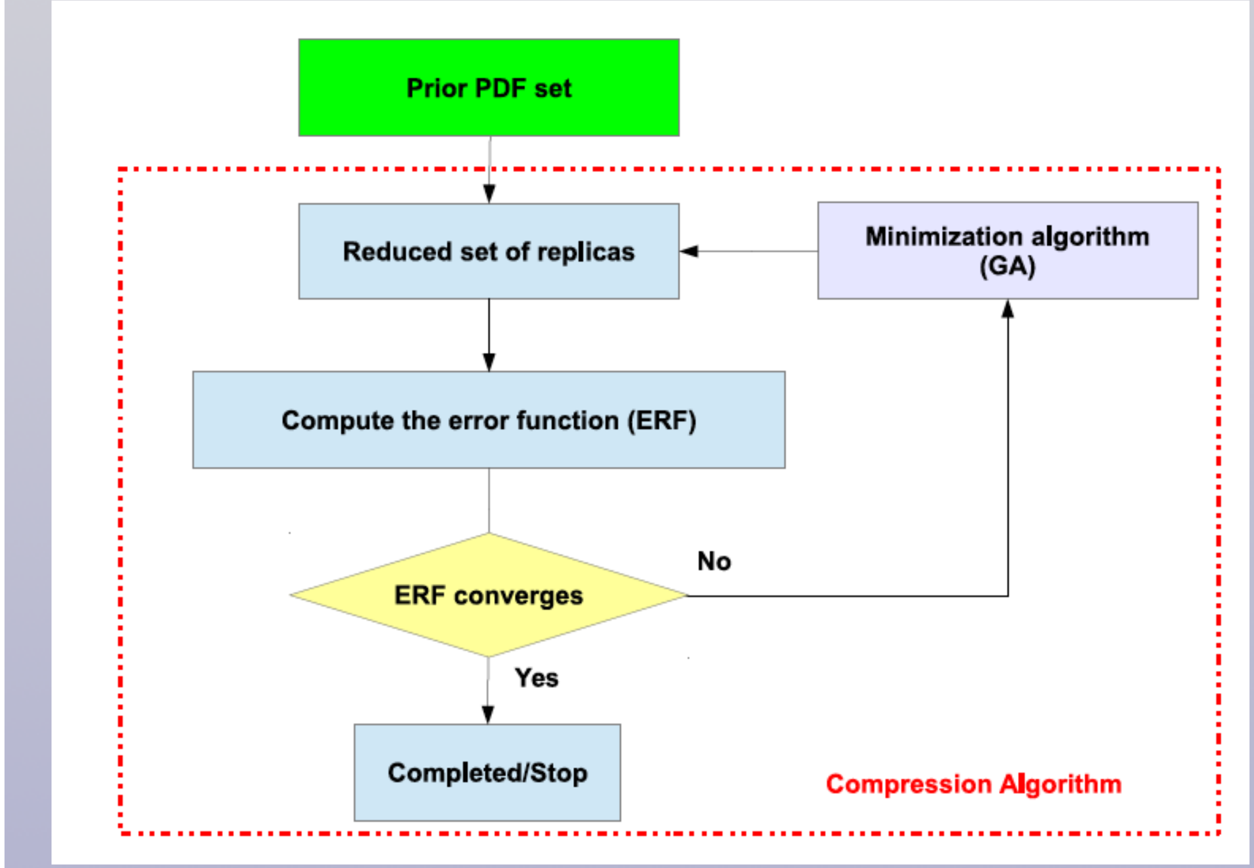
CMC-PDFs

in collaboration with **S. Carrazza, J. I. Latorre** and **G. Watt**
arXiv:1504.06459

Compressed Monte Carlo PDFs

- Start from **combined MC set:** statistically sensible combination, but **too large number of replicas**

- **Compress the original probability distribution** to one with a **smaller number of replicas**, in a way that all the relevant estimators (mean, variances, correlations etc) for the PDFs are reproduced



- Minimise **distance between prior and compressed sets** using Genetic Algorithms

$$\mathrm{ERF} = \sum_k \frac{1}{N_k} \sum_i \left(\frac{C_i^{(k)} - O_i^{(k)}}{O_i^{(k)}} \right)^2 ,$$

- Various options about how the **error function** to be minimised can be defined, ie., to reproduce central values add a term

$$ERF_{CV} = \frac{1}{N_{\mathrm{CV}}} \sum_{i=-n_f}^{n_f} \sum_{j=1}^{N_x} \left(\frac{f_i^{\mathrm{CV}}(x_j,Q) - g_i^{\mathrm{CV}}(x_j,Q)}{g_i^{\mathrm{CV}}(x_j,Q)} \right)^2$$

$$f_i^{\mathrm{CV}}(x_j,Q) = \frac{1}{N_{\mathrm{rep}}} \sum_{r=1}^{N_{\mathrm{rep}}} f_i^r(x_j,Q)$$

- Same for **variances, correlations and higher moments**

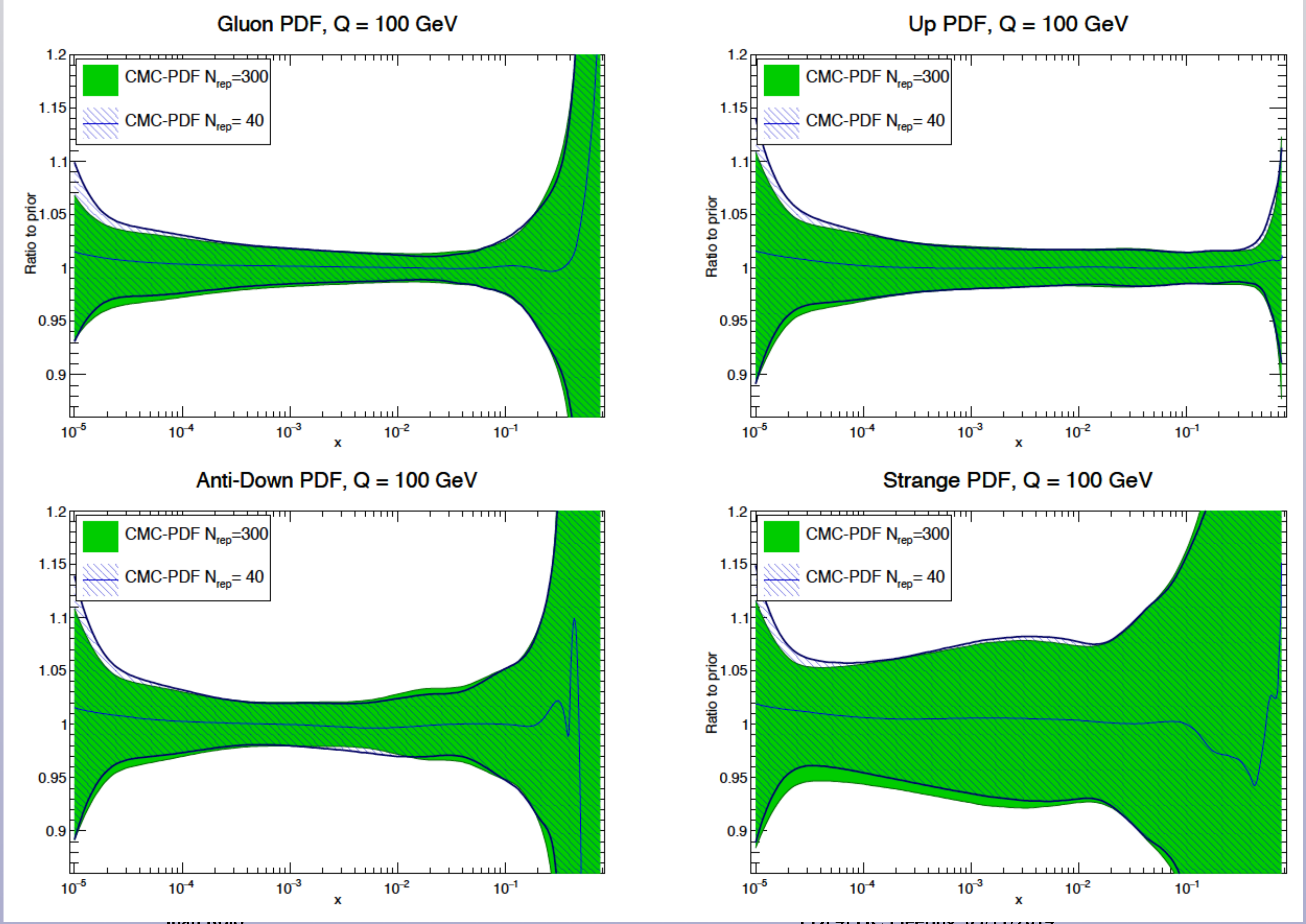
$$ERF_{KOL} = \frac{1}{N_{KOL}} \sum_{i=-n_f}^{n_f} \sum_{j=1}^{N_x} \sum_{k=1}^{6} \left(\frac{F_i^k(x_j,Q) - G_i^k(x_j,Q)}{G_i^k(x_j,Q)} \right)^2$$

$$N_{KOL} = \frac{1}{N_{\mathrm{rand}}} \sum_{d=1}^{N_{\mathrm{rand}}} \sum_{i=-n_f}^{n_f} \sum_{j=1}^{N_x} \sum_{k=1}^{6} \left(\frac{R_i^k(x_j,Q) - G_i^k(x_j,Q)}{G_i^k(x_j,Q)} \right)^2$$

- The algorithm also minimises the **Kolmogorov distance** between the original and compressed distributions

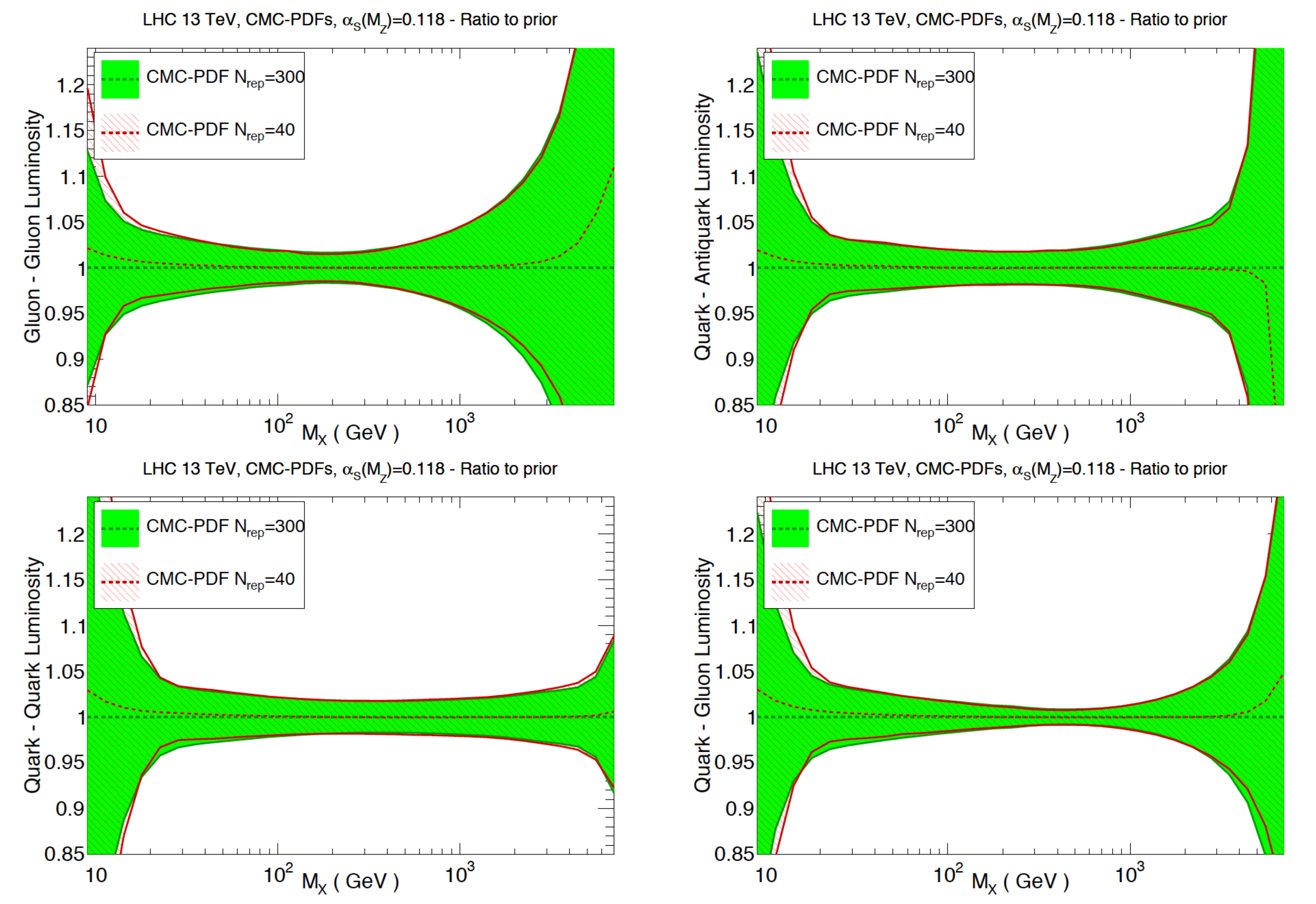
Results of the compression

- For example, for **N$_{\text{rep}}$=40** replicas the **compressed and the original PDFs** are virtually identical



