Author(s): Casuga, Carlisle; Karhunen, Mikko; Mäntysaari, Heikki

Title: Inferring the initial condition for the Balitsky-Kovchegov equation

Please cite the original version:

Casuga, C., Karhunen, M., & Mäntysaari, H. (2024). Inferring the initial condition for the Balitsky-Kovchegov equation. Physical Review D, 109(5), Article 054018. https://doi.org/10.1103/PhysRevD.109.054018

Inferring the initial condition for the Balitsky-Kovchegov equation

Carlisle Casuga[*], Mikko Karhunen,[†] and Heikki Mäntysaari[‡]

Department of Physics, University of Jyväskylä, P.O. Box 35, 40014 University of Jyväskylä, Finland
and Helsinki Institute of Physics, P.O. Box 64, 00014 University of Helsinki, Finland

We apply Bayesian inference to determine the posterior likelihood distribution for the parameters describing the initial condition of the small-x Balitsky-Kovchegov evolution equation at leading logarithmic accuracy. The HERA structure function data is found to constrain most of the model parameters well. In particular, we find that the HERA data prefers an anomalous dimension $\gamma \approx 1$ unlike in previous fits where $\gamma > 1$ which led to, e.g., the unintegrated gluon distribution and the quark-target cross sections not being positive definite. The determined posterior distribution can be used to propagate the uncertainties in the nonperturbative initial condition when calculating any other observable in the color glass condensate framework. We demonstrate this explicitly for the inclusive quark production cross section in proton-proton collisions and by calculating predictions for the nuclear modification factor for the F_2 structure function in the EIC and LHeC/FCC-he kinematics.

I. INTRODUCTION

Understanding the high-energy structure of protons and nuclei is one of the central goals of the next-generation Electron-Ion Collider (EIC) [1]. Especially the fact that one can perform nuclear-DIS experiments for the first time in collider kinematics is intriguing: at high center-of-mass energies it is possible to access the high-density part of the nuclear wave function where nonlinear QCD dynamics with emergent saturation phenomena is expected to play a major role.

Precise theory calculations taking into account saturation effects are necessary to probe in detail the nonlinear QCD dynamics in hadronic collisions at RHIC and at the LHC, as well as in accurate nuclear-DIS experiments [2] at the EIC or at the proposed future LHeC/FCC-he collider [3]. The color glass condensate (CGC) effective theory of QCD [4] provides a natural framework to describe gluon saturation at high energies. In recent years there has been a rapid progress toward next-to-leading order accuracy in CGC calculations, see e.g. Refs. [5–15]. In addition to higher order corrections, precise theoretical predictions require a well constrained nonperturbative input describing the

proton or nuclear structure at moderately small-x, which is an initial condition to the perturbative evolution equations such as the Balitsky-Kovchegov (BK) [16,17] equation resumming contributions enhanced by large logarithms of center-of-mass energy.

The initial condition for the BK evolution has been fitted to the HERA structure function data [18,19] at leading [20–23] and recently also at next-to-leading [24,25] order accuracy (see also Ref. [26] for a complementary approach to determine the initial condition from the large-x structure of the proton). This initial condition is a necessary input when calculating any other scattering process at small-x, such as particle production in proton-nucleus collision at the LHC, see, e.g., Refs. [8,21,27,28].

All current fits to the DIS data extracting the initial condition for the BK evolution lack an uncertainty analysis, i.e., do not provide any method to propagate the uncertainties in the initial condition parametrization to the calculated observables. When developing the CGC calculations to the precision level, a statistically rigorous treatment of these uncertainties can be seen to be of equal importance as the higher order corrections when comparing predictions to current and future experimental data.

In this work we extract, for the first time with an uncertainty estimate, the nonperturbative initial condition for the BK evolution from the HERA structure function data [19]. This is achieved by employing a Bayesian inference setup extensively used in the field, recently when extracting, e.g., properties of the quark gluon plasma [29–34] and the event-by-event fluctuating proton geometry [35]. The determined posterior distribution for the initial condition parameters allows for a rigorous propagation of initial

[*]carlisle.doc.casuga@jyu.fi
[†]mikko.a.karhunen@student.jyu.fi
[‡]heikki.mantysaari@jyu.fi

condition parametrization uncertainties to calculations of all other observables in the CGC framework. In this work we present results based on a leading order analysis but note that the developed computationally efficient setup can be extended to the numerically heavy next-to-leading order case, where it has been recently shown that combined analyses including both total cross section and charm production data are feasible [24].

This paper is structured as follows. In Sec. II we review the calculation of the proton structure functions at high energy in the CGC approach. The Bayesian inference setup is described in Sec. III. The determined initial condition for the BK evolution with an uncertainty estimate is presented in Sec. IV. To illustrate the propagation of uncertainties to calculations of various observables, we calculate in Sec. V inclusive quark production in proton-proton collisions and the nuclear modification factor for the structure function F_2 which will be measured at the EIC but for which no predictions that include the BK evolution consistently with the HERA data exist in the literature (see however Ref. [36]). Conclusions are presented in Sec. VI, and the importance of including correlations among the experimental uncertainties is analyzed in the Appendix.

II. DEEP INELASTIC SCATTERING AT HIGH ENERGY

The total cross section in lepton-nucleus scattering is typically expressed in terms of the reduced cross section defined as

$$\sigma_r(x_{\mathrm{Bj}}, y, Q^2) = F_2(x_{\mathrm{Bj}}, Q^2)$$
$$- \frac{y^2}{1 + (1 - y)^2} F_L(x_{\mathrm{Bj}}, Q^2). \quad (1)$$

Here $y = Q^2/(sx)$ is the inelasticity, x_{Bj} is the Bjorken-x, Q^2 the photon virtuality and $\sqrt{s}$ the electron-proton center-of-mass energy. The structure functions F_2 and F_L are related to the total virtual photon-nucleus cross section as

$$F_2 = \frac{Q^2}{4\pi^2 \alpha_{\mathrm{em}}} \left(\sigma_T^{\gamma^* A} + \sigma_L^{\gamma^* A} \right) \quad (2)$$

$$F_L = \frac{Q^2}{4\pi^2 \alpha_{\mathrm{em}}} \sigma_L^{\gamma^* A}. \quad (3)$$

Here the subscripts T and L refer to the transverse and longitudinal virtual photon polarization states, respectively.

In the dipole picture the total cross section for the virtual photon-nucleus scattering at high energy factorizes to a product of the virtual photon wave function $\psi^{\gamma^* \to q\bar{q}}$ describing $\gamma^* \to q\bar{q}$ fluctuation, and the dipole-target scattering amplitude N as [37]

$$\sigma_{T,L}^{\gamma^* p} = 2 \sum_f \int d^2\mathbf{b} \, d^2\mathbf{r} \frac{dz}{4\pi} |\psi^{\gamma^* \to q\bar{q}}(\mathbf{r}, Q^2, z)|^2$$
$$\times N(\mathbf{r}, \mathbf{b}, x_{\mathrm{Bj}}). \quad (4)$$

Here z is the fraction of the photon longitudinal momentum carried by the quark, $\mathbf{r}$ is the transverse size of the $q\bar{q}$ dipole (higher Fock states would enter beyond leading order) and $\mathbf{b}$ is the center-of-mass of the dipole. We sum over light quarks $f = u, d, s$, as it has been shown in Ref. [22] that in a leading order calculation with the energy dependence obtained from the BK evolution it is not possible to simultaneously describe the total and charm production cross section data.

Following Refs. [20–22], when calculating the virtual photon-proton cross section we neglect the impact parameter dependence in the dipole-proton amplitude N and replace

$$\int d^2\mathbf{b} \to \frac{\sigma_0}{2}. \quad (5)$$

Here the proton transverse area $\sigma_0/2$ is taken to be a free parameter.

