# Reweighting PDFs

... or how to get rid of PDF fitters (almost)

**Alberto Guffanti**

Niels Bohr International Academy & Discovery center
Niels Bohr Institute - Copenhagen

Terascale School, "Proton Structure in the LHC Era"
DESY, 22/10/2012

# Reweighting PDFs

## Why?

- **Including new data** in global PDF fits is usually a job for a restricted set of people who have access to **complicated**, **massive** and (mostly) **private codes**
- This is especially true when considering **hadronic processes**, for which the computations of **higher order corrections** is **computationally very intensive**
- Wouldn't it be nice to have a **method to quickly assess the impact of new data** on a PDF determination, which is based only on **public tools** and is **as fast as producing a** theory-experiment comparison **plot** for the dataset under study?
- **Bayesian reweighting** provides such a method ... so let's have a closer look!

# Bayesian Reweighting

## Brief history

- First implementation suggested by Giele and Keller, who thought of it as method for producing new PDF fits

  [W. Giele & S. Keller, hep-ph/9803393]

- Reformulated in the context of the NNPDF fits (based on Monte Carlo methodology for uncertainties estimation) and applied for the first time to include data in a global PDF fit (**NNPDF2.2**)

  [R.D. Ball et al., arXiv:1012.0836]
  [R.D. Ball et al., arXiv:1108.1758]

- Recently extended by G. Watt and R. Thorne to PDF based on the Hessian method for estimation of uncertainties (see also LHCb studies, De Lorenzi et al., arXiv:1011.4260)

  [G. Watt & R.S. Thorne, arXiv:1205.4024]

# Bayesian Reweighting

## The Idea

- The $N_{rep}$ **replicas** of a Monte Carlo PDF set provide a s**ampling of the probability density** in the space of parton distribution functions

- **Expectation values for observables** which depend on PDFs are obtained by taking the **average** for a given observable **over the replica set**

$$\langle \mathcal{F}[f_i(x,Q^2)] \rangle = \frac{1}{N_{rep}} \sum_{k=1}^{N_{rep}} \mathcal{F}\left( f_i^{(net)(k)}(x,Q^2) \right)$$

... with corresponding expressions for variances, correlations, etc.

- The **central idea** of Bayesian reweighting is to **assess the impact of** including **new data** in a PDF determination by **updating the probability** density of PDFs **without performing a complete refit**

# Bayesian Reweighting

## The weights formula

- We can apply **Bayes Theorem** to determine the conditional probability of the PDF upon inclusion of the new data

$$\mathcal{P}_{\rm new}(\{f\}) = \mathcal{N}_\chi \mathcal{P}(\chi^2|\{f\})\mathcal{P}_{\rm init}(\{f\}), \quad \mathcal{P}(\chi^2|\{f\}) = [\chi^2(y,\{f\})]^{\frac{n_{dat}-1}{2}} e^{-\frac{\chi^2(y,\{f\})}{2}}$$

- Averages over the sample are no **weighted sums**

$$\langle \mathcal{F}[f_i(x,Q^2)]\rangle = \sum_{k=1}^{N_{rep}} w_k \mathcal{F}\left(f_i^{(net)(k)}(x,Q^2)\right)$$

and the **weights** are given by

$$w_k = \frac{[\chi^2(y,f_k)]^{\frac{n_{dat}-1}{2}} e^{-\frac{\chi^2(y,f_k)}{2}}}{\sum_{i=1}^{N_{\rm rep}} [\chi^2(y,f_i)]^{\frac{n_{dat}-1}{2}} e^{-\frac{\chi^2(y,f_i)}{2}}}$$

# Bayesian Reweighting

## How constraining are the new data?

- The original sample of replica was constructed through importance sampling of the old probability distribution and it is thus **maximally efficient** (i.e. **all replicas are equiprobabile** and it gives the **best representation** of the probability density **for a given number replicas**)
- After reweighting the new replicas set will not give anymore a **maximally efficient** representation of the new probability distribution
- This loss of efficiency can be quantified using the **Shannon Entropy**

$$N_{\text{eff}} \equiv \exp\left\{\frac{1}{N}\sum_{k=1}^{N} w_k \ln(N/w_k)\right\}$$

  to estimate the "**effective number of replicas**" left after reweighting
- The **smaller** the **Shannon entropy** the **more constraining** the new **data** are

# Bayesian Reweighting

## How compatible are the new data with the old ones?

- If the value of the **Shannon entropy** obtained after reweighting a prior set with a given dataset becomes **too small** the **reweighting** procedure becomes **unreliable**

- There are **two reasons** why that can happen
  - the **new data contain a lot of information** on PDFs not present in the prior fit (⇒ **refit**)
  - the **new data** are **incompatible** with data included in the prior PDF set

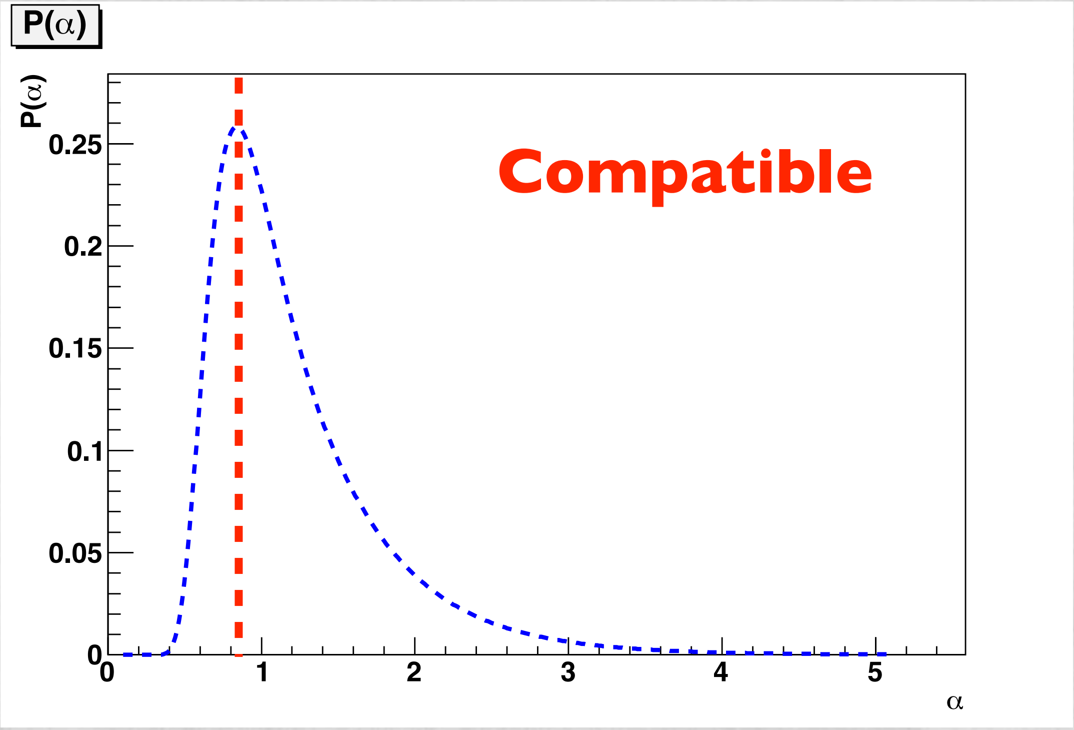
- We can distinguish the two cases by looking at the probability density of the **nuisance parameter** ($\alpha$), defined as a **rescaling factor** for the **uncertainties** on the new data