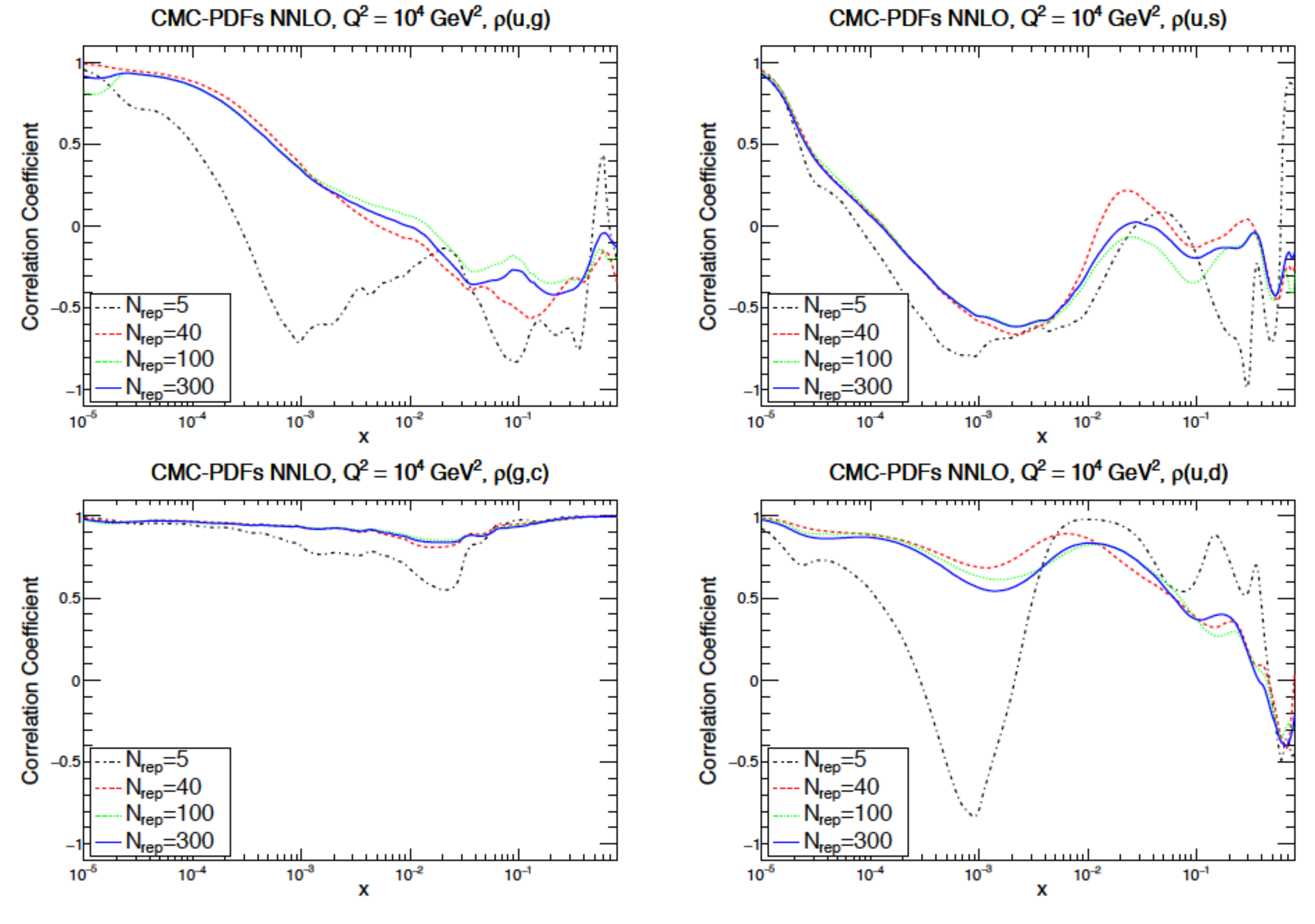
Results of the compression

- For example, for N_{rep}=40 replicas the compressed and the original PDFs are virtually identical



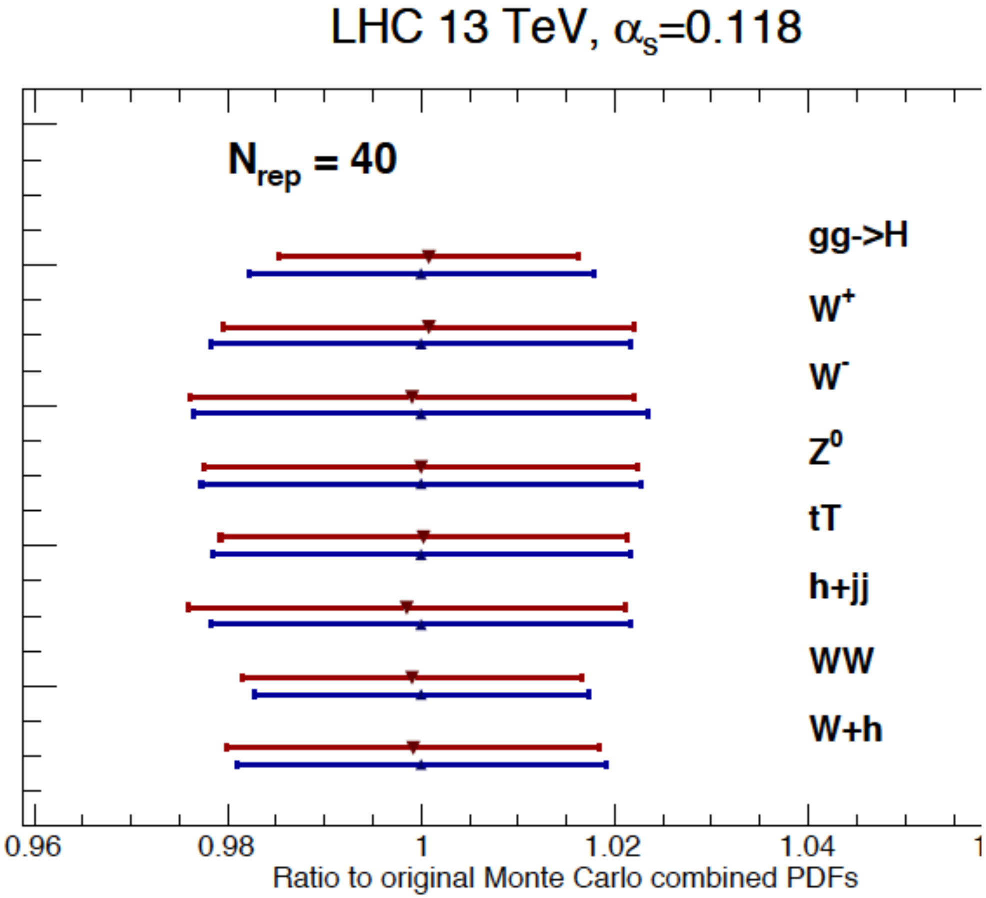
Results of the compression

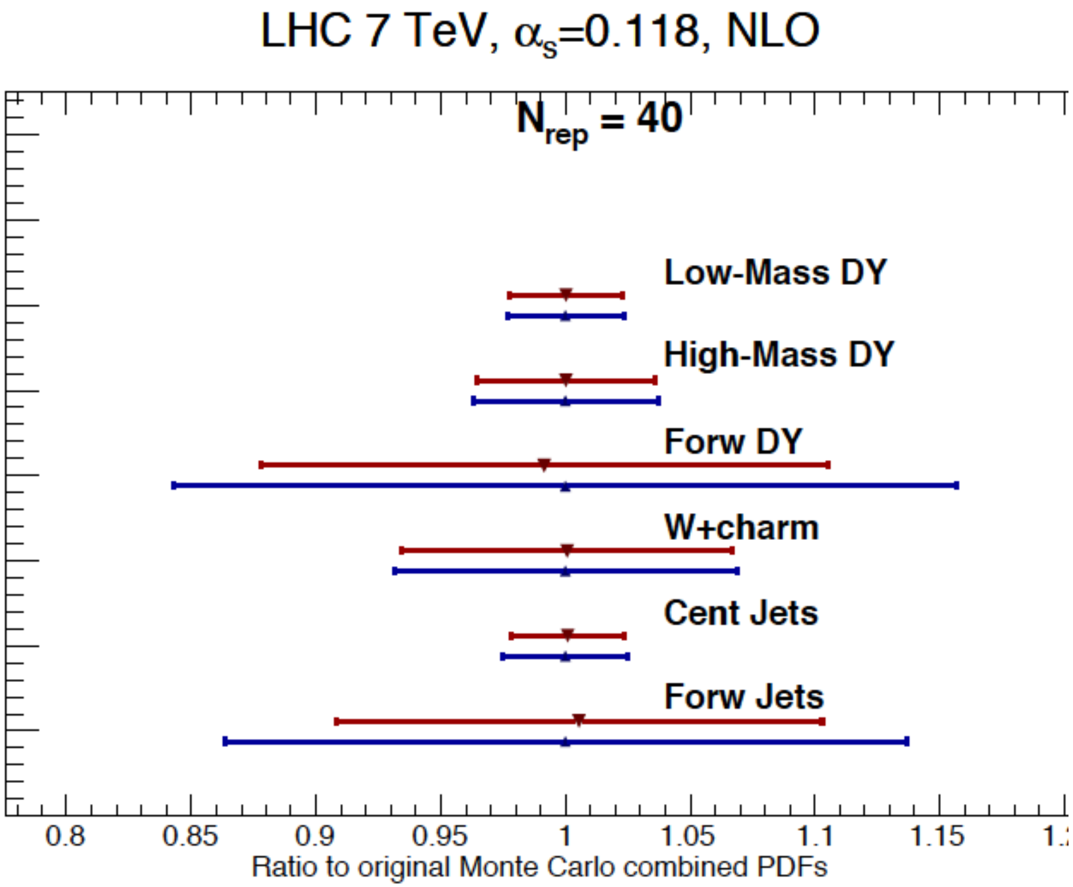
- **Correlations between PDFs are also nicely reproduced**



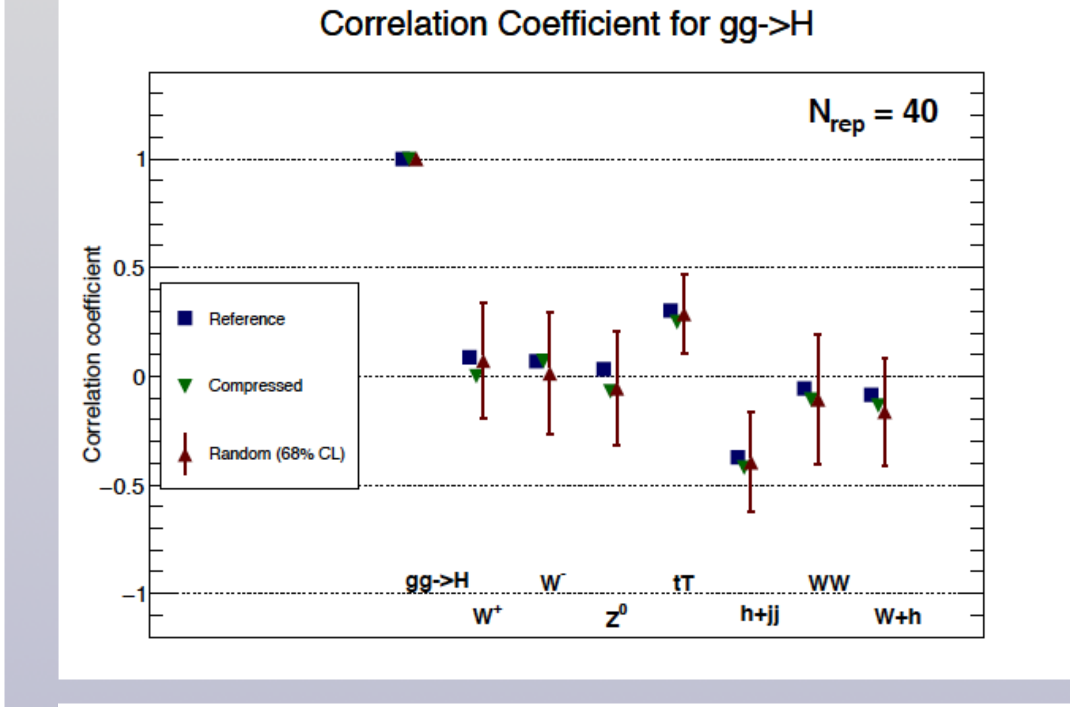
Phenomenology

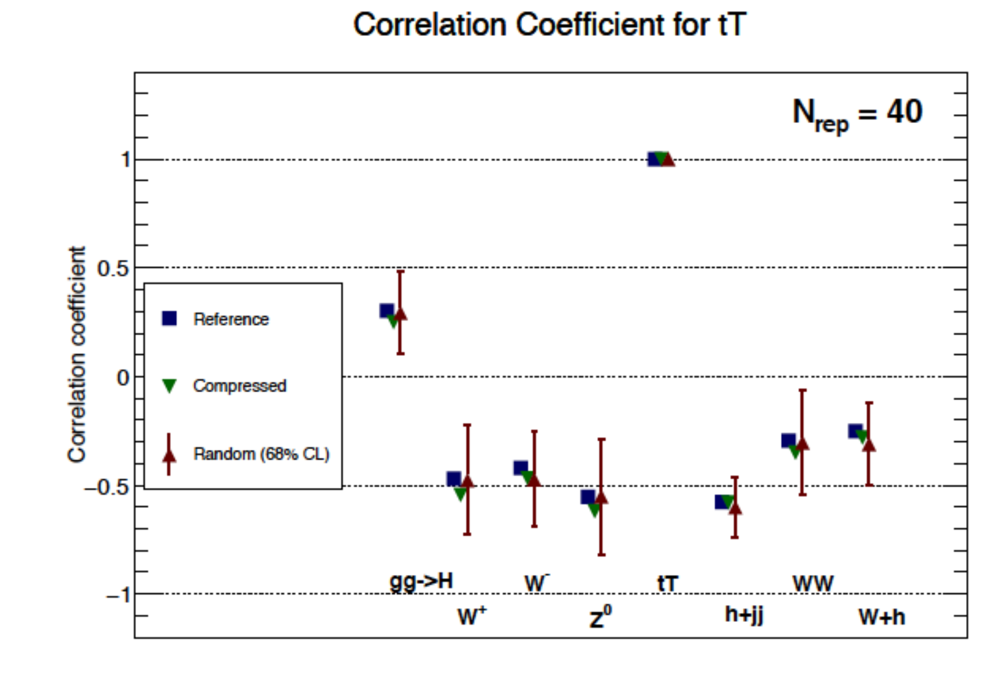
- As expected from the good agreement at the PDF level, CMC-PDFs are also validated at the level of **LHC inclusive cross-sections and differential distributions**

