

Longitudinal dynamics of large and small systems from a 3D Bayesian calibration of RHIC top-energy collision data

A. Mankolli^{1,*}, C. Shen,^{2,3} M. Luzum,⁴ J.-F. Paquet,¹ M. Singh,¹ J. Velkovska,¹ S. A. Bass,⁵ C. Gale,⁶ G. A. C. da Silva,⁴ L. Du,^{7,8,6} L. Kasper,¹ G. S. Rocha,^{9,1} D. Soeder,⁵ S. Tuo,¹ G. Vujanovic,¹⁰ X. Wu,⁶ W. Zhao,^{7,8} M. Chartier,¹¹ Y. Chen,¹ R. Datta,² R. Dolan,² R. Ehlers,^{7,8} H. Elfner,^{12,13,14} R. J. Fries,^{15,16} D. A. Hangal,^{17,2} B. V. Jacak,^{7,8} P. M. Jacobs,^{7,8} S. Jeon,⁶ Y. Ji,¹⁸ F. Jonas,^{7,8} M. Kordell, II,^{15,16} A. Kumar,¹⁰ R. Kunnawalkam-Elayavalli,¹ J. Latessa,¹⁹ Y.-J. Lee,^{20,21} A. Majumder,² S. Mak,¹⁸ C. Martin,²² H. Mehryar,¹⁹ T. Mengel,²² C. Nattrass,²² J. Norman,¹¹ M. Ockleton,¹¹ C. Parker,^{15,16} J. H. Putschke,² H. Roch,^{23,24} G. Roland,^{20,21} B. Schenke,²⁵ L. Schwiebert,¹⁹ A. Sengupta,^{15,16} C. Sirimanna,⁵ R. A. Soltz,^{17,2} I. Soudi,²⁶ Y. Tachibana,²⁷ X.-N. Wang,^{28,7,8} and J. Zhang¹

(JETSCAPE Collaboration)

¹Department of Physics and Astronomy, *Vanderbilt University*, Nashville, Tennessee 37235, USA

²Department of Physics and Astronomy, *Wayne State University*, Detroit, Michigan 48201, USA

³RIKEN BNL Research Center, *Brookhaven National Laboratory*, Upton, New York 11973, USA

⁴Instituto de Física, *Universidade de São Paulo*, C.P. 66318, 05315-970 São Paulo, SP, Brazil

⁵Department of Physics, *Duke University*, Durham, North Carolina 27708, USA

⁶Department of Physics, *McGill University*, Montréal, Quebec H3A 2T8, Canada

⁷Department of Physics, *University of California*, Berkeley, California 94720, USA

⁸Nuclear Science Division, *Lawrence Berkeley National Laboratory*, Berkeley, California 94720, USA

⁹Instituto de Física, *Universidade Federal Fluminense*, Niterói, Rio de Janeiro, 24210-346, Brazil

¹⁰Department of Physics, *University of Regina*, Regina, Saskatchewan S4S 0A2, Canada

¹¹Oliver Lodge Laboratory, *University of Liverpool*, Liverpool, United Kingdom

¹²GSI Helmholtzzentrum für Schwerionenforschung, 64291 Darmstadt, Germany

¹³Institute for Theoretical Physics, *Goethe University*, 60438 Frankfurt am Main, Germany

¹⁴Frankfurt Institute for Advanced Studies, 60438 Frankfurt am Main, Germany

¹⁵Cyclotron Institute, *Texas A&M University*, College Station, Texas 77843, USA

¹⁶Department of Physics and Astronomy, *Texas A&M University*, College Station, Texas 77843, USA

¹⁷Lawrence Livermore National Laboratory, *Livermore*, California 94550, USA

¹⁸Department of Statistical Science, *Duke University*, Durham, North Carolina 27708, USA

¹⁹Department of Computer Science, *Wayne State University*, Detroit, Michigan 48202, USA

²⁰Laboratory for Nuclear Science, *Massachusetts Institute of Technology*, Cambridge, Massachusetts 02139, USA

²¹Department of Physics, *Massachusetts Institute of Technology*, Cambridge, Massachusetts 02139, USA

²²Department of Physics and Astronomy, *University of Tennessee*, Knoxville, Tennessee 37996

²³University of Jyväskylä, *Department of Physics*, P.O. Box 35, FI-40014 University of Jyväskylä, Finland

²⁴Helsinki Institute of Physics, P.O. Box 64, FI-00014 University of Helsinki, Finland

²⁵Physics Department, *Brookhaven National Laboratory*, Upton, New York 11973, USA

²⁶Fakultät für Physik, *Universität Bielefeld*, D-33615 Bielefeld, Germany

²⁷Akita International University, Yuwa, Akita-city 010-1292, Japan

²⁸Key Laboratory of Quark and Lepton Physics (MOE) and Institute of Particle Physics, *Central China Normal University*, Wuhan 430079, China



(Received 10 February 2026; accepted 28 April 2026; published 15 July 2026)

A comprehensive Bayesian analysis of the three-dimensional (3D) dynamics of high-energy nuclear collisions is presented. We perform a systematic model-to-data comparison using simulations of large and small collision systems, and a broad range of measurements from the PHENIX, STAR, PHOBOS, and BRAHMS collaborations spanning nearly two decades of operations at the BNL Relativistic Heavy Ion Collider (RHIC). In particular, we perform fully 3D multistage simulations including rapidity-dependent energy deposition with global energy conservation using the 3D Glauber model, along with relativistic viscous hydrodynamics with MUSIC. We calibrate the model on rapidity- and p_T -differential observables and analyze the respective constraints on initial state and transport properties they provide. We emphasize the additional constraints provided by rapidity-dependent measurements, the differences in large and small system calibrations, and the tension exhibited by particular observables. We use our calibrated model to make predictions of observables in p -Au and ^3He -Au collisions.

*Contact author: andi.mankolli@vanderbilt.edu

Furthermore, we facilitate direct comparison of experimental measurements by highlighting the dependence of flow measurements on the rapidity of the regions of interest and reference, as well as the importance of the centrality selection. In particular, we examine the apparent differences between the STAR and PHENIX v_2 and v_3 measurements in small systems.

DOI: [10.1103/7kpw-vfgp](https://doi.org/10.1103/7kpw-vfgp)

I. INTRODUCTION

The strong nuclear force described by quantum chromodynamics (QCD) can, under extreme conditions of energy density, give rise to a deconfined state of quarks and gluons called the quark-gluon plasma (QGP). The study of the QGP has been the principal goal of the relativistic nuclear collision programs at the Large Hadron Collider (LHC) at CERN and the Relativistic Heavy Ion Collider (RHIC) at Brookhaven National Laboratory. Measurements spanning over two decades at these colliders have confirmed the hydrodynamic nature of the QGP expansion [1–5]. Phenomenological studies using portions of this wealth of data in a Bayesian inference framework have made significant progress in constraining the transport properties of strongly coupled QCD matter [6–29], which remain difficult to constrain with lattice QCD calculations.

The experimental programs have, in addition, provided evidence for the possible emergence of QGP in smaller collision systems [30–37]. In particular, studies at RHIC have measured hydrodynamic flow signals in a variety of small systems and their strong dependence on the initial state of the collisions [38,39]. Furthermore, the orientation of the event planes and the hydrodynamic response to the initial geometry have been understood to have particularly strong longitudinal dependence in smaller collisions [39–41].

The aim of this study is to exploit the large number of rapidity-dependent measurements at the RHIC top energy in tandem with a $(3 + 1)$ -dimensional $(3 + 1)$ D model of nuclear collisions to better constrain the properties of the QGP formed in heavy-ion collisions, to gain insight about the nature of the longitudinal dynamics in these collisions, and to explore any tension between large and small systems within a hydrodynamic description. This information can be extracted only thanks to our use of a fully $(3 + 1)$ D model that not only describes the finite rapidity region explicitly and makes possible direct comparisons with forward and backward measurements but also allows us to incorporate the finite rapidity effects embedded in midrapidity measurements.

In Sec. II we describe in detail the theoretical model of nuclear collisions we employ in our study. Section III covers the techniques we use to emulate our model's predictions and gives an overview of our Bayesian analysis procedure and choices. In Sec. IV, we list and describe the experimental data we use in our calibrations and discuss our selections. In Sec. V, we show the results of Bayesian calibrations for our default dataset as well as several topical subsets and highlight what can be learned from them; in that Section, we also discuss the tension which different observables place on our model. In Sec. VI we use our calibrated model to provide a general description of the longitudinal structure of collectivity across collision systems and to make additional predictions for measurements, including in p -Au and ^3He -Au collisions.

II. THE $(3 + 1)$ D MODELING OF NUCLEAR COLLISIONS

Bulk observables in high-energy nuclear collisions, such as the number of produced particles and their azimuthal distributions, have long been well-described by multistage models that include an initial-state energy-deposition phase followed by a hydrodynamic evolution and (often) hadronic transport [42–46] thereafter. Typically, the models, using a variety of prescriptions, define an initial energy and momentum density distribution for a locally thermally equilibrated hydrodynamic medium, which is then evolved for several fm/c as an expanding relativistic fluid. Then, as the systems become increasingly dilute, individual fluid cells are converted into hadrons that scatter with each other and decay into stable states. In the high-energy limit, longitudinal boost-invariance is a good approximation for midrapidity observables [47]. It reduces computational cost in event-by-event simulations, enabling comprehensive Bayesian inference with $(2 + 1)$ D models and midrapidity data. A number of such studies in the last decade have been shown to provide constraints on the transport coefficients of QGP [6,7,10–14,16–20]. A few studies have performed Bayesian calibrations with full $(3 + 1)$ D simulations and deduced additional constraints from experimental data in forward- and backward-rapidity regions [8,9,15,21–24]. Some of these studies have included models that differ substantially from ours, particularly in the prescription of the initial state [8,9,15,24]. Others have used the same overarching models, although with different physics assumptions and scope [21–23]. In this study, we use the three-dimensional (3D) Glauber model [48–50] to describe the initial state, followed by $(3 + 1)$ D viscous relativistic hydrodynamics for the plasma phase [51–53], and a hadronic afterburner to model the hadronic interactions of final-state particles [54,55].

A. The 3D initial state

The initial state of heavy-ion collisions, particularly its longitudinal dependence, is still not well understood in terms of a first-principles QCD description [56–59]. However, models that sample nucleons from density distributions and accordingly deposit energy in the transverse plane of the collision have been successful when coupled to hydrodynamics in describing the experimental data [60,61]. One such model is the Monte Carlo Glauber model [62–65]. In recent years, efforts have been made to extend the description of some such models to the energy deposited in the longitudinal direction [15,48,49,66]. The 3D Monte Carlo Glauber model, which we use for this study, has been a result of such efforts and has been used for a number of years to successfully describe a range of heavy-ion collision data, particularly at energies probed in the RHIC Beam Energy Scan program [48,49].

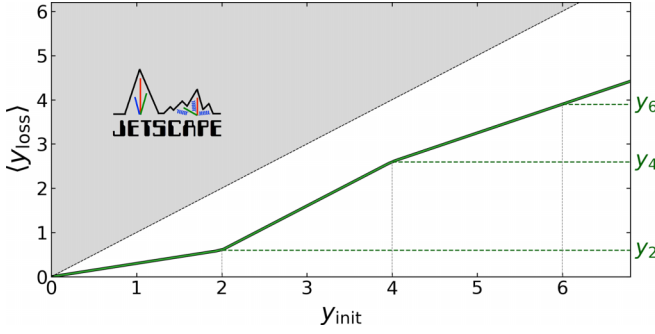


FIG. 1. Average rapidity loss parametrization as a function of incoming beam rapidity y_{init} given by Eq. (2). The gray area is restricted because y_{loss} must be less than y_{init} .

For a given impact parameter in a collision, nucleons are first sampled according to a 3D Woods-Saxon distribution (such that collisions may take place at finite z positions to an extent modulated by the collision energy). Any given nucleon is modeled as a collection of three valence quark hotspots and one soft gluon hotspot, each of which may lose and deposit energy in a collision [49,67]. The longitudinal momenta related to the valence quark hotspots are sampled from the nucleons' parton distribution functions.¹ Their spatial positions are sampled from a 3D Gaussian with width $\sqrt{B_G}$, a parameter we vary in the Bayesian analysis to constrain subnucleonic fluctuations,

$$P(\vec{r}) \propto e^{-r^2/(2B_G)}. \quad (1)$$

For producing the bulk medium, not all binary collisions between incoming nucleons in the heavy-ion collision participate independently. We model this many-body effect with a parameter $\alpha_{\text{shadowing}}$ that can vary between 0 and 1 to control the probability of string production in individual nucleon-nucleon collisions.

In a nucleon-nucleon collision, one or multiple strings may be formed between pairs of hotspots. These hotspots lose energy and decelerate along the longitudinal direction with a constant string tension [48]. The lost energy is distributed in the longitudinal direction and is eventually deposited into hydrodynamic fields as source terms after a formation time τ_{form} in the collision-pair rest frame. We calibrate the string formation time in the Bayesian analysis.

The amount of energy lost by colliding nucleons is modeled according to an average rapidity loss $\langle y_{\text{loss}} \rangle$. This is a quantity generally modeled to describe collisions at different energies [or, equivalently, incoming beam rapidities $y_{\text{init}} = y_{\text{beam}} = \text{arccosh}(\sqrt{s_{\text{NN}}}/(2m_N))$]. In the analysis, we vary its value at three points of incoming beam rapidity, namely, y_2 (at $y_{\text{init}} = 2$), y_4 (at $y_{\text{init}} = 4$), and y_6 (at $y_{\text{init}} = 6$), in a piece-wise parametrization that in principle allows for constraining the $\sqrt{s_{\text{NN}}}$ dependence of the rapidity loss. Illustrated in Fig. 1,

the value of the function at every point of y_{init} is thus given by

$$\langle y_{\text{loss}} \rangle = \begin{cases} (y_2)(y_{\text{init}})/2, & 0 < y_{\text{init}} \leq 2 \\ y_2 + (y_4 - y_2)\frac{y_{\text{init}} - 2}{2}, & 2 < y_{\text{init}} \leq 4 \\ y_4 + (y_6 - y_4)\frac{y_{\text{init}} - 4}{2}, & 4 < y_{\text{init}}. \end{cases} \quad (2)$$

Since in this study we restrict ourselves to experimental data collected at $\sqrt{s_{\text{NN}}} = 200$ GeV, corresponding to incoming beam rapidity $y_{\text{beam}} = 5.36$, we expect to have strong sensitivity to the latter two parameters, and limited sensitivity to y_2 .² The event-by-event fluctuations on rapidity loss are introduced by a restricted Gaussian distribution with mean $\langle y_{\text{loss}} \rangle$ and variance $\sigma_{y_{\text{loss}}}$ [49].

The wounded nucleons lose part of their energy and momentum during the collision. The remainder are treated as collision remnants. These remnants do not have partners to form a string, yet they also deposit energy and momentum into the fluid fields according to a scaled down rapidity loss relative to that of the collision strings. The proportionality constant is given by the parameter α_{rem} such that $\langle y_{\text{loss}}^{\text{rem}} \rangle = \alpha_{\text{rem}} \langle y_{\text{loss}} \rangle$ with $\alpha_{\text{rem}} \in [0, 1]$. If $\alpha_{\text{rem}} = 0$ the remnants deposit no energy.

The energy and momentum from the string-forming hotspots and from the wounded nucleons form a source current, J^ν , whose spatial distribution is smeared according to Gaussian profiles in the transverse plane and a flux-tube-like profile in the longitudinal direction with corresponding widths in the transverse and longitudinal directions (σ_x and σ_η , respectively). An additional parameter, α_{shift} , modulates the longitudinal dependence of the string transverse profiles [49].

Including energy-momentum source terms from both collision strings and remnants ensures strict energy-momentum conservation when mapping the initial state of the collision to the subsequent hydrodynamic phase, a central feature of this model [48,49,66]. All the energy and momentum present in the incoming beam are thus accounted for throughout the collision evolution.

B. Relativistic hydrodynamics

The contributed energy from strings and collision remnants as described above is ultimately deposited as a source term, J^ν , to the energy-momentum tensor at a given time τ_{form} :

$$\partial_\mu T^{\mu\nu} = J^\nu. \quad (3)$$

We then evolve the system with Israel-Stewart relativistic viscous hydrodynamics in 3 + 1 dimensions solved using the MUSIC code [51–53] in the iEBE-MUSIC framework [22]. The

¹Nuclear parton distribution modifications are also included in this treatment (EPS09) [68].

²At the same time, the parametrization allows for posterior degeneracy between two parameters if the constraining region falls, as it does for our $\sqrt{s_{\text{NN}}} = 200$ GeV collisions, at an intermediate incoming rapidity between the two parameters. We take the step of imposing on top of this parametrization a monotonicity condition at the level of the Bayesian likelihood that forces $y_4 \leq y_6$ (and $y_2 \leq y_4$). This step also ensures that the y_2 posterior, due to the correlation with y_4 , will not be uniform.

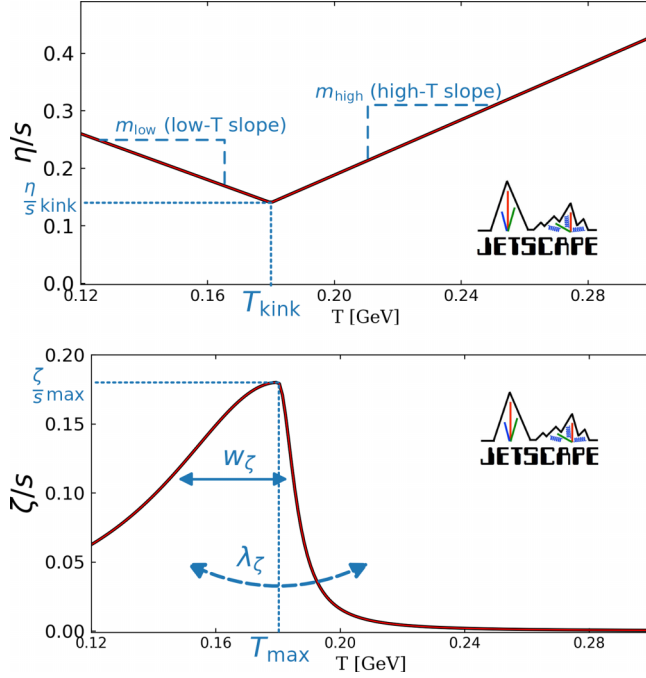


FIG. 2. Viscosity parametrization as a function of temperature for the specific shear viscosity given by Eq. (4) (top) and specific bulk viscosity given by Eq. (5) (bottom).

hydrodynamic properties that we aim to constrain with the data are the first-order transport coefficients η/s and ζ/s , the specific (scaled by the entropy density) shear and bulk viscosity, respectively. They are understood to have a nontrivial temperature dependence which is parametrized in this work, as was done in recent analyses [10,11,19,20], in a way that incorporates a few basic assumptions about the behavior near the deconfinement transition. These include the specific shear viscosity featuring a change in slope in the deconfinement region, and the specific bulk viscosity decreasing to zero at large temperature. Each transport coefficient is encoded by four parameters, illustrated in Fig. 2. For the specific shear viscosity we have:

$$\frac{\eta}{s}(T) = \begin{cases} \frac{\eta}{s}_{\text{kink}} + (m_{\text{low}})(T_{\text{kink}} - T), & T < T_{\text{kink}} \\ \frac{\eta}{s}_{\text{kink}} + (m_{\text{high}})(T - T_{\text{kink}}), & T > T_{\text{kink}}, \end{cases} \quad (4)$$

where T_{kink} gives the temperature of the kink and $(\eta/s)_{\text{kink}}$ its value. The parameters m_{low} and m_{high} give the slopes in the two temperature regions. On top of this parametrization we also enforce positivity [$\eta/s = \max(0, \eta/s)$] to reject unphysical viscosities. For the specific bulk viscosity we use a skewed Cauchy distribution:

$$\frac{\zeta}{s}(T) = \frac{(\zeta/s)_{\text{max}} \Lambda^2}{\Lambda^2 + (T - T_{\text{max}})^2}, \quad (5)$$

with

$$\Lambda = w_{\zeta}[1 + \lambda_{\zeta} \text{sign}(T - T_{\text{max}})], \quad (6)$$

where T_{max} gives the temperature of the maximum and $(\zeta/s)_{\text{max}}$ its value. The parameter w_{ζ} encodes the width of the distribution and λ_{ζ} encodes the asymmetry around T_{max} .

The parametrizations result in correlations between the values of the coefficients at different temperatures. These correlations are ultimately embedded in the analysis posteriors. A number of recent analyses have used different parametrizations for the viscosity coefficients, although no detailed studies on the impact of the parametrization on the posteriors have yet been done (a brief discussion of this question can be found in Ref. [29]).