The squared photon wave functions summed over quark helicities read [37]

$$|\psi_T(r, z, Q^2)|^2 = \frac{2N_c}{\pi} \alpha_{\mathrm{em}} e_q^2 \{[z^2 + (1-z)^2]\varepsilon^2 K_1^2(\varepsilon r) + m_f^2 K_0^2(\varepsilon r)\} \quad (6)$$

and

$$|\psi_L(r, z, Q^2)|^2 = \frac{8N_c}{\pi} \alpha_{\mathrm{em}} e_q^2 Q^2 z^2 (1-z)^2 K_0^2(\varepsilon r), \quad (7)$$

with $\varepsilon^2 = z(1-z)Q^2 + m_q^2$. Here $m_q = 0.14$ GeV is the light quark mass and e_q is the quark charge. One could also in principle take the light quark mass to be a fit parameter as, e.g., in Refs. [22,38]. However these previous analyses have found only a small sensitivity on m_q and as such we choose a fixed value that is compatible with previous fits.

The Bjorken-x dependence of the dipole amplitude is given by the Balitsky-Kovchegov equation:

$$\frac{\partial N(\mathbf{x}_{01})}{\partial Y} = \int d^2\mathbf{x}_2 K_{\mathrm{BK}}(\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2)$$
$$\times [N(\mathbf{x}_{02}) + N(\mathbf{x}_{12}) - N(\mathbf{x}_{01})$$
$$- N(\mathbf{x}_{02})N(\mathbf{x}_{12})]. \quad (8)$$

Here $\mathbf{x}_{ij} = \mathbf{x}_i - \mathbf{x}_j$ and $Y = \ln x_0/x$ (see also Ref. [39] for a detailed discussion of the evolution variable). In this work we initiate the BK evolution at $x_0 = 0.01$. When the

running coupling corrections are included following the Balitsky prescription [40] the kernel reads

$$K_{\text{BK}}(\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2) = \frac{N_c \alpha_s(\mathbf{x}_{01}{}^2)}{2\pi^2} \left[\frac{\mathbf{x}_{01}{}^2}{\mathbf{x}_{12}{}^2 \mathbf{x}_{02}{}^2} \right.$$
$$+ \frac{1}{\mathbf{x}_{02}{}^2} \left(\frac{\alpha_s(\mathbf{x}_{02}{}^2)}{\alpha_s(\mathbf{x}_{12}{}^2)} - 1 \right)$$
$$\left. + \frac{1}{\mathbf{x}_{12}{}^2} \left(\frac{\alpha_s(\mathbf{x}_{12}{}^2)}{\alpha_s(\mathbf{x}_{02}{}^2)} - 1 \right) \right]. \quad (9)$$

In this work we present a leading-order analysis. We note that in addition to next-to-leading order corrections [6,7,13,14], it is also possible to resum corrections enhanced by large transverse logarithms to the BK equation following Refs. [41–43] (that in practice approximate the full NLO BK equation accurately [44,45]). Furthermore, it would be worthwhile to use the recently developed resummation procedure from Ref. [39] where the evolution is formulated in terms of the target rapidity instead of the projectile rapidity. These improvements we leave for future work.

For the initial condition of the BK evolution we use a McLerran-Venugopalan model [46] inspired parametrization as, e.g., in Ref. [21]

$$N(\mathbf{x}_{ij}, x = x_0) = 1 - \exp\left[-\frac{(\mathbf{x}_{ij}{}^2 Q_{s,0}^2)^\gamma}{4} \right.$$
$$\left. \times \ln\left(\frac{1}{|\mathbf{x}_{ij}|\Lambda_{\text{QCD}}} + e_c \cdot e \right) \right]. \quad (10)$$

Here the free parameters are $Q_{s,0}^2$ related to the initial saturation scale, the anomalous dimension γ controlling the shape of the dipole amplitude at small $|\mathbf{x}_{ij}|$ and the infrared regulator e_c. The Bayesian inference setup discussed in Sec. III can efficiently handle multidimensional parameter space, which allows us to use this more flexible parametrization for the initial condition compared to functional forms used in previous fits (e.g., in Ref. [21] either γ or e_c was fixed).

The strong coupling constant in Eq. (9) depends on the transverse distance scale. Following again Refs. [20,21] we write the coordinate space strong coupling as

$$\alpha_s(\mathbf{r}^2) = \frac{12\pi}{(33 - 2n_f)\ln\left(\frac{4C^2}{\mathbf{r}^2\Lambda_{\text{QCD}}^2}\right)}. \quad (11)$$

Here we take $n_f = 3$ as we only consider light quarks in this work, and $\Lambda_{\text{QCD}} = 0.241$ GeV. In the infrared region the coupling is frozen to $\alpha_s = 0.7$. The parameter C^2 controls the strong coupling scale in the coordinate space, and based on Refs. [40,47] its expected value is $C^2 = e^{-2\gamma_E} \approx 0.32$. However we consider C^2 to be a free parameter to control the uncertainty of the strong coupling

scale and to absorb some missing higher order corrections, for example the fact that the next-to-leading order evolution is known to result in slower evolution speed compared to the leading order case [45]. Previous leading order fits have also been found to prefer a much larger $C^2 = 7\ldots15$ [21].

III. BAYESIAN ANALYSIS SETUP

The nonperturbative parameters describing the initial condition for the BK evolution, the proton transverse area and the scale of the coordinate space running coupling can be estimated using Bayesian parameter inference. In this work we use a similar inference procedure as, e.g., in Refs. [29,33,35]. The output is a multidimensional likelihood distribution that comprehensively describes the parameter uncertainties and correlations based on prior beliefs and experimental data.

Bayesian inference is based on Bayes' theorem:

$$P(\mathbf{y}(\boldsymbol{\theta})|\mathbf{y}_{\text{exp}}) \propto P(\mathbf{y}_{\text{exp}}|\mathbf{y}(\boldsymbol{\theta}))P(\boldsymbol{\theta}). \quad (12)$$

Here the model output vector with given parametrization $\boldsymbol{\theta} = (Q_{s,0}^2, \gamma, e_c, C^2, \sigma_0/2)$ over N kinematical points $\mathbf{x}_i = (x_{\text{Bj}}, Q^2, y)_i$ is represented by $\mathbf{y}(\boldsymbol{\theta}) = (\mathbf{y}(\boldsymbol{\theta})_{\mathbf{x}_1}, \mathbf{y}(\boldsymbol{\theta})_{\mathbf{x}_2}, \ldots, \mathbf{y}(\boldsymbol{\theta})_{\mathbf{x}_N})$. The posterior function, represented by $P(\mathbf{y}(\boldsymbol{\theta})|\mathbf{y}_{\text{exp}})$, is the probability of $\mathbf{y}(\boldsymbol{\theta})$ being true given the experimental data, $\mathbf{y}_{\text{exp}}$, as evidence.

The probability represented by $P(\mathbf{y}_{\text{exp}}|\mathbf{y}(\boldsymbol{\theta}))$ is a measure of the agreement between the observation and the model calculation at the kinematical point $\mathbf{x}$ given the model parameters $\boldsymbol{\theta}$. We use a multivariate normal distribution in which case the logarithm of the likelihood function reads

$$\log P(\mathbf{y}_{\text{exp}}|\mathbf{y}(\boldsymbol{\theta}))$$
$$= -\frac{1}{2}[\Delta\mathbf{y}(\boldsymbol{\theta})^T\Sigma^{-1}(\boldsymbol{\theta})\Delta\mathbf{y}(\boldsymbol{\theta}) + \log(2\pi\det\Sigma)]. \quad (13)$$

Here $\Sigma = \Sigma_{\text{model}}(\boldsymbol{\theta}) + \Sigma_{\text{exp}}$ is a matrix sum of the model, or emulator as discussed shortly, and the experimental covariance matrices and the difference $\Delta\mathbf{y}(\boldsymbol{\theta}) = \mathbf{y}(\boldsymbol{\theta}) - \mathbf{y}_{\text{exp}}$ is a vector with length N. The experimental data considered in this work is the reduced cross section data measured by the H1 and ZEUS collaborations [19], where the published dataset also includes the full covariance matrix describing correlations among the 162 sources of systematic uncertainty. Furthermore we include as correlated uncertainties the 7 different procedural uncertainties originating from the combination of the H1 and ZEUS datasets. These correlations between different systematic uncertainties have not been included in previous dipole model fits, and the effect of these correlations on the final posterior distribution is illustrated in the Appendix. The uncorrelated systematic uncertainties and statistical uncertainties are added in quadrature.