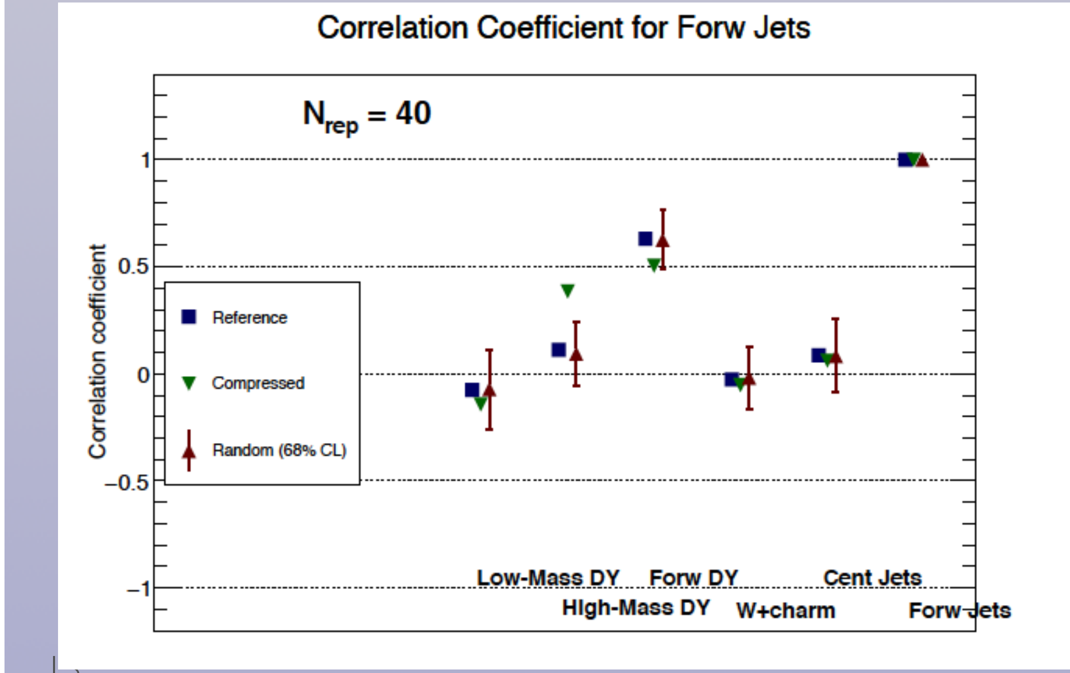


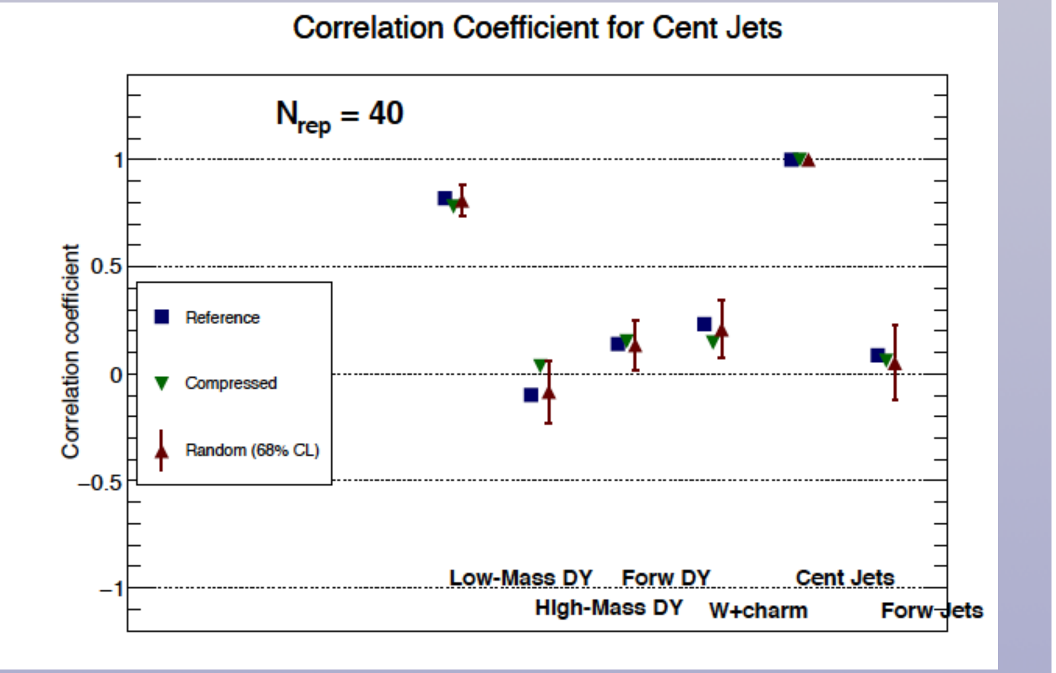


Correlations of LHC xsecs


Correlation Coefficient for gg->H

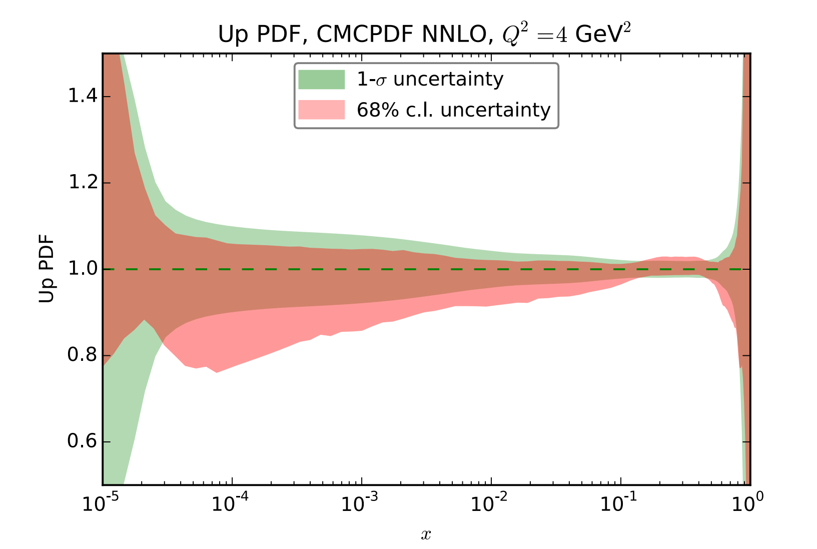

Correlation Coefficient for tT

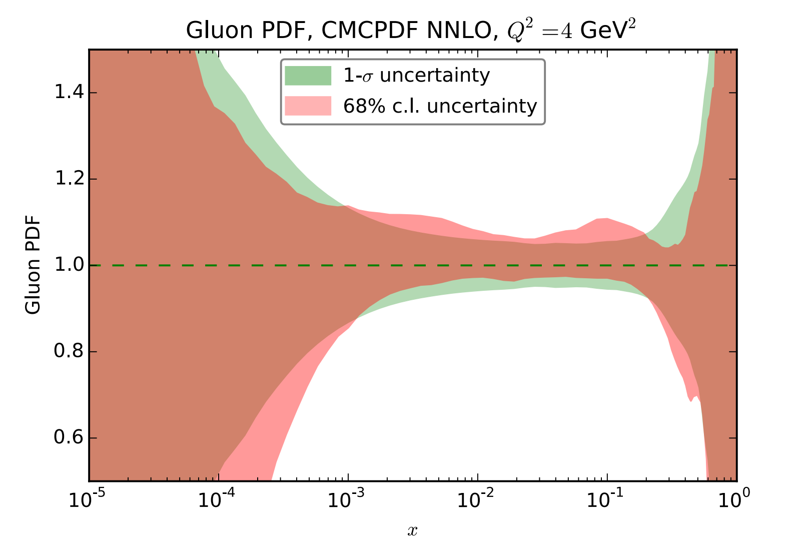

Correlation Coefficient for Forw Jets


Correlation Coefficient for Cent Jets

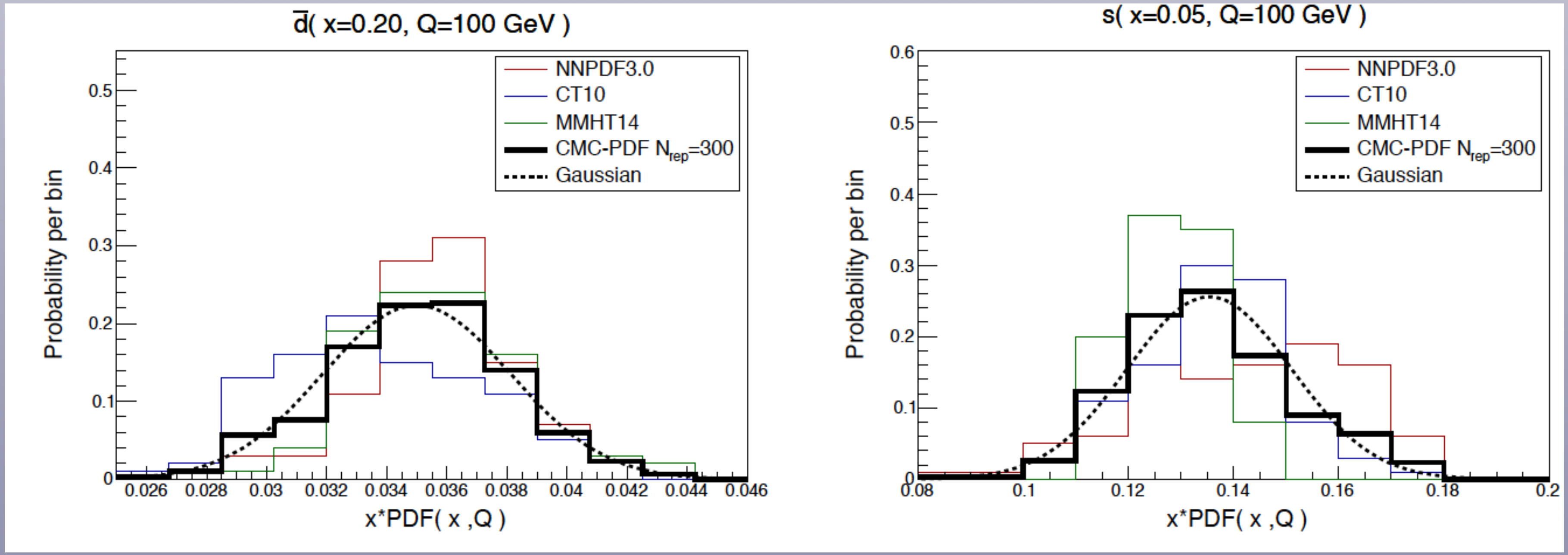
Non-gaussianities in the combined set

- One intrinsic limitation of **any Hessian representation** is that, to begin with, the **Monte Carlo combination of PDF sets is not Gaussian.** Compare **68% CL and one-sigma intervals for CMC300**





- For some PDF flavours and regions of **x, a Gaussian** does not reproduce **the full probability density**



CMC-PDF summary

- The Monte Carlo combination of PDFs results in a **non-Gaussian distribution**

- The **CMC-PDF algorithm** successfully reproduces **all the statistical properties of the prior distribution**, including **higher moments and correlations**, while reducing substantially the number of replicas

- However, one might argue that in regions where the **Gaussian approximation** is reasonable, one should use a Hessian representation. However, we want to do this **without the need to use any fixed functional form** and thus avoid the usual bias

- MC2hessian: **using the MC replicas themselves as basis of linear expansion**

- **Having CMC-PDFs** both in the MC and Hessian representation allows to **determine precisely where the latter approximation breaks down** - typically in regions with few data, like large-x, that are crucial for searches mc^{2}hessian

MC2Hessian

in collaboration with **S. Carrazza**, **S. Forte**, **K. Zassabov** and **J. I. Latorre**
arXiv:1505.06736