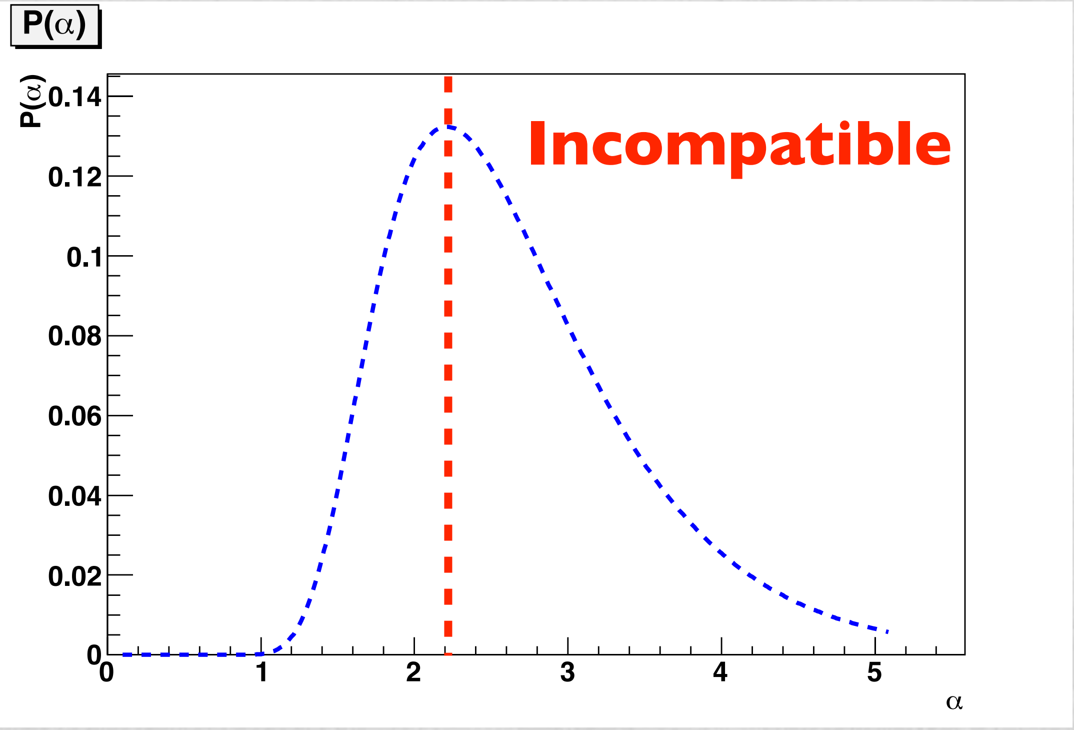
$$\mathcal{P}(\alpha) \propto \frac{1}{\alpha} \sum_{k=1}^{N} w_k(\alpha)$$

# Bayesian Reweighting

## How compatible are the new data with the old ones?

- If the **probability density peaks close to one** (or below one) the **new data** are **compatible** with the data included in the prior fit
- If the **probability density peaks far above one** the **uncertainties** on the new data are probably **underestimated** and these data are thus **incompatible** with the data included in the prior fit



# Unweighting PDFs

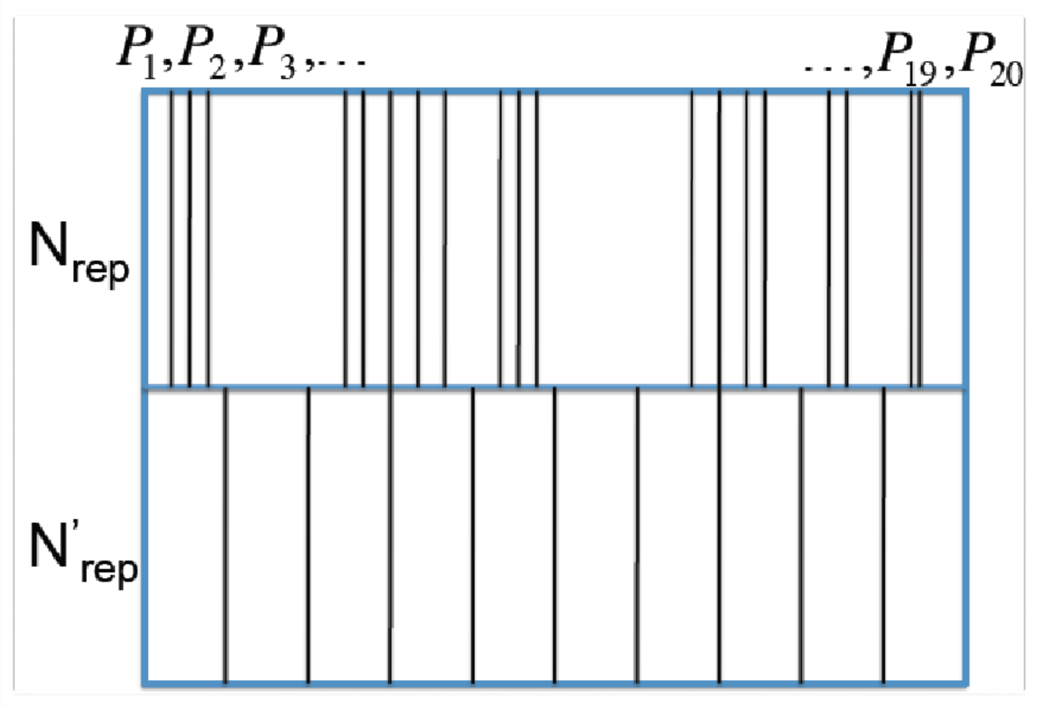
## Obtaining a standard PDF set from a reweighted one