For the equation of state, we use the common prescription of lattice calculations at high temperature matched to a hadron gas calculation at low temperature [69]. We assume $\mu_B = 0$. The equation of state, particularly in the transition region, is not explicitly varied and represents one source of uncertainty not accounted for by the analysis. We choose to take the equation of state as a given and to focus on constraining the initial state and the transport coefficients of QCD (as was done in most recent analyses, including ones focusing on the finite- μ_B region).

C. Particization and the hadronic phase

As the fluid cools down and crosses the deconfinement transition region, the relevant microscopic degrees of freedom become once again the confined hadrons. To describe hadronic interactions that take place at this stage, and to ultimately compare with experimental data, we convert the fluid to particles using the standard Cooper-Frye prescription [70] implemented in the iSS package [71]. This is done on a hypersurface of constant energy density given by e_{switch} , the value of which is varied in the analysis.

Furthermore, we use Grad's approximation [72–76] for the viscous corrections to the equilibrium momentum distribution of particles in Cooper-Frye. In the JETSCAPE two-dimensional (2D) analysis [10], the theoretical uncertainties related to the choice of viscous corrections were explored in detail. In this study we do not explicitly account for this uncertainty, although we use the model which the 2D analysis showed was most favored by the data. We checked the impact of particization model on various observables with our calibrated model and find little sensitivity.

The hadronic phase, then, is described using UrQMD [54,55], which simulates hadronic scatterings and decays until interactions cease between particles; that is, until “kinetic freeze-out.”

III. MODEL EMULATION AND THE BAYESIAN COMPARISON WITH DATA

We work in a Bayesian statistical framework in order to perform a systematic and rigorous comparison of the model with experimental data. Following Bayes' Theorem of conditional probability,

$$\mathcal{P}(\theta|D) = \frac{\mathcal{P}(D|\theta)\mathcal{P}(\theta)}{\mathcal{P}(D)}, \quad (7)$$

where $\mathcal{P}(\theta)$ denotes the prior knowledge about the model parameters θ , $\mathcal{P}(D|\theta)$ the likelihood that the data D are consistent with a given set of parameters, and $\mathcal{P}(D) = \int d\theta \mathcal{P}(D|\theta) \mathcal{P}(\theta)$ the normalizing Bayes evidence. The posterior

$\mathcal{P}(\theta|D)$ provides quantitative probabilistic constraints on the parameters.

The posterior is a probability distribution whose dimensionality is set by the number of parameters. It is a complicated function that must be sampled numerically. Probing accurately the posterior through sampling and marginalization is challenging, since it requires obtaining the model calculations of all calibration observables—quantities averaged over many simulated events—at a large number of parameter sets. The required number of posterior samples at different parameter sets strongly depends on the number of model parameters—about 10 million samples for 20 model parameters, in our case.

From the standpoint of direct model calculations, this task was already computationally insurmountable even for the analyses exploiting boost invariance to model the physics using $(2+1)$ D hydrodynamic simulations. It remains very much the case for our full $(3+1)$ D simulations. To get around this challenge, Gaussian process emulators (GPs) have been used as a fast surrogate for the model's predictions in all large-scale Bayesian analyses of heavy-ion collisions. In this analysis, we use a very similar emulator to the one used in the JETSCAPE 2D Bayesian calibration. Details of the emulation are described in Sec. III B and more thoroughly laid out in Ref. [10].

A. Priors

We choose, as in previous analyses, to use uniform distributions as priors for our parameters [10,11]. This includes the parameters used to vary the transport coefficients and their temperature dependence [Eqs. (4) and (5)].

It is important to note that using uniform priors does not necessarily mean using weak priors, and that the prior ranges encode important prior information. For example, uniform priors of the viscosity parameters still encode a significant amount of information about their temperature dependence. Our approach in this analysis is to use wide ranges for the priors, generally consistent with the priors of previous analyses, to help understand the impact of rapidity-dependent observables. We underscore that we do not incorporate the posteriors obtained in previous heavy-ion calibrations for common parameters in the choice of our priors due to overlapping calibration datasets and differences in modeling. The 20 model parameters and their prior ranges are listed in Table I.

The observable prior is the representation of the assumed prior knowledge about the model parameters at the level of individual observable probability distributions. We show an approximation³ of this in Fig. 3 for a representation of all⁴ the observables considered in our two calibration systems

³Putting aside the statistical uncertainty on individual model calculations, the figure is also an approximation of the prior distribution in that it is sampled at only about 400 Latin hypercube design points.

⁴This figure, as well as later figures in the same style showing the observable posteriors, will include a representative subset of the full observable set considered in the analysis, omitting additional centrality classes of a number of observables. However, every observable

TABLE I. All 20 model parameters described in Sec. II and varied in the analysis along with the corresponding stage of the collision and the prior range.

Parameter	Collision stage	Prior range
y_2	Initial state	[0,2]
y_4	Initial state	$[y_2, 4]$
y_6	Initial state	$[y_4, 6]$
σ_{loss}	Initial state	[0,1]
α_{rem}	Initial state	[0,1]
$\alpha_{\text{shadowing}}$	Initial state	[0,1]
B_G [GeV ⁻²]	Initial state	[2,25]
σ_x [fm]	Initial state	[0.1,0.5]
σ_η	Initial state	[0.1,0.8]
α_{shift}	Initial state	[0,1]
τ_{form} [fm]	Hydrodynamization	[0.2,1]
$(\eta/s)_{\text{kink}}$ [GeV]	Hydro	[0.13,0.3]
m_{low} [GeV ⁻¹]	Hydro	$[-2,1]$
m_{high} [GeV ⁻¹]	Hydro	$[-1,2]$
$(\eta/s)_{\text{kink}}$	Hydro	[0.01,0.2]
$(\zeta/s)_{\text{max}}$	Hydro	[0.01,0.2]
$(\zeta/s)_{\text{Tmax}}$ [GeV]	Hydro	[0.12,0.3]
w_ζ [GeV]	Hydro	[0.025,0.15]
λ_ζ	Hydro	$[-0.8,0.6]$
e_{switch} [GeV/fm ³]	Particlization	[0.1,0.6]

(including those observables not included in calibrations for technical reasons discussed below, but with which the model will still be compared). The plot includes a 90% credible interval for each observable bin predicted in the Latin hypercube sample (described below) of model simulations. We see a generally wide observable prior that covers essentially all the data within experimental uncertainties.

B. The model emulator

The emulator acts as a fast surrogate for the model and estimates the calculations' statistical and interpolation uncertainties (including correlated uncertainties). To ensure the most efficient emulator training, the chosen prior ranges for the parameters coincide with the bounds of the parameter space sampled to train the emulator. The training space is sampled using a space-filling, Latin hypercube algorithm, as was done in previous analyses [19,20]. Exceptionally, for two parameters, y_4 and y_6 , the emulator training parameter ranges were wider than the Bayesian prior ranges ($[0,4]$ and $[0,6]$ instead of $[y_2, 4]$ and $[y_4, 6]$, respectively). This was due to the need to enforce a physical, monotonic behavior to the rapidity loss as a function of incoming rapidity and the particular choice of our parametrization, while avoiding these correlations in the Latin hypercube sampling of the training space.

The model is used to simulate approximately 400 parameter sets, or design points, spanning the training space for

type is represented. For a full summary of the data points used in calibration, refer to Table II.

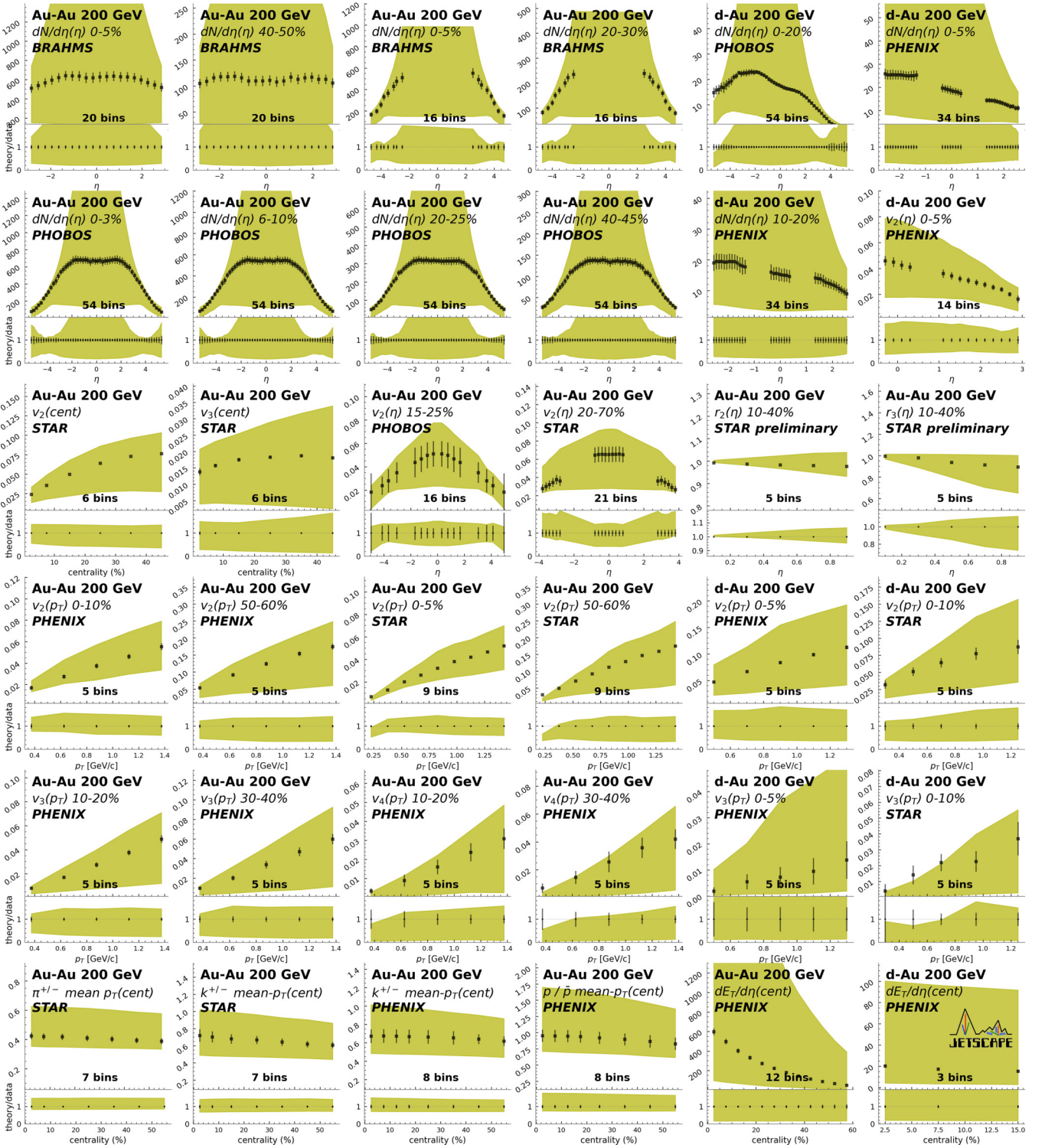


FIG. 3. The observable prior for the data included in the calibration, as well as some not included. The solid markers are the experimental data, while the bands show a 90% interval of the prior calculated using the full model at the design points. Each plot is labeled with the number of data bins corresponding to each observable in a given plot.

each of the two collision systems (d -Au and Au-Au). We train Gaussian Processes for each of 10 principal components (PCs) per collision system, into which the normalized observable data have been decomposed. The principal component analysis (PCA) rewrites the observables as a linear combination of

“projections” along the principal components. A significantly smaller number of these are used in the analysis compared with the number of original observable bins (roughly a reduction by two orders of magnitude), after truncating the linear combination past a reasonable number of terms. The reduction

allows for significantly more efficient emulation and, in all cases in this analysis, the principal components that are used collectively account for well over 99% of the variance in the observables across the prior parameter space. This is possible due to the high degree of redundancy in the information content of observables.

The principal component analysis transformation is performed once at training time, such that the emulators are trained on the principal components. The inverse transformation is performed each time the emulators are prompted to predict the model calculations at a given parameter point, meaning that the emulator output is rotated back to the original observable space before comparing with data. The analysis choices related to PCA are described in additional detail in Appendix A.

In addition, we perform two other transformations in observables and parameter space to improve the performance of the emulator. The emulator is trained using transformed parameters of the viscosity parametrization. Instead of training on the four shear and four bulk parameters, we train on 10 values of the viscosity as a function of temperature for each, respectively.⁵ We also transform (before principal component analysis) observables which feature a relatively large dynamical range, namely, the charged hadron multiplicity and the transverse energy, training on their logarithm. These procedures are further discussed in Appendix B, where we also show emulator validation done with a subset of the simulated design points.

C. Comparison with data and parameter estimation

To efficiently sample the parameter space, we use a Markov chain Monte Carlo (MCMC) algorithm implemented in the EMCEE package [77]. A given number of MCMC “walkers” (in our analysis, 200 of them as default) are initialized by uniformly sampling the 20 model parameters. The likelihood function, shown below, and the Bayesian posterior probability are computed at each step for each walker to determine which walkers in the ensemble are updated to the proposed position and which remain unchanged. Through many iterations, the MCMC chain eventually converges to a representative sample of the posterior distribution. In principle, the representative sample approaches and becomes the exact probability distribution in the limit of infinite steps; however, we arbitrarily terminate the algorithm at a practical stopping point. The MCMC phase up to convergence is called the “burn-in.”

The optimal length of MCMC chains may be determined by directly checking for convergence in the probability distributions of parameters. In Appendix C, we show the convergence checks performed for the default calibration of this analysis, for which we ultimately use 20 000 burn-in steps. We

see that after roughly that many steps, the distributions for all parameters converge.

To quantify the agreement with data of any given observable prediction of the model (or, in our case, its surrogate), we define the likelihood function as the following multivariate normal distribution:

$$\mathcal{P}(D|\theta) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} \exp \left\{ -\frac{1}{2} \Delta Y^T \Sigma^{-1} \Delta Y \right\}, \quad (8)$$

where n is the number of data points, $\Delta Y = Y_{\text{expt}} - Y_{\text{theor}}(\theta)$, the difference between the experimental observables and the model prediction at a parameter point θ , and Σ , the covariance matrix. It is a standard likelihood choice used in previous Bayesian analyses. It incorporates the uncertainties on the model and experiment in the form of covariance matrix $\Sigma = \Sigma_{\text{expt}} + \Sigma_{\text{theor}}$. The covariance matrix includes along its diagonal the uncertainties on any given observable bin. It also encodes correlations in the uncertainties between different observable bins in its off-diagonal elements.

The model portion of the uncertainty in our analysis is comprised of systematic and statistical uncertainties in the model predictions, as well as emulation uncertainty. Since the emulator is trained on model calculations limited by finite statistical uncertainty, the emulator’s estimate of its uncertainty includes the statistical model uncertainty. We do not separately estimate the theoretical systematic uncertainties that remain uncaptured by our calibrated model. As a consequence, the posteriors in this analysis should be interpreted as indicating the theoretical systematic uncertainty only to the extent that it is probed by the varied model parameters.

To construct the experimental covariance matrix, which is correspondingly the sum of systematic and statistical uncertainties, we have very limited input from measurements on any correlations in the uncertainty across observable bins. Therefore, we make the assumption of uncorrelated experimental uncertainties—that is, a diagonal experimental covariance matrix. Following previous analyses, we rely on the emulator to estimate the model correlations comprising the theoretical contribution to the covariance. We show an example of these emulator-estimated correlations in Appendix D.

As mentioned in the previous section, before emulator training, we perform a transformation of certain observables—namely, the multiplicity and the transverse energy—by taking the logarithm in order to improve the emulator performance. This yields emulator predictions in the transformed “log space,” with a corresponding covariance matrix also in such space. Given the options of transforming these predictions and covariance matrices to the original observable space or transforming the experimental data and its diagonal covariance matrix to log space, we choose the latter and ultimately compute the likelihood in log space.

D. Sampling the posterior and making predictions

Having obtained the parameter posterior distribution, we can sample it and construct the observable posterior, the representation of our knowledge at the level of observables, after comparing with data. While we can make predictions using any one parameter sample from the posterior—indeed, for

⁵The sampling of the two viscosity coefficients is still done according to the distributions of the eight (four for each) original parameters in Eqs. (4) and (5). For any given sample, before the emulator is called, the values of those parameters are converted to 20 (10 for each) values of the viscosity, spanning the temperature range uniformly according to the corresponding parametrization.

example, using the Maximum *A Posteriori* (MAP), i.e., most likely parameter set—the width of the observable posterior is a meaningful measure of the model uncertainty. This variation within the posterior in many cases exceeds the statistical uncertainty that can be reasonably achieved on the predictions of any given observable with the emulator for a particular parameter set. Therefore, in our predictions with the calibrated model, we emphasize using many samples of the posterior to account for this uncertainty.

We use a large number of samples of the posterior—around 1000—evaluated by the emulator, to construct a 90% credible interval for all of the observables in the calibration systems. This gives an efficient estimate that can be readily used for any of a number of posteriors from calibrations using different datasets, for example.

In addition, in order to obtain more precise calculations independent of the emulator uncertainty, we simulate 10 000 collisions with the full model for each of ten posterior samples from our default calibration. We do this for each of our two calibration systems, Au-Au and *d*-Au, as well as for *p*-Au and ³He-Au. The results in Sec. VI are all obtained using these ten samples, five of which we plot individually in the figures, as described in that Section. The values of the parameters for each of the five plotted samples are shown in Appendix K.

E. Closure tests

An essential check of any Bayesian comparison with data is to use, in place of the data, the model predictions for a given parameter set not used in training. The so-called “pseudo-data” may be assigned arbitrary “experimental” uncertainty, and varying the magnitude of this uncertainty can help gauge the sensitivity of constraints. If the uncertainty is set to be relatively small, the closure is limited by the emulation uncertainty. For the closure tests in this analysis, we assign the true experimental uncertainty for the corresponding observables in order to make the most direct analogy to the real calibration. We performed closure tests, shown in Appendix E, for a number of parameter sets from the simulated design points that were not used to train the emulator. The posterior distributions for the parameters are obtained with an identical procedure to that used for the default calibration against real data, but since the truth parameter values are a priori known, they may be compared with the posterior constraints to gauge the performance of the analysis. In Appendix E, we show a selection of marginalized posteriors for three closure points. We see generally consistent posterior credible intervals with the truth values of parameters across different closure points, offering confidence in the robustness of the analysis. We additionally see the sensitivity of particular parameters to the observables reflected in the closure test posteriors, with the rapidity loss parameters in particular yielding strongly peaked distributions.

IV. THE EXPERIMENTAL DATA

A critical component of the Bayesian analysis is curation of the experimental data. Our analysis used a wide range of measurements from RHIC’s BRAHMS, STAR, PHENIX,

TABLE II. Listing of all the experimental data used in this work for the default calibration. For each observable, we indicate the collision system, the experimental collaboration, the number of centrality bins used, and the total number of data points used to calculate the likelihood.

Observable	System	Experiment	Cent. bins	Data points
$dN_{\text{ch}}/d\eta(\eta)$	Au-Au	PHOBOS [78]	11	594
$v_2(\eta)$	Au-Au	PHOBOS [79]	3	48
$v_2(\eta)$	Au-Au	STAR [80]	1	21
$v_2(p_T)$	Au-Au	PHENIX [81]	6	30
$v_2(p_T)$	Au-Au	STAR [80]	7	56
$v_2(\text{cent.})$	Au-Au	STAR [80]	6	6
$v_3(\text{cent.})$	Au-Au	STAR [82]	6	6
$\pi^{+/-} \langle p_T \rangle (\text{cent.})$	Au-Au	STAR [83]	7	7
$K^{+/-} \langle p_T \rangle (\text{cent.})$	Au-Au	STAR [83]	7	7
$K^{+/-} \langle p_T \rangle (\text{cent.})$	Au-Au	PHENIX [84]	8	8
$p/\bar{p} \langle p_T \rangle (\text{cent.})$	Au-Au	PHENIX [84]	8	8
$dN_{\text{ch}}/d\eta(\eta)$	<i>d</i> -Au	PHENIX [85]	3	102
$v_2(\eta)$	<i>d</i> -Au	PHENIX [85]	1	14
$v_2(p_T)$	<i>d</i> -Au	PHENIX [38]	1	5
$v_2(p_T)$	<i>d</i> -Au	STAR [86]	1	5
Total				917

and PHOBOS Collaborations. For pragmatic reasons, such as limitations of our model or of our emulators, or due to limitations in experimental data reporting (see Appendix F), we did not include all curated data in our Bayesian analysis. We divided the measurements into a calibration set, listed in Table II, and a separate set used for additional comparisons, listed in Table III. We explain below how the classification into the two categories was made.