The basic strategy

- We want to construct a Hessian (linear) representation of MC PDFs, without resorting to any ad-hoc meta-parametrization: **use the same MC replicas as expansion basis!**

$$f_\alpha^{(k)} \approx f_{H,\alpha}^{(k)} \equiv f_\alpha^{(0)} + \sum_{i=1}^{N_{\rm eig}} a_i^{(k)} (\eta_\alpha^{(i)} - f_\alpha^{(0)}) \, , \quad k = 1, \ldots, N_{\rm rep} \, ,$$

- The **values of the expansion coefficients,** and their range of variation, is determined from a figure of merit in the space of PDF

$$\chi^{2(k)}_{\rm pdf} \equiv \sum_{i,j=1}^{N_x} \sum_{\alpha,\beta=1}^{N_{\rm f}} \left(\left[f_{H,\alpha}^{(k)}(x_i,Q_0^2) - f_\alpha^{(k)}(x_i,Q_0^2) \right] \cdot \left({\rm cov}^{\rm pdf} \right)^{-1}_{ij,\alpha\beta} \cdot \left[f_{H,\beta}^{(k)}(x_j,Q_0^2) - f_\beta^{(k)}(x_j,Q_0^2) \right] \right) .$$

- Then the **usual Hessian technology** can now be applied to diagonalize eigenvectors and compute PDF uncertainties using now an orthogonal basis

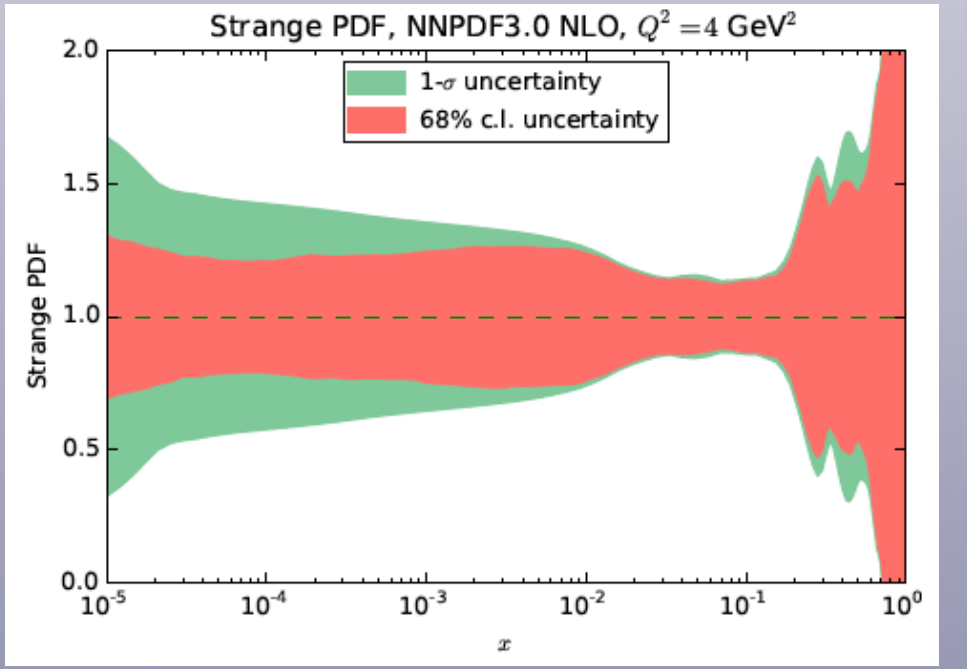
$$\sigma^{\rm PDF}_{H,\alpha}(x,Q^2) = \sqrt{ \sum_{i=1}^{N_{\rm eig}} \left[\sum_{j=1}^{N_{\rm eig}} \frac{v_{ij}}{\sqrt{\lambda_i}} \left(\eta_\alpha^{(j)}(x,Q^2) - f_\alpha^{(0)}(x,Q^2) \right) \right]^2 } \, ,$$

$$\sigma^{\rm PDF}_{H,\alpha}(x,Q^2) = \sqrt{ \sum_{i=1}^{N_{\rm eig}} \left(\widetilde{f}_\alpha^{(i)}(x,Q^2) - f_\alpha^{(0)}(x,Q^2) \right)^2 }$$

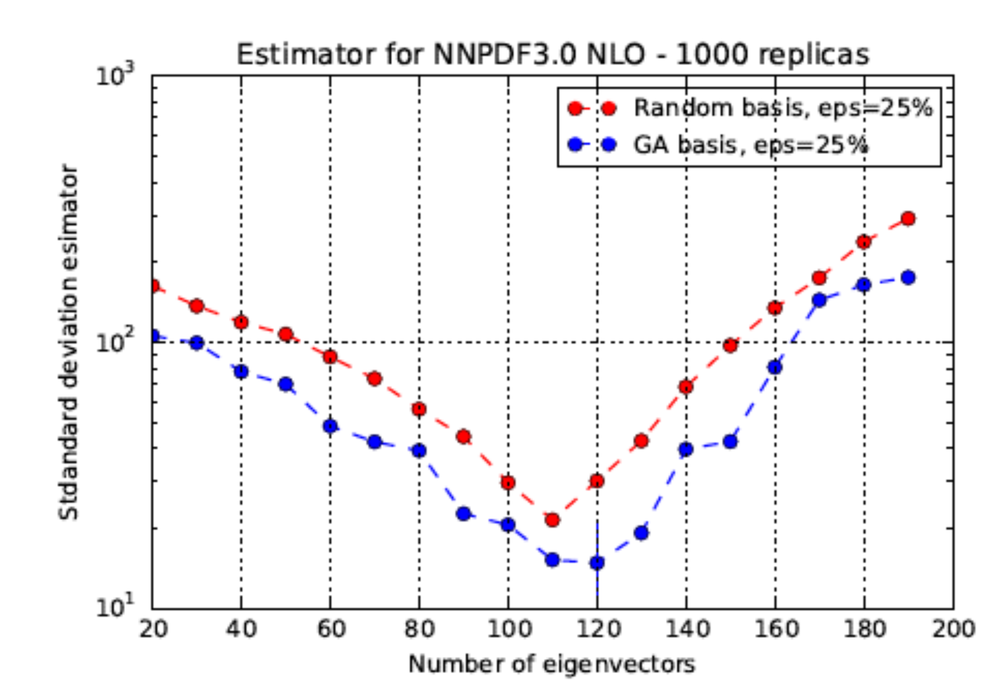
Basis selection

- A key ingredient of the method is the selection of the expansion basis
- Use **Genetic Algorithms** to select the optimal replicas for the expansion
- Include in the GA only those p**oints in x and PDF flavour where 1-sigma and 68% CL intervals are not too different:** we want to construct a Hessian **only where the underlying PDF distribution is Gaussian**
- The **redundancy of a MC PDF set** implies that there is an optimal number of eigenvectors that minimises the difference between Hessian and MC

1-sigma vs 68% CL for NNPDF3.0

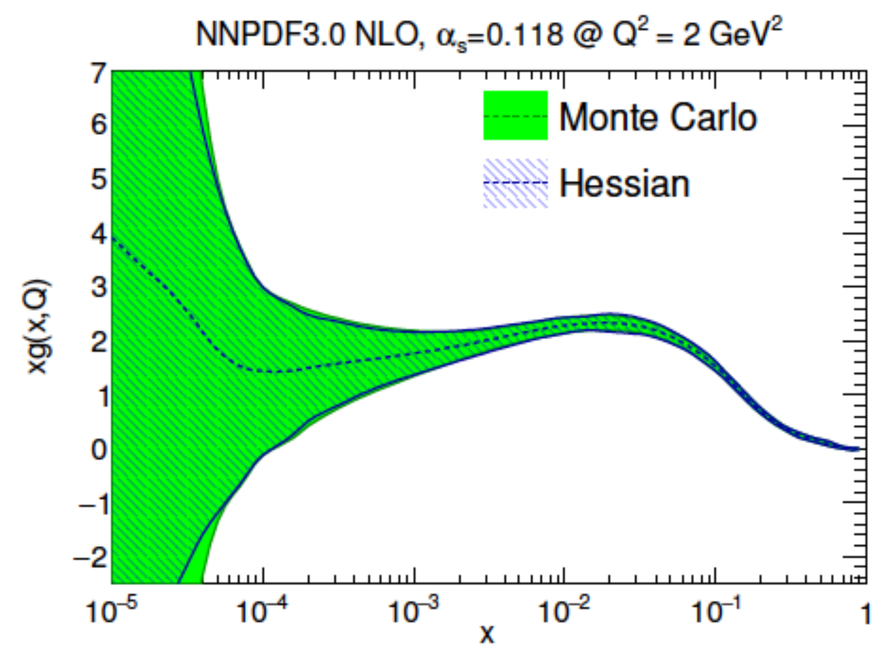


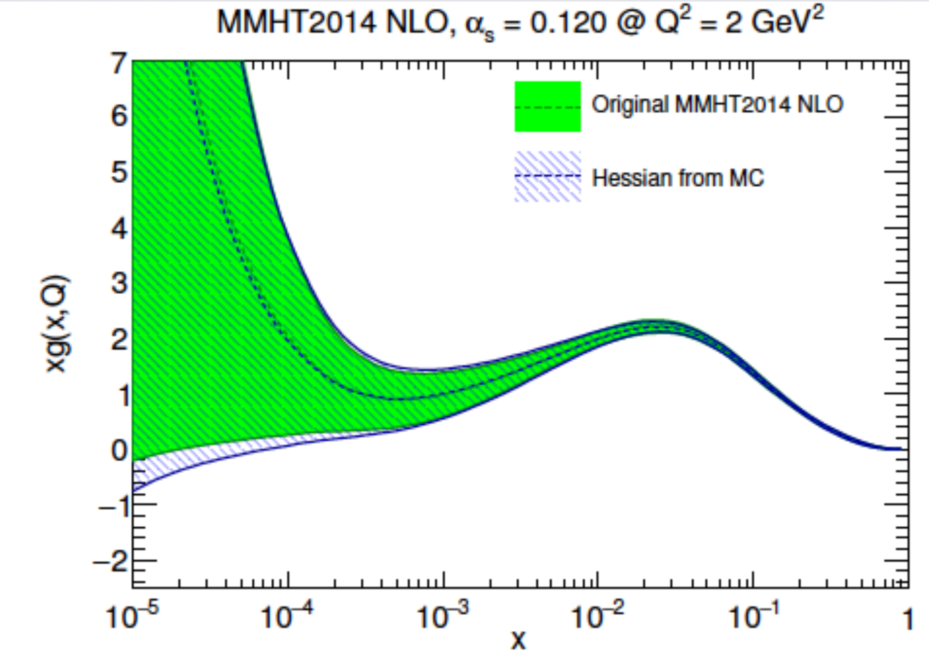
Determination of optimal Neig



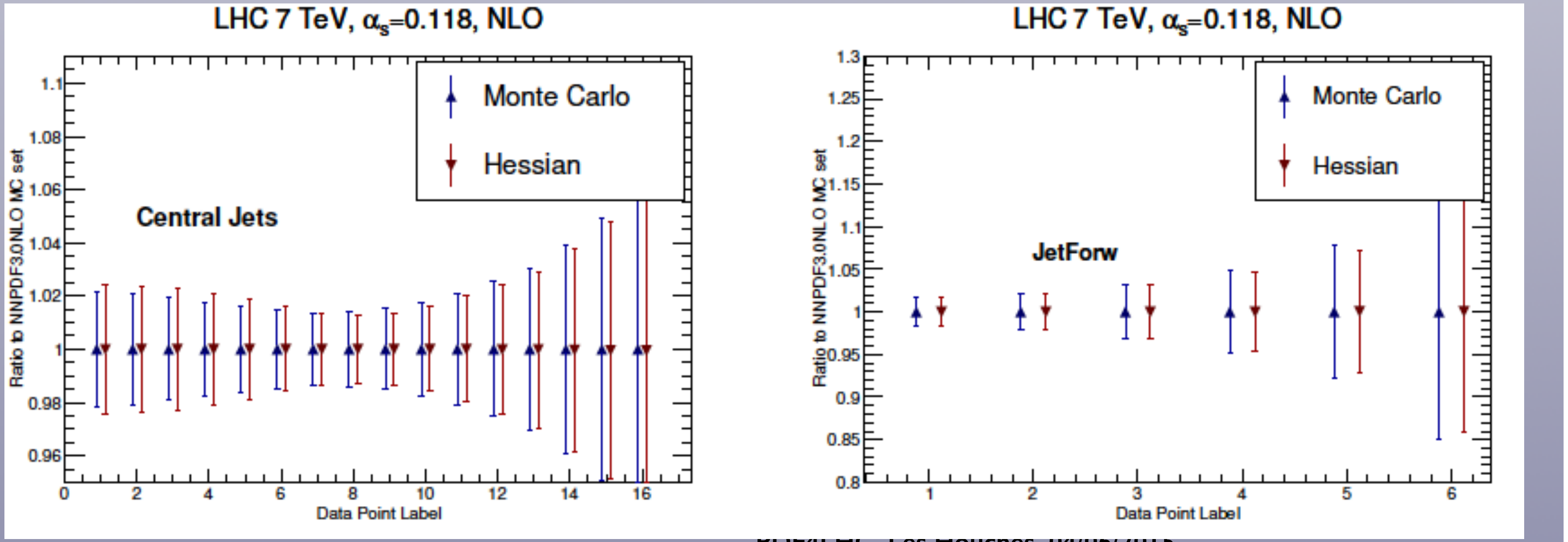
Results: NNPDF3.0 and MMHT14

- The MC2Hessian algorithm succeeds in **reproducing a native MC set** (NNPDF3.0) and a **native Hessian set converted to the MC representation** (MMHT14)