TABLE I.   MAP and median values for the 4- and 5-parameter setups. The uncertainty estimates in the 95% credible intervals are also shown.

Parameter description	Prior range	4-parameter case		5-parameter case	
		Median	MAP	Median	MAP
Initial scale, $Q_{s,0}^2$ [GeV^2]	[0.04, 0.11]	$0.061^{+0.014}_{-0.007}$	0.060	$0.067^{+0.041}_{-0.015}$	0.077
Infrared regulator, e_c	[0.5, 60.0]	$35.3^{+23.3}_{-23.7}$	38.9	$27.5^{+31.0}_{-23.0}$	15.6
Running coupling scale, C^2	[2.0, 10.0]	$4.97^{+2.22}_{-1.40}$	4.60	$4.72^{+2.52}_{-1.58}$	4.47
Proton transverse area, $\sigma_0/2$ [mb]	[12.0, 18.0]	$14.1^{+1.1}_{-1.0}$	13.9	$14.0^{+1.3}_{-1.2}$	13.9
Anomalous dimension, γ	[0.9, 1.1]	1 (fixed)	1 (fixed)	$1.01^{+0.04}_{-0.03}$	1.01
Saturation scale, Q_s^2, at $x_{Bj} = 0.01$ [GeV^2]		0.279	0.280	0.288	0.289
χ^2/d.o.f.		1.013	1.011	1.016	1.012

The prior, represented by $P(\boldsymbol{\theta})$ in Eq. (12), encodes initial knowledge on the range and shape of the probability distribution of the model parameters. Not much information is known about the true distribution of the model parameters other than estimates of previous fits from Refs. [20–22]. In this study, the prior is chosen to be a flat multivariate distribution with bounds corresponding to guesses of possible parameter ranges, outside of which the posterior is set to zero. Initially, the width of the posterior distribution is unknown, so the study conservatively begins with a wide allowable extent and is later narrowed down. The prior ranges used in the final analysis are shown in Table I.

In order to effectively sample the multidimensional parameter space, we employ Gaussian process emulators (GPEs) [48] as computationally efficient surrogates for the actual theory calculation, and the Markov Chain Monte Carlo (MCMC) sampling method. The MCMC importance sampling is implemented via the EMCEE package [49] that uses an affine-invariant ensemble sampler. This sampler uses an algorithm similar to the Metropolis-Hastings (MH) where each step of a walker from $A \rightarrow B$ is accepted with a probability of $q \sim \frac{P(\theta_B)}{P(\theta_A)}$. Unlike the MH algorithm, EMCEE employs an ensemble of k *walkers* where the advancement of the chain of one walker is also based on the current positions of the other walkers; therefore, it converges much faster than MH. Each step is then appended to the MCMC chain to, eventually, form the target posterior distribution.

The GPEs are trained at different points in the parameter space sampled from a latin hypercube which ensures that the design points are optimally distributed in the parameter space. This significantly reduces the number of training points needed to get accurate estimates from GPEs, and in this analysis we use 500 training parametrizations. We include data in the region 2.0 GeV^2 < Q^2 < 50.0 GeV^2 where there are 403 datapoints. The low-Q^2 cut ensures the presence of a perturbative scale, and the upper limit removes contribution from the region dominated by DGLAP dynamics not included in the present analysis.

We represent the HERA data in this kinematical domain by 6 principal components that capture 99.99999% variance of the model output within the prior. The GP output at a certain $(\mathbf{x}, \boldsymbol{\theta})$ is a mean estimate and a covariance matrix that describes the emulator uncertainties and is used to form the matrix Σ_{model}.

The GPEs are validated by comparing their predictions with the actual model calculations of the reduced cross section. The left panel of Fig. 1 show the comparison between the model and the emulator output at all kinematical (x_{Bj}, Q^2, y) points in the HERA data computed using separate validation sets of model parameters not included in the training of the GPEs. The emulator accuracy is found to very good, with the average relative difference being 0.047% in our standard setup where $Q_{s,0}^2, e_c, C^2, \sigma_0/2$ and γ are free parameters. The right-hand side panel of Fig. 1 shows the corresponding z-score defined as

$$z = \frac{\text{GPE} - \text{model}}{\sigma_{\text{GPE}}}, \qquad (14)$$

where σ_{GPE} is the uncertainty estimate of the GPE. The width of the z distribution is less than unity which means that the emulator uncertainties are typically slightly overestimated.
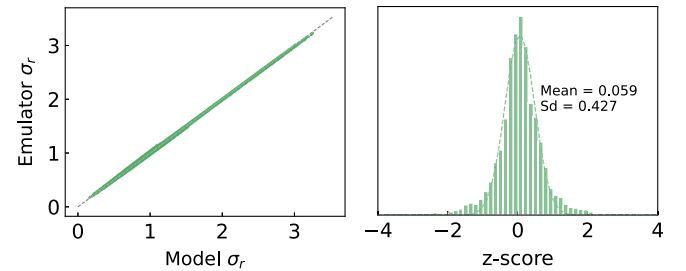

FIG. 1.   Validation for the 5-parameter Bayesian fit (left) comparing the calculated reduced cross section to the emulator prediction. The z-score plot (right) is constructed with 500 validation points over 403 kinematics. The dashed line shows a Gaussian fit to the z-score histogram, for which the mean and the standard deviation (SD) are given in the figure.

However, it is the experimental uncertainty that dominates in the likelihood function.

IV. INFERRED INITIAL CONDITION FOR THE BK EVOLUTION

We determine the posterior distribution of model parameters in two separate setups. First, in our main setup we consider all parameters discussed above ($Q_{s,0}^2$, e_c, C^2, $\sigma_0/2$ and γ) to be free and determine the corresponding posterior distribution. We will refer to this setup as the 5-parameter case. For comparison we also consider the case where we set $\gamma = 1$ as in the standard MV model, and refer to this as

the 4-parameter setup. This parametrization corresponds to the "MV^e" one used in Ref. [21].

The obtained posterior distributions are shown in Fig. 2 for both the 4- and 5-parameter cases. The plots along the diagonal present the 1D projections of the posterior showing the likelihood distributions for individual parameters. The off-diagonal plots are 2D histograms that illustrate the correlations between the pairs of model parameters.