Another key component of the analysis is writing and validating analysis scripts to calculate each observable, matching the experimental definition. This includes reproducing as closely as possible each measurement’s centrality determination, kinematic cuts, and (in the case of particle correlation observables) the reference regions; we discuss below some challenges we faced in doing so. In general, (3 + 1)D simulations, as compared with boost-invariant (2 + 1)D simulations, allow us to match experimental definitions more precisely for nearly all measurements.

In a final section, we compare the magnitude of the emulator uncertainty to the experimental uncertainty, using a type of visualization that we believe would have great benefits to be adopted in the community.

A. Data selection

In selecting the data, we make conservative choices with regard to the ability of our model to describe the physics. That is, we limit the measurements up to 60% centrality in Au-Au (with the exception of the 20%–70% centrality STAR v_2 measurement) and up to 20% in *d*-Au. We also set an upper bound of 1.5 GeV for the p_T -dependent observables.

Given our assumption of vanishing net baryon density, we choose to focus on charged hadrons and do not in general include measurements of identified particle spectra or

TABLE III. Listing of the experimental data not used in calibration, but nevertheless computed and compared with in this analysis. For each observable, we indicate the collision system, the experimental Collaboration, the number of centrality bins used, and the total number of data points considered.

Observable	System	Experiment	Cent. bins	Data points
$dN_{ch}/d\eta(\eta)$	Au-Au	BRAHMS [87]	6	216
$v_3(p_T)$	Au-Au	PHENIX [81]	6	30
$v_4(p_T)$	Au-Au	PHENIX [81]	5	25
$p/\bar{p} \langle p_T \rangle (\text{cent.})$	Au-Au	STAR [83]	7	7
$\pi^{+/-} \langle p_T \rangle (\text{cent.})$	Au-Au	PHENIX [84]	8	8
$dE_T/d\eta(\text{cent.})$	Au-Au	PHENIX [88]	12	12
$dN_{ch}/d\eta(\eta)$	^3He -Au	PHENIX [85]	1	26
$v_2(\eta)$	^3He -Au	PHENIX [85]	1	24
$v_2(p_T)$	^3He -Au	STAR [86]	1	7
$v_2(p_T)$	^3He -Au	PHENIX [38]	1	8
$v_3(p_T)$	^3He -Au	STAR [86]	1	7
$v_3(p_T)$	^3He -Au	PHENIX [38]	1	8
$dN_{ch}/d\eta(\eta)$	d -Au	PHOBOS [89]	1	54
$dE_T/d\eta(\text{cent.})$	d -Au	PHENIX [88]	3	3
$v_3(p_T)$	d -Au	PHENIX [38]	1	5
$v_3(p_T)$	d -Au	STAR [86]	1	5
$r_2(p_T)$	d -Au	STAR [40]	1	5
$r_3(p_T)$	d -Au	STAR [40]	1	5
$dN_{ch}/d\eta(\eta)$	p -Au	PHENIX [85]	1	26
$v_2(\eta)$	p -Au	PHENIX [85]	1	24
$v_2(p_T)$	p -Au	STAR [86]	1	7
$v_2(p_T)$	p -Au	PHENIX [38]	1	8
$v_3(p_T)$	p -Au	STAR [86]	1	7
$v_3(p_T)$	p -Au	PHENIX [38]	1	8
Total				535

anisotropies. We make an exception to this rule to include the mean transverse momentum of identified particles from STAR and PHENIX [83,84], due to the lack of a corresponding charged hadron measurement and the known sensitivity of the transport coefficients to these observables. Specifically, we use the STAR $\pi^{+/-}$ and $K^{+/-}$ mean p_T , and the PHENIX $K^{+/-}$ and $p/\bar{p}$ mean p_T .⁶

For these mean- p_T measurements, we averaged the data values and the uncertainties between particles and antiparticles. Specifically, since the systematic uncertainty dominates in these measurements from both STAR and PHENIX, we assume the errors to be completely correlated when combining them.

On the measurements broadly considered within the model's scope, we made a further selection about whether to use them for calibration or to simply compare them with the posterior. The factors going into the selection were specific to

each measurement, although we did make a general assessment of emulator accuracy across all observables.

Through careful emulator validation, we determined that our emulator's accuracy varied significantly across observables. Using a set of validation points, we performed emulator validation by comparing the emulator's prediction of its uncertainty to its actual performance in describing different observables. We used the following guiding principle: when the emulator significantly underestimated the true error in predicting an observable, the emulator was deemed insufficiently accurate to describe that observable for the purposes of parameter estimation, and the observable was excluded from the calibration. For more details on this procedure, see Appendix B.

As a result of this emulator validation, we had to exclude many higher-order harmonic measurements, including differential v_3 measurements in both Au-Au and d -Au, as well as differential v_4 (and potentially higher orders) in Au-Au. We also did not include in calibration the transverse energy, $dE_T/d\eta(\text{cent.})$, in either Au-Au or d -Au. This was purely due to a technical reason: our calculation—configured at run-time across large-scale simulations—included only the transverse energy of charged hadrons, while the PHENIX [88] definition included the total hadronic transverse energy. The difference between the two definitions is expected to be around 3/2, given that pions tend to dominate. We choose not to include the observable in the calibration, but we do scale our calculations by the 3/2 correction factor to compare our calibrated analysis with the PHENIX measurements.

Another observable that we compared with, but did not calibrate on, is STAR's de-correlation coefficients r_n [40], since the data are preliminary.

The exclusion of ^3He -Au and p -Au measurements from calibration was due uniquely to a limited computational budget (and, thus, ultimately, emulation accuracy). Since each calibration system's design points had to be simulated separately, we opted to maximize the number of design points for a single large and a single small system, given finite computational resources. For this reason, all model predictions for ^3He -Au and p -Au observables in this paper are done using full model simulations and do not rely explicitly on model emulation.

Despite observables being excluded from calibration, we do compare with them on multiple occasions, sometimes using emulation,⁷ and often using direct calculations from the model.

B. Matching calculations to data

For all the measurements included in this analysis, the centrality determination, kinematic cuts, and other specifications of the experimental measurements were matched when computed. We give additional detail below regarding the centrality

⁶The STAR $p/\bar{p}$ and the PHENIX $\pi^{+/-}$ measurements did not remove the weak decay contribution, while in our training simulations we do not include this contribution. Although the effect is estimated to be within a few percent, we err on the side of accuracy by not including these two measurements in calibration.

⁷Even observables for which we did not deem the emulator precise enough for calibration are still, in most cases, sufficiently well-emulated to offer some qualitative insight in comparisons of the posterior with data.

selection and the calculation of momentum anisotropies, and explain the approach we used when insufficient information was available from the experimental publications.

a. Centrality determination. As pointed out throughout the paper, the precise determination of centrality in accordance with each measurement's procedure is one of the key capabilities of a 3D analysis. For every measurement we compare with, the rapidity window matched that used in experiment. Additionally, the relevant quantity used for sorting the events measured in that region—whether the multiplicity or the transverse energy—correspondingly matched experiment as well.

The transverse momentum (or energy) cut for centrality determination is a detail not always stated in the papers we considered. When mentioned, it is often the case that no cut was implemented. Consequently, as a generic choice across all observables we calculate, we do not include any momentum or energy cut. We verified (not shown) that the effect from reasonable cuts on transverse momentum or energy on the multiplicity- or E_T -based sorting of events, and ultimately on the observable, is generally small ($\approx 0.1\%$) across observables.

b. v_n observables. For all calculations of v_n in this analysis, we use the scalar product method.⁸ The rapidity cuts and resulting gaps follow the experimental choices regardless of method. For all measurements we considered except one, the scalar product method is expected to be an accurate match to the measurement. The one exception is the Au-Au PHOBOS v_2 [79] measured with the event-plane method. The discrepancy cannot necessarily be evaluated precisely, although, informed by previous estimates [90], we attempt to account for it by increasing in the analysis the experimental uncertainty for the PHOBOS Au-Au $v_2(\eta)$ by 10% of the mean value of the observable. For small systems with low multiplicity, the effect is expected to be small [90], and we do not increase the uncertainty.

We encountered other challenges with the STAR v_n measurements from Ref. [80], from which we used multiple datasets. The analysis presented a large number of related azimuthal anisotropy measurements, cross-checks, and systematic studies performed using a variety of methods and kinematic cuts, with in-text measurement details not always unambiguously identified to a particular plot. For the $v_2(\eta)$ measurement, one source of confusion concerned the regions of interest and reference, as it was not entirely clear whether the points at midrapidity and those at forward or backward rapidity, which were measured using different detectors, shared a reference region. For our calculation, we assumed that they did share a reference. The largest potential source of uncertainty, however, concerned the averaging over the large centrality bin of 20%–70%. Indeed, the extension to peripheral collisions up to 70% centrality was a singular exception in our analysis. The ambiguity was in large part related to weights, or lack thereof, applied to smaller centrality

bins in constructing the observable in the large bin. We were unable to obtain a conclusive prescription. Moreover, when comparing the midrapidity bins of this measurement to the integrated midrapidity v_2 measurement presented in that same paper [80], we find an apparent inconsistency between them, regardless of the weighting scheme used to average smaller centrality bins. As a result, we use this dataset cautiously, and we increase in the analysis the experimental uncertainties by 12% of the mean observable value, just enough to allow for apparent consistency with the integrated dataset. (Scaling the midrapidity values of the rapidity-dependent measurement by a factor of 1.12 matches them to the centrality average of the integrated, midrapidity measurement.) We decided in favor of its inclusion in the calibration largely on the basis of the general scarcity of $v_2(\eta)$ RHIC data. We discuss the inter-relationship between these datasets and the rest of the measurements in the analysis in greater detail in Sec. V A 2.

The measurement of $v_2(p_T)$ in Au-Au from STAR [80] presented a unique challenge for the analysis: the lowest p_T bin (0.15 to 0.3 GeV/ c) was significantly outside of our observable prior. All combinations of model parameters that we investigated (our prior) predicted a v_2 smaller than the STAR measurement, especially for peripheral events, where the prior is multiple σ away from the measurement. Given that this is the lowest bin of the measurement, we believe the tension may arise from missing physics in our model or some issue with the measurement. As a result of these considerations, this bin was ultimately not included in calibration. It is still shown in all the observable posterior comparison plots in the paper and, as expected, rather poorly described by the constrained model.

C. Emulator and experimental uncertainties

Whenever available and reported separately by experiments, the statistical and systematic uncertainties were added in quadrature. Whenever this breakdown was not reported in experimental papers, on the HEPData repository, or on individual experiment websites, we made reasonable use of what was available.

In Fig. 4, we show a 2D comparison plot of the emulation and experimental variance, respectively, scaled by the mean prediction and measurement, for all observable bins ultimately included in the analysis (Table II). The multiplicities, in particular in Au-Au, and their corresponding uncertainties are in log-space and tend to populate the bottom-left corner as a result. We mark in color and label three example observables that feature in Figs. 25 and 26 in the Appendix D discussion on the covariance matrix and correlations between observables.

We see that the uncertainties on most points are rather balanced between the emulator uncertainty and the experimental uncertainty. The location of points on the 2D plane relative to the dashed diagonal line indicates the extent to which the information content in experimental data is exploited in the analysis: datasets far above the diagonal have much larger emulation uncertainty than experimental uncertainty, and thus the analysis is not using efficiently the information they contain. Since the emulator uncertainty contains the statistical uncertainty, one cannot differentiate in Fig. 4 if the uncertainty

⁸See Appendix M for a definition of the scalar product method used in the analysis, and for a discussion of potential prediction discrepancies with slightly different definitions.

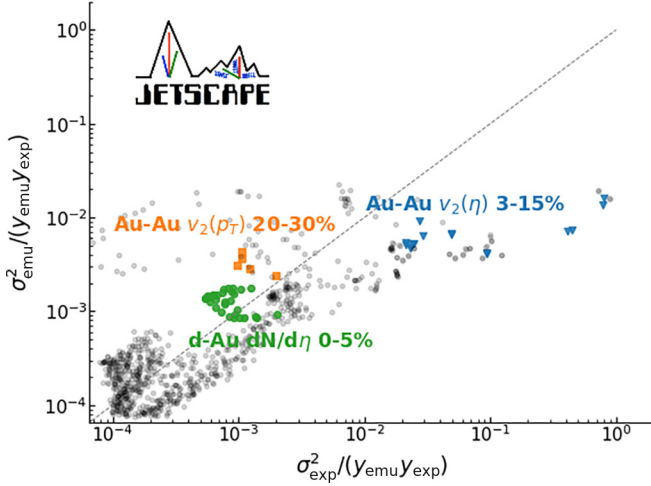


FIG. 4. Comparison of experimental and emulation uncertainties for all observables in the calibration. Each point represents one observable bin, with coordinates given by the scaled experimental variance (horizontal axis) and scaled emulation variance (vertical axis). The variances (diagonal elements of the covariance matrices) are scaled by the product of the mean experimental measurement y_{exp} and the mean emulator prediction y_{emu} . The emulation uncertainties combine interpolation and statistical uncertainties, evaluated at the maximum *a posteriori* (MAP) parameter point. Points near the diagonal indicate balanced contributions from experimental and emulation uncertainties, while points above the diagonal are dominated by a combination of interpolation and statistical uncertainty from the emulator. The three observables shown in Figs. 25 and 26 are indicated in color and labeled.

is large because of statistical uncertainty (a larger number of events required to reduce uncertainty), or if the emulator is struggling to interpolate the observable in the parameter space. We point out that this situation is not unique to the present Bayesian analysis, but it has generally not been illustrated like this in the past.

V. BAYESIAN POSTERIORS

A. The “default” multisystem calibration

The default calibration includes all the selected experimental data within the physics scope of the analysis, which the model is expected to reasonably describe. We include *d*-Au data in this calibration under the hydrodynamic assumption for small systems.⁹ A detailed overview of the data and the selection criteria is provided in the previous section (Sec. IV). We show in Fig. 5 the observable posterior, indicating with solid colored bands the 90% credible interval computed with the emulator using one hundred thousand samples of the posterior. The bands are plotted in green for data included in calibration; for data not included in the calibration, olive bands

(lighter color) are used. The experimental data and associated total uncertainties are plotted as black markers. In the same figure, in the lower section of each panel, we show the theory-to-data ratios (that is, posterior-band to experimental points), which more readily allow a visual comparison of the fit across observables. The uncertainties on the ratio plot still indicate the width of the posterior, correspondingly scaled by the data.

We see a good fit to data across experiments, systems, and observables, showing the descriptive range of the model. In particular, we see largely consistent fits to the multiplicity measurements both in Au-Au and *d*-Au and extending to large pseudorapidity. Notably, the width of the posterior is larger around midrapidity compared with finite rapidity, reflecting the corresponding difference in the emulator uncertainty in those two regions. The mean transverse momentum shows a good fit to the data for protons, kaons, and pions, although slightly overestimated for the protons. The fits to the v_2 and v_3 , integrated as well as p_T - and η -differential, are overall good, although they put the model under some tension. Specifically, the simultaneous p_T and centrality dependence of the v_2 in Au-Au presents a challenge for the model. In addition, the rapidity dependence of the STAR v_2 is poorly reproduced, even though the analogous PHOBOS measurement is well-described. We explore the origins of some of this tension in more detail in Sec. V A 2.

Remarkably, the fit to a host of measurements not included in calibration is quite good: specifically, the $v_3(p_T)$ and $v_4(p_T)$ in Au-Au from PHENIX, even though no other p_T -differential measurements of v_3 and no measurements of v_4 at all are included in calibration. In addition, the transverse energy $dE_T/d\eta$ as a function of centrality is well described for both Au-Au and *d*-Au. Lastly, even though the posterior is relatively wide, the PHENIX and STAR data for $v_3(p_T)$ in *d*-Au are close to the model prediction. We come back to this small system data in Sec. VIB, where we place it in the context of other small-system measurements.

The posteriors of the viscosity parameters are shown as a function of temperature in Fig. 6 (in green), where we also show the posteriors for the midrapidity-only calibration (in orange). The midrapidity-only calibration features a proper subset of the default dataset, including all integrated midrapidity observables, as well as the midrapidity bins (defined in general as $|\eta| < 1$, with a few exceptions extending to $|\eta| = 2$ for Au-Au measurements) of observables measured as a function of η . In the default calibration, we find a preference for finite shear and bulk viscosity, particularly at low temperature. The inclusion of the forward and backward regions, which correspond to a relatively larger portion of the lifetime spent at lower temperatures, produces a stronger effect in that region. The preference for both finite shear and finite bulk viscosity is consistent with findings in the 2D JETSCAPE calibration, which combined RHIC and LHC data [10]. However, the RHIC-only results of that calibration showed much weaker support for large viscosity than the combined result, suggesting a selection of midrapidity RHIC data does not show a significant preference for a viscous hydrodynamics in at least one other model. We expand on these conclusions and the effect of rapidity-dependent data in the next section.

⁹If the present calibration is used for future studies of heavy-ion physics, the Au-Au-only tune, also analyzed here, may be preferred depending on the scope and assumptions of the studies.

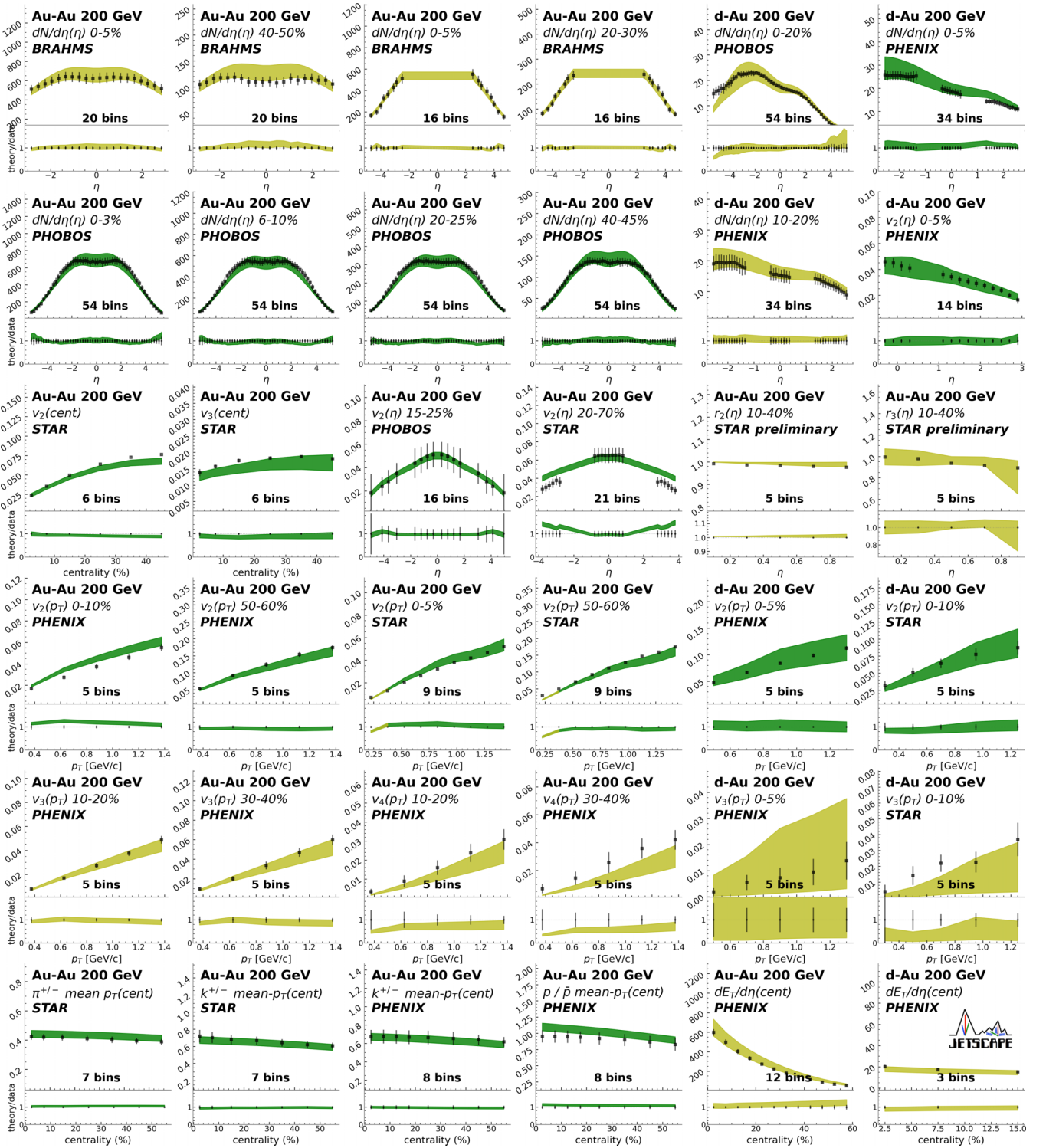


FIG. 5. The observable posterior for the data included in the calibration, plotted in green, as well as predictions for some not included, plotted in olive. The solid markers are the experimental data, while the bands show a 90% interval of the posterior calculated using the emulator. Each plot is labeled with the number of data bins corresponding to each observable in a given plot.