- We want to define a procedure to generate an **unweighted** PDF set (i.e. a set in which all replicas are equiprobable) starting from one we constructed via reweighting
- To do that we define the probability cumulants of the original fit as

$$P_k \equiv P_{k-1} + p_k = \sum_{j=0}^{k} p_j$$

- The weights in the new, unweighted, set are the defined as

$$w'_k = \sum_{j=1}^{N'_{\rm rep}} \theta\Big(\frac{j}{N'_{\rm rep}} - P_{k-1}\Big)\theta\Big(P_k - \frac{j}{N'_{\rm rep}}\Big)$$




This is equivalent to the **graphical procedure** of picking a number of replicas from the original reweighted set which is equal to the number of lower segments whose right edge is contained in the upper segment corresponding to the specific replica

# Validating Bayesian reweighting

## Validating the reweighting procedure

- We now validate the **reweighting procedure** checking that our methodology yields results which satisfy a number of **consistency tests**
- **Including a given dataset** in a prior fit by **reweighting** or **refitting** should yield **statistically equivalent results**
- If we include **two or more datasets** we can choose to include them in a **single step** (as a single dataset) **or in successive ones**: the two procedures should **yield statistically equivalent results**
- When **including sets in successive steps** results should not depend on the order in which the reweighting is performed
- We will assess the **statistical equivalence of** different **fits** by looking at PDFs and at **distances between central values and uncertainties**

# Interlude: PDF distances

## Quantitavely measuring equivalence of PDF sets

- We define the **distances between central values** of two PDF sets as

$$d(q_j) = \sqrt{\left\langle \frac{\left(\langle q_j \rangle_{(1)} - \langle q_j \rangle_{(2)}\right)^2}{\sigma_1^2[q_j] + \sigma_2^2[q_j]} \right\rangle_{N_{\rm part}}}$$

  and similarly for the distances between Standard deviations

- We compute **averages over samples of 100 replicas** picked from the Monte Carlo ensembles **allowing for repetitions** (and we average over 1000 such samples to tame statistical fluctuations)

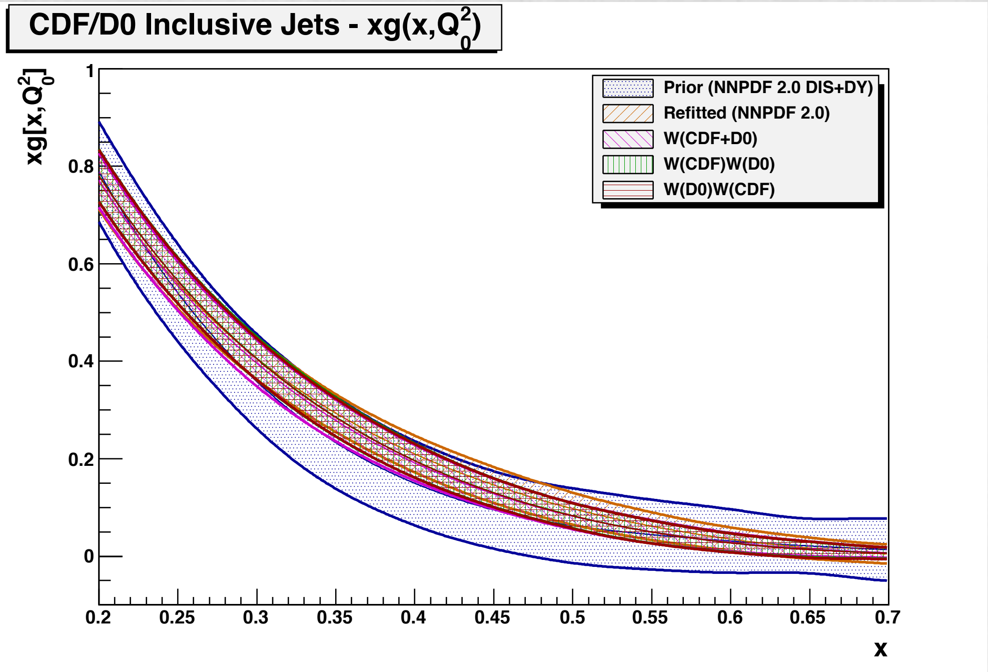
- Two **statistically equivalent** fits have **distances** $\sim$**1**

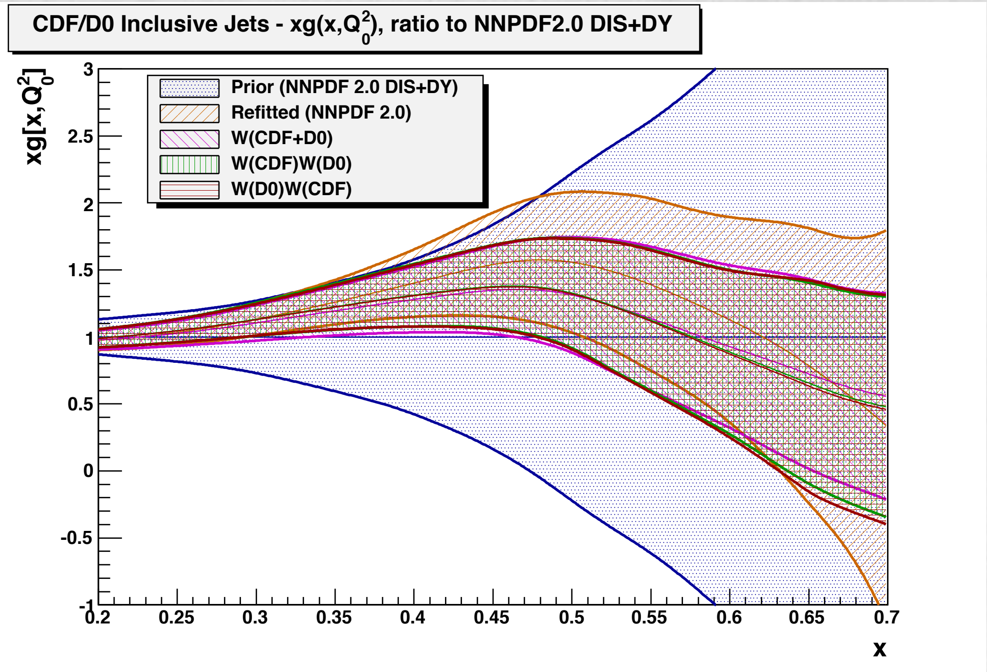
- If two PDFs have a **distance of 10** they differ by **1-sigma**

# Validating Bayesian reweighting

## Inclusive jet production at the Tevatron

- Start from **NNPDF2.0 DIS+DY** fit as a prior fit
- Add CDF and D0 inclusive jet data through **refitting** (**NNPDF2.0**)
- Add **CDF & D0** jet data via reweighting as a **single dataset**
- Add **CDF** data then **unweight** and then add **D0** data
- Add **D0** data then **unweight** and then add **CDF** data
- All procedures yield **statistically equivalent results**