The HERA data is found to constrain most of the model parameters well, except the infrared regulator e_c for which the likelihood distribution extends to large values of e_c in both setups. The broad distribution is not unexpected as the
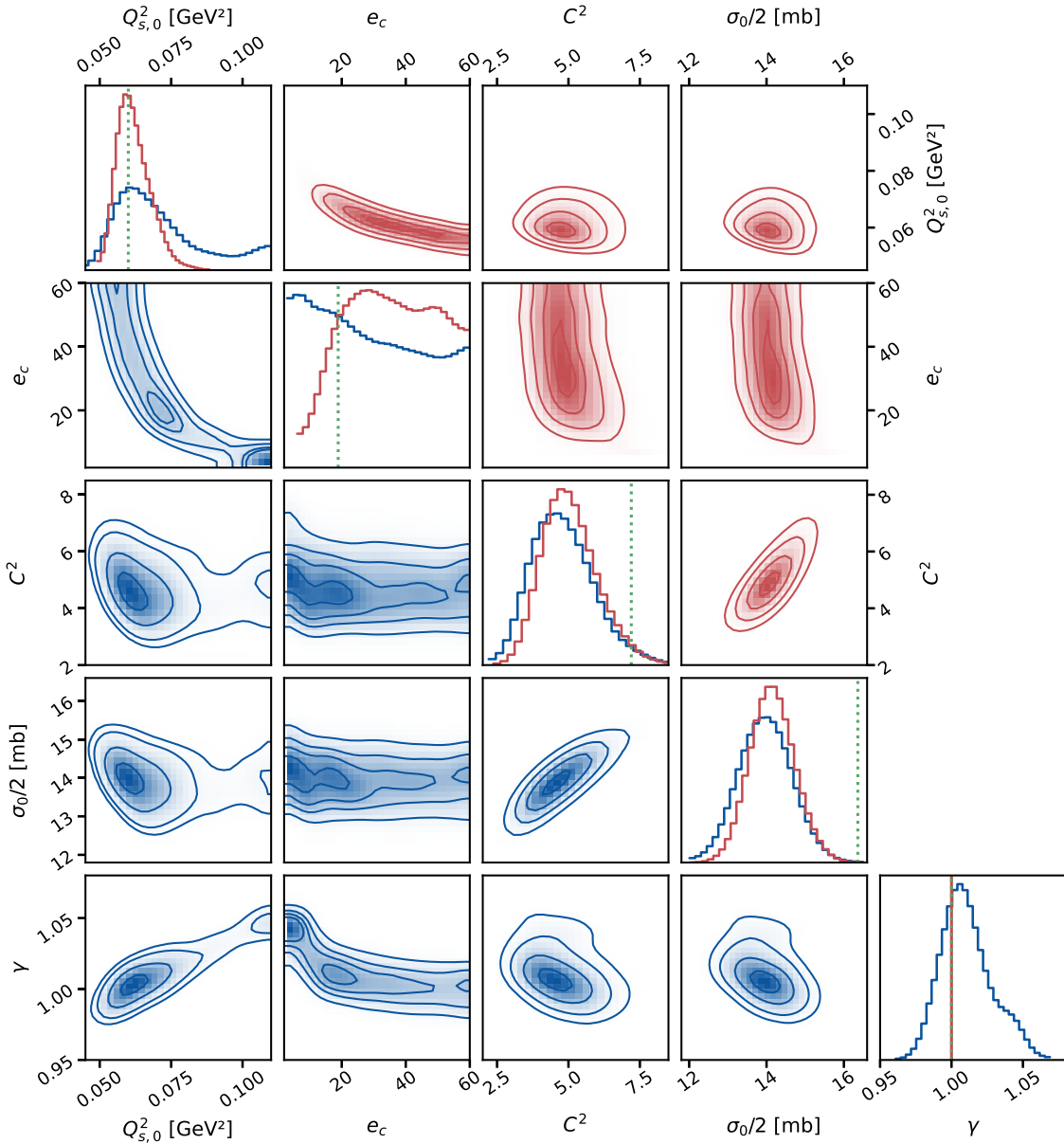

FIG. 2.   1- and 2-dimensional projections of the posterior probability distributions for both the 4-parameter (red) and 5-parameter case (blue). The dotted green line shows the best fit values from the MV^e fit of Ref. [21].

dipole amplitude depends on e_c weakly and only in the region where $r \gtrsim 1/\Lambda_{\rm QCD}$, see Eq. (10). Also, in the 5-parameter case, large values for $Q_{\rm s,0}^2$ are allowed and compensated by small e_c, and in both setups we obtain a clear negative correlation between $Q_{\rm s,0}^2$ and e_c.

In the 5-parameter case where γ is a free parameter, we find that the HERA data prefers $\gamma = 0.95...1.05$. This is in contrast to previous fits performed in Refs. [20–22] where the Q^2-dependence of the HERA data was shown to prefer $\gamma \approx 1.1...1.2$ (in fits with fixed $e_c = 1$). As we will discuss in more detail in Sec. V, having $\gamma \lesssim 1$ is advantageous e.g. when calculating inclusive particle production cross section in proton-nucleus collisions. The most likely values for the individual parameters in the 4-parameter case are similar to what is reported in the previous leading order fit [21].

In addition to the negative $Q_{\rm s,0}^2 - e_c$ correlation discussed above, we also find a clear negative correlation between $Q_{\rm s,0}^2$ and $\sigma_0/2$ especially in the 5-parameter case. This is because, in the region where the virtual photon-proton cross section is dominated by small dipoles, one has

$$\sigma^{\gamma^* p} \sim \frac{\sigma_0}{2} N(r) \sim \frac{\sigma_0}{2} \left(\frac{Q_{\rm s,0}^2}{Q^2}\right)^\gamma, \qquad (15)$$

where we use the fact that $r^2 \sim 1/Q^2$ (except in the aligned jet limit where $z \approx 0$ or $z \approx 1$). This dependence also explains the negative correlation seen between γ and $\sigma_0/2$.

As stated earlier, $Q_{\rm s,0}$ parametrizes the saturation scale at initial $x_{\rm Bj}$. The saturation scale, $Q_{\rm s}$, determines the scale at which nonlinear dynamics becomes important. Independent of the functional form of the BK initial condition, we can define (following, e.g., [21]) $Q_{\rm s}^2$ as the solution to $N(\mathbf{r}^2 = 2/Q_{\rm s}^2, x_{\rm Bj}) = 1 - e^{-1/2}$. The determined saturation scales at initial $x_{\rm Bj}$ in both 4- and 5-parameter setups are shown in Table I, and are found to be identical in both cases. Furthermore, Fig. 3 shows the $x_{\rm Bj}$ dependence of $Q_{\rm s}^2$, which is identical in both setups despite the differences in the posterior distributions.

At large $Q_{\rm s,0}^2$, the typical running coupling $\alpha_{\rm s}(r) \sim \alpha_{\rm s}(1/Q_{\rm s})$ is smaller. Consequently in that case there is no need for as large C^2 to get a slow enough evolution speed compatible with the x-dependence of the HERA data. Hence the negative correlation between C^2 and $Q_{\rm s,0}^2$ is obtained, and consequently a positive correlation between C^2 and $\sigma_0/2$. These correlations appear very clearly in both setups. Consequently, it is essential to take into account the correlations between the model parameters when estimating uncertainties when calculating different cross sections.

The median values for the parameters are presented in Table I along with the *maximum a posteriori* (MAP) estimates corresponding to the maximum of the posterior distribution.

The compatibility with the HERA data is quantified by $\chi^2/{\rm d.o.f.}$ defined as

$$\frac{\chi^2}{\rm d.o.f.} = \frac{1}{N-p} \Delta \mathbf{y}(\boldsymbol{\theta})^T \Sigma_{\rm exp}^{-1} \Delta \mathbf{y}(\boldsymbol{\theta}), \qquad (16)$$

where $N = 403$ is the number of experimental datapoints in the considered kinematical domain and $p = 4$ or $p = 5$ is the number of free model parameters. The $\chi^2/{\rm d.o.f.} \approx 1.01$ shows that an excellent description of the precise HERA data is obtained, similarly as in previous leading order fits. By calculating the average cross section over many samples from the posterior and not using a single parametrization (e.g., median or MAP), we get $\chi^2/{\rm d.o.f.} = 1.02$ in both the 4- and 5- parameter cases, showing that both the MAP and median parametrizations are good estimates when calculating the proton structure functions.

The good agreement with the HERA data is also illustrated in Fig. 4 where we show a comparison to the reduced cross section data in a few selected virtuality bins. The central lines are the average values obtained by computing the cross section using many posterior samples, and the uncertainty band corresponds to two standard deviation variation. To allow for asymmetric uncertainty estimates, in this work we always calculate separately the 2 standard deviation (2σ) uncertainty band above and below the mean value. We note that the reduced cross section is generically underestimated at low Q^2. However, we still get a good χ^2 when the correlated systematic uncertainties in the HERA data are taken into account. If the statistical and systematical uncertainties in the HERA data are simply added in quadrature, we get a somewhat larger $\chi^2/{\rm d.o.f.} = 2.2$ in both the 4- and 5-parameter case. See also discussion in the Appendix for the effect of the correlated uncertainties on the posterior distribution.