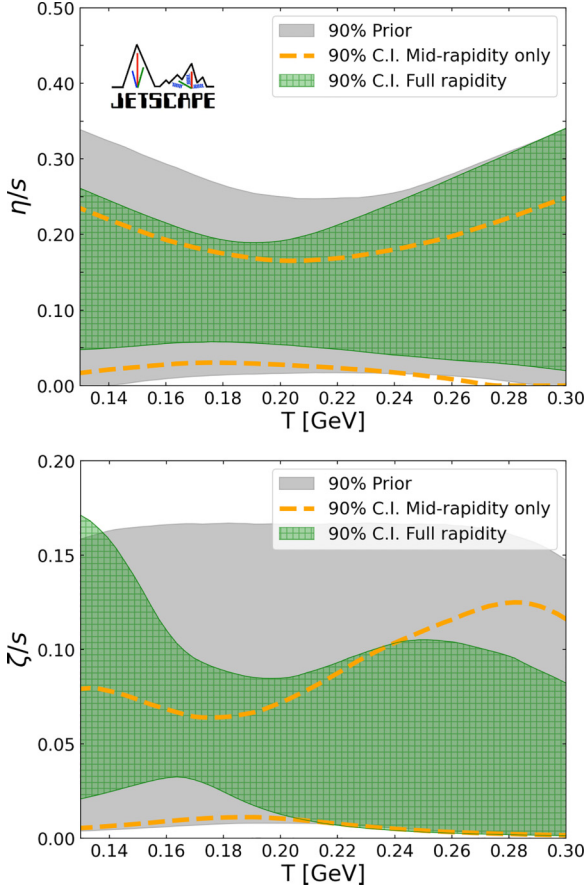


FIG. 6. The posterior for the specific shear (top) and bulk (bottom) viscosity as a function of temperature for the default calibration (in green) and the calibration using only the midrapidity data (in orange).

The parameter posteriors from the calibration are shown in a corner plot in Fig. 7 for nine of the initial state and hydrodynamization parameters.¹⁰ Across the diagonal, we show the marginalized distributions for each of the parameters, while in the off-diagonal region, we show the 2D-correlations between parameters. In each plot, we indicate with a dashed red line the median values of each marginalized one-dimensional (1D) distribution, for which we also plot in dashed green lines the 90% bounds of the posterior and, above, the numerical values of the median and those bounds. We note that most of these parameters show significant constraints relative to the flat priors, with some especially so: notably, the rapidity loss and shadowing parameters. All posteriors from the default calibration—which again includes all of the selected data considered here from both Au-Au and *d*-Au at the RHIC top energy listed in Table II—are consistent with those found in a Beam Energy Scan calibration of this model on Au-Au data (including at 200 GeV collision energy) [21] for all comparable parameters between the two studies.

¹⁰The corner plot featuring all 20 varied parameters is included in Appendix G.

1. The constraints from forward and backward rapidity

A key finding from this analysis is that forcing our 3D model to fit, simultaneously, the experimental data at all rapidities yields different constraints on model parameters than if comparing only to midrapidity. We have already seen this in the viscosity posteriors in Fig. 6. We might consider that we could readily glean conclusions from looking at the comparison between the data at all rapidities with predictions from the model constrained on only the midrapidity data. This is shown in Fig. 8, again computed with the emulator using one hundred thousand samples of the posterior.¹¹ If introducing finite rapidity measurements introduces a tension in the model, we might expect that the calibration using both midrapidity and finite rapidity data would describe the midrapidity points more poorly. We do not see clear visual evidence of that. That said, if examining more closely, we can still note differences. In particular, the width of the posterior, both at midrapidity and at finite rapidity, is significantly reduced when using the full dataset. This is especially noticeable for the multiplicity in *d*-Au and the $v_2(\eta)$ in Au-Au. These differences, which we might otherwise miss in “chi-by-eye” fits, are exactly those the Bayesian calibration allows us to extract quantitative information from, in the form of the parameter posteriors.

As with the viscosity, we see quantitative confirmation when we examine the individual parameter posteriors for a selection of the initial-state parameters. In Fig. 9, we show again the marginalized posteriors and 2D correlation plots of the default calibration from Fig. 7 in green, but additionally overlay in orange the corresponding posterior information for the midrapidity-only calibration. While the posteriors are always consistent between the two calibrations, we see significant differences for some of the parameters, most notably for α_{rem} , the energy loss of the collision remnants, and $\alpha_{\text{shadowing}}$, the nuclear shadowing factor. The posteriors for the rapidity loss (y_4 and y_6) are similar between the two calibrations (though narrower for the default calibration), likely due to those parameters determining the energy density profile near midrapidity. On the other hand, α_{rem} contributes more directly to energy deposition (and thus ultimately multiplicity) away from midrapidity and is constrained only when including the forward and backward data. Similarly, we see a broad distribution of the $\alpha_{\text{shadowing}}$ parameter when using only midrapidity data. In the midrapidity calibration, we can also see a clear positive correlation between this parameter and the y_6 , due to their opposing effects on the midrapidity multiplicity, which are shown in sensitivity plots in Appendix H. The correlation represents a degeneracy in the model that is only resolved with the inclusion of forward and backward rapidity data. We note that for the default calibration

¹¹While we exclude in this “subcalibration” any data points explicitly measured at forward or backward rapidities, as discussed in previous sections, a number of the rapidity-integrated measurements implicitly contain forward or backward rapidity information, which we also incorporate in our calculations. In other words, our midrapidity calibration is not “purely” a midrapidity one in the sense of previous 2D calibrations, which lacked access to this information.

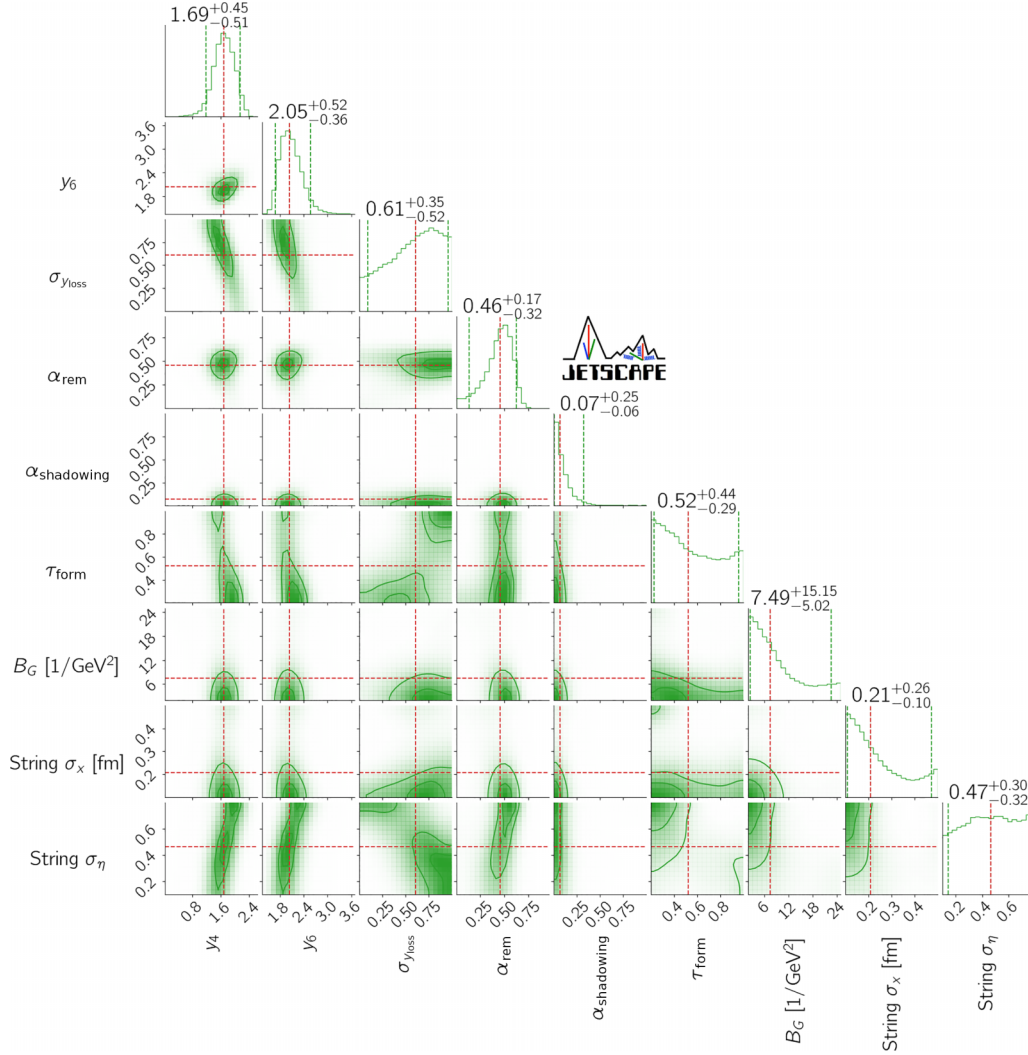


FIG. 7. The parameter posteriors for the default calibration for the nine parameters related to the initial state and hydrodynamization. The diagonal plots show the one-dimensional marginalized posteriors for each of the parameters. The off-diagonal plots show the two-dimensional correlations between pairs of parameters. The dashed red line indicates the median values of each marginalized 1D distribution, and the dashed green lines indicate the corresponding 90% bounds.

this degeneracy is no longer prominent: the 2D y_6 - $\alpha_{\text{shadowing}}$ posterior is localized with both parameters relatively well-constrained.

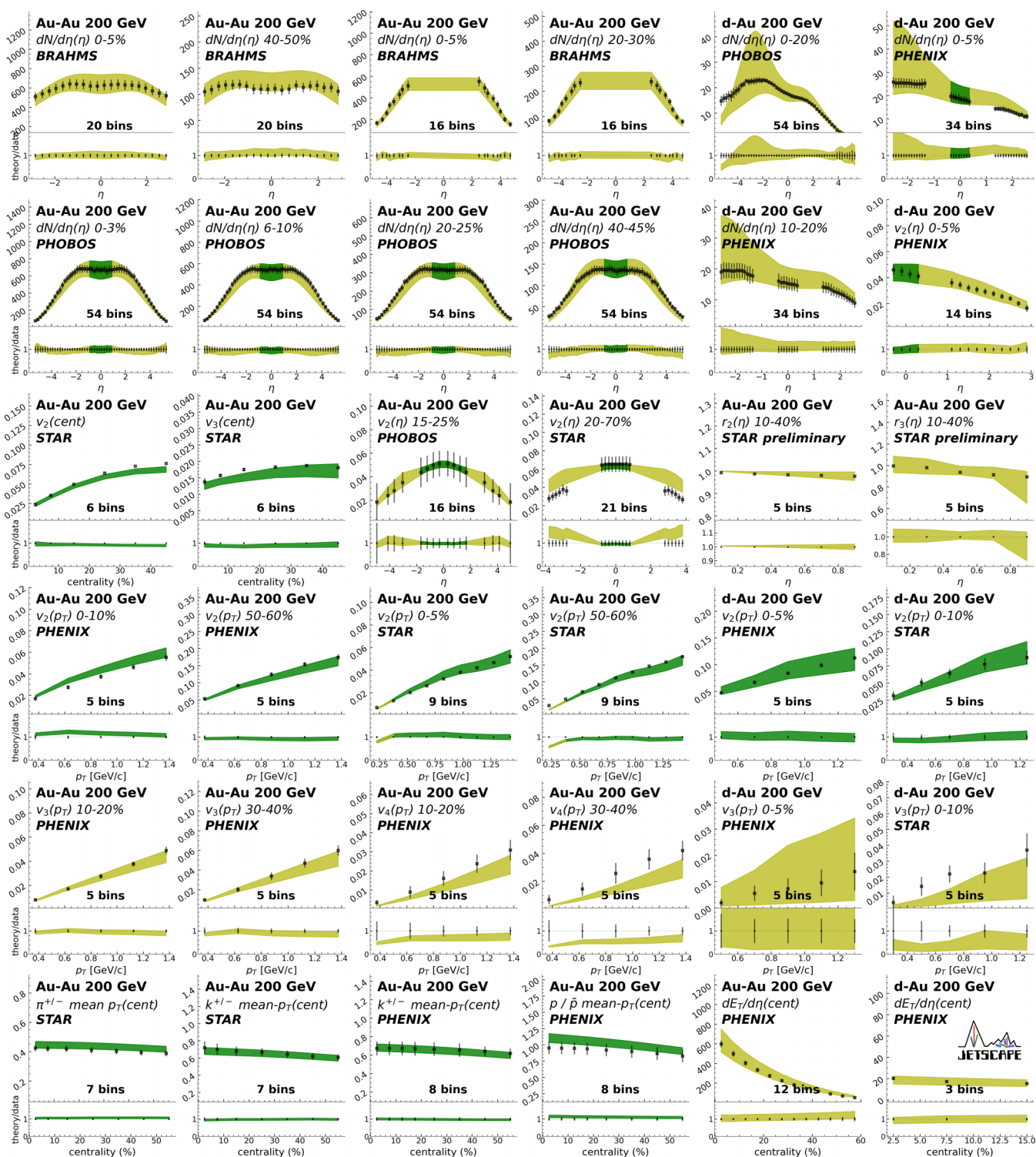
2. Observable tension

It is clear from the comparison between data and the observable posterior in Fig. 5 that the default dataset places our model under tension. In particular, the tension may be readily seen in that figure in the plots of the STAR $v_2(\eta)$ in Au-Au, and some of the lower- p_T bins of the $v_2(p_T)$ in Au-Au for both STAR and PHENIX measurements. For all remaining data bins used in calibration, some part of the 90% band of the posterior is always in close proximity (and in most cases on top) of the measurements.

In Figs. 10 and 11, we examine the origin of some of this tension by calibrating on various subsets of observables. For each of the cases mentioned above, we aim in this section to

identify which additional observables in our default dataset are chiefly responsible for forcing the model into a preferred region of parameter space at the expense of accurately describing said dataset.

First, we look at the STAR measurement of $v_2(\eta)$ in Au-Au. As we noted in Sec. IV, this dataset presents several challenges in matching precisely to the experimental procedure, particularly from the aggregation into a large centrality bin (20%–70%). Nevertheless, as noted in that section, we carefully attempted to account for any methodological uncertainty related to it, and therefore expect our model to fit the data within uncertainties. We can see in the top plot of Fig. 10 the 90% interval of the posterior for this observable compared with the data for three minimally restrictive calibrations. In green, we show the posterior if using only the STAR $v_2(\eta)$ measurement itself in constraining the model and note that we can describe all the points with accuracy over the full range in pseudorapidity. In subsequent bands, we show the posterior



when adding (indicated by a “+”) various other observables to the calibration in addition to the one plotted. The description remains accurate, then, if we pair that measurement with the

integrated v_2 as a function of centrality, also measured by STAR. That posterior (shown in red) prefers a larger value at midrapidity (where the integrated measurement is taken),

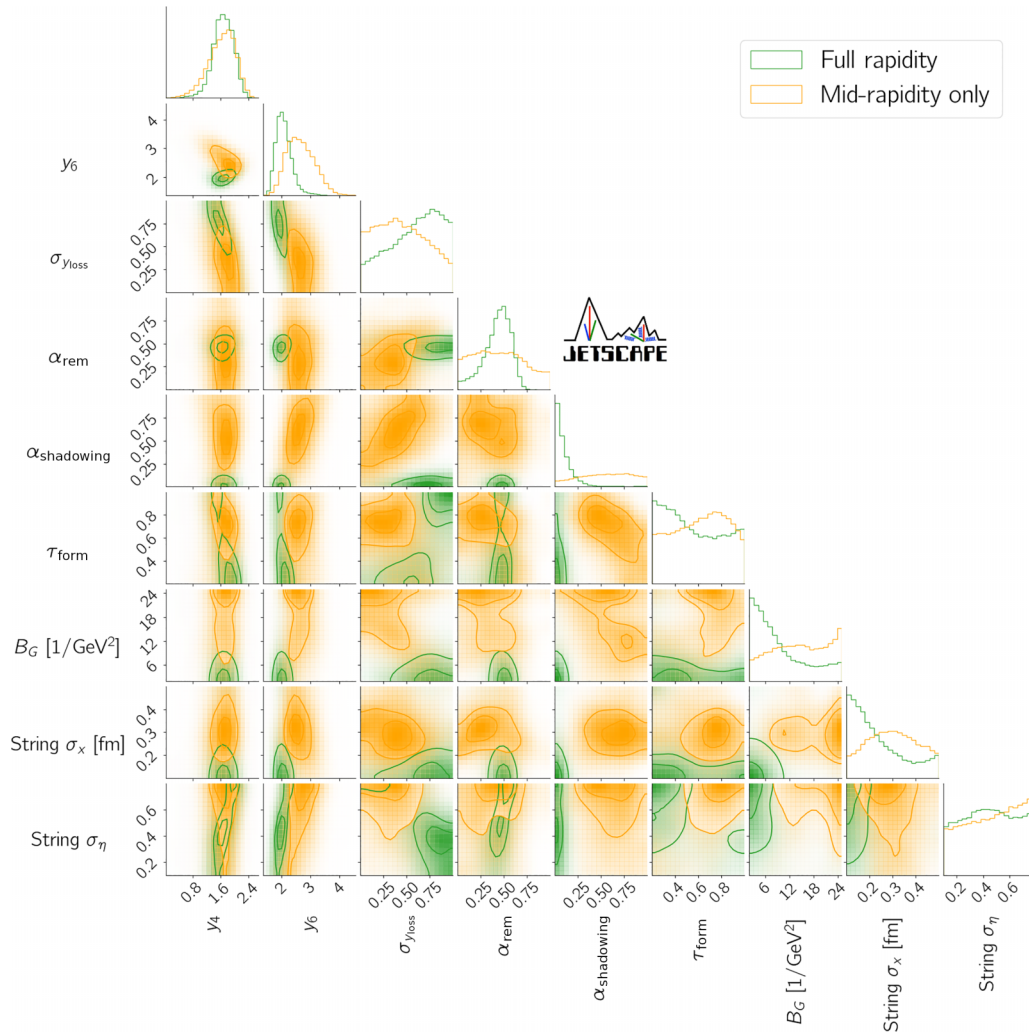


FIG. 9. The parameter posteriors for the default (green) and the midrapidity-only (orange) calibrations for the nine parameters related to the initial state and hydrodynamization. The data included in the midrapidity-only calibration are the subset of the default calibration including the midrapidity points of the rapidity-dependent measurements, and all rapidity-integrated observables. The diagonal plots show the one-dimensional marginalized posteriors for each of the parameters. The off-diagonal plots show the two-dimensional correlations between pairs of parameters.

while a calibration using the STAR $v_2(\eta)$ and the corresponding $v_2(\eta)$ from PHOBOS (shown in purple), yields only mild differences. We conclude that the model is not placed under significant tension from the $v_2(\eta)$ measurements alone, or from the integrated v_2 .

However, once we include multiplicity measurements extending to large pseudorapidity and from multiple centralities, as shown in the bottom plot of that figure, the constraints tighten and the description of the STAR $v_2(\eta)$ worsens, particularly at large $|\eta|$. Once the multiplicity is included, the description at large $|\eta|$ tends to remain poor, regardless of additional calibration observables, as shown in the last two bands. Adding the integrated v_2 or the $v_2(p_T)$ from STAR will modulate the midrapidity value but will not improve the forward and backward regions. Additional centrality bins of these observables not included in these minimal calibrations reinforce these effects. There appears, then, a tension between the model and this ensemble of

data, highlighted by the difficulty in describing the correlation between v_2 and $dN/d\eta$ across pseudorapidity for all v_2 measurements. While the posteriors for the corresponding d -Au observables do not clearly exhibit this same tension, higher precision in measurements or emulation may bring it to the fore.

Another observable we highlight is the central bin (0%–10%) of the PHENIX $v_2(p_T)$ measurement in Au-Au. This measurement is overestimated for a few of the p_T bins by the default posterior. In Fig. 11, we show that the model tension here is specifically between the centrality bins of $v_2(p_T)$. First, in the top plot of that figure, we show in green the posterior when calibrating on only this central bin. In the same plot, we see that adding the most peripheral bin (the 50%–60%) in that set of measurements to the calibration immediately shows a strong effect on the 0%–10% bin, leaving most of the posterior to overestimate the experimental uncertainty markers for the intermediate p_T bins.