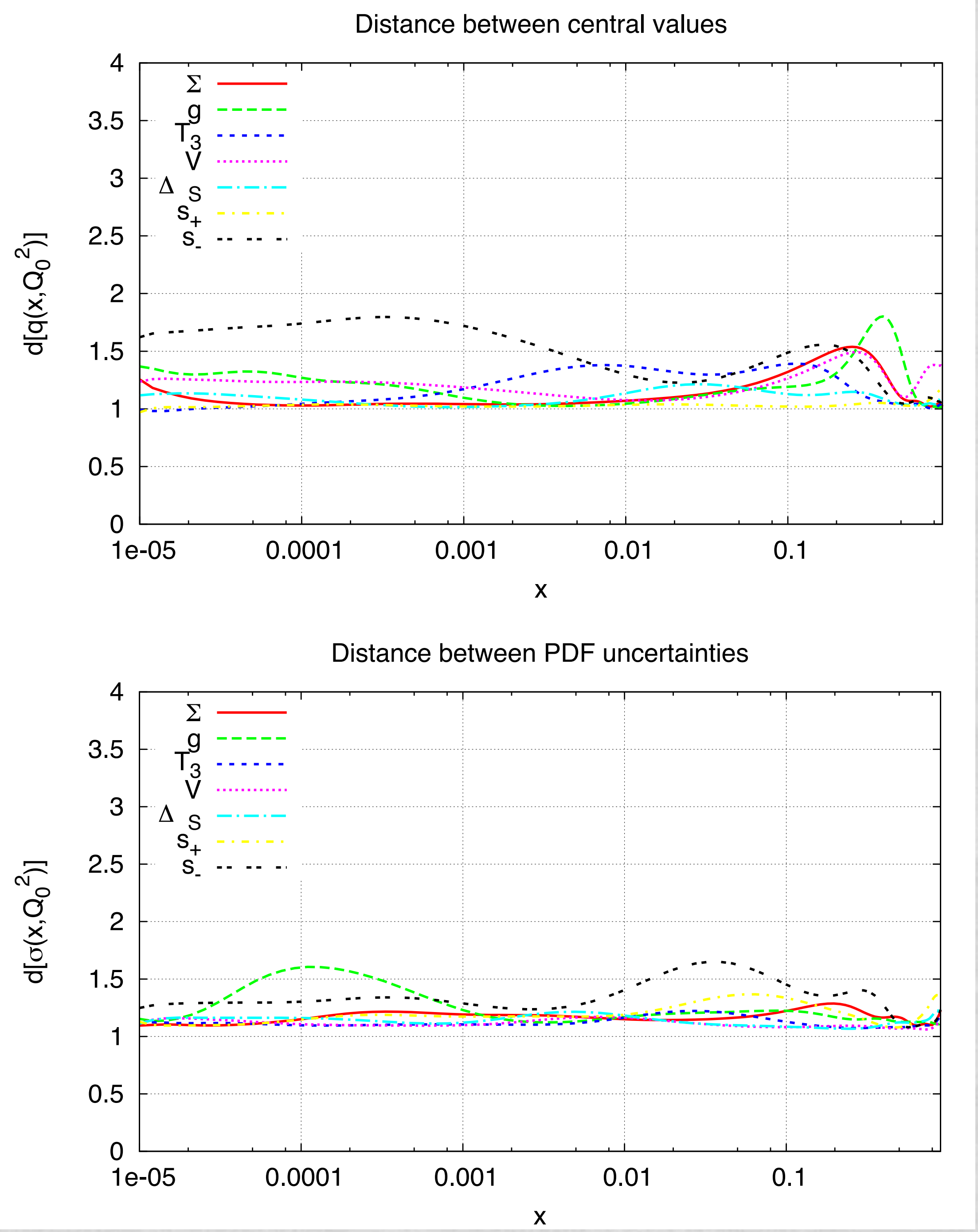


# Validating Bayesian reweighting

## Inclusive jet production at the Tevatron

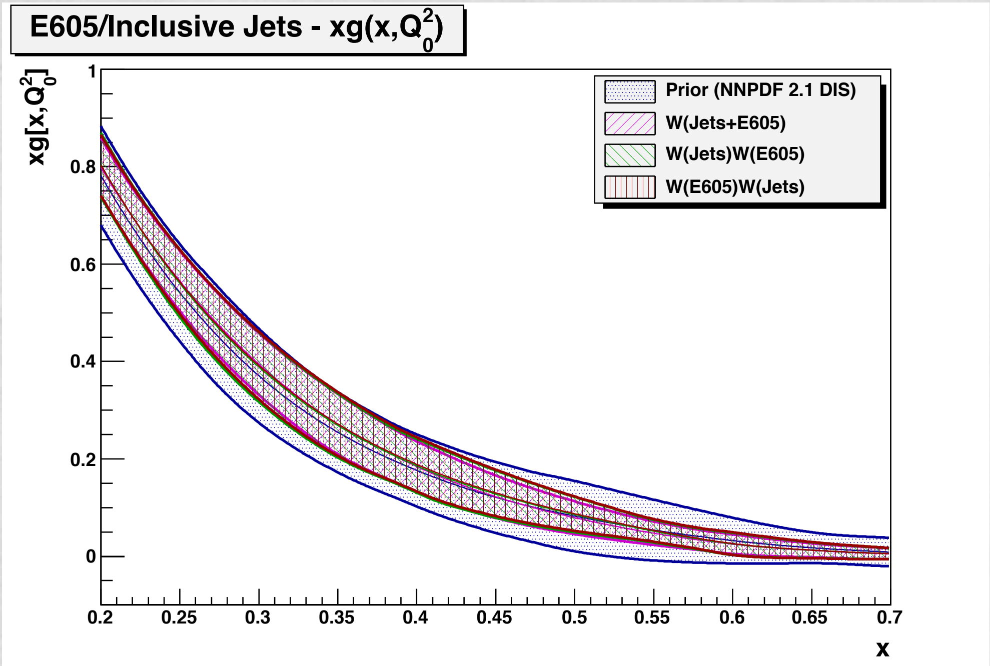
- Start from **NNPDF2.0 DIS+DY** fit as a prior fit
- Add CDF and D0 inclusive jet data through **refitting** (**NNPDF2.0**)
- Add **CDF & D0** jet data via reweighting as a **single dataset**
- Add **CDF** data then **unweight** and then add **D0** data
- Add **D0** data then **unweight** and then add **CDF** data
- All procedures yield **statistically equivalent results**