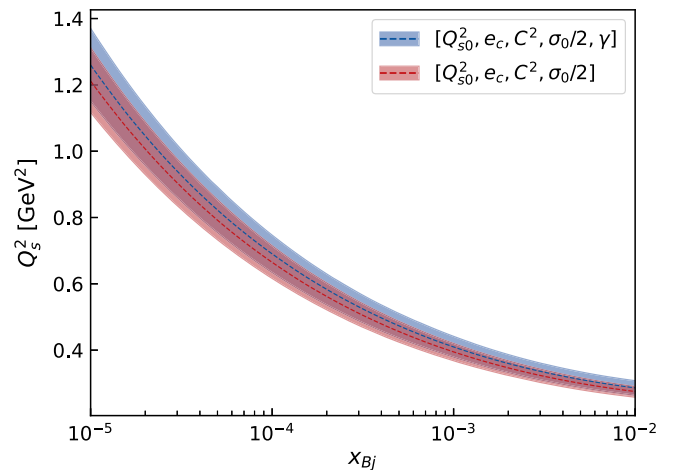The initial condition for the dipole amplitude $N(r, x_{\rm Bj} = 0.01)$ and the evolved amplitude at a much smaller

FIG. 3. The saturation scale as a function of $x_{\rm Bj}$, defined as the solution to $N(\mathbf{r}) = 1 - e^{-1/2}$ when $\mathbf{r}^2 = 2/Q_s^2$. This is shown for both the 4- and 5- parameter cases with 2σ uncertainty bands.
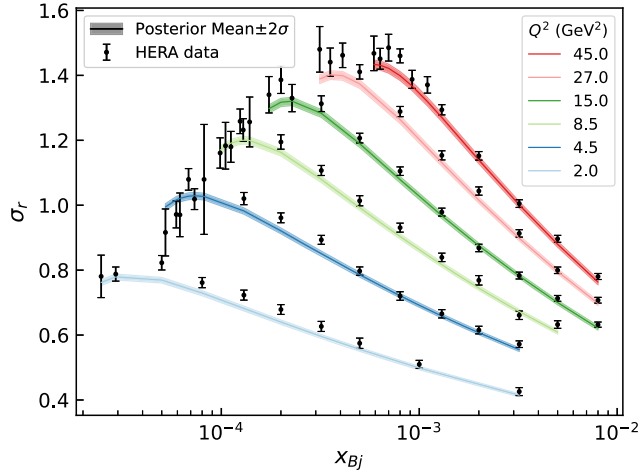
FIG. 4. The reduced cross section as a function of $x_{\rm Bj}$ and the 2σ uncertainty band obtained from the posterior samples in selected Q^2 bins compared with the HERA data at $\sqrt{s} = 318$ GeV [19].

$x_{\rm Bj} = 10^{-5}$ are shown in Fig. 5. Note that the initial dipole amplitude only depends on $Q_{s,0}$, e_c and γ, and furthermore the evolution speed is also sensitive to C^2. The central value of the dipole amplitude is obtained as an average of the dipole amplitudes computed using different samples from the posterior distribution, and the uncertainty estimate is a 2 standard deviation band. We show the dipole amplitudes both from the 4- and 5-parameter setups, and, as expected, the uncertainty band in the 5-parameter case is slightly larger due to a broader posterior distribution. However the actual (2 standard deviation) uncertainty in the dipole amplitude is small, typically a few percent and maximally $\approx 10\%$.
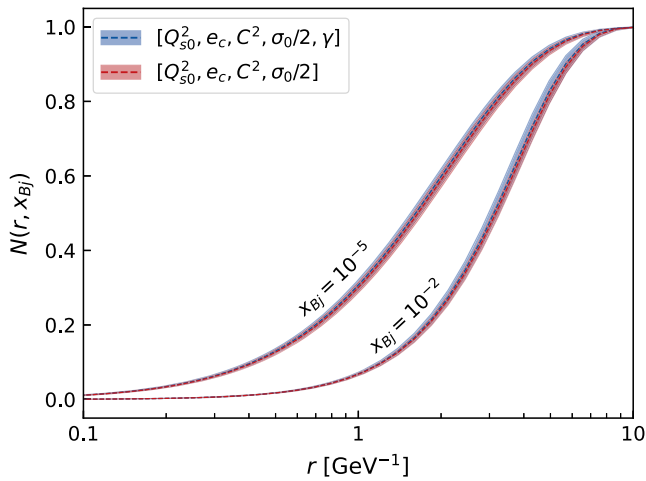

FIG. 5. The dipole amplitude at the initial condition and evolved down to $x_{\rm Bj} = 10^{-5}$ calculated in the 4- and 5- parameter cases. The band shows a 2σ uncertainty.

V. APPLICATIONS

As the dipole-target scattering amplitude is a convenient degree of freedom at small-x (see e.g. Ref. [50]), the determined posterior distribution is a necessary input to calculations of all observables. The fact that the full posterior distribution, and not just the best fit values, are available, also enables us to rigorously take into account the uncertainties in the nonperturbative parameters describing the initial condition for the BK evolution. In this Section, we demonstrate how the uncertainties in the extracted dipole amplitude propagate to the inclusive quark-target cross section and to the nuclear modification factor for the F_2 structure function that will be measured at the EIC.

A. Inclusive quark production

Let us first consider quark-target scattering which is the relevant subprocess in inclusive forward hadron production in pp or pA collisions. The inclusive forward quark production cross section in proton-proton collisions is directly proportional to the two-dimensional Fourier transform of the initial dipole-amplitude [21,27,51,52]:

$$\frac{{\rm d}\sigma^{q+A \to q+X}}{{\rm d}y {\rm d}^2\mathbf{k}} = \frac{\sigma_0/2}{(2\pi)^2} xq(x,\mu^2) \tilde{S}_p(\mathbf{k}), \qquad (17)$$

where the produced quark transverse momentum and rapidity are $\mathbf{k}$ and y, respectively. The dipole amplitude is evaluated at $x = |\mathbf{k}|/\sqrt{s}e^{-y}$ where $\sqrt{s}$ is the center-of-mass energy. Furthermore, $xq(x,\mu^2)$ is the quark parton distribution function evaluated at scale μ^2, and

$$\tilde{S}_p(\mathbf{k}) = \int {\rm d}\mathbf{r}\, e^{-i\mathbf{k}\cdot\mathbf{r}}[1 - N(\mathbf{r})]. \qquad (18)$$

The two-dimensional Fourier transform of the dipole amplitude scaled by $\sigma_0/2$ at the initial $x_{\rm Bj} = 0.01$ is calculated in both the 4- and 5-parameter case, and the results are shown in Fig. 6. Although the median γ obtained is close to 1, the posterior covers parametrizations where anomalous dimensions slightly larger than unity (up to $\gamma \approx 1.05$) are encountered, resulting in negative values of $\tilde{S}_p(\mathbf{k})$ in the large k region. This is because the Fourier transform is not positive definite with $\gamma > 1$ [53]. At smaller x after the BK evolution the Fourier transform would be positive definite as the BK evolution drives the anomalous dimension toward an asymptotic value $\gamma = 0.6...0.8$.

Additionally we observe that although the uncertainties in the determined dipole amplitude are typically around 5% (see Fig. 5), uncertainties in the quark-target cross section (in the $\tilde{S}_p \times \sigma_0/2$) can be significantly larger. Even in the 4-parameter case where the Fourier transform is positive definite, e.g., at $k = 2$ GeV the upper limit of the 2σ uncertainty band is around 150% larger than the mean
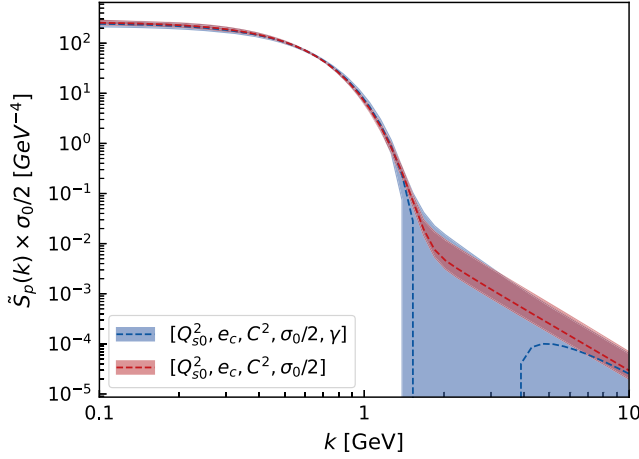
FIG. 6. 2D Fourier transform of the dipole-proton amplitude scaled by $\sigma_0/2$ at the initial $x = 0.01$ in the 4- and 5-parameter cases.

value. This highlights the importance of properly taking into account the parametrization uncertainties when applying the dipole amplitude determined from the HERA data when describing particle production processes at the LHC as e.g. in Ref. [21]. At next-to-leading order [9,10] the possibility to emit a gluon changes the kinematics from $1 \rightarrow 1$ at LO to $1 \rightarrow 2$ at NLO, which can have a significant effect on the p_T distribution (e.g., one can obtain a power-law tail at high p_T even with a Gaussian dipole). The importance of properly take into account uncertainties in the initial dipole-target scattering amplitude when calculating inclusive spectra at NLO has been recently emphasize in Ref. [28].