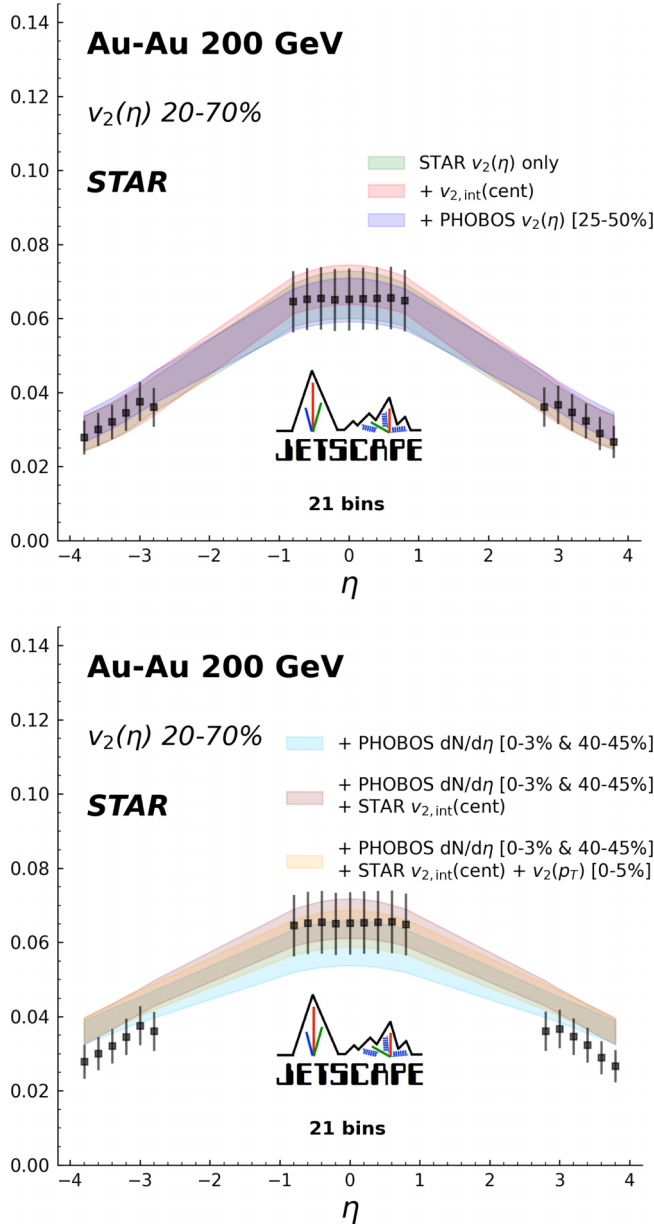


FIG. 10. The comparison of the STAR $v_2(\eta)$ posterior with the measurement, showing calibrations with various subsets of the data and highlighting the model tension between certain observables and the fit to the STAR measurement. The top plot shows calibrations using subsets of v_2 data only. The bottom plot shows calibrations that include subsets of multiplicity data. All data in these calibrations are from Au-Au measurements.

Then, in the bottom plot of that figure, we show in blue the joint calibration with the most central bins of the $v_2(p_T)$ measured by STAR and PHENIX. While the measurements differ in their kinematics, we show that the model has no trouble describing the PHENIX measurement when constrained also on the STAR data. The crucial effect comes again from the peripheral bins. In the bottom plot, we show the effect of the STAR peripheral bin of that set of measurements being included in calibration. In much the same way as in the top plot, this calibration shows a strong preference for larger values of

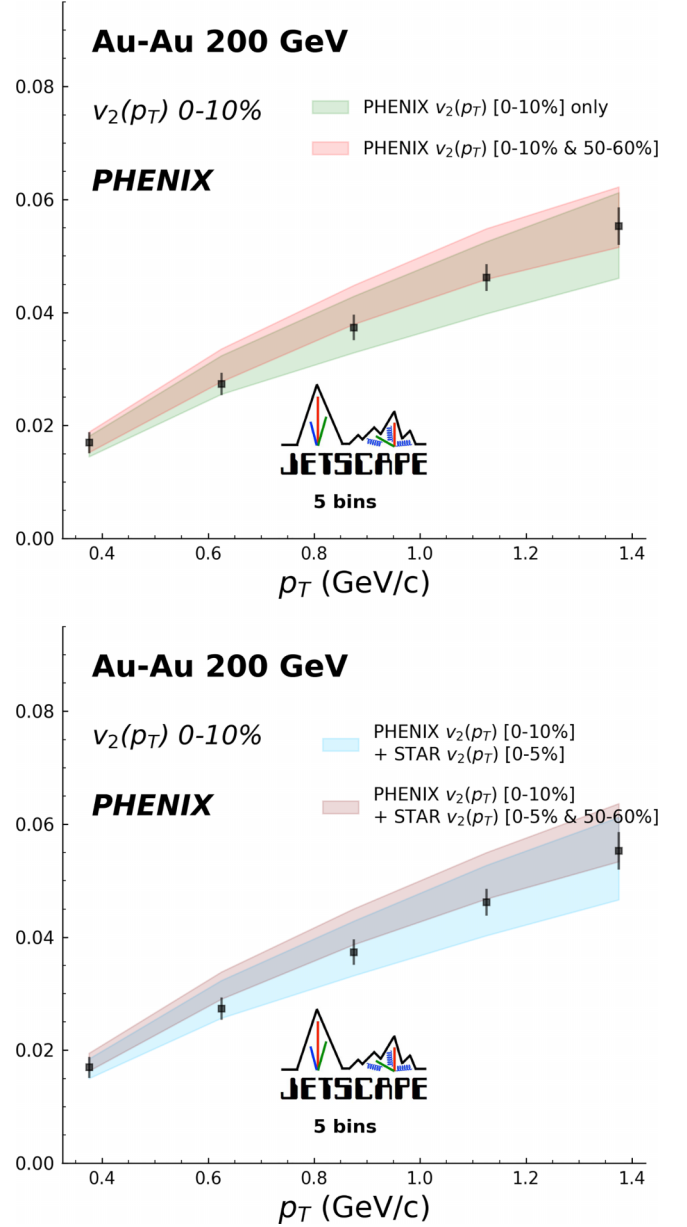


FIG. 11. The comparison of the PHENIX $v_2(p_T)$ posterior for central (0%–10%) Au-Au collisions with the measurement, showing calibrations with various subsets of the data and highlighting the model tension between different bins of v_2 data and the fit to the PHENIX measurement. The top plot shows calibrations using subsets of PHENIX v_2 data only. The bottom plot shows calibrations that include subsets of STAR data. All data in these calibrations are from Au-Au measurements.

the PHENIX observable, indicating once more that the model is placed under significant tension by the simultaneous p_T and centrality dependence of v_2 in Au-Au. This behavior may be a consequence of the centrality dependence of the initial-state eccentricity of our initial condition model (see, e.g., Fig. 2 of Ref. [44] for a discussion of this effect in 2D initial condition models).

What we have shown in this section is the tendency of the model to struggle to describe certain groups of observables

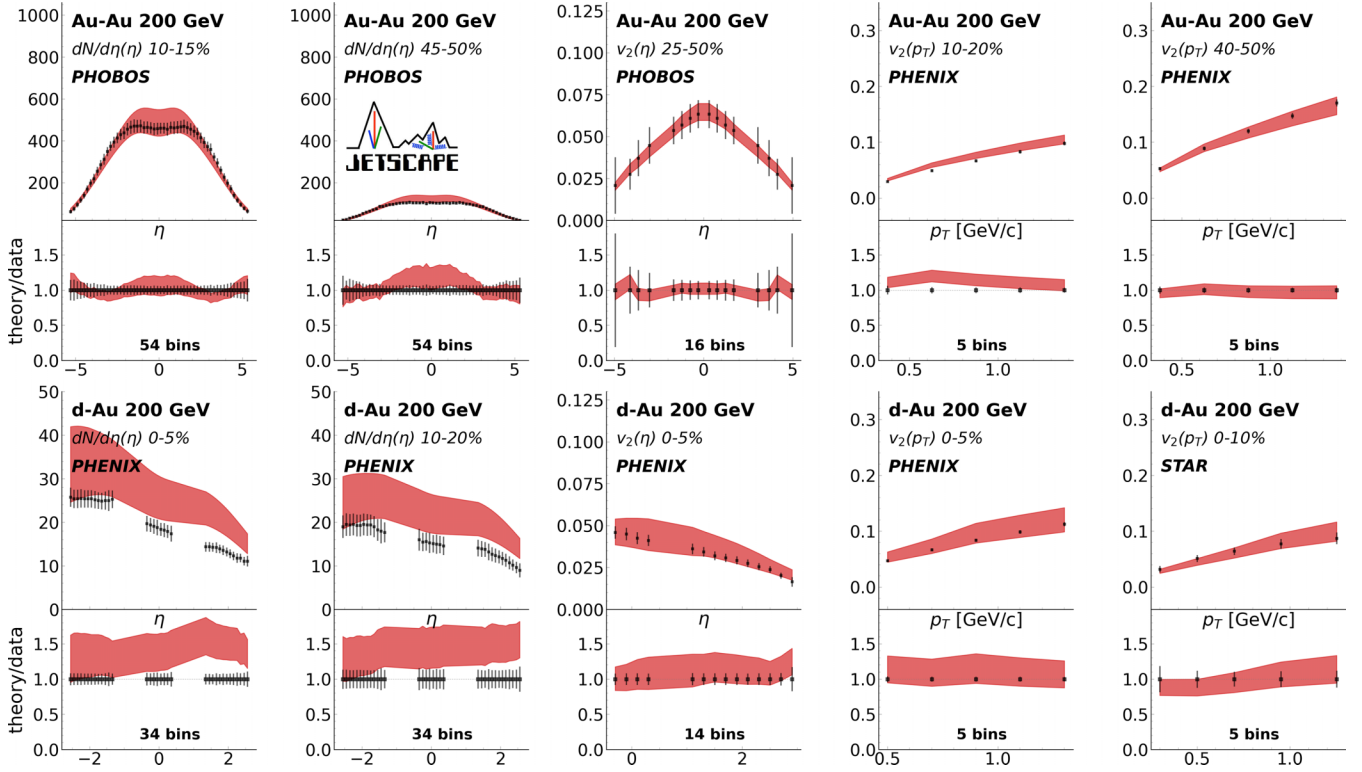


FIG. 12. The model posterior for select observables constrained using the Au-Au subset of the default observable set. The band represents the 90% credible interval of the posterior, and the bottom panel in the figures shows the model posterior scaled by the values of experimental data in each bin.

simultaneously. We have demonstrated how the tension apparent in the many-observable default calibration can be roughly pinned down to a few observables by means of successive calibrations with subsets of the data.

B. Comparison of independent Au-Au and *d*-Au calibrations

In this section we show how our model performs when calibrated with data from systems of different size and how the parameter posteriors change. We answer two overarching, related questions. First, how well can the model, calibrated on data from only a small (large) system, describe various measurements from the corresponding larger (smaller) system? Second, which parameters show similar constraints, and which show the most difference? If we assume the same model—in particular, a hydrodynamic QGP—describes physics for both systems (and that the differences among the two in the initial state, size of the system, etc., are well-described), constraining data from one system should yield a consistent (though not necessarily precise) description of data from the other. It could, of course, be the case that the differences between the systems outside of the hydrodynamic description are not fully captured by the model. The conclusions from this comparison, and indeed throughout the paper, are always drawn with that fact in mind.

We show in Figs. 12 and 13 (using one hundred thousand samples of the respective posteriors and the trained emulator), illustrative comparisons to a select few measured observables

for each of the independent system calibrations (we show a more complete picture of the observable posterior for each calibration in Appendix I). We can see from the top row of Fig. 12 that a model constrained on only the Au-Au data shows a level of consistency with that system's data (top row) similar to the default calibration. The predictions it makes for *d*-Au observables, however, are in some cases off the mark, as shown in the bottom row of that figure. While the *d*-Au v_2 measurements, both as a function of η and p_T , and for both PHENIX and STAR, are consistent with predictions, the multiplicity, particularly on the deuteron-going side, is significantly overestimated by the 90% posterior band. This result highlights the necessity of calibrating the model on small system data in order to provide a consistent description of observables in such systems.

Meanwhile, in Fig. 13, we show the posterior for select observables for a model constrained on only the *d*-Au data. Here, on the top row, we see that its predictions of Au-Au observables are, remarkably, generally consistent with measurements within uncertainties, including for the multiplicity. Nevertheless, they are significantly different compared with the predictions of the default or Au-Au-only calibrations in that the posteriors are much wider for most observables. For example, much of the posterior band includes midrapidity multiplicity predictions underestimating the data. In both the default and the Au-Au calibrations, such multiplicity values are nearly absent from the narrower posterior. Looking at the *d*-Au observables in the bottom row of that figure, we see that predictions for all observables are now consistent with

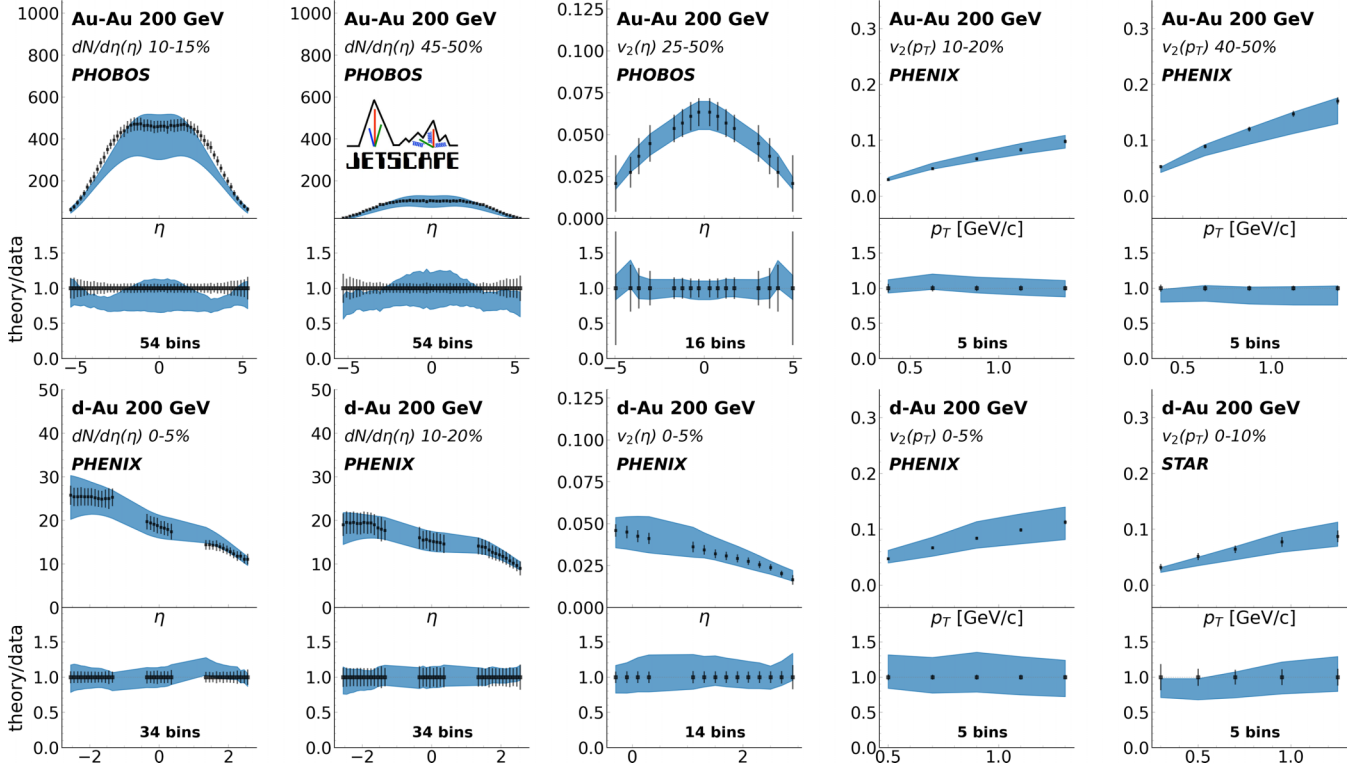


FIG. 13. The model posterior for select observables constrained using the d -Au subset of the default observable set. The band represents the 90% credible interval of the posterior, and the bottom panel in the figures shows the model posterior scaled by the values of experimental data in each bin.

the data. However, we note that the posterior bands for the v_2 observables are not significantly narrower compared with the Au-Au-only calibration (though we can still discern some slight differences). That is, whether we calibrate on Au-Au or on d -Au data exclusively, we obtain a similar description of the d -Au v_2 , despite the significant difference in the corresponding multiplicity descriptions. While this result gives credence to the assumption of describing both systems hydrodynamically, we also note that the relatively large emulation uncertainty for d -Au observables, and in particular the d -Au v_n , does limit the information that can be extracted from the data.

The posteriors of twelve of the model parameters for each of the calibrations discussed in this section are shown in Fig. 33 in Appendix J. We summarize here that the individual system posteriors are generally consistent with each other and with the default posterior for all parameters. For some parameters, like the y_6 and the $\alpha_{\text{shadowing}}$, the particular system constraints reinforce each other, leading to narrower posteriors for the combined default calibration. Notably, the posterior for τ_{form} is narrower for the Au-Au-only calibration than for the combined default, a case where the particular system constraints appear to be in tension.

In Fig. 14, we highlight the posteriors of the viscosity parameters as a function of temperature for the individual system calibrations, as well as the default result. The Au-Au-only calibration shows a preference for both finite shear and finite bulk viscosity. This is consistent with the preference for finite shear and bulk viscosity shown in the RHIC calibration of the

2D JETSCAPE analysis, which used only Au-Au STAR data [10]. Of course, in our Au-Au calibration, we include the finite rapidity data, which we showed in Sec. V A 1 to be largely responsible for this preference. The d -Au-only posterior for the shear viscosity is similar to the Au-Au, although it (and ultimately the combined posterior) prefers a slightly lower shear viscosity at low temperature. The bigger difference is in the bulk viscosity posteriors, where the two calibrations are only partly overlapping from about 190 MeV and below. The d -Au posterior is narrower than the Au-Au in this region; it prefers a low value and does not exclude a zero bulk viscosity. Meanwhile, the Au-Au posterior prefers larger values. The combined calibration posterior ultimately covers both tails and is narrowest around the point of minimal overlap of the individual system posteriors. As in previous studies of bulk viscosity with sufficiently flexible ansatz [10,11,19,20], constraints here tend to be stronger in the deconfinement temperature range predicted by Lattice QCD. The tightness of constraints is enhanced in the combined calibration due to some degree of tension between the individual d -Au and Au-Au results. The differences between these individual system posteriors may be due to a number of contributing factors. For one, the Au-Au calibration is constrained on mean p_T data of identified hadrons, including protons, which show significant sensitivity to the bulk viscosity, and in particular a preference for larger values, as highlighted in the plots in Appendix H. In addition, the Au-Au calibration has the benefit of constraining data from a wide range of centralities, and thus temperature profile evolutions, as well as, in general, lower

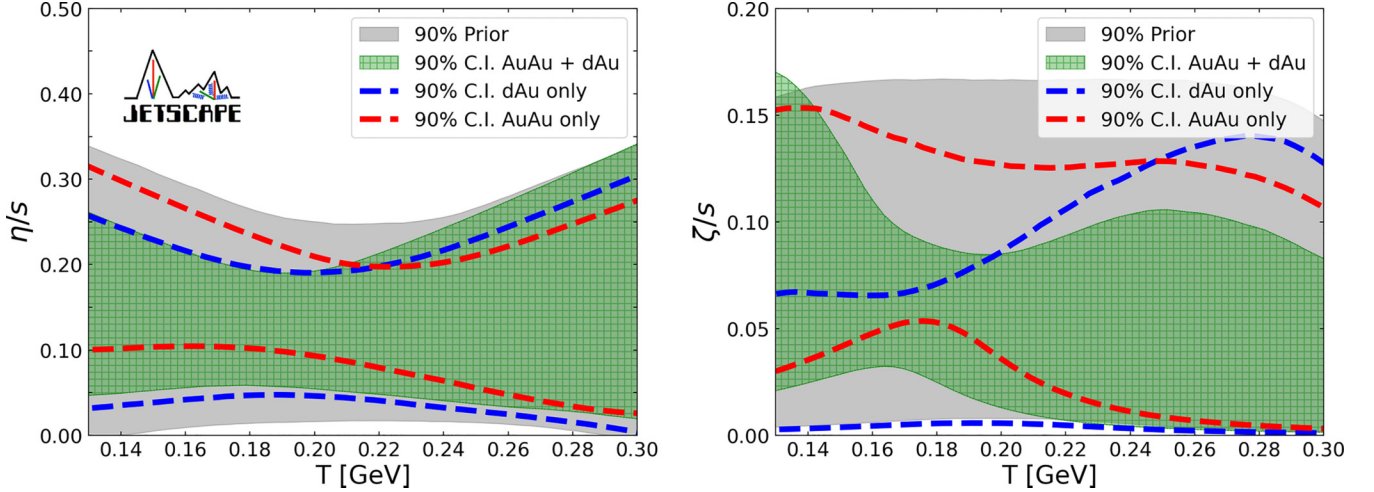


FIG. 14. The posterior for the specific shear (left) and bulk (right) viscosity as a function of temperature for three calibrations: the default (green), the Au-Au subset (red), and the d -Au subset (blue).

uncertainties on both the experimental data and the model emulation.