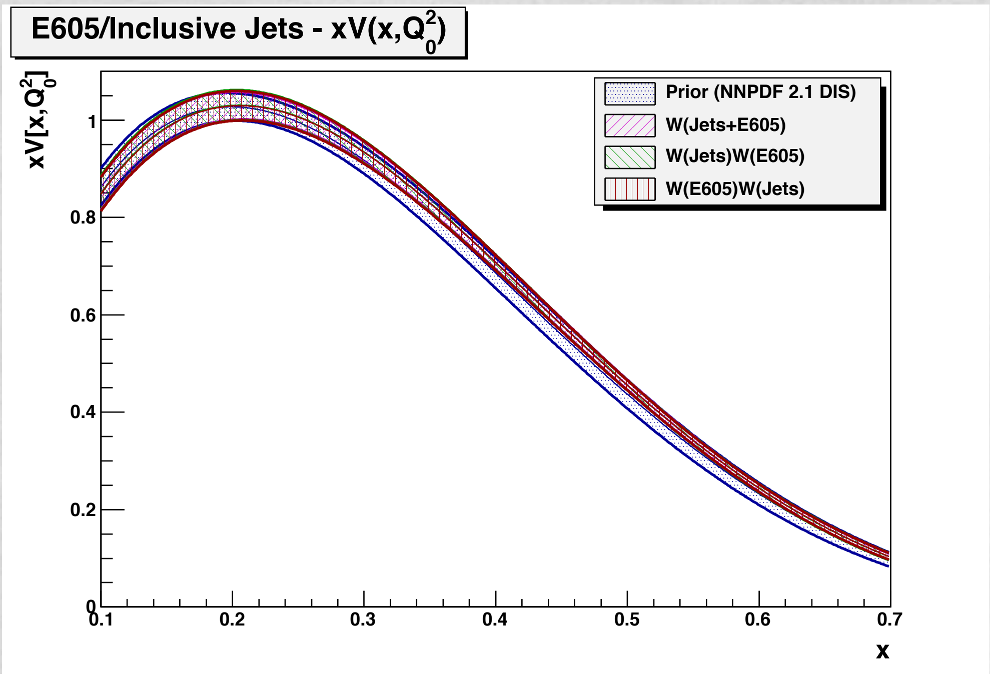


# Validating Bayesian reweighting

## Inclusive jet production and fixed target Drell-Yan

- Start from **NNPDF2.1 DIS** fit as a prior fit
- Add **E605 Drell-Yan & Tevatron jet** data via reweighting as a **single dataset**
- Add **Tevatron jet** data then **unweight** and then add **E605 Drell-Yan** data
- Add **E605 Drell-Yan** data, **unweight** and then add **Tevatron jet** data
- All procedures yield **statistically equivalent results**

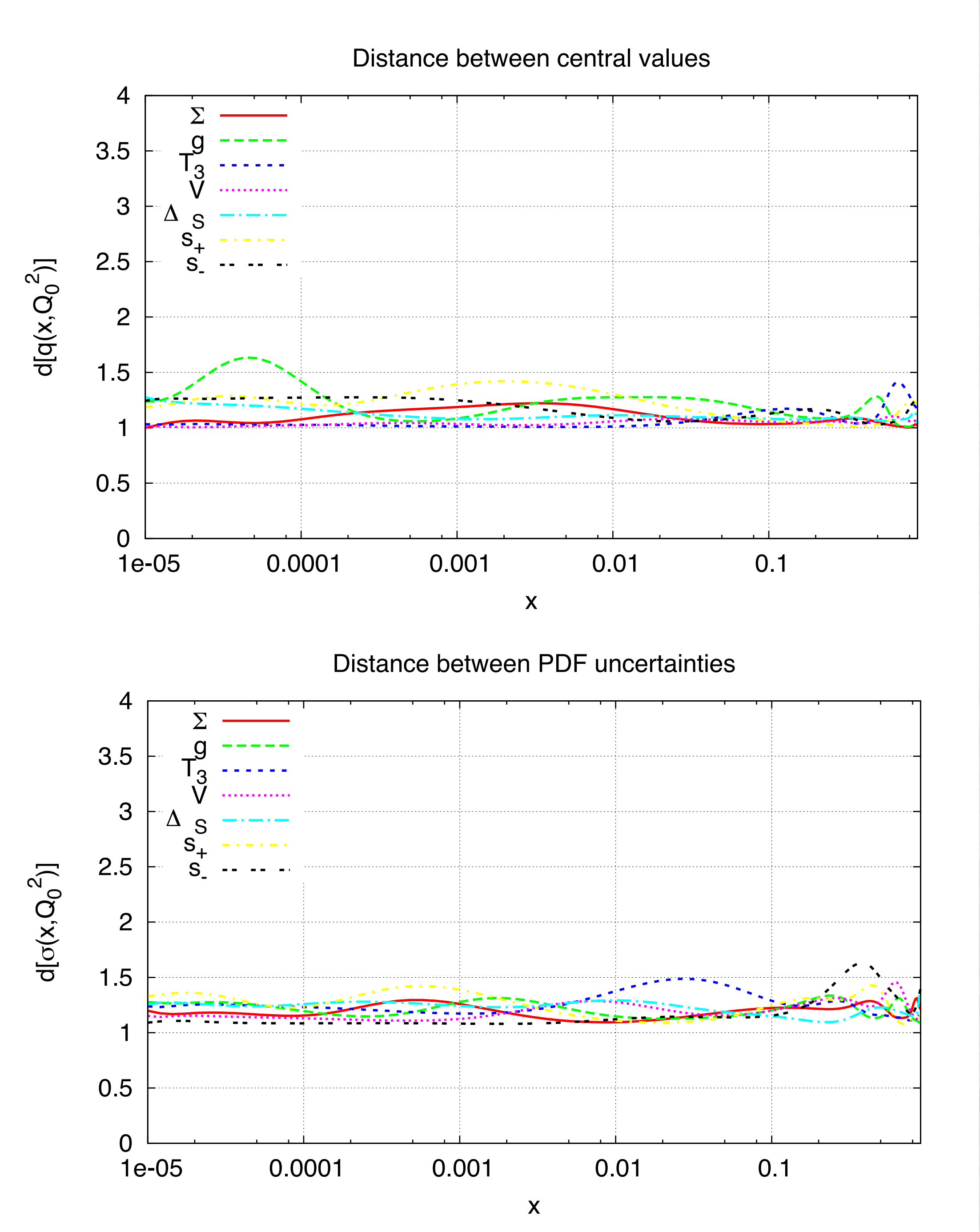


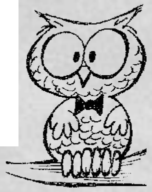



# Validating Bayesian reweighting

## Inclusive jet production and fixed target Drell-Yan

- Start from **NNPDF2.1 DIS** fit as a prior fit
- Add **E605 Drell-Yan & Tevatron jet** data via reweighting as a **single dataset**
- Add **Tevatron jet** data then **unweight** and then add **E605 Drell-Yan** data
- Add **E605 Drell-Yan** data, **unweight** and then add **Tevatron jet** data
- All procedures yield **statistically equivalent results**