B. Nuclear effects on F_2

Much more pronounced saturation effects can be expected to seen in scattering processes where the target is a heavy nuclei instead of the proton, as the saturation scale in the nucleus scales roughly as $\sim A^{1/3}$. The dipole-proton scattering amplitude determined in this work can be generalized to the dipole-nucleus case, e.g., by employing the optical Glauber model. Following Ref. [21], we write the initial condition for the BK evolution of the dipole-nucleus scattering amplitude at fixed impact parameter **b** as

$$N_A(\mathbf{r}, \mathbf{b}, x = x_0) = 1 - \exp\left[-AT_A(\mathbf{b})\frac{\sigma_0}{2}\frac{(\mathbf{r}^2 Q_{s,0}^2)^\gamma}{4}\right.$$
$$\left.\times \ln\left(\frac{1}{|\mathbf{r}|\Lambda_{\text{QCD}}} + e_c \cdot e\right)\right], \quad (19)$$

where $T_A(\mathbf{b})$ is the transverse thickness function of the nucleus of mass number A. The dipole-nucleus amplitudes at fixed **b** are then evolved to small-x by solving the BK equation without the impact parameter dependence. This thickness function is obtained by integrating the Woods-Saxon distribution

$$\rho_A(\mathbf{b}, z) = \frac{n}{1 + \exp\left[\frac{\sqrt{\mathbf{b}^2 + z^2} - R_A}{d}\right]} \quad (20)$$

over z, where $d = 0.54$ fm and $R_A = (1.12A^{1/3} - 0.86A^{-1/3})$ fm. The normalization condition $\int d^2\mathbf{b}T_A(\mathbf{b}) = 1$ fixes the overall constant n. In the large $|\mathbf{b}| > b_{\text{max}}$ region where $AT_A(\mathbf{b})\sigma_0/2 < 1$, the saturation scale of the nucleus falls below that of the proton. In order to avoid an unphysically rapid growth of the nuclear size in this region we write, following again Ref. [21], $N_A(\mathbf{r}, \mathbf{b}, x) = AT_A(\mathbf{b})\frac{\sigma_0}{2}N(\mathbf{r}, x)$, i.e., use a dipole-proton scattering amplitude scaled such that all nuclear effects vanish.

We compute the nuclear modification factor for the structure function F_2 defined as

$$R_{eA} = \frac{F_{2,A}}{AF_{2,p}}, \quad (21)$$

where $F_{2,A}$ is the F_2 structure function for the nucleus with mass number A. In this work we consider gold, i.e. $A = 197$, for which this modification factor will be measured at the EIC. The obtained R_{eA} as a function of x_{Bj} is shown in Fig. 7 and as a function of Q^2 in Fig. 8.

We predict a significant nuclear suppression in the F_2 structure function already in the EIC kinematics. By construction $R_{eA} \rightarrow 1$ at the initial $x = 0.01$ in the limit $Q^2 \rightarrow \infty$, and a stronger suppression is seen both toward small-x_{Bj} and low-Q^2. Unlike above when studying the inclusive quark production cross section, in the case of R_{eA} the parametrization uncertainties effectively cancel in the structure function ratio and the 2σ uncertainty is at 1–2% level. The uncertainty grows slightly toward small-x_{Bj} where also the C^2 parameter controlling the evolution speed become important.
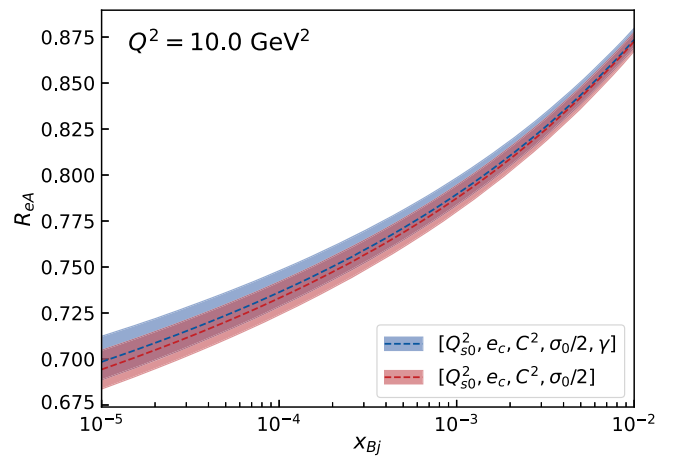

FIG. 7. Nuclear modification factor for the structure function F_2 as a function of x_{Bj} at fixed $Q^2 = 10$ GeV^2.
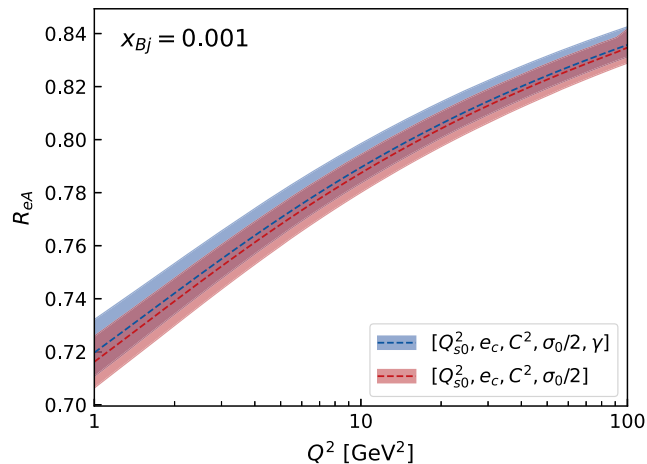
FIG. 8. Nuclear modification factor for the structure function F_2 as a function of photon virtuality Q^2 at fixed $x_{\rm Bj} = 0.001$.

VI. CONCLUSIONS AND OUTLOOK

We have determined the posterior distribution for the nonperturbative parameters describing the dipole-proton scattering amplitude at the initial condition of the small-x BK evolution. This is a necessary input to calculations of all scattering processes at high energy in the color glass condensate framework. We have also determined, for the first time, uncertainty estimates for these nonperturbative parameters. The obtained posterior distribution makes it possible to rigorously take into account the uncertainties in the initial condition parametrization when looking for signals of gluon saturation from various different observables. The posterior distributions determined in this work can be found from the Supplemental Material [54].

We have shown that the HERA structure function data constrains most of the model parameters well and the remaining uncertainties in the obtained dipole amplitude are at a few percent level. The flexible Bayesian inference setup allows us to include more freedom in the initial dipole-proton amplitude than in the previous fits, and in particular we find that the HERA data does not require a large anomalous dimension $\gamma \approx 1.1\ldots1.2$ unlike in the analyses presented in Refs. [20–22]. Instead we get $\gamma \approx 1$ which is more natural considering that $\gamma \leq 1$ is required to obtain a positive definite quark production cross section and unintegrated gluon distribution [53].

We demonstrate the uncertainty propagation by calculating the inclusive quark production cross section in proton-proton collisions and the nuclear modification factor R_{eA} for the structure function F_2 to be measured at the EIC and at the LHeC/FCC-he. Significant uncertainties up to ~100% were found in the case of the inclusive quark production in proton-proton collisions. On the other hand, the uncertainties were found to mostly cancel in the nuclear modification factor R_{eA} for the F_2 structure function. As such it can be crucial to properly propagate

the uncertainties in the nonperturbative input when comparing the CGC predictions describing the gluon saturation effects to, e.g., LHC inclusive spectra.