VI. PREDICTIONS WITH THE CALIBRATED MODEL

In the previous section, we have seen the fits to experimental data and constraints on parameters they provide. In Fig. 5, for example, we showed the fits to Au-Au and d -Au measurements, both explicitly used in the calibration, and those not included. However, we need not restrict our model calculations to measurements in the calibration systems, or even to observables that have been measured at all. The calibrated model is a powerful tool that may be used to provide a wide variety of predictions, and we show in this section some of the results we can obtain with it. For all of the results in this Section, we rely on the ten high-statistics (ten-thousand-event) posterior samples introduced in Sec. III D to represent the spread of the posterior for any given observable. The predictions from the first five of these samples¹² are individually plotted in Figs. 15–21, and 34. In addition, the ten-sample ensemble is used to estimate—via linear interpolation—a 90% credible interval band representative of the posterior, again shown in all the same Figures. This procedure introduces a relatively small uncertainty in the estimation of the 90% band. However, it strikes a reliable compromise between showing the behavior of individual samples and showing a representation of the posterior without explicit emulation uncertainty—all clearly on the same plot—given limited visual and computational resources.

We choose to show predictions this way, rather than with a single MAP parameter set, which, although the best fit to calibration data, necessarily fails to capture the uncertainties quantified in the Bayesian analysis. We do not plot the statistical uncertainty on any of the individual sample predictions as it is always negligible compared with

the width of the posterior itself. We underscore that using high-statistics model simulations means that the predictions are not explicitly affected by the emulator uncertainty, thus

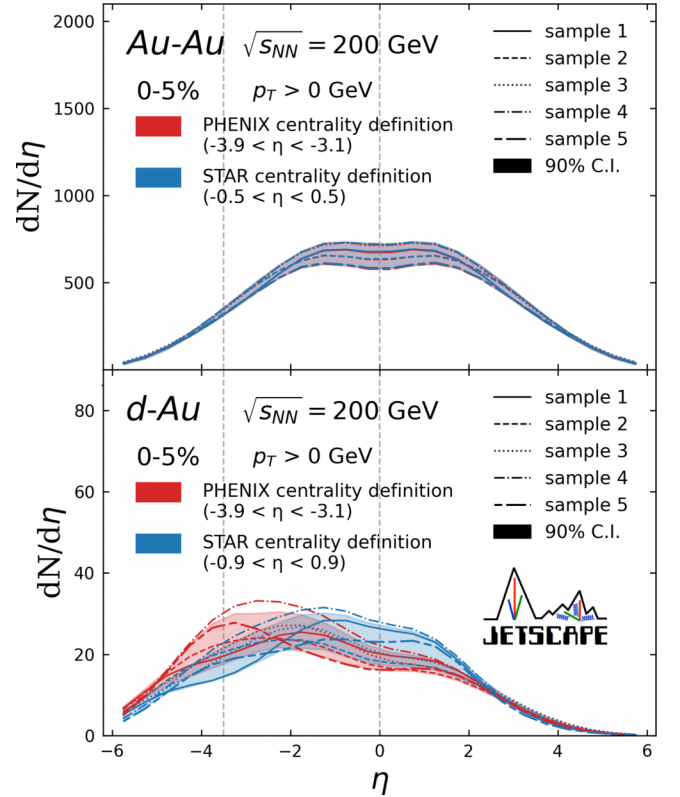


FIG. 15. The calibrated model calculations for $dN/d\eta$ in Au-Au (top) and d -Au (bottom) for two generic choices of centrality selection. The blue calculations use the midrapidity region (often used in STAR measurements) for centrality selection, and the red calculations use a backward region (the Beam-Beam Counter acceptance used in many PHENIX measurements). The calculations are plotted for five individual parameter samples of the posterior simulated for ten thousand events each. In addition, a 90% credible interval band, estimated from ten such high-statistics random samples, is shown.

¹²The parameter values for each of the five individually plotted samples are provided in Appendix K.

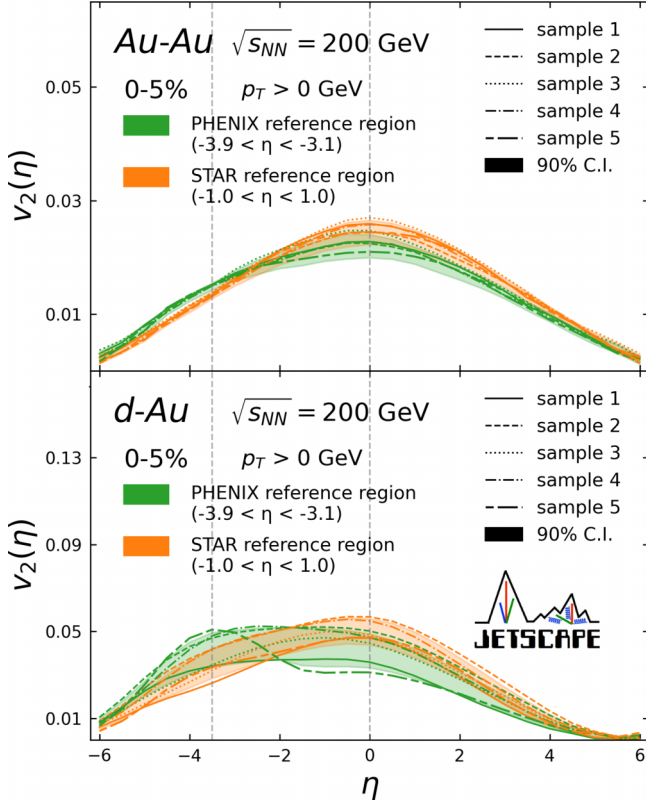


FIG. 16. The calibrated model calculations for $v_2(\eta)$ in Au-Au (top) and d -Au (bottom) for two generic choices of the reference rapidity region. The orange calculations use the midrapidity region (often used in STAR measurements) and the green calculations use a backward region (the Beam-Beam Counter acceptance used in many PHENIX measurements). The calculations are plotted for five individual parameter samples of the posterior simulated for ten thousand events each. In addition, a 90% credible interval band, estimated from ten such high-statistics random samples, is shown.

allowing us to explore observables that have larger emulator uncertainties.

A. Observables in calibrated systems and the longitudinal structure of collectivity

We first show the dependence of observables on the centrality selection method. When selecting centrality, experiments usually separate the kinematic region used to sort the events based on their multiplicity (and thereby centrality) and the region used to make the measurement itself. This separation is enforced to avoid correlation biases that arise if the two are done in overlapping regions. However, selection on any particular acceptance region may still bias measurements, especially when the multiplicity in the selection region is low. We show in Fig. 15, by means of individually plotted posterior samples, as well as an estimated 90% credible interval band, that the effect on the measured multiplicity is small in Au-Au but significant in d -Au when comparing two generic methods of centrality selection: by sorting at midrapidity (as in some STAR measurements) and at backward rapidity (as typically in PHENIX measurements, using the Beam-Beam

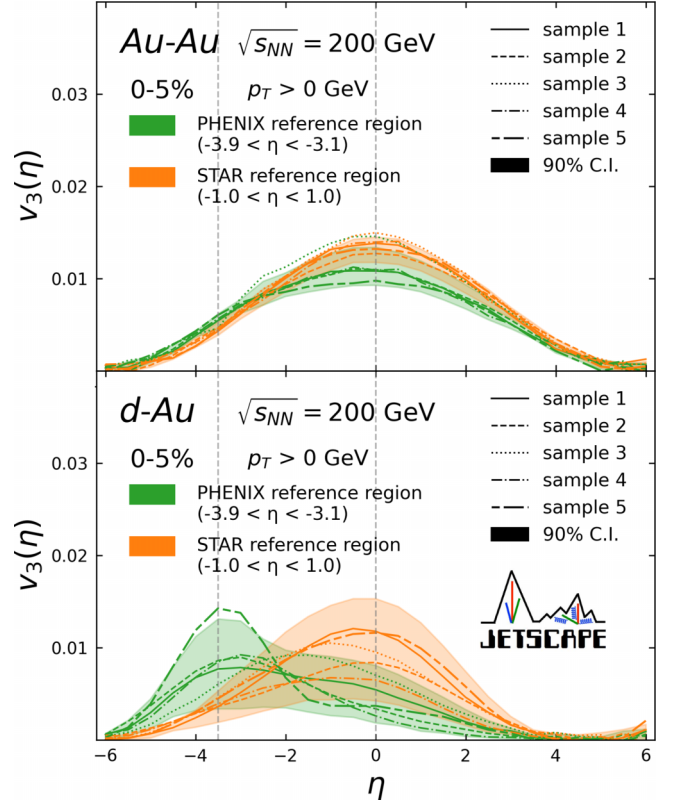


FIG. 17. The calibrated model calculations for $v_3(\eta)$ in Au-Au (top) and d -Au (bottom) for two generic choices of the reference rapidity region. The orange calculations use the midrapidity region (often used in STAR measurements) and the green calculations use a backward region (the Beam-Beam Counter acceptance used in many PHENIX measurements). The calculations are plotted for five individual parameter samples of the posterior simulated for ten thousand events each. In addition, a 90% credible interval band, estimated from ten such high-statistics random samples, is shown.

Counter acceptance rapidity).¹³ The multiplicity is shown to be larger in a rapidity region if the centrality selection was performed in the same or a nearby region, an effect long understood and demonstrated in minor apparent discrepancies between experimental measurements over the years, particularly in small systems.

Furthermore, we show in a series of figures (Figs. 16–18) the calibrated model's description of the longitudinal structure of collectivity—that is, the dependence of collective flow signals on the pseudorapidity and associated gaps between particles. We again plot the five simulated posterior samples, along with the 90% credible interval ten-sample estimate, to represent the spread of the full posterior. In some cases, such as for the d -Au $v_2(\eta)$, the posteriors are wide relative

¹³While the choices of generic ranges for the centrality selection and (as in the following paragraphs) the reference region of flow measurements do not reflect a particular experimental Collaboration's choices for all measurements, a roughly common range allows for easier comparison of our own calculations across systems and observables.

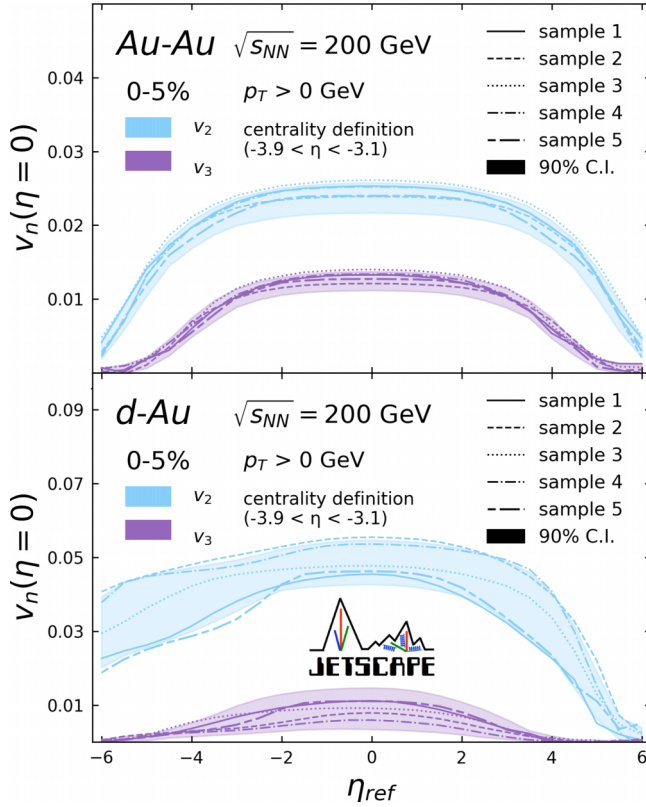


FIG. 18. The calibrated model calculations for flow coefficients in Au-Au and d -Au at midrapidity as a function of the reference rapidity region. The calculations are plotted for five individual parameter samples of the posterior simulated for ten thousand events each. In addition, a 90% credible interval band, estimated from ten such high-statistics random samples, is shown.

to the differences between the probed observables, and thus overlap. However, even in these cases, the differences on a sample-by-sample basis are often significant.

In measurements of azimuthal anisotropy, the methods typically involve particles from two (or more) regions in the kinematic acceptance, often separated in pseudorapidity: the region of interest, where the measurement is quoted to have been made, and the reference region, used to pin down the orientation of the event, which of course varies event-by-event. In experiment, where the background due to jets, decays, or other correlations risks contaminating the flow signal (particularly in small systems), rapidity gaps are effective in removing these nonflow contributions. However, the use of rapidity gaps in these measurements also influences the flow signal itself, due to the nontrivial longitudinal structure of collectivity, leading to effects such as the decorrelation of the event planes in rapidity [40,41,50,86,91–98].

In Fig. 16, we show the $v_2(\eta)$ in Au-Au and in d -Au for two generic reference flow regions commonly used in the PHENIX or STAR experiments. In Fig. 17, we show similarly the $v_3(\eta)$ in Au-Au and in d -Au for, like in Fig. 16, two generic reference flow regions commonly used in the PHENIX or STAR experiments. The calibrated model shows a strong dependence of the flow, both $v_2(\eta)$ and $v_3(\eta)$, on

the rapidity region used as a reference, such that the signal increases with the proximity of the reference region. The effect is discernible even when the posterior is relatively wide (d -Au). Indeed, in the small system, the difference between the average values of our simulated samples is larger. The effect is also larger for v_3 than for v_2 in both systems.

We see a similar trend in Fig. 18, where we show the dependence of the midrapidity value of v_2 and v_3 as a function of the reference rapidity region, in the full range from -6 to 6 units of pseudorapidity, in bins of 0.5 . The dependence is again stronger in d -Au than in Au-Au and in v_3 than in v_2 . We include a slightly different version of the plot in Appendix L to better highlight the effect, where we show the fractional value of the observable scaled by the midrapidity-reference benchmark value.

In this section, we have examined a set of results showing the nontrivial dependence of flow measurements on the full 3D structure of the collisions. In particular, we reiterate that a 3D analysis allows us to avoid the potential ambiguity faced by 2D analyses when attempting to describe flow measurements that depend on the relative rapidity of reference particles and particles of interest. In much the same way, the 3D description is crucial in describing asymmetric small system collisions, for which we make some more wide-ranging predictions in the next section.

B. Extending to p -Au and ^3He -Au

In this section, we show predictions with the calibrated model for a number of small system measurements by STAR and PHENIX and compare them with data. Having calibrated the model on both a large and a small system, we gauge whether the model has gained sufficient insight to describe differences among small systems, namely, p -Au, d -Au, and ^3He -Au. We compare with two categories of observables, first with the rapidity-dependent measurements from PHENIX of the $dN/d\eta$ and $v_2(\eta)$ in the three systems and then with the p_T -dependent measurements from both STAR and PHENIX of v_2 and v_3 in those same three systems. While the two experiments measured nominally the same observables, they used distinct kinematic cuts, centrality selection, and reference regions. We see how much of these differences the calibrated model can account for.

In Fig. 19, we see the comparison with PHENIX measurements of $dN/d\eta$. All of the d -Au data points were used in calibration, and they are plotted in green. The data for the remaining systems are plotted in olive. The posterior as a whole (represented again by the same five high-statistics samples and the ten-sample estimate of the 90% band in use throughout our results) is consistent with the data in all three systems. Individual samples, however, especially in p -Au and ^3He -Au, can still be far from the data. Similarly, in Fig. 20, we see that the comparison between the calibrated model and the $v_2(\eta)$ in the three systems is good. Here, only the forward and midrapidity points from the d -Au data were used in calibration, due to experimental uncertainties from the nonflow contribution—larger in small systems—in the backward region (proximity and overlap to the reference and centrality selection region would enhance the nonflow). Nevertheless,

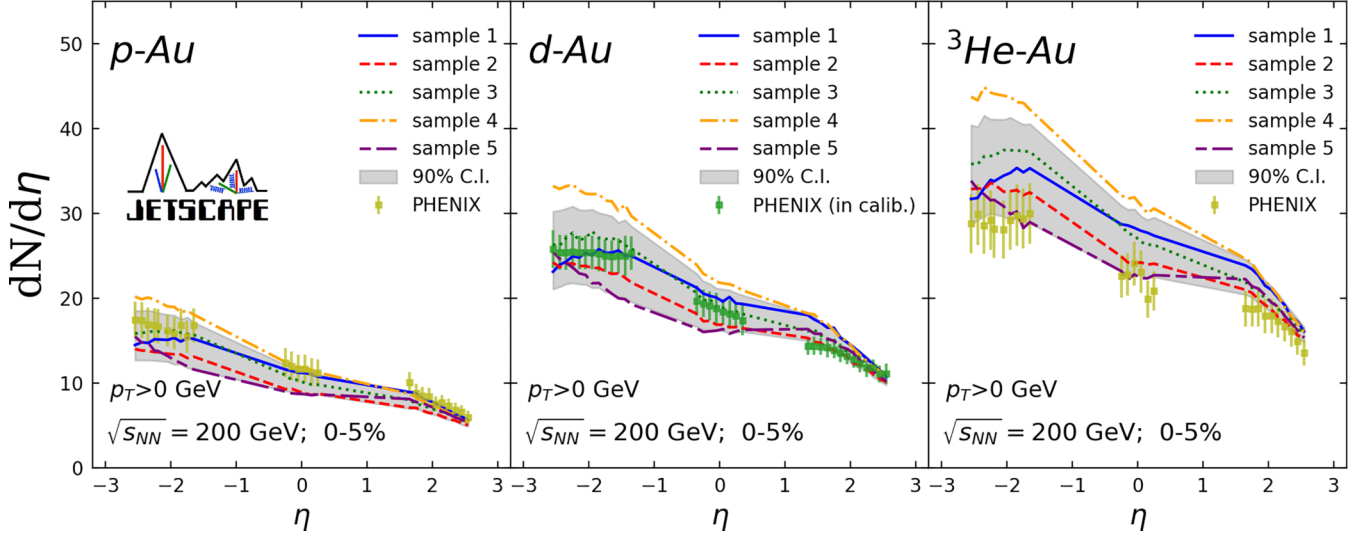


FIG. 19. The PHENIX measurements for the $dN/d\eta$ in p -Au, d -Au, and ^3He -Au compared with a sampling of the posterior calibrated on the default dataset. The calibration dataset includes the dark green data points (d , Au) and does not include the olive green data points (p , Au, and ^3He -Au). The calculations are plotted for five individual parameter samples of the posterior simulated for ten thousand events each. In addition, a 90% credible interval band, estimated from ten such high-statistics random samples, is shown.

we include those points in this comparison figure, as well as the points in that region measured in the other systems. We expect nonflow contribution in the data to enhance the flow signal in that region, an effect that should decrease with system size. In accordance with that expectation, we find the model calculation underestimating the data (though still consistent within the large experimental uncertainties) only in the backward region of the p -Au measurement—the place of maximum nonflow—and in agreement with the data everywhere else.