Extensive validation at the level of **PDF correlations** and **LHC inclusive xsec and differential distributions** as well



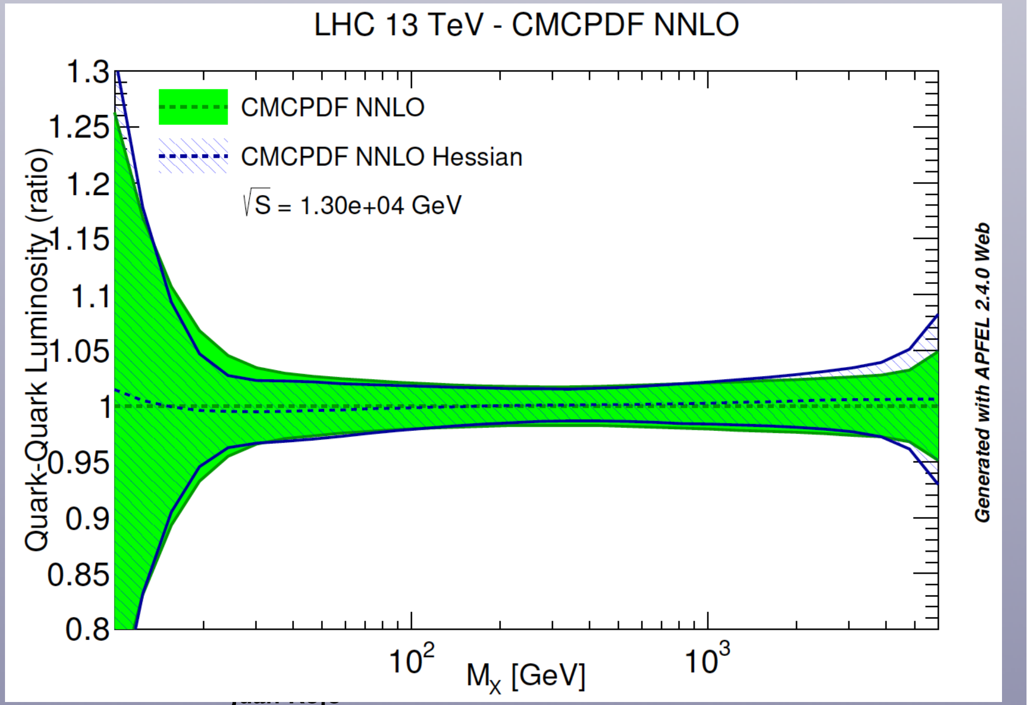
CMC-H PDFs

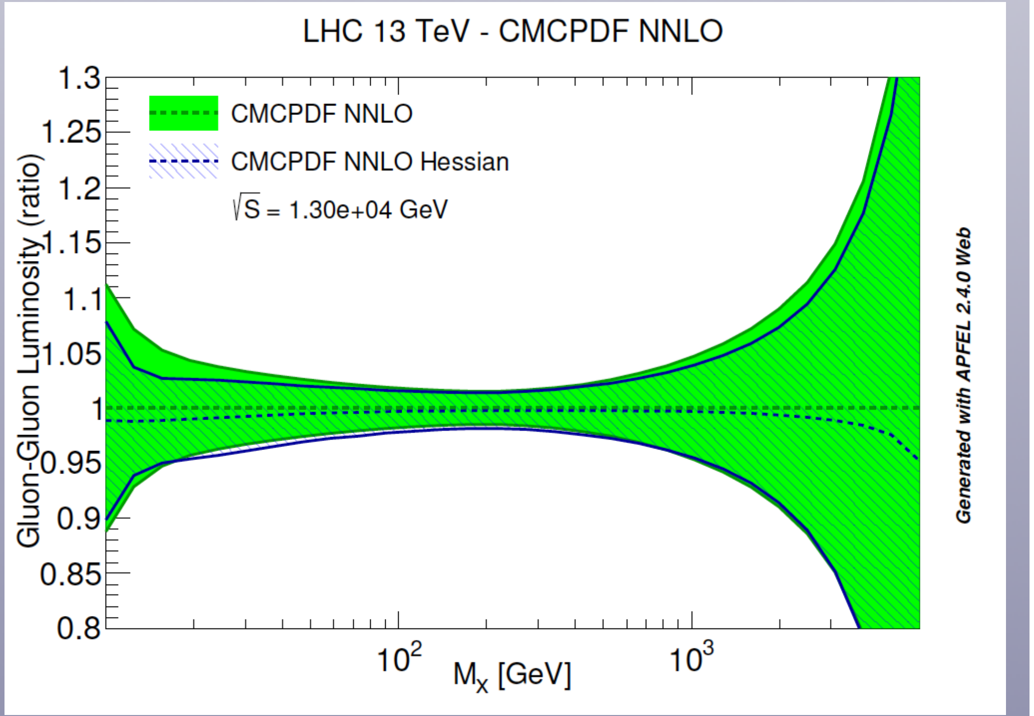
- The **MC2Hessian algorithm** can also be applied to the original **300 replicas** of the combination of NNPDF3.0, CT14 and MMHT14 to obtain **a Hessian version of the combined PDF set**

- Note that **no compression** is required here: the **Hessian representation effectively performs a compression**, since from **300 replicas** we reduce to **90 symmetric eigenvectors**

- When applied to CMC-PDF, the MC2Hessian is exactly the **same idea as the Meta-PDFs** but avoiding the use of a potentially biased ad-hoc **meta-parametrization**

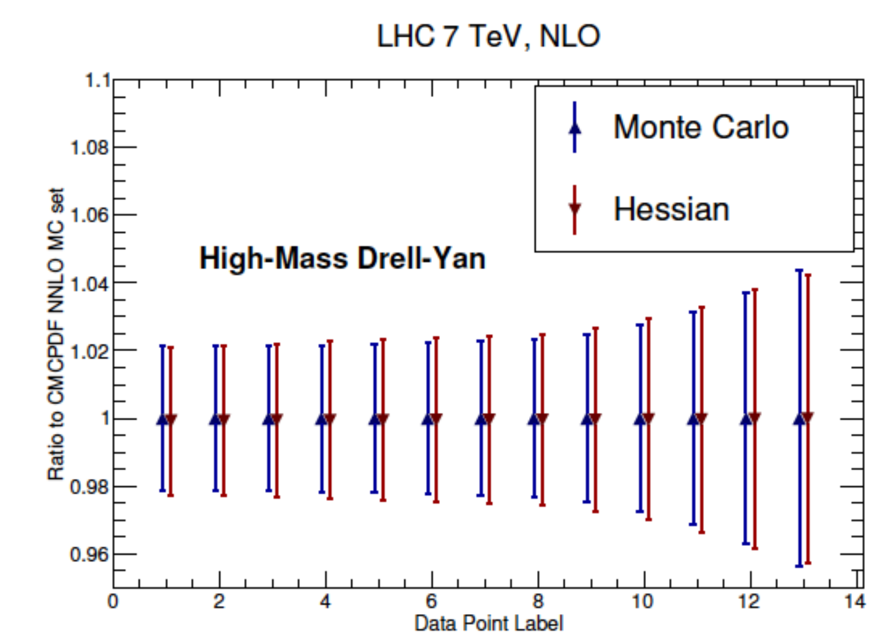
- **Reasonable agreement** with original MC combination, except of course in regions where to begin with the **prior distribution is non-Gaussian** (like large invariant masses)

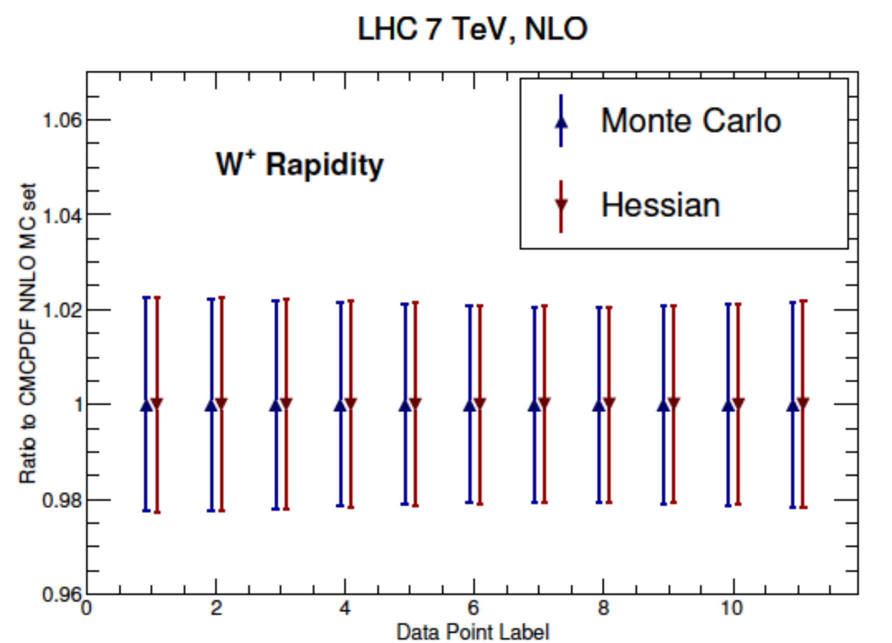


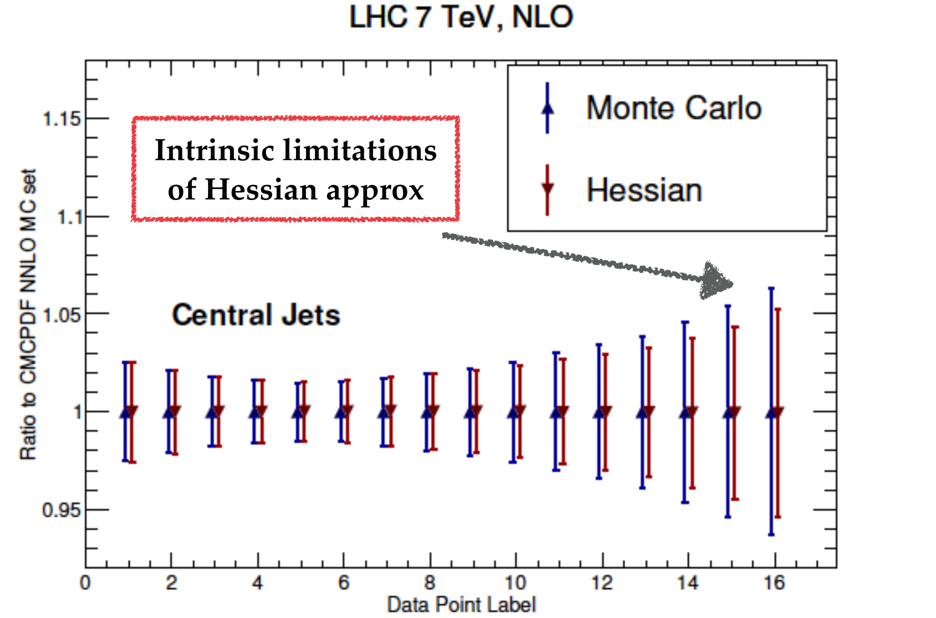


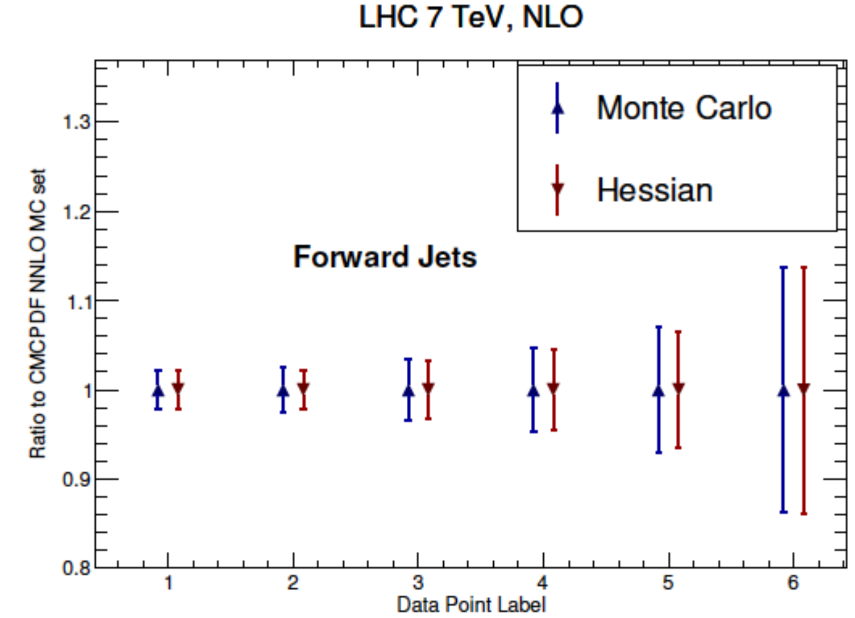
CMC-H PDFs - LHC phenomenology

Extensive validation of the CMC-H PDFs for a **wide range of LHC differential distributions**






Intrinsic limitations of Hessian approx



MC2Hessian summary

- The Monte Carlo combination of PDFs results in a **non-Gaussian distribution,** in general: it might be dangerous to adopt a Hessian representation as baseline.