In the future, we will apply the developed computationally efficient setup to determine the initial condition for the BK evolution at next-to-leading order accuracy. This is especially intriguing as it has been recently shown that at NLO global analyses including both total and heavy quark production cross section data simultaneously become feasible [24,25]. We expect the proper treatment of initial condition uncertainties to be of equal importance as higher-order corrections when the color glass condensate calculations are promoted to the precision level.

APPENDIX: TREATMENT OF EXPERIMENTAL CORRELATED SYSTEMATIC UNCERTAINTIES

In this work we have included, for the first time, the correlations between the different sources of systematical uncertainty in the HERA structure function data when determining the initial condition for the BK evolution. In order to quantify the effect these previously neglected correlations have on the determined initial condition, we also extract the initial condition for the BK evolution by neglecting all correlations in the experimental data. Instead, we choose the covariance matrix to be diagonal with $\Sigma_{\rm exp} = {\rm diag}(\sigma_1^2, \sigma_2^2, \sigma_3^2, \ldots, \sigma_N^2)$ where the statistical and both correlated and uncorrelated systematical uncertainties are added in quadrature: $\sigma_i^2 = \sigma_{\rm sys,uncorr}^2 + \sigma_{\rm stat,uncorr}^2 + \sigma_{\rm corr}^2$.

The posterior distribution obtained with and without the correlated uncertainties in the experimental data in the 5-parameter case are shown in Fig. 9. Overall we find quite similar posterior distributions, except that when the correlations between the experimental uncertainties are known there is more flexibility for the model calculations to agree with the data and we typically obtain broader distributions
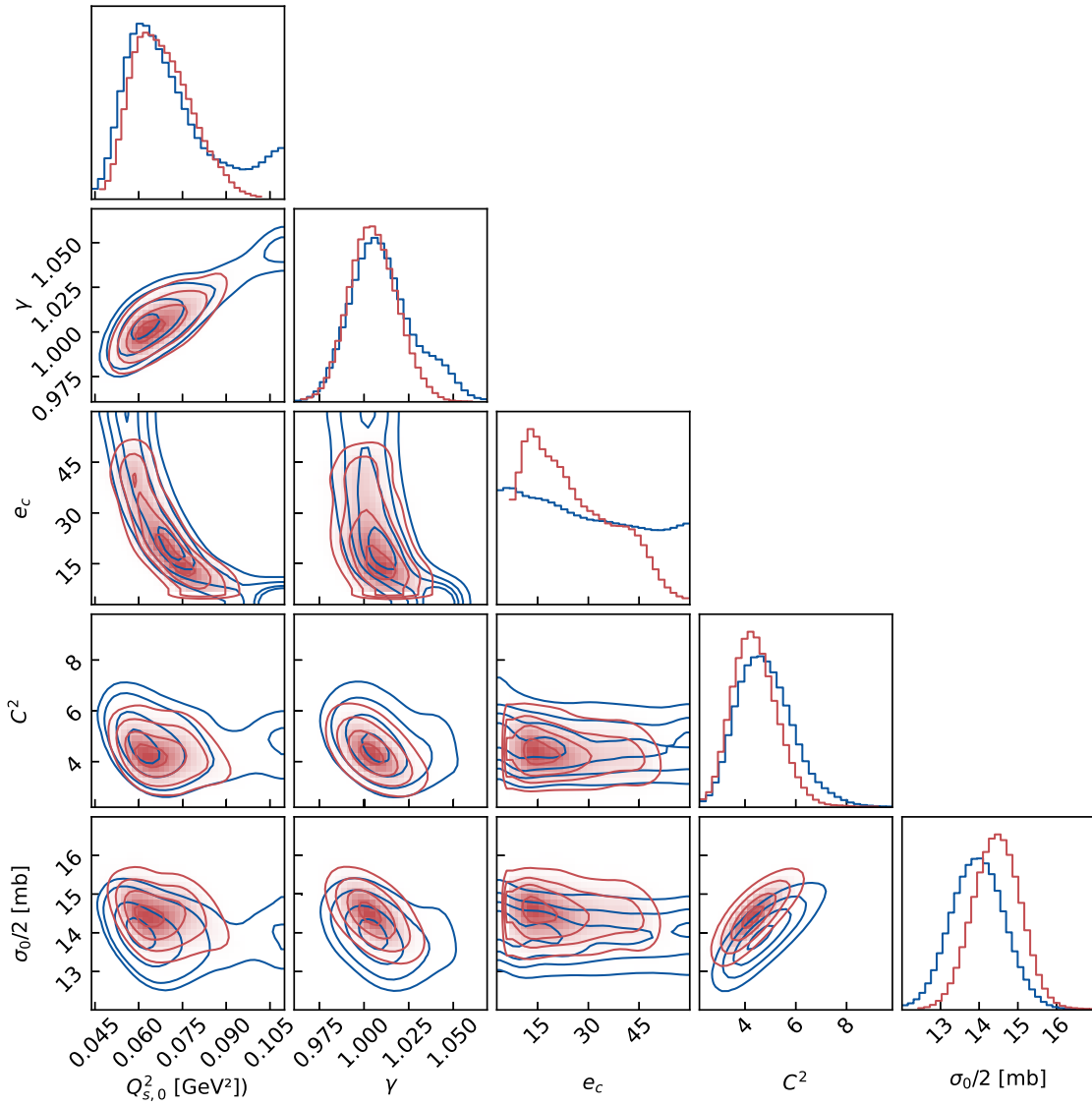
FIG. 9. Posterior distribution in the 5-parameter case obtained with (blue) and without (red) including the correlations among the experimental uncertainties.

for the parameters. Additionally the preferred values for the C^2 (controlling the evolution speed) and $\sigma_0/2$ (proton size) are also slightly shifted by the inclusion of the experimental covariance matrix.

[1] R. Abdul Khalek *et al.*, Science requirements and detector concepts for the electron-ion collider: EIC Yellow Report, Nucl. Phys. **A1026**, 122447 (2022).

[2] A. Morreale and F. Salazar, Mining for gluon saturation at colliders, Universe **7**, 312 (2021).

[3] P. Agostini *et al.* (LHeC, FCC-he Study Group Collaboration), The large hadron-electron collider at the HL-LHC, J. Phys. G **48**, 110501 (2021).

[4] F. Gelis, E. Iancu, J. Jalilian-Marian, and R. Venugopalan, The color glass condensate, Annu. Rev. Nucl. Part. Sci. **60**, 463 (2010).

[5] R. Boussarie, A. V. Grabovsky, D. Y. Ivanov, L. Szymanowski, and S. Wallon, Next-to-leading order computation of exclusive diffractive light vector meson production in a saturation framework, Phys. Rev. Lett. **119**, 072002 (2017).

[6] G. Beuf, T. Lappi, and R. Paatelainen, Massive quarks at one loop in the dipole picture of deep inelastic scattering, Phys. Rev. Lett. **129**, 072001 (2022).

[7] G. Beuf, T. Lappi, and R. Paatelainen, Massive quarks in NLO dipole factorization for DIS: Transverse photon, Phys. Rev. D **106**, 034013 (2022).

[8] Y. Shi, L. Wang, S.-Y. Wei, and B.-W. Xiao, Pursuing the precision study for color glass condensate in forward hadron productions, Phys. Rev. Lett. **128**, 202302 (2022).

[9] G. A. Chirilli, B.-W. Xiao, and F. Yuan, Inclusive hadron productions in pA collisions, Phys. Rev. D **86**, 054005 (2012).

[10] A. M. Stasto, B.-W. Xiao, and D. Zaslavsky, Towards the test of saturation physics beyond leading logarithm, Phys. Rev. Lett. **112**, 012302 (2014).

[11] P. Caucal, F. Salazar, B. Schenke, T. Stebel, and R. Venugopalan, Back-to-back inclusive dijets in DIS at small x: Complete NLO results and predictions, Phys. Rev. Lett. **132**, 081902 (2024).