That said, we are interested in comparing with small system data from multiple experiments, and both STAR and PHENIX have published p_T -dependent v_n measurements in these three systems [38,86]. We note that the large “discrepancy” between the STAR and PHENIX v_3 measurements in d -Au has been a particular source of debate [86]. An earlier prediction of these two measurements using a hand-tuned version of our model was shown in Ref. [94], highlighting that the discrepancy can largely be accounted for by the experimental details in each measurement. In our results, we note that the

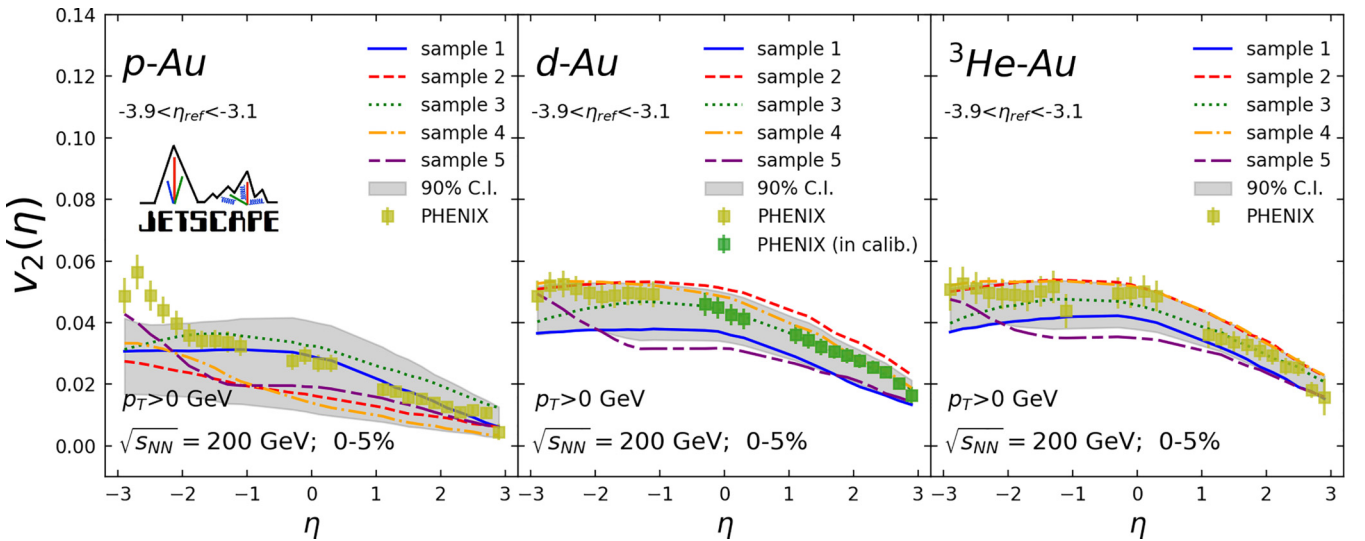


FIG. 20. The PHENIX measurements for the $v_2(\eta)$ in p -Au, d -Au, and ^3He -Au compared with a sampling of the posterior calibrated on the default dataset. The calibration dataset includes the dark green data points (d , Au) and does not include the olive green data points (p , Au, and ^3He -Au). The calculations are plotted for five individual parameter samples of the posterior simulated for ten thousand events each. In addition, a 90% credible interval band, estimated from ten such high-statistics random samples, is shown.

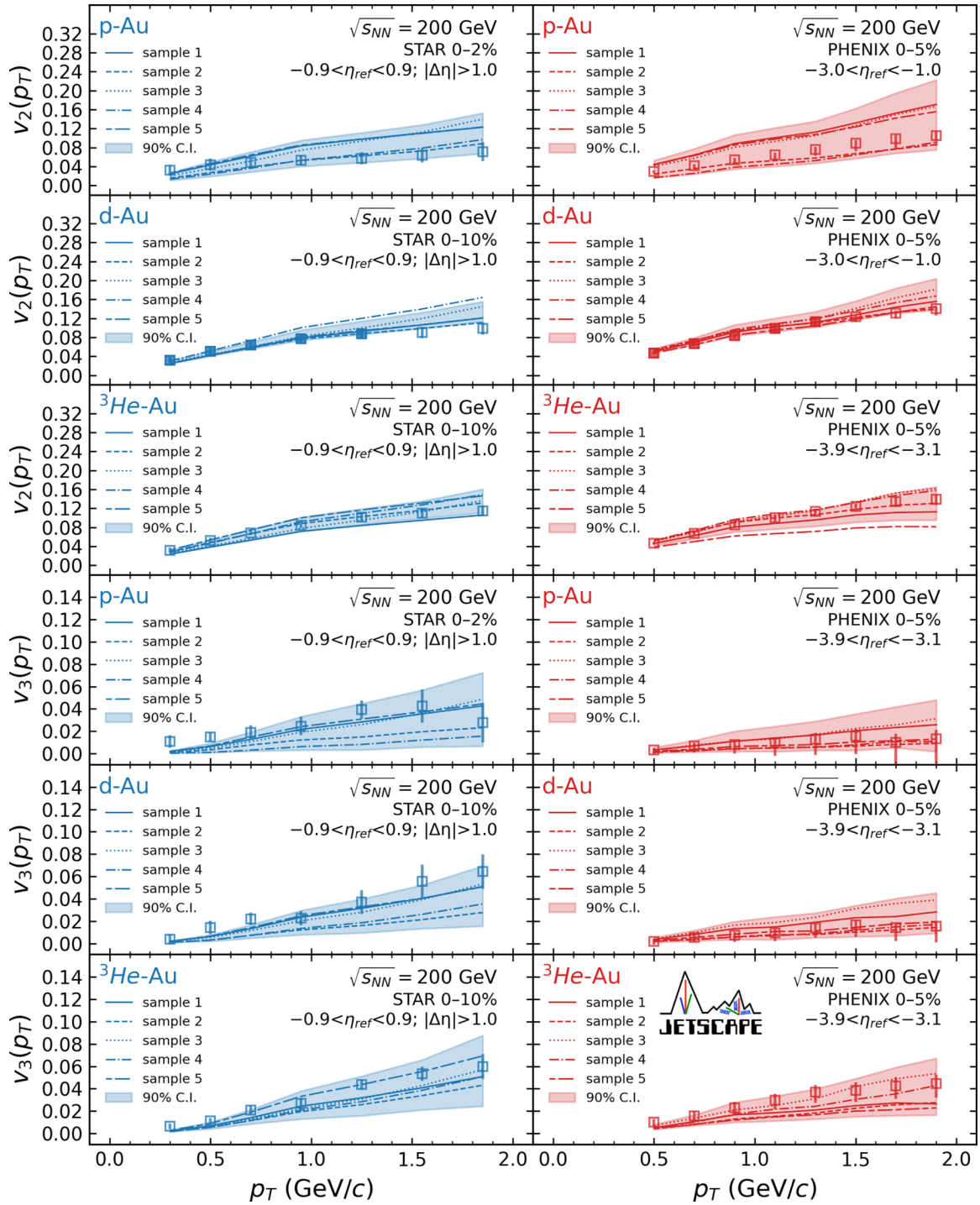


FIG. 21. The PHENIX (red) and STAR (blue) measurements for the $v_2(p_T)$ and $v_3(p_T)$ in p -Au, d -Au, and ^3He -Au compared with a sampling of the posterior calibrated on the default dataset and calculated according to the analysis details of the two experiments. The calibration dataset includes the filled square data points (d , Au v_2 only) and does not include the open square data points. The calculations are plotted for five individual parameter samples of the posterior simulated for ten thousand events each. In addition, a 90% credible interval band, estimated from ten such high-statistics random samples, is shown.

posterior (in the fifth row of Fig. 21) is largely consistent with both measurements, although it more closely describes PHENIX data.

As we can see in Fig. 21, the posterior is indeed overall consistent with the data from STAR and PHENIX across all measurements. It is relatively wide compared with experi-

mental uncertainties, especially for the v_2 in p -Au from both experiments, the v_2 in ^3He -Au from PHENIX, and the v_3 in p -Au from STAR. Even though none of these p_T -dependent datasets in the additional small systems is clearly at odds with the predictions, the width of the posteriors holds promise for obtaining additional constraints by including the datasets in future calibrations.

VII. SUMMARY

In this work, we presented a comprehensive $(3+1)\text{D}$ Bayesian calibration of high-energy nuclear collision dynamics at the RHIC top energy, jointly analyzing large (Au-Au) and small (d -Au, with extensions to p -Au and ^3He -Au) systems. Leveraging a selection of over two decades of measurements from PHENIX, STAR, PHOBOS, and BRAHMS, the study constrains both the longitudinal structure of the initial state and the transport properties of the QGP. The model combines a 3D Glauber initial condition with rapidity-dependent energy deposition and global energy conservation, relativistic viscous hydrodynamics using MUSIC, and a hadronic scattering phase using UrQMD.

The Bayesian analysis features emulation-facilitated model-to-data comparison across the prior range of a 20-dimensional parameter space. A broad dataset is used to calibrate the model, as displayed in Table II, including η -differential, p_T -differential, and integrated observables from the full range in rapidity and across multiple collision systems. The calibration dataset includes all observables that the model can reasonably describe and which can be reliably emulated. We use an additional set of measurements to compare with our calibrated model. The exclusion of the measurements in this additional dataset from calibration is due, in most cases, to the difficulty in emulating them. Recent developments in machine-learning methods present great potential for enhancing emulator performance with precise uncertainty quantification. Recent work [99–101] has demonstrated that such methods, when appropriately implemented, can provide significant improvements for emulating nuclear collision observables. Integrating these methods in our Bayesian workflow in future calibrations could allow for extending the calibration dataset to many more observables, and to more fully exploiting those already included.

We find that compared with midrapidity data alone, rapidity-dependent measurements provide additional constraints on the longitudinal profile of the initial-state energy deposition, the transport coefficients, and the resulting longitudinal structure of correlations. In particular, their inclusion shows a preference for a larger specific shear and bulk viscosity, as shown in Fig. 6.

The individual system calibrations show largely consistent constraints but do indicate some degree of model tension between Au-Au and d -Au. Remarkably, as shown in Fig. 13, the posterior from a calibration using only d -Au observables gives a reasonably good description of the Au-Au data. On the other hand, the Au-Au calibration, which includes the majority of all data points and a greater variety of observables, struggles to describe the $dN/d\eta$ in d -Au. These and other sources of

tension are addressed in the analysis by performing additional calibrations using targeted subsets of the default dataset as well as by examining the local sensitivity of observables to changes in the parameters.

The decorrelation of event planes and differences in the strength of correlation observables as a function of rapidity (and relative rapidity) have been a topic of recent theoretical and experimental exploration [41,50,86,97,102]. Using high-statistics nonemulated posterior samples, we demonstrate throughout Sec. VIA the strong dependence of flow coefficients on the rapidity of both the region of interest and the reference, as well as the impact of the centrality selection region, essential considerations for consistent comparisons with experiments. The 3D analysis allows us to access this information embedded in rapidity-dependent data. It also enables direct comparisons of experimental measurements which differ nontrivially due to their incorporation of forward or backward information in the form of the relative rapidity and the centrality selection. We show the rapidity-dependent effects to be generally stronger in d -Au than in Au-Au, as perhaps best illustrated in Fig. 34.

We make further predictions with the calibrated model (again without the need for explicit emulation) for additional smaller systems (p -Au, ^3He -Au). We find a relatively good match across systems, demonstrating the capability of the model to discern the differences among small systems. In particular, we find the posterior to be consistent with rapidity- and p_T -differential PHENIX and STAR data in those systems, as shown in Fig. 21. We also see a reasonable agreement in, notably, d -Au, where the differences between the STAR and PHENIX measurements have long sparked fruitful discussion about the origin of collectivity in small systems.

While the results in this work are derived from a comprehensive modeling of heavy-ion collisions and systematic comparison with experimental data, they may be further extended in several directions. In terms of theoretical modeling, uncertainties related to the choice of initial state and pre-equilibrium models, of viscous corrections at particlization, or the parametrization of viscosity coefficients may be explored. As previously mentioned, improvements in model emulation can allow for more extensive comparisons and more precise and accurate constraints. Furthermore, the modeling can be extended to include hard processes such that higher transverse momentum data can be compared with in conjunction with the low momentum datasets. Regarding experimental data, the model shows promise in describing measurements at LHC energies, and the calibration could additionally benefit from comparison with the abundant and precise data collected at the LHC. Furthermore, the correlations present in experimental uncertainties could be treated more realistically, either through data-driven estimation techniques or by direct experimental input.

The results underscore that $(3+1)\text{D}$ calibration with rapidity-dependent data is sensitive to initial-state longitudinal structure and QGP transport properties. Our calibrated model provides a tuned baseline for studying the longitudinal structure of correlations and guiding future measurements, both at RHIC and LHC, that can further improve constraints on high-energy nuclear collision models. It also serves

as a calibrated 3D background for jet energy-loss studies, paving the way for joint hard-soft modeling of heavy-ion collisions.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation (NSF) within the framework of the JETSCAPE collaboration, under Grant No. OAC-2004571 (CSSI:X-SCAPE) and OAC-2514008 (CSSI:C-SCAPE). It was also supported under NSF awards PHY-2111568 and PHY-2413003 (R.J.F., M.K., C.P., and A.S.), and Award No. DMS-2316012 (S.M.); it was supported in part by the US Department of Energy, Office of Science, Office of Nuclear Physics under grant/contract numbers DE-AC02-05CH11231 (X.-N.W. and W.Z.), DE-AC52-07NA27344 (D.A.H. and R.A.S.), DE-SC0013460 (A.M., I.S., R.Da. and R.Do.), DE-SC0021969 (C.Sh. and W.Z.), DE-SC0024232 (C.Sh. and H.R.), DE-SC0012704 (B.S.), DE-FG02-92ER40713 (J.H.P.), DE-FG02-05ER41367 (S.A.B.), DE-SC0024660 (R.K.E.), DE-SC0024347 (J.-F.P., G.S.R. and M.S.) and DE-FG05-92ER40712 (J.V., S.T. and A.Ma.). The work was also supported in part by the National Science Foundation of China (NSFC) under Grants No. 11935007, No. 11861131009, and No. 11890714 (X.-N.W.), by the Natural Sciences and Engineering Research Council of Canada under reference numbers SAPIN-2020-00048, SAPIN-2024-00026, and SAPIN-2023-00029 (C.G., S.J., and G.V.), by the University of Regina President's Tri-Agency Grant Support Program (G.V.), and by the Canada Research Chair program (G.V. and A.K.) Reference No. CRC-2022-00146, by JSPS KAKENHI Grants No. 22K14041 and No. 25K07303 (Y.T.), by the Research Council of Finland, the Centre of Excellence in Quark Matter, by the European Research Council [ERC, Grant Agreements No. ERC-2023-101123801 GlueSatLight and No. ERC-2018-ADG-835105 YoctoLHC (H.R.) and No. ERC-2018-ADG-835105 YoctoLHC (I.S.)], and by the São Paulo Research Foundation (FAPESP) under projects 2017/05685-2, 2018/24720-6, and 2023/13749-1 (M.L.). C.Sh., J.-F.P., and R.K.E. acknowledge a DOE Office of Science Early Career Award. This work used the Anvil HPC facility at Purdue University and Stampede2 at Texas Advanced Computing Center, through allocations PHY210068 and PHY230071 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program [104], which is supported by U.S. National Science Foundation Grants No. 2138259, No. 2138286, No. 2138307, No. 2137603, and No. 2138296. This research also used services provided by the OSG Consortium [105–108], which is supported by the National Science Foundation Awards No. 2030508 and No. 2323298.

The content of this article does not reflect the official opinion of the European Union and responsibility for the information and views expressed therein lies entirely with the authors.

DATA AVAILABILITY

The posterior chains used to obtain results from the four main calibration datasets (default, midrapidity, Au-Au, and d -Au) discussed in this analysis are made publicly available for community use [103].

APPENDIX A: PRINCIPAL COMPONENTS

As described in the main text, our emulators are trained to reproduce the values of a number of principal components of the linear decomposition of the model calculations. Each time the emulator is invoked to make a prediction, the principal components are transformed back to observable space.

A subset of principal components collectively describes a well-defined finite portion of the variance in the training model calculations. In principle, the inclusion of additional principal components will always cover a greater portion of the variance. However, we find that the returns diminish quickly, and only a few principal components are necessary to describe an overwhelming portion of the variance. Furthermore, the computational benefits of using PCA at all are reduced if too many principal components are used. We therefore find a balance between these opposing effects by considering the portion of variance covered, the effects on emulation performance in plots like those in Appendix B, and closure tests in order to select the appropriate number of principal components. Ultimately, we use 10 PCs for each of the Au-Au and d -Au emulators, accounting for more than 99% of the variance of the training calculations in both systems.

We “truncate” the rest of the principal components, which leads to an uncertainty in the predictions due to this truncation that is not already accounted for by the emulator. We estimate the truncation error using a separate white-noise kernel and add it to the emulator covariance to explicitly account for the uncertainty [109].

Lastly, before and during PCA, we engage in “postprocessing” of the model calculations. This includes transforming some of the observables by taking the logarithm of the multiplicity and transverse energy, as described in the main text (Sec. III B), and transforming all observables to have zero mean and unit variance across the training dataset. In addition, we examine the PCA distribution to identify and remove outliers in 2D plots of any pairs of principal components in the training dataset, since PCA yields linearly uncorrelated principal components. This step was shown to reduce emulation uncertainty as well as lead to improved closure. In this process of design curation, we remove about 2% of the training design points from each system's emulator training dataset.

APPENDIX B: EMULATOR VALIDATION

As a prerequisite to this analysis, we perform a thorough validation of the model emulation. We examine a number of quantifiers, such as the shape of the prediction distri-

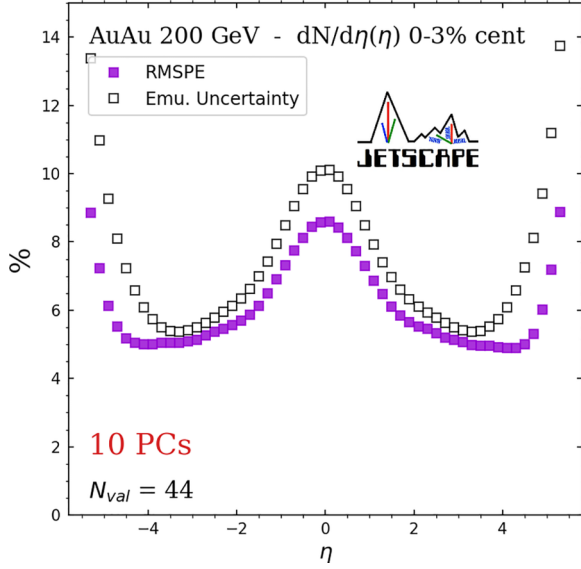


FIG. 22. The root mean square percent error (RMSPE, filled squares) in percent between model calculations and emulator predictions averaged across 44 validation points for each η bin of the Au-Au $dN/d\eta$ in 0%–3% central collisions as in the measurement by PHOBOS. The empty squares show the emulator’s estimate of the uncertainty in percent, also averaged over all design points for each bin.

butions around the mean, the dependence of the residuals on the magnitude of the mean, and the root mean square percent error (RMSPE) between the model predictions and the model calculations in the validation set. We use 44 of the design points as our validation set, keeping them separate from the training designs. We also evaluate how the RMSPE compares with the emulator’s estimate of its uncertainty.

We use these quantifiers, and in particular the RMSPE, to determine whether the emulator performance is sufficient for observables to be included in the analysis. We perform an observable-by-observable assessment (where we define an “observable” as all the bins of a differential observable—in η or p_T —in a given centrality class, or, for integrated observables, all the centrality bins). While we do make some choices to exclude from the analysis certain experimental centrality or p_T bins “within” an observable due to uncertainties related to the model scope, as described in Sec. IV, all decisions stemming from the emulation performance are done at the level of the observable as a whole, and we do not, in general, discriminate between individual bins of a given observable.

We make these assessments after having taken several steps to optimize the emulator performance, such as the discretization of the viscosity parameters on a grid of temperature, the log-transformation of the multiplicity and transverse energy observables, and the choice of the number of principal components, among others. We pay particular attention to ensure that the emulator’s estimate of its uncertainty does not underestimate significantly the uncertainty we estimate “empirically” as the RMSPE obtained from our validation points. We aim

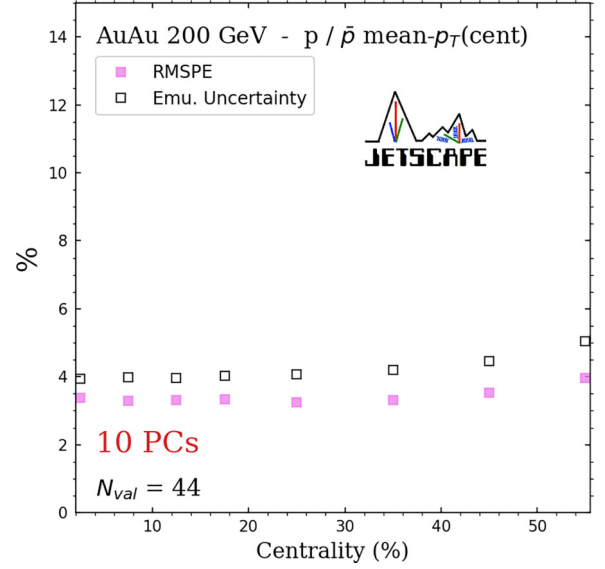


FIG. 23. The root mean square percent error (RMSPE, filled squares) between model calculations and emulator predictions averaged across 44 validation points for each centrality bin of the Au-Au proton $\langle p_T \rangle$ as in the measurement by PHENIX. The empty squares show the emulator’s estimate of the uncertainty in percent, also averaged over all design points for each bin.

to keep the relative emulation uncertainty to no more than 15%–20% as a bin-average across each observable, although for most observables it is significantly smaller.