- In those regions where the Gaussian approximation is reasonable, one can construct a Hessian representation **without the need to use any fixed functional form** and thus avoid the usual bias, **by using the MC replicas themselves as basis of linear expansion**

- **The MC2Hessian algorithm** successfully reproduces both native MC sets (NNPDF3.0) and MC sets obtained from native Hessian sets (MMHT14)

- When applied to the combined MC PDF set, **no need to perform the compression step:** the algorithm manages to perform an efficient reduction from 300 replicas to 90 symmetric eigenvectors

- This latter point is essentially **same idea as META-PDF**s but without the need of introducing an ad-hoc meta-parametrization

- One might still complain that Hessian PDF sets are great, but that for **specific applications** (Higgs, top quad physics, W mass, …) one should use s**pecialised minimal sets with only a small number of eigenvectors** relevant for these processes -> thus we have developed the **Specialized Minimal PDFs**

SMPDF

Specialized Minimal PDFs

in collaboration with **S. Carrazza, S. Forte and K. Zassabov**

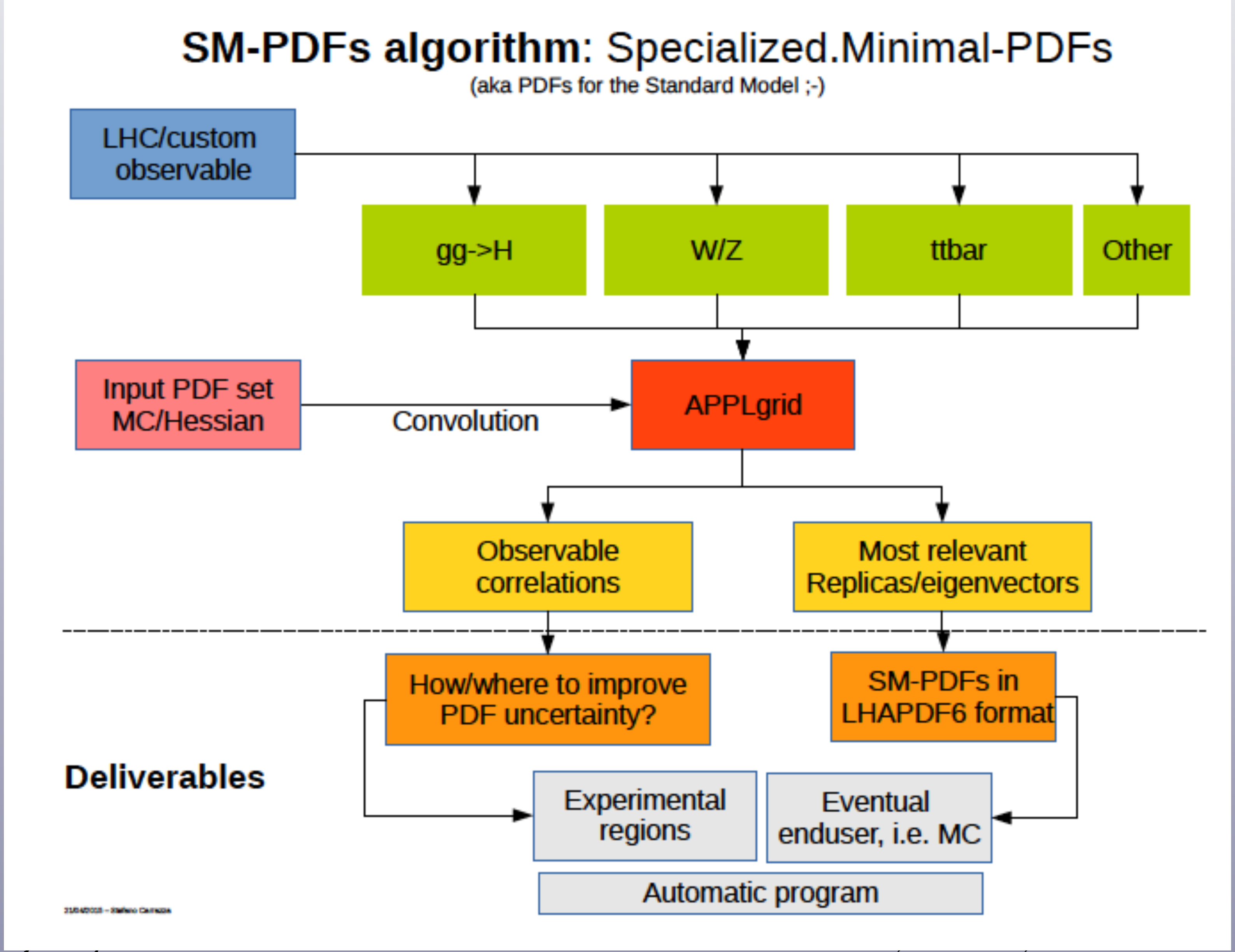
arXiv:1506.aaaaa

Motivation

- **PDF sets specialised to a particular set of processes** might be useful to reduce the CPU burden of theory calculations, and facilitate the **treatment of PDF uncertainties and PDF-induced correlations**
- One clear example are **Higgs analysis in the HXS WG**, but similar **Specialized Minimal PDFs** can be useful in the **TOP4LHC WG (if restricted to top physics)** and for the **W mass measurement (if restricted to W, Z production)**
- In native Hessian sets, one can use the **Dataset Diagonalization method** of J. Pumplin (**arxiv:0904.2425**), but this method has some limitations, including:
  - Relies on the existence of a **relatively simple PDF parametrisation**
  - Selection of relevant eigenvectors performed at the **level of selected cross-section**s, not of the underlying PDF regions of x and flavours
- Using the techniques developed with the MC2Hessian algorithm, we have found a new strategy to produced **Specialized Minimal PDFs (SM-PDFs)** for specific applications

General strategy

Basic idea: select which observables are relevant for some specific application, generate APPLgrids for these, determine the region where the correlations between PDFs and these observables is large, and perform the MC2Hessian transformation only for these



SM-PDFs algorithm: Specialized.Minimal-PDFs

(aka PDFs for the Standard Model ;-)

LHC/custom observable

gg->H

W/Z

ttbar

Other

Input PDF set MC/Hessian

Convolution

APPLgrid

Observable correlations

Most relevant Replicas/eigenvectors

Deliverables

How/where to improve PDF uncertainty?

SM-PDFs in LHAPDF6 format

Experimental regions

Eventual enduser, i.e. MC

Automatic program

General strategy (II)

- Define the **observables one wants to reproduce with the specialised PDFs**: total xsecs, differential distributions etc. For example, for **Higgs: ggH inclusive xsec, pt dist, rap dist, same for VFB Higgs etc**

- Compute APPLgrids for these observables using **MadGraph5_aMC@NLO** with **aMCfast**

- Compute, for all these selected observables, correlations with the input PDFs:

$$\rho_{\rm MC}^{\alpha\sigma}(x,Q^2)=\frac{N_{\rm rep}}{N_{\rm rep}-1}\left(\frac{\left\langle f_\alpha^{(k)}(x,Q^2)\sigma^{(k)}\right\rangle_{\rm rep}-\left\langle f_\alpha^{(k)}(x,Q^2)\right\rangle_{\rm rep}\left\langle\sigma^{(k)}\right\rangle_{\rm rep}}{s_\alpha^{\rm PDF}(x,Q^2)\cdot s_y}\right)$$

- **Keep only those points in *x* and PDF flavours** for which the correlation between PDFs and selected cross-sections is above some certain threshold, for all processes considered:

$$X_{\alpha\sigma}=\left\{x\in[0,1]:\rho_{\rm MC}^{\alpha\sigma}(x,Q^2)>t\rho_{\rm max}\right\}$$

$$X_{\alpha\Sigma}=\bigcup_{\sigma\in\Sigma}X_{\alpha\sigma}$$

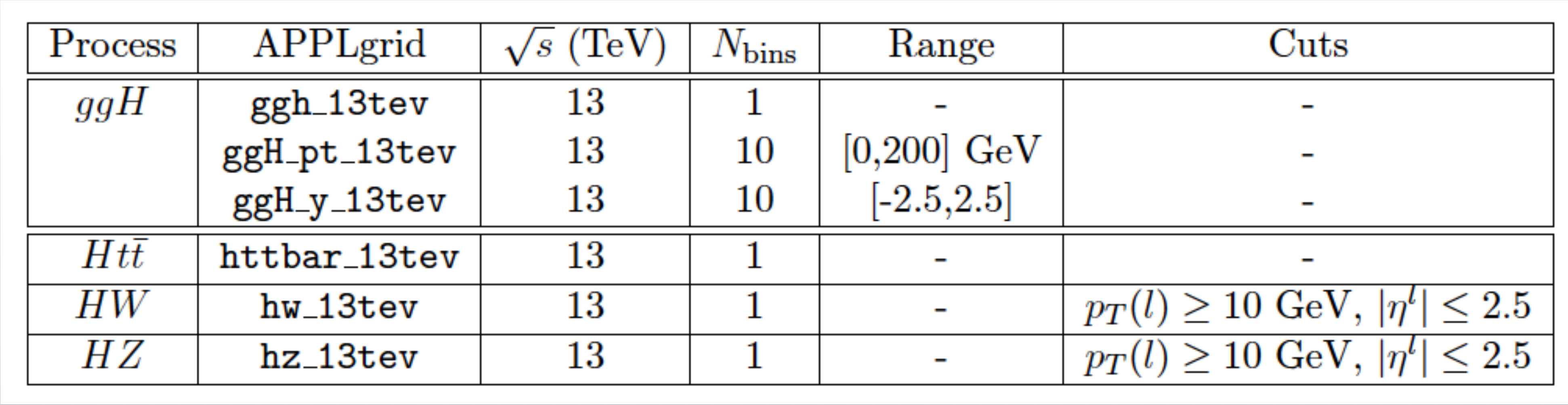
$$\Sigma=\left\{\sigma_1\ldots\sigma_{N_{\rm pred}}\right\}$$

- Apply **MC2hessian** algorithm for only **points in *x* and PDF flavour**s which have been selected above.

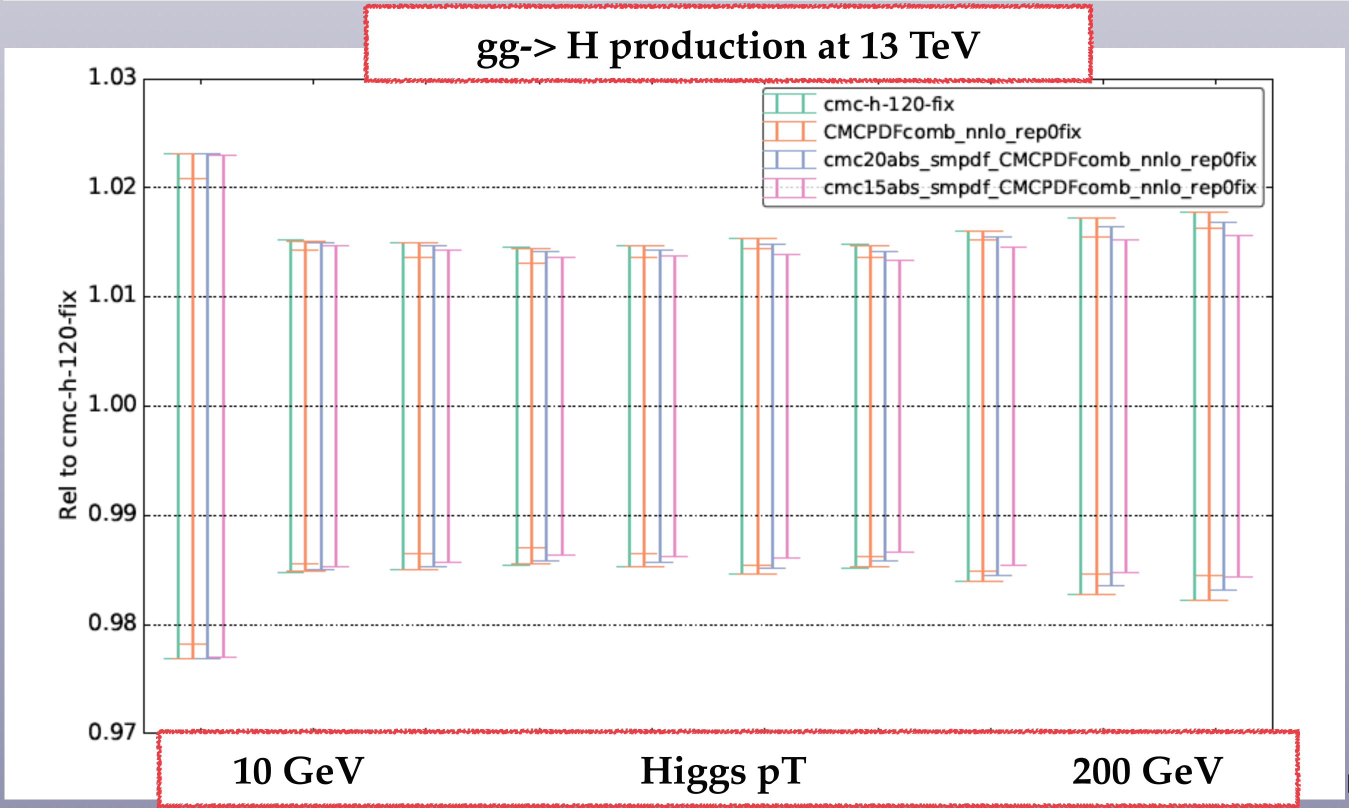
- In the limit of more and more added processes, the **method converges adiabatically to the standard MC2Hessian algorithm**