[12] F. Bergabo and J. Jalilian-Marian, Dihadron production in DIS at small x at next-to-leading order: Transverse photons, Phys. Rev. D **107**, 054036 (2023).

[13] B. Ducloué, H. Hänninen, T. Lappi, and Y. Zhu, Deep inelastic scattering in the dipole picture at next-to-leading order, Phys. Rev. D **96**, 094017 (2017).

[14] G. Beuf, Dipole factorization for DIS at NLO: Combining the $q\bar{q}$ and $q\bar{q}g$ contributions, Phys. Rev. D **96**, 074033 (2017).

[15] H. Mäntysaari and J. Penttala, Complete calculation of exclusive heavy vector meson production at next-to-leading order in the dipole picture, J. High Energy Phys. 08 (2022) 247.

[16] Y. V. Kovchegov, Small x F_2 structure function of a nucleus including multiple pomeron exchanges, Phys. Rev. D **60**, 034008 (1999).

[17] I. Balitsky, Operator expansion for high-energy scattering, Nucl. Phys. **B463**, 99 (1996).

[18] F. D. Aaron et al. (H1, ZEUS Collaborations), Combined measurement and QCD analysis of the inclusive $e^{\pm}p$ scattering cross sections at HERA, J. High Energy Phys. 01 (2010) 109.

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

[20] J. L. Albacete, N. Armesto, J. G. Milhano, and C. A. Salgado, Non-linear QCD meets data: A global analysis of lepton-proton scattering with running coupling BK evolution, Phys. Rev. D **80**, 034031 (2009).

[21] T. Lappi and H. Mäntysaari, Single inclusive particle production at high energy from HERA data to proton-nucleus collisions, Phys. Rev. D **88**, 114020 (2013).

[22] J. L. Albacete, N. Armesto, J. G. Milhano, P. Quiroga-Arias, and C. A. Salgado, AAMQS: A non-linear QCD analysis of new HERA data at small-x including heavy quarks, Eur. Phys. J. C **71**, 1705 (2011).

[23] B. Ducloué, E. Iancu, G. Soyez, and D. N. Triantafyllopoulos, HERA data and collinearly-improved BK dynamics, Phys. Lett. B **803**, 135305 (2020).

[24] H. Hänninen, H. Mäntysaari, R. Paatelainen, and J. Penttala, Proton structure functions at next-to-leading order in the dipole picture with massive quarks, Phys. Rev. Lett. **130**, 192301 (2023).

[25] G. Beuf, H. Hänninen, T. Lappi, and H. Mäntysaari, Color glass condensate at next-to-leading order meets HERA data, Phys. Rev. D **102**, 074028 (2020).

[26] A. Dumitru, H. Mäntysaari, and R. Paatelainen, High-energy dipole scattering amplitude from evolution of low-energy proton light-cone wave functions, Phys. Rev. D **107**, 114024 (2023).

[27] J. L. Albacete and C. Marquet, Single inclusive hadron production at RHIC and the LHC from the color glass condensate, Phys. Lett. B **687**, 174 (2010).

[28] H. Mäntysaari and Y. Tawabutr, Complete next-to-leading order calculation of single inclusive π^0 production in forward proton-nucleus collisions, Phys. Rev. D **109**, 034018 (2024).

[29] J. E. Bernhard, J. S. Moreland, S. A. Bass, J. Liu, and U. Heinz, Applying Bayesian parameter estimation to relativistic heavy-ion collisions: Simultaneous characterization of the initial state and quark-gluon plasma medium, Phys. Rev. C **94**, 024907 (2016).

[30] J. E. Bernhard, J. S. Moreland, and S. A. Bass, Bayesian estimation of the specific shear and bulk viscosity of quark-gluon plasma, Nat. Phys. **15**, 1113 (2019).

[31] G. Nijs and W. van der Schee, Hadronic nucleus-nucleus cross section and the nucleon size, Phys. Rev. Lett. **129**, 232301 (2022).

[32] D. Soeder, W. Ke, J. F. Paquet, and S. A. Bass, Bayesian parameter estimation with a new three-dimensional initial-conditions model for ultrarelativistic heavy-ion collisions, arXiv:2306.08665.

[33] J. E. Parkkila, A. Onnerstad, S. F. Taghavi, C. Mordasini, A. Bilandzic, M. Virta, and D. J. Kim, New constraints for QCD matter from improved Bayesian parameter estimation in heavy-ion collisions at LHC, Phys. Lett. B **835**, 137485 (2022).

[34] D. Everett et al. (JETSCAPE Collaboration), Phenomenological constraints on the transport properties of QCD matter with data-driven model averaging, Phys. Rev. Lett. **126**, 242301 (2021).

[35] H. Mäntysaari, B. Schenke, C. Shen, and W. Zhao, Bayesian inference of the fluctuating proton shape, Phys. Lett. B **833**, 137348 (2022).

[36] N. Armesto, T. Lappi, H. Mäntysaari, H. Paukkunen, and M. Tevio, Signatures of gluon saturation from structure-function measurements, Phys. Rev. D **105**, 114017 (2022).

[37] Y. V. Kovchegov and E. Levin, *Quantum Chromodynamics at High Energy* (Cambridge University Press, Cambridge, England, 2012), Vol. 33.

[38] H. Mäntysaari and P. Zurita, In depth analysis of the combined HERA data in the dipole models with and without saturation, Phys. Rev. D **98**, 036002 (2018).

[39] B. Ducloué, E. Iancu, A. H. Mueller, G. Soyez, and D. N. Triantafyllopoulos, Non-linear evolution in QCD at high-energy beyond leading order, J. High Energy Phys. 04 (2019) 081.

[40] I. Balitsky, Quark contribution to the small-x evolution of color dipole, Phys. Rev. D **75**, 014001 (2007).

[41] E. Iancu, J. D. Madrigal, A. H. Mueller, G. Soyez, and D. N. Triantafyllopoulos, Resumming double logarithms in the QCD evolution of color dipoles, Phys. Lett. B **744,** 293 (2015).

[42] E. Iancu, J. D. Madrigal, A. H. Mueller, G. Soyez, and D. N. Triantafyllopoulos, Collinearly-improved BK evolution meets the HERA data, Phys. Lett. B **750,** 643 (2015).

[43] G. Beuf, Improving the kinematics for low-x QCD evolution equations in coordinate space, Phys. Rev. D **89,** 074039 (2014).

[44] I. Balitsky and G. A. Chirilli, Next-to-leading order evolution of color dipoles, Phys. Rev. D **77,** 014019 (2008).

[45] T. Lappi and H. Mäntysaari, Next-to-leading order Balitsky-Kovchegov equation with resummation, Phys. Rev. D **93,** 094004 (2016).

[46] L. D. McLerran and R. Venugopalan, Computing quark and gluon distribution functions for very large nuclei, Phys. Rev. D **49,** 2233 (1994).

[47] Y. V. Kovchegov and H. Weigert, Triumvirate of running couplings in small-x evolution, Nucl. Phys. **A784,** 188 (2007).

[48] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. **12,** 2825 (2011), https://www.jmlr.org/papers/v12/pedregosa11a.html.

[49] D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman, EMCEE: The MCMC hammer, Publ. Astron. Soc. Pac. **125,** 306 (2013).

[50] F. Dominguez, C. Marquet, B.-W. Xiao, and F. Yuan, Universality of unintegrated gluon distributions at small x, Phys. Rev. D **83,** 105005 (2011).

[51] A. Dumitru and J. Jalilian-Marian, Forward quark jets from protons shattering the colored glass, Phys. Rev. Lett. **89,** 022301 (2002).

[52] A. Dumitru, A. Hayashigaki, and J. Jalilian-Marian, The color glass condensate and hadron production in the forward region, Nucl. Phys. **A765,** 464 (2006).

[53] B. G. Giraud and R. Peschanski, Fourier-positivity constraints on QCD dipole models, Phys. Lett. B **760,** 26 (2016).

[54] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevD.109.054018 for Parameter vectors sampled from the Posterior.