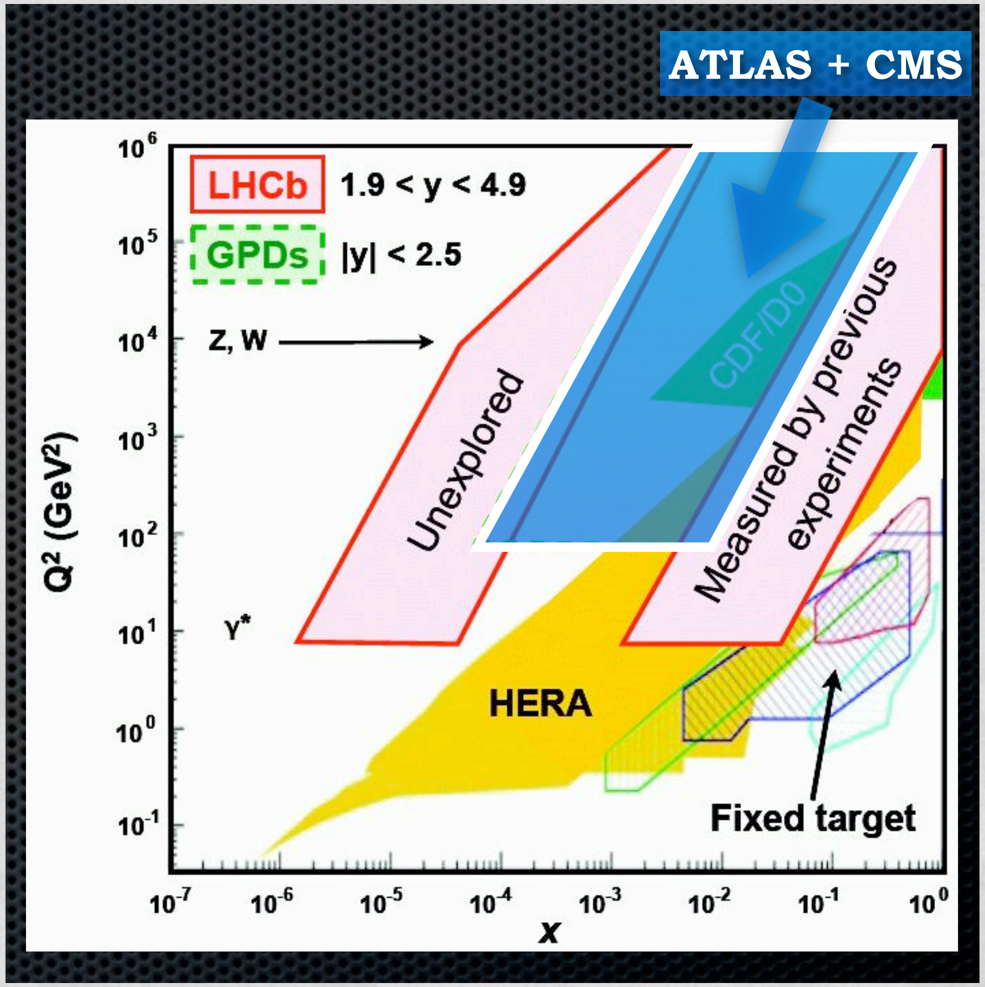
# Bayesian reweighting at work

## The NNPDF2.2 PDF set

- In 2011 the first data from the LHC providing new constraints for PDF fits appeared
- In particular the W lepton asysmmetry measurements from ATLAS and CMS, provided new constraints on light flavour separation

$$A_W^l \sim \frac{u(x_1, M_W^2)\bar{d}(x_2, M_W^2) - d(x_1, M_W^2)\bar{u}(x_2, M_W^2)}{u(x_1, M_W^2)\bar{d}(x_2, M_W^2) + d(x_1, M_W^2)\bar{u}(x_2, M_W^2)}$$

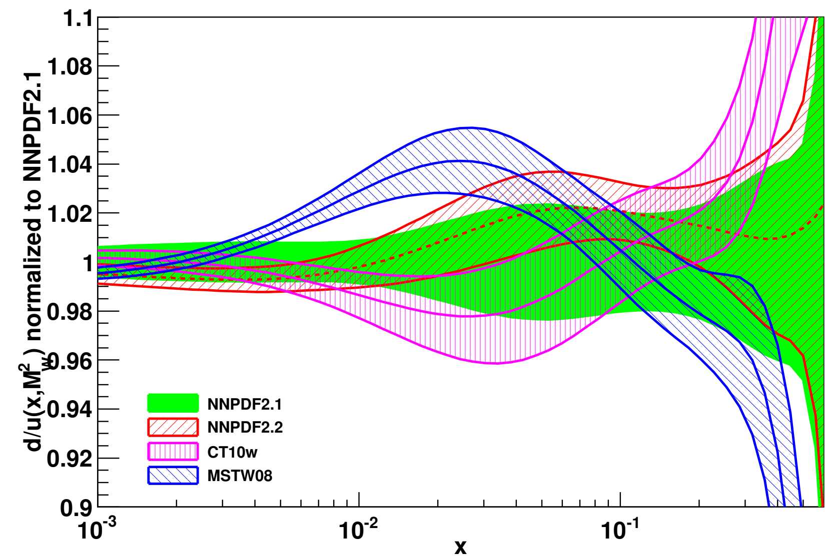
- NNPDF was the first collaboration to include these data in a global fit using the Bayesian reweighting technique