SM-PDFs for Higgs physics

- First exercise: construct **Specialised Minimal PDFs for Higgs physics:**



Process	APPLgrid	$\sqrt{s}$ (TeV)	N_{bins}	Range	Cuts
ggH	ggh_13tev	13	1	-	-
	ggH_pt_13tev	13	10	[0,200] GeV	-
	ggH_y_13tev	13	10	[-2.5,2.5]	-
$Ht\bar{t}$	httbar_13tev	13	1	-	-
HW	hw_13tev	13	1	-	$p_T(l) \geq 10$ GeV, $\lvert\eta^l\rvert \leq 2.5$
HZ	hz_13tev	13	1	-	$p_T(l) \geq 10$ GeV, $\lvert\eta^l\rvert \leq 2.5$



gg-> H production at 13 TeV

10 GeV **Higgs pT** **200 GeV**

- Use the **CMC-H PDF** as input

- We check that with **15-20 (symmetric) eigenvectors** we can **reproduce all Higgs xsecs and differential distributions**

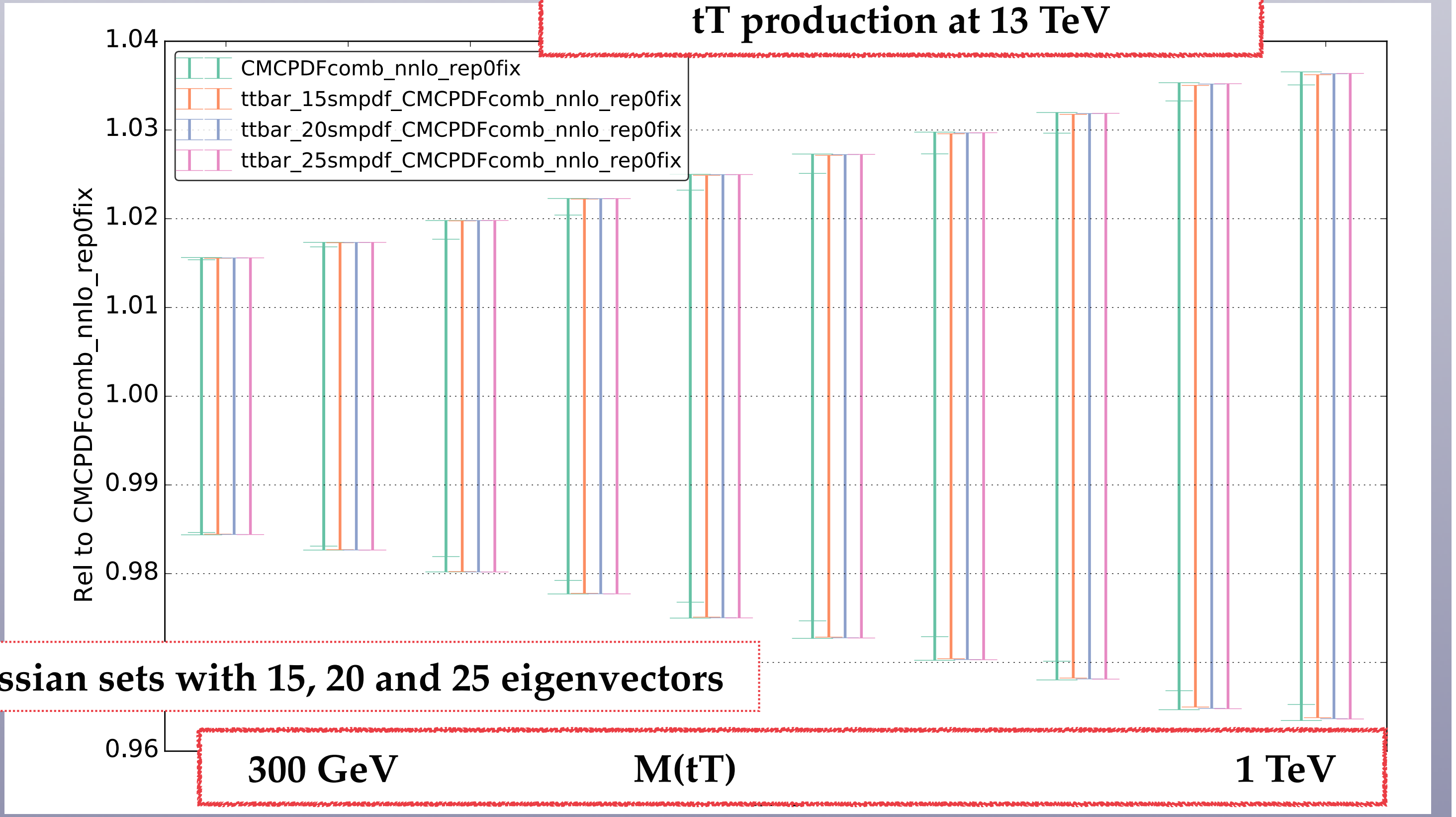
- This means that of the starting 120 eigenvectors, > **100 are not needed for Higgs physics**

SM-PDFs for Top physics

- Then construct **Specialised Minimal PDFs for Top quark physics.** Specific request from **TOPLHC WG**.
- **Included the following set of distributions** computed with aMC@NLO and aMCfast

ttbar_13tev.root: inclusive cross-section, 1 bin
ttbar_tbarpt_13tev.root: tx pT distribution, 10 bins, 40 to 400 GeV
ttbar_tbary_13tev.root: tx rapidity, 10 bins, -2.5 to 2.5
ttbar_tpt_13tev.root: t pT distribution, 10 bins, 40 to 400 GeV
ttbar_ttbarinvmass_13tev.root: pair invariant mass, 10 bins, 300 to 1000
ttbar_ttbarpt_13tev.root: pair pT, 10 bins, 20 to 200 GeV
ttbar_ttbary_13tev.root: pair rapidity, 12 bins, -3 to 3
ttbar_ty_13tev.root: t rapidity, 10 bins, -2.5 to 2.5

- With **only 15 eigenvectors,** we can reproduce **all relevant distributions in top quark** pair production
- Additional processes (like single top) can be easily added if needed



tT production at 13 TeV

Original CMC300 vs reduced Hessian sets with 15, 20 and 25 eigenvectors

300 GeV M(tT) 1 TeV

SM-PDFs for W mass determination

- Then construct **Specialised Minimal PDFs for W mass determination studies quark physics. Should be very useful for onginh ATLAS and CMS analysis**

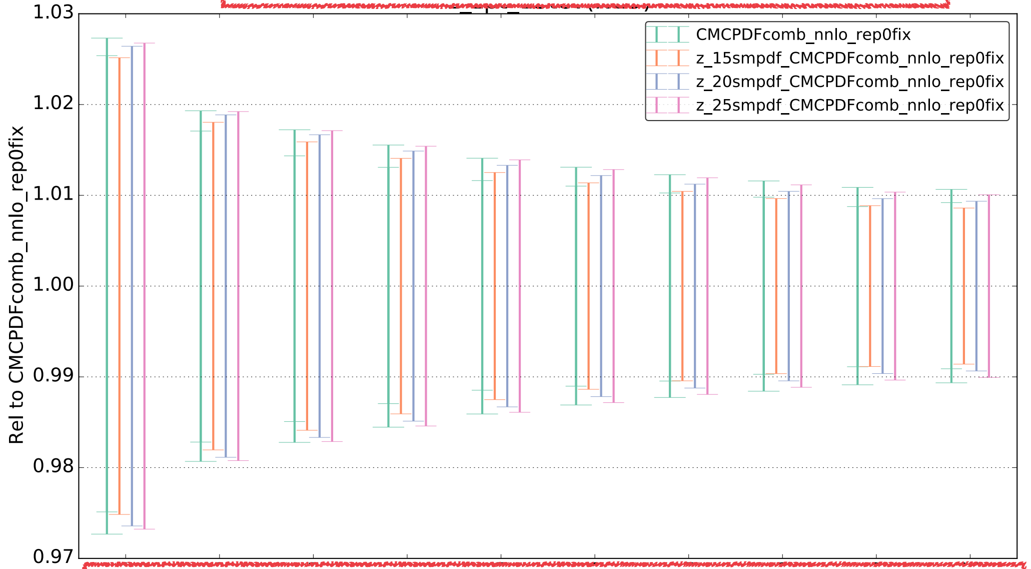
- **Included the following set of distributions:**

w_13tev.root: inclusive cross-section, 1 bin
w_cphi_13tev.root: cos(phi), 10 bins, -1 to 1
w_etmiss_13tev.root: missing Et, 10 bins, 0 to 200 GeV
w_lpt_13tev.root: lepton pT, 10 bins, 0 to 200 GeV
w_ly_13tev.root: lepton rapidity, 10 bins, -2.5 to 2.5
w_mt_13tev.root: transverse mass, 10 bins, 0 to 200 GeV
w_wpt_13tev.root: W pT, 10 bins, 0 to 200 GeV
w_wy_13tev.root: W rapidity, 10 bins, -4 to 4
z_13tev.root: inclusive cross-section, 1 bin
z_lmpt_13tev.root: lepton- pT distribution, 10 bins, 0 to 200 GeV
z_lmy_13tev.root: lepton- rapidity, 10 bins, -2.5 to 2.5
z_lplminvmass_13tev.root: pair invariant mass, 10 bins, 50 to 130 GeV
z_lplmpt_13tev.root: lepton pair pT, 10 bins, 0 to 200 GeV
z_lppt_13tev.root: lepton+ pT distribution, 10 bins, 0 to 200 GeV
z_lpy_13tev.root: lepton+ rapidity, 10 bins, -2.5 to 2.5
z_zpt_13tev.root: z pt, 10 bins, 0 to 200 GeV
z_zy_13tev.root: z rapidity, 5 bins, -4 to 4

SM-PDFs for W mass determination

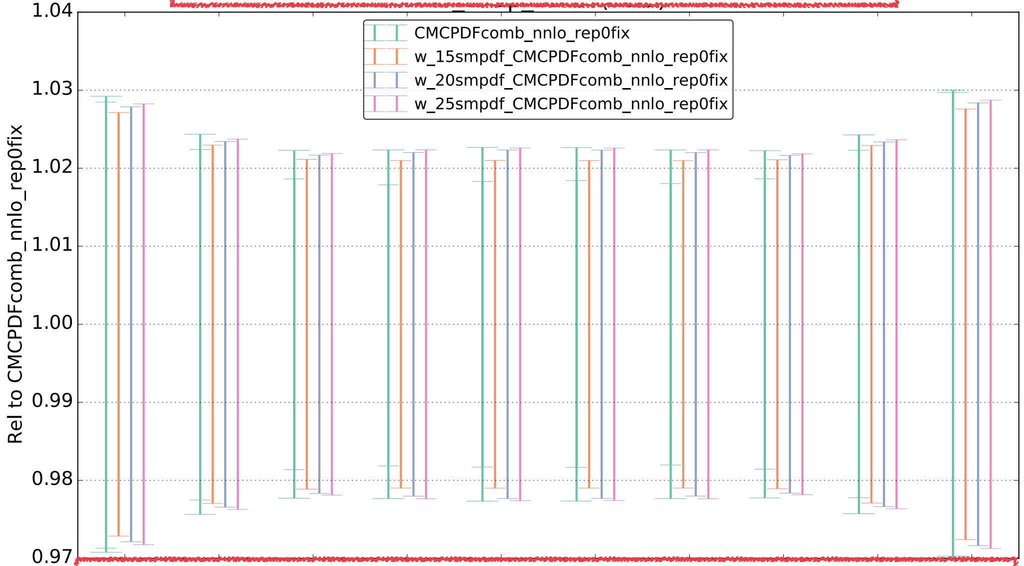
- Then construct **Specialised Minimal PDFs for W mass determination studies. Should be useful for ongoing ATLAS and CMS analysis**
- A somewhat **larger number of eigenvectors selected** here, since W,Z production involve all PDF flavours in a wide range of x),
- Even in this case **with only 25 eigenvectors** we can reproduce the original MC combination result
- Again, adding additional process upon request in this specialised set is straightforward

Inclusive Z production at 13 TeV



10 GeV Z pT 200 GeV

Inclusive W production at 13 TeV



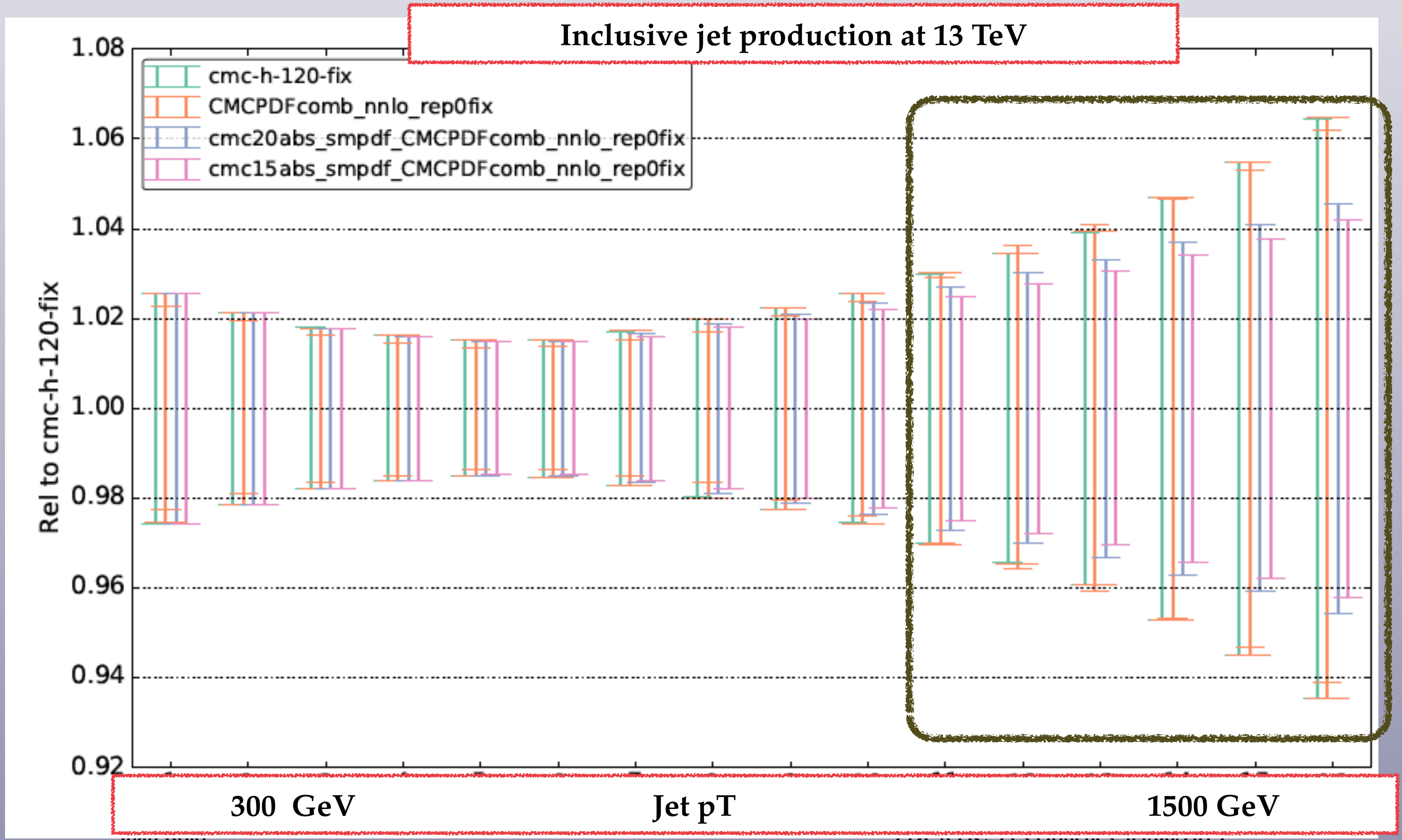
-2.5 W rapidity 2.5

Original CMC300 vs reduced Hessian sets with 15, 20 and 25 eigenvectors

Uses and misuses SM-PDFs

- One major complain about the use of specialised PDFs is that they can be **misused** on calculations other than those for which they were **originally designed**

- In the SM-PDF approach, since we **select eigenvectors based on PDF correlations**, this problem is greatly reduced. For instance, using the reduced set for Higgs physics, predictions for **inclusive jet** production give reasonable results, except at large jet pt (large-x quarks, do not enter Higgs physics)



Inclusive jet production at 13 TeV

300 GeV **Jet pT** **1500 GeV**

PDF uncertainties at the LHC Run II

- In just a few months, **motivated by the many excellent discussions within the PDF4LHC working group**, we have started (and now completed!) an extensive program to **provide optimal tools for the usage of PDFs at Run II:**

☑ **CMC-PDFs:** statistically robust combination of PDF sets: only **40 replicas required to compute PDF uncertainties to arbitrary processes in full generality**

☑ **CMC-H PDFs:** Hessian representation of the MC combination of PDF sets: with only **90 eigenvectors** we can describe PDF uncertainties (where the underlying distribution is Gaussian)

☑ **SM-PDFs:** Using MC2Hessian algorithm but only in restricted regions of x and PDF flavours, identified upon s**election of a range of LHC processes**. Can do all of **Higgs physics with 15-20 eigenvectors**, all of top quark physics with **15 eigenvectors** and W,Z physics with **25 eigenvectors**

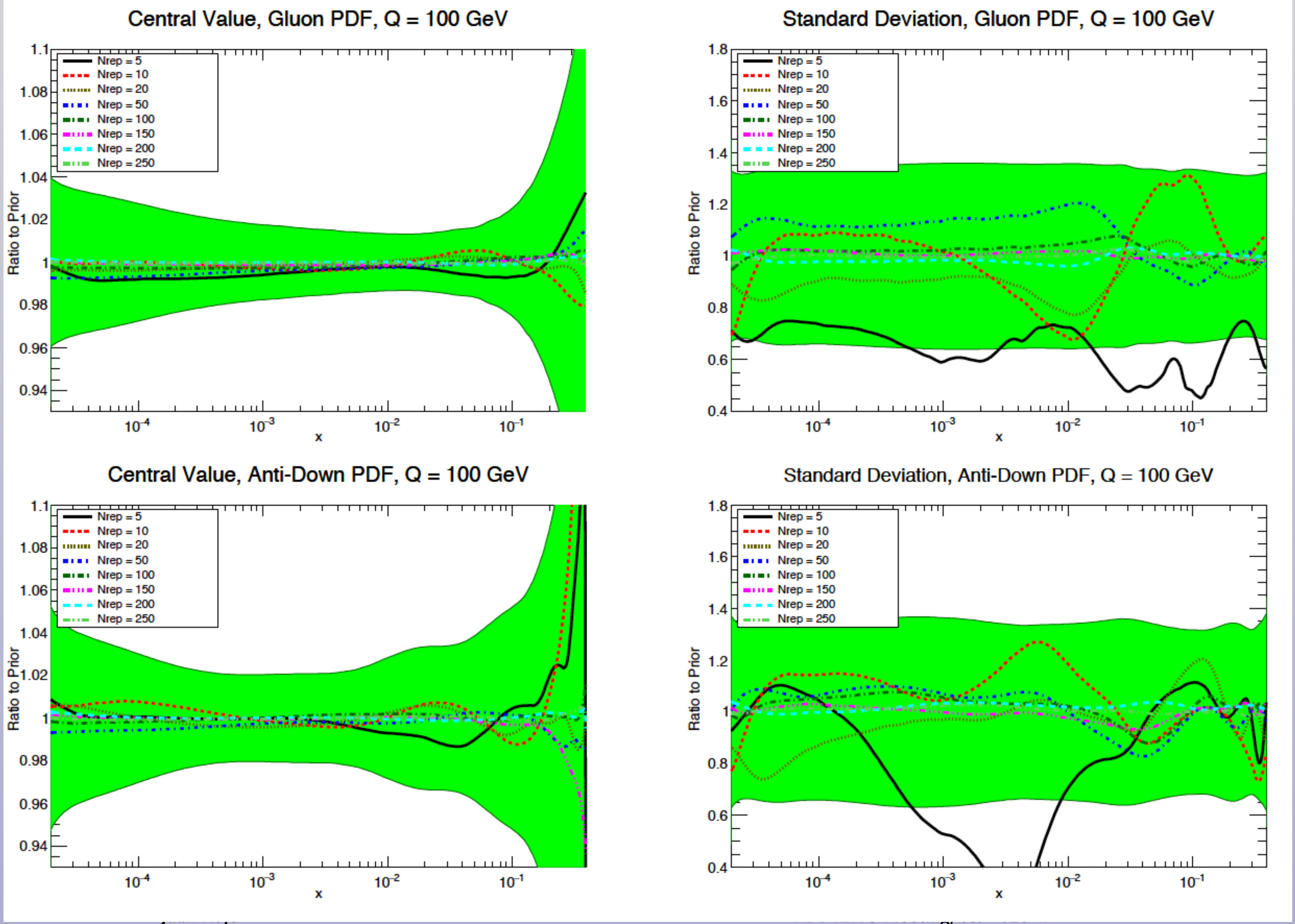
All the tools needed for the updated PDF4HC recommendation are now available

An extensive benchmark comparison between CMC and META approaches in Pavel's slides

Extra Material

Results of the compression

- For example, for **N$_{\mathbf{rep}}$=40** replicas the compressed and the original PDFs are virtually identical



Central Value, Gluon PDF, Q = 100 GeV

Standard Deviation, Gluon PDF, Q = 100 GeV

Central Value, Anti-Down PDF, Q = 100 GeV

Standard Deviation, Anti-Down PDF, Q = 100 GeV

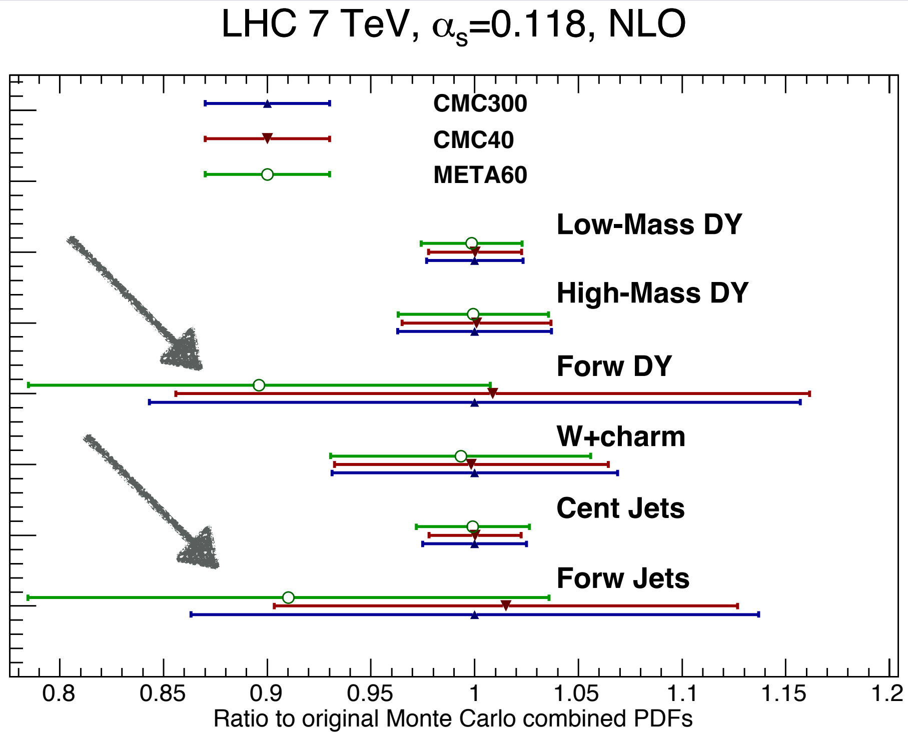
Meta-PDFs: dependence on parametrization

- Unless the **meta-parametrization is suitably tuned and adjusted**, results can be inaccurate, and this will be only discovered when looking at cross-sections

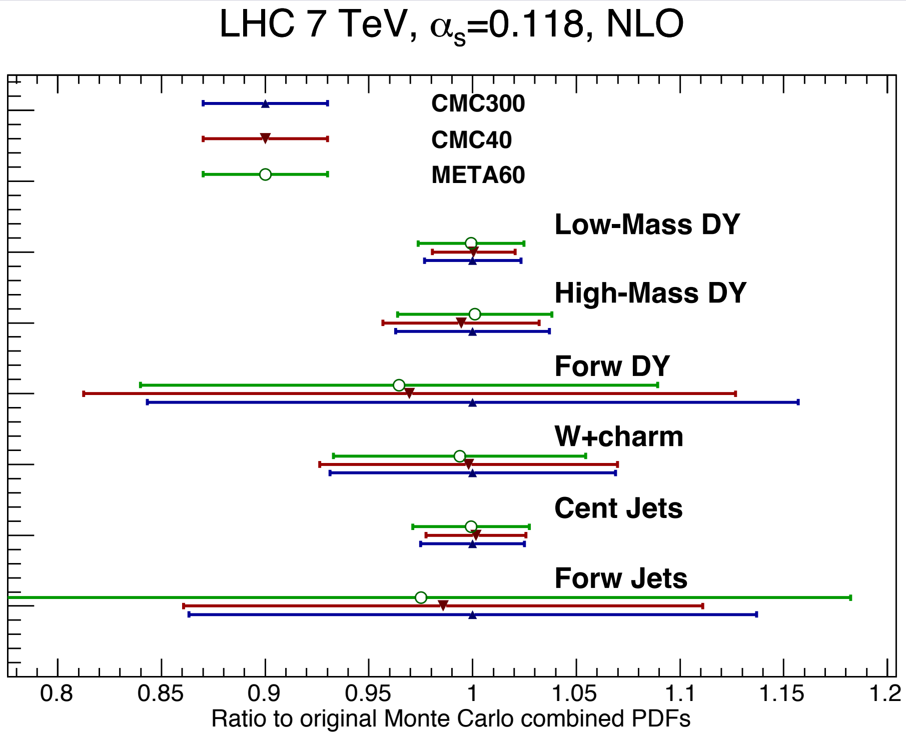
- When benchmarking CMC and META, we discovered that the original META parametrisation lead to incorrect results for forward observables

- Of course, one can always tune to minimise (but not eliminate) the problem, but the bottom line is that achieving **good reproduction requires careful validation** (in CMC this is automatic by construction!)

Param 13/04



Param 28/04


LHC 7 TeV, α_s=0.118, NLO
CMC300
CMC40
META60
Low-Mass DY
High-Mass DY
Forw DY
W+charm
Cent Jets
Forw Jets
Ratio to original Monte Carlo combined PDFs