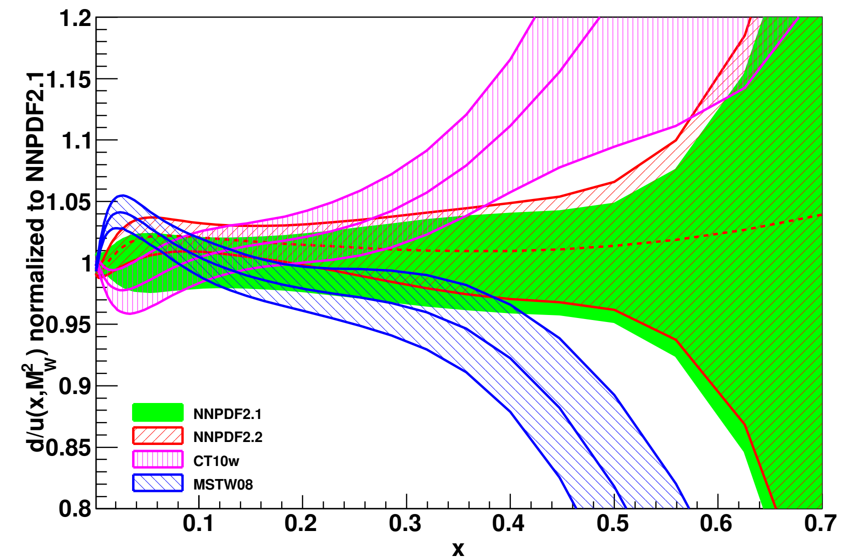
# Bayesian reweighting at work

## The NNPDF2.2 PDF set

LOG scale



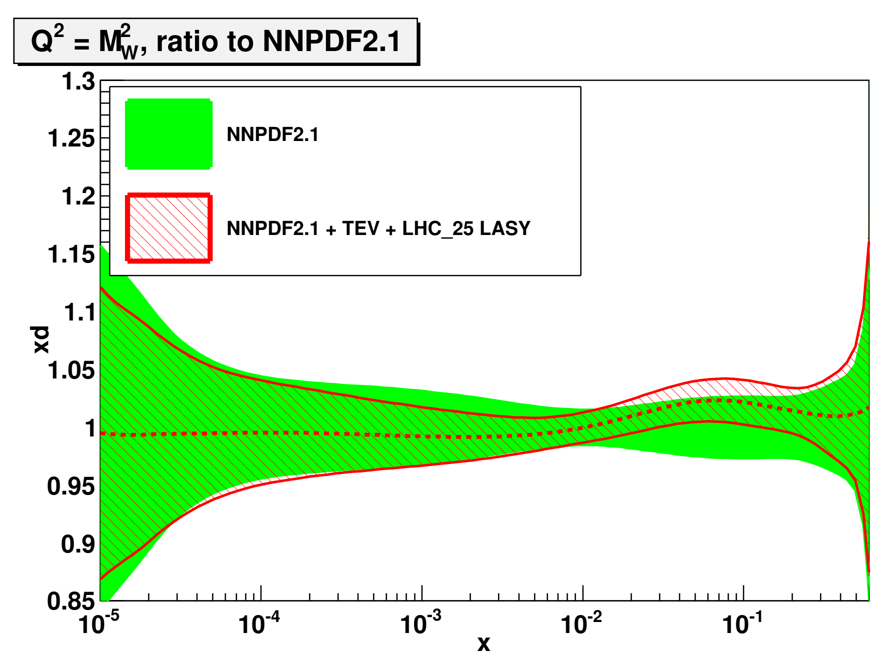

LIN scale




**ATLAS+CMS(25)+D0$_\mu$+D0$_e$(20)**

$N_{\rm eff}$ = 196, $\chi^2_{\rm d.o.f.}$ : 2.18 $\to$ 0.86

- Uncertainty reduction medium-small $x$ and shift in central value driven by the LHC data
- Uncertainty reduction at medium-large $x$ driven by Tevatron data



# Reweighting for Hessian sets

## The Thorne-Watt methodology

[G. Watt & R.S. Thorne, arXiv:1205.4024]

- For **Hessian sets** (assuming symmetric uncertainties) the **uncertainties on a given observables** can be computed as

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

- A set of Monte Carlo replicas can then be generated according to

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

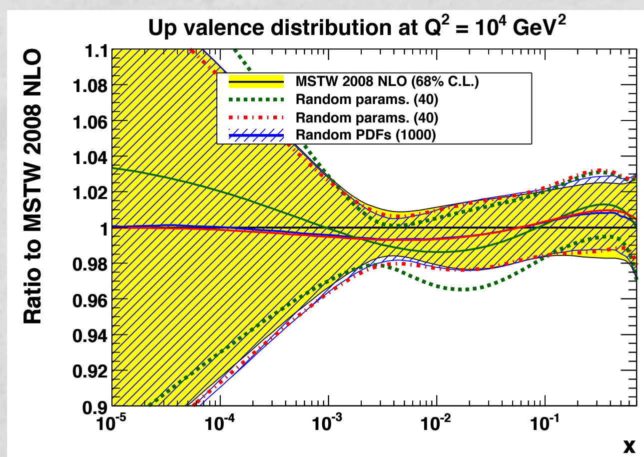
where $S_0$ is the “central set” and $R_{jk}$ are gaussianly distributed random numbers

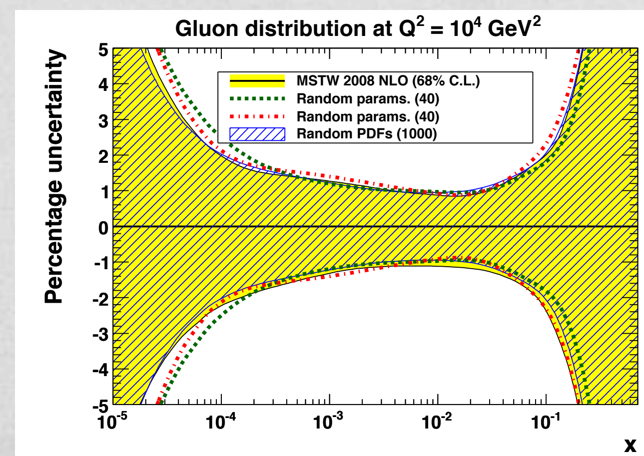
# Reweighting for Hessian sets

## The Thorne-Watt methodology

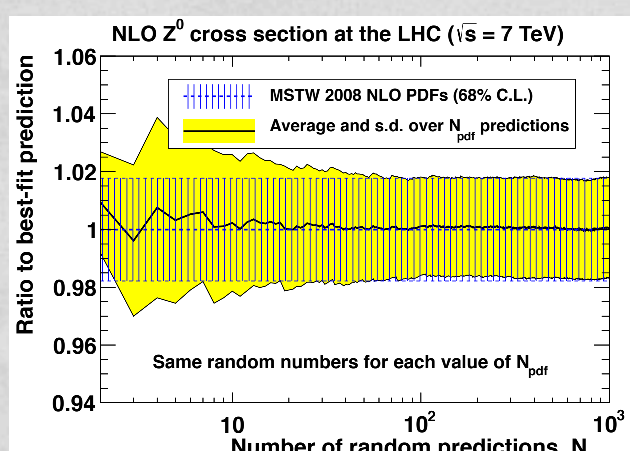
[G. Watt & R.S. Thorne, arXiv:1205.4024]

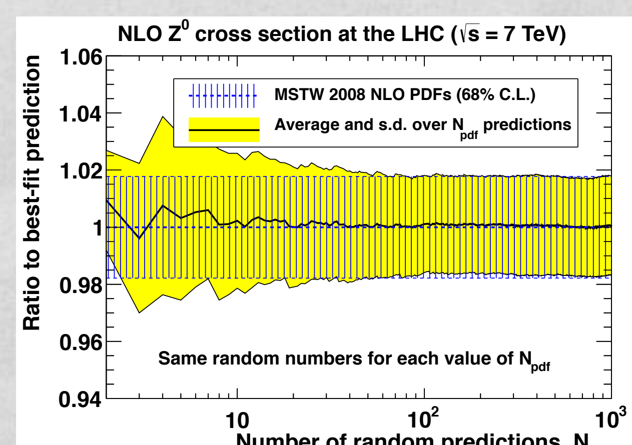
- How well does the Monte Carlo PDF ensemble we generated reproduce the original probability distribution of PDFs given by the Hessian eigenvectors?




... and what about observables?



# Reweighting for Hessian sets

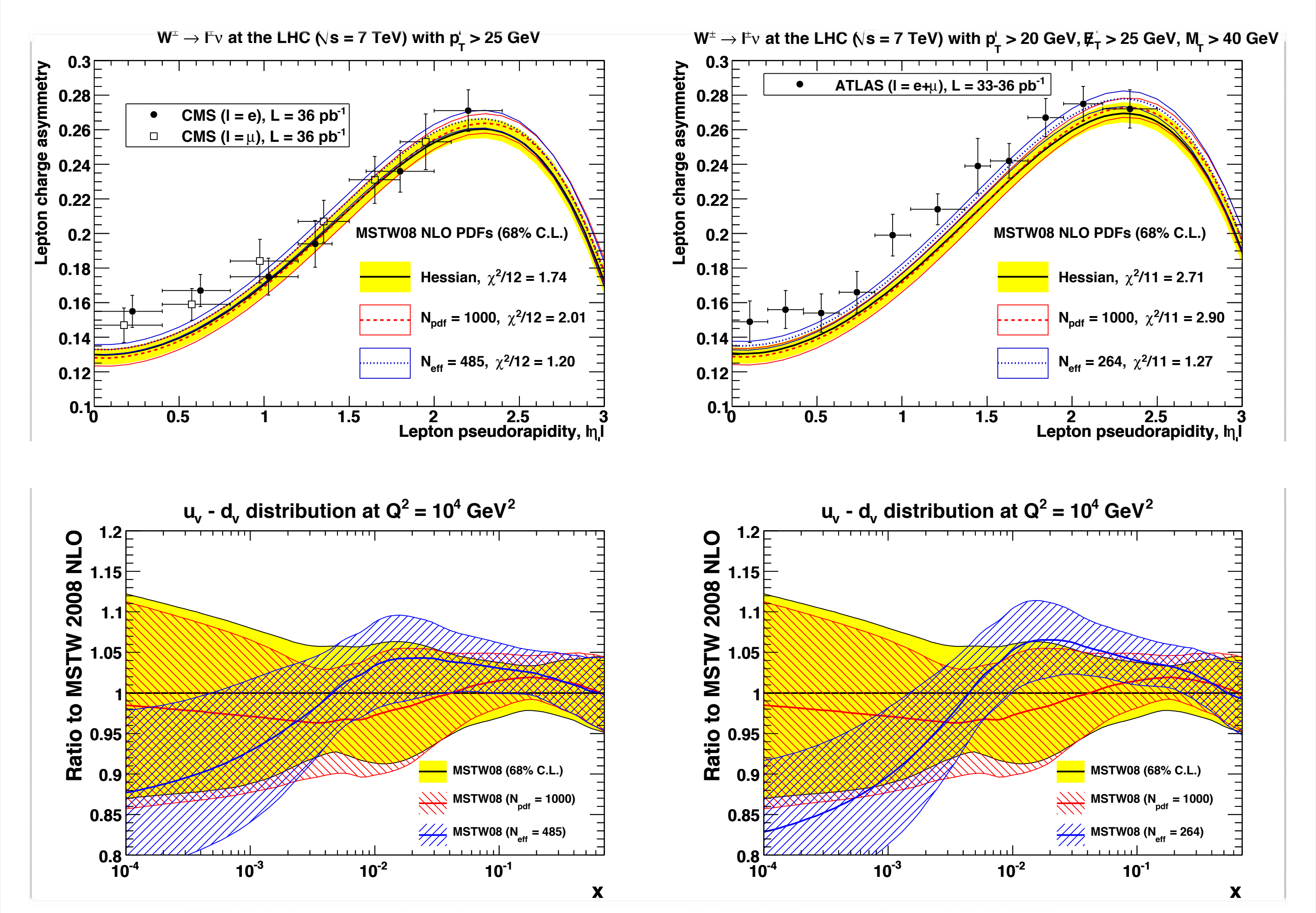
## The Thorne-Watt methodology

[G. Watt & R.S. Thorne, arXiv:1205.4024]

- Once replicas are generated according to the recipe given, the very same Bayesian reweighting methodology described for NNPDF can be used to compute weights and generate a reweighted PDF set including new data

- Used to assess the impact of LHC W lepton asymm. on MSTW08 PDFs

- Valence quark distribution mostly affected in the medium-/small-x region



# A public NNPDF reweighting tool

## The nnpdfrw code

- (Soon to be) Publicly available from the NNPDF Collaboration webpage

http://nnpdf.hepforge.org/

- Included as a module in the **HERAfitter** package (maintainers K. Lohwasser, AG)
- It is a C++ code, compiled against ROOT libraries for plotting features (soon to be changed)
- Takes as input a prior PDF set (Monte Carlo replicas) and a given dataset (input allowed as theory predictions and Covariance Matrix or $\chi^2$ values)
- Return an LHAPDF compatible file (unweighted PDF set) and debugging plots ($P(\alpha)$, $\chi^2$ distribution per replica, etc.)

# Conclusions & Outlook

- **Inclusion of new data** in global PDF fits is usually performed by PDF fitting collaborations using **complicated**, **massive** and (mostly) **private codes**
- **Bayesian reweighting** provides a **method to quickly assess the impact of new data** on a PDF determination and is based on the use of **public tools**
- **Bayesian reweighting** was initially developed for Monte Carlo sets but the same techniques have recently been extended to be used with Hessian sets
- **Bayesian reweighting** proved effective in studying the impact of Tevatron and LHC data on PDF fits and even producing a new global PDF fits which incorporates these data (**NNPDF2.2**)
- A **public too**l (nnpdfrw) to perform Bayesian reweighting analyses is (soon to be) **available from the NNPDF website**
- nnpdfrw (allowing use of MC and Hessian sets) is also **available as a module** of the **HERAfitter** package