In Sec. III B, we mention the discretized grid used to redefine the specific shear and bulk viscosity parameters for emulator training. We use ten uniformly distributed points across the prior range to serve as the new parameters, replacing the original four parameters determining the functional form. The equivalent replacement is done for both the shear and bulk viscosity, totaling 20 new parameters used in emulator training due to viscosity discretization in place of the original eight. We find that an even finer grid, whether enhancing the density in a localized temperature region or uniform across the full range, does not significantly improve the emulation beyond this point.

In Figs. 22 and 23, we show the validation-set-average RMSPE for two sample observables (the Au-Au multiplicity for a central bin from PHOBOS, and the mean p_T of protons as a function of centrality from PHENIX) as well as the emulator’s estimate of the uncertainty across the range of pseudorapidity and centrality bins, respectively. We see that the $dN/d\eta$ RMSPE depends on pseudorapidity, being largest at midrapidity as well as at large values of $|\eta|$, and that it approximates decently well the emulator error across the range. Meanwhile, the RMSPE of the integrated proton mean p_T has little dependence on centrality, is in general low (below 5%), and tracks well with the emulator uncertainty. The mean p_T is an example of an observable that we are able to emulate well in the analysis, while some η bins of the multiplicity, particularly for more peripheral centrality classes, or without optimization procedures like taking the logarithm, are more uncertain.

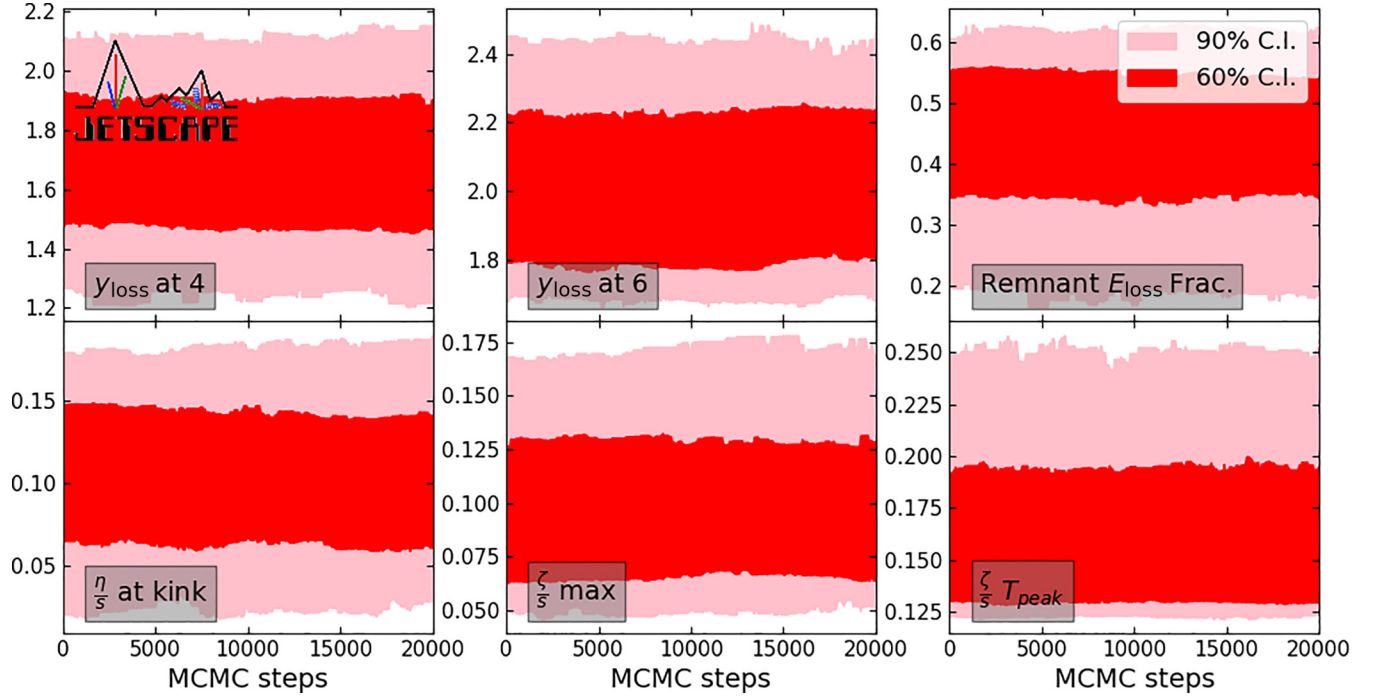


FIG. 24. The distributions of six sample parameters among walkers at particular steps of the MCMC chain following the burn-in phase. All the distributions remain stable after—and in some cases converge well before—the conclusion of the burn-in.

APPENDIX C: MCMC CONVERGENCE

To ensure the convergence of the MCMC algorithm to the posterior distribution, we use a burn-in phase of 20 000 steps. We verify the convergence by showing in Fig. 24 the distribution of individual select parameters within the MCMC chain as a function of steps undertaken after the burn-in has concluded. The bands at any given step indicate the distribution of the walkers for the respective parameters at that step. We can see that the distributions of the parameters have reached sufficient convergence by the end of the burn-in phase and remain stable as we sample the posterior.

APPENDIX D: COVARIANCE MATRIX

As described in Sec. III, the covariance matrix contains, in principle, information about the uncertainties from both experimental data and the model. For an aggregation of the modeling uncertainty, we rely on the emulator’s estimation, including for correlations between observables. The experimental uncertainties that enter into the diagonal of the covariance matrix come directly from experimental publications (for a select few observables, we adjust these values as described in Sec. IV to account for methodological discrepancies). Previous works have found that prescriptions for the nondiagonal components of the covariance matrix can have counterintuitive effects [110]. We leave to future work a systematic examination of the benefits and pitfalls of various approximation schemes of experimental uncertainty correla-

tions across rapidity, transverse momentum, centrality, and more.

As some examples of the relative contributions to the total covariance matrix from emulation and experimental uncertainties in our analysis, we show in Fig. 25 the relevant diagonal quantities for three observables. In the central bin of the d -Au $dN/d\eta$ from PHENIX, we have an example of an observable for which the emulation or experimental uncertainties are relatively balanced. The 3%–15% centrality bin of the Au-Au $v_2(\eta)$ from PHOBOS is an example of an observable where the experimental uncertainties dominate, while the 20%–30% Au-Au $v_2(p_T)$ from PHENIX shows an observable dominated by the emulation uncertainty. We also show the strength of correlations in the emulation predictions for the same three observables in Fig. 26. Due to significant breaks in the continuity of the pseudorapidity binning, we plot the η -differential observables with two symmetric white gaps to indicate these breaks. The gaps are not drawn to scale and serve only as reminders. Within the sections that remain “together,” the bins are uniform in width. All three observables discussed in this section are also included in the aggregator plot in Fig. 4, where the relative theoretical and experimental contributions to uncertainty are shown for all observables used in calibration.

APPENDIX E: CLOSURE TEST POSTERiors

In this Appendix, we show, as seen in Fig. 27, the posteriors for select parameters from three closure-test parameter sets and compare them with the truth values. We see in general good closure across parameters, as well as clear differences

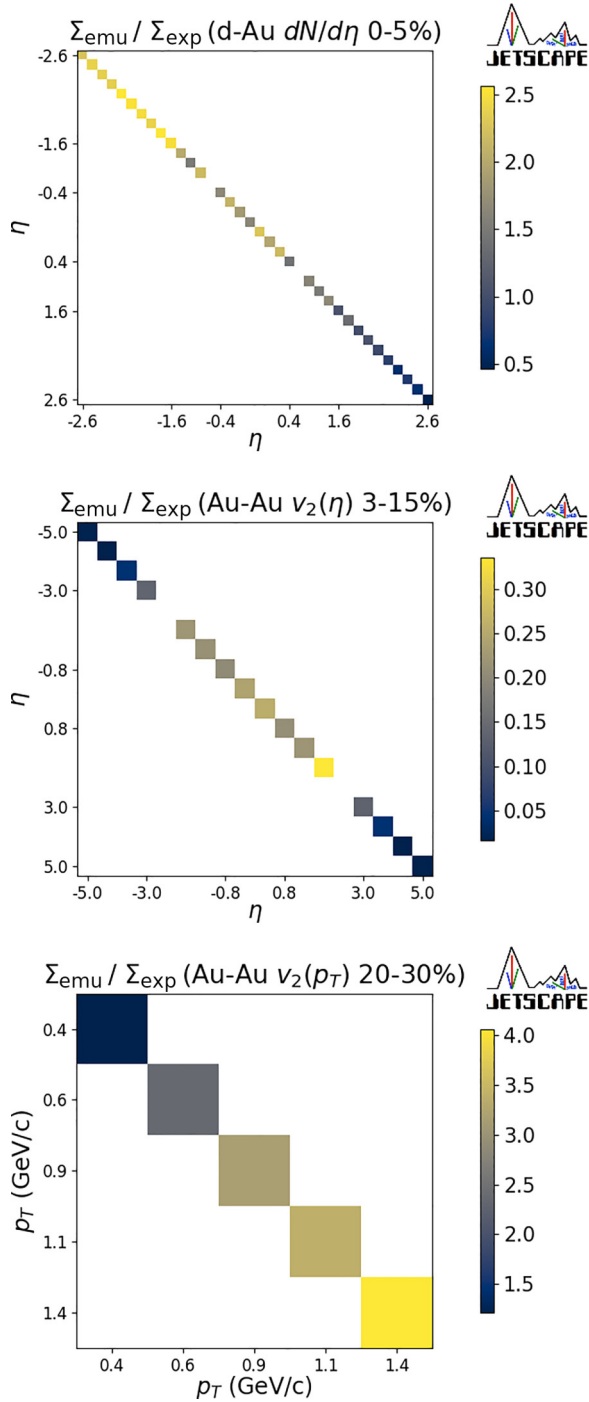


FIG. 25. The ratio between the diagonal of the emulation uncertainty matrix and the experimental (assumed diagonal) covariance for the 0%–5% centrality class of the d -Au multiplicity from PHENIX (top), the 3%–15% centrality class of the Au-Au $v_2(\eta)$ from PHOBOS (middle), and the 20%–30% centrality class of the Au-Au $v_2(p_T)$ from PHENIX (bottom), highlighting three different cases of the relative contributions from each source of uncertainty.

in the constraining power of pseudo-data between different parameters. We note that we see the best closure for the most sensitive parameters, in particular the rapidity loss pa-

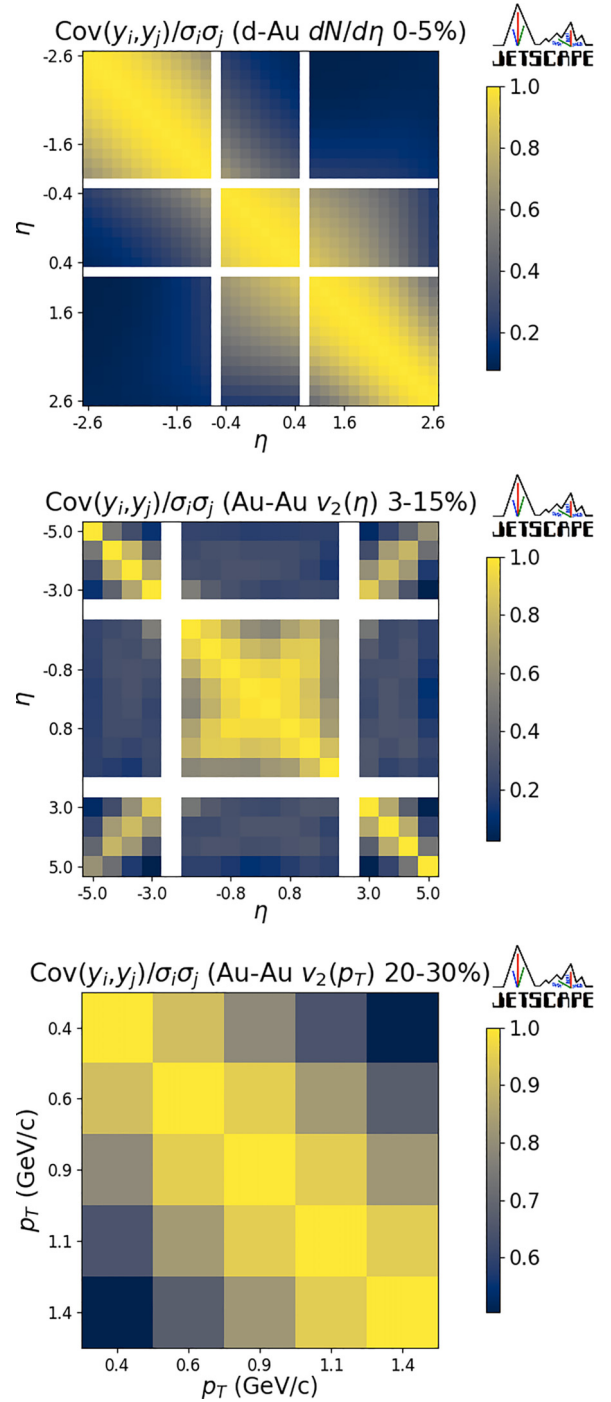


FIG. 26. The emulator uncertainty correlation matrix, including correlations between different η and p_T bins for the same observables shown in Fig. 25: the 0%–5% centrality class of the d -Au multiplicity from PHENIX (top), the 3%–15% centrality class of the Au-Au $v_2(\eta)$ from PHOBOS (middle), and the 20%–30% centrality class of the Au-Au $v_2(p_T)$ from PHENIX (bottom).

rameters, while the viscosity parameters often show weaker constraints. Nevertheless, even the constraints on viscosity are significant, and we furthermore see additional indications of sensitivity in closure tests (not shown) where only a single parameter from a set (the width of the bulk viscosity function,

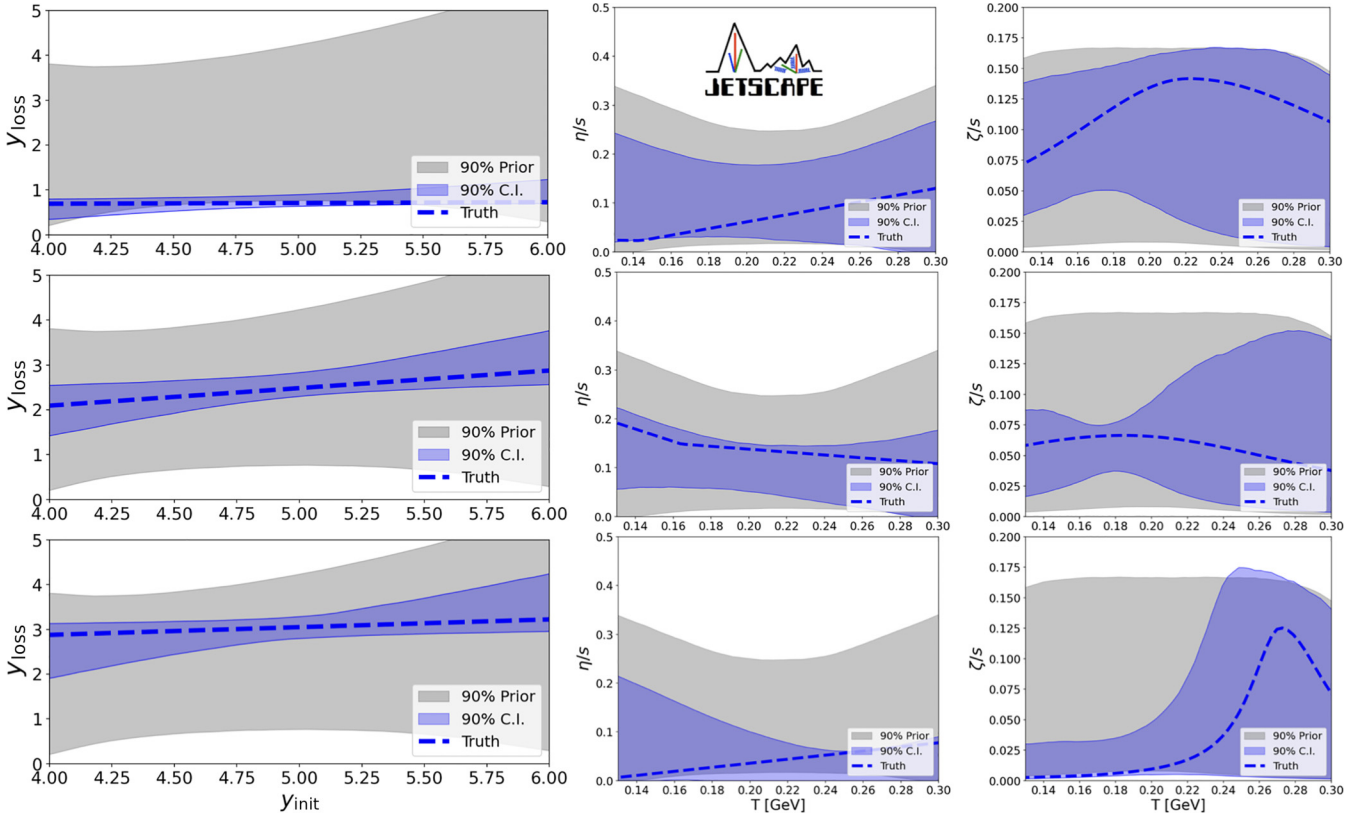


FIG. 27. The closure test posteriors for the rapidity loss as a function of incoming beam rapidity, and the shear and bulk viscosity as a function of temperature, from runs with the model calculations of three validation parameter sets used as pseudo-data.

for example) is varied. The posteriors in such emulator-assisted tests are responsive to these small changes in a single parameter.

Since we use the true experimental uncertainties for the pseudo-data at any given closure point, we do, in some cases, have a smaller pseudo-data uncertainty than the corresponding statistical uncertainty of the mean model calculations for that closure point. While the pseudo-data uncertainty may be arbitrarily small in general, if the statistical model uncertainty in generating the predictions for it is sufficiently large, closure tests with small uncertainties fail to constrain the model. Although we do not strictly enforce this requirement in our closure tests, we do restrict the analysis as a whole to observables we can compute with relatively low statistical uncertainty.

APPENDIX F: A NOTE ON THE EXPERIMENTAL DATA

In this Appendix, we point out some additional technical details regarding accessing the data reported by select experimental papers and the decisions we made about how to include them in the analysis.

The BRAHMS $dN/d\eta$ measurement in Au-Au reported only statistical uncertainties in HEPData, but showed “total” uncertainties for three representative positive rapidity bins in a table in the paper [87]. Given that the total uncertainties were much larger than the statistical ones reported, we used

the total uncertainties from the three bins to interpolate the values of the uncertainty for the fine-binned measurement. We reflected the uncertainty values to populate the negative-rapidity bins. This was the case for both the middle and large rapidity measurements reported in that paper, which we used to compare with our calibrated model. For this reason, as well as some apparently inconsistent binning between the report on HEPData and the points in the paper, we did not include the measurement in the calibration set.

In the case of the STAR measurement of $v_2(\eta)$ in Au-Au [80], the data, which was reported on the experiment website, did not show a breakdown of statistical and systematic uncertainties, and did not specify any details about the reported uncertainties. In this and any such cases, we used the reported uncertainties as one standard deviation total uncertainties.

In the case of the PHOBOS measurement of that same observable in Au-Au [79], the PHOBOS website reported “90% confidence intervals,” so those were accordingly scaled to obtain the 1σ values. For some of the bins, the uncertainties were asymmetric. In any such cases, we chose to use the maximum value between the upper and lower bounds as the σ for a symmetric error, since we do not, in general, incorporate asymmetric uncertainties in the Bayesian analysis. We do, however, use asymmetric uncertainties as reported by experiment in plots of observables with which we only compare.

The measurement of $v_2(\eta)$ by PHOBOS [79] presented an additional challenge. For a few points, there appeared to be a significant discrepancy between the plot in the paper and the reported observable value on the PHOBOS website. For all bins in that analysis, we decided to use the points from the plot in the paper by digitally extracting them. The reported uncertainties between the plot and the website appeared consistent, so we used them as reported.

We digitized data from a plot for at least one additional measurement: the integrated STAR v_3 in Au-Au [82], although, in this case, it was the uncertainties that showed a possible discrepancy. In the HEPData entry, only statistical uncertainties were reported; however, the plotted error bars in the paper indicated that the systematic uncertainties were significant. We again gave precedence to the plot and used the uncertainties in the figure as the total uncertainties.

We made efforts over the course of this study to reach out to experimental collaborators, including those from all four of the long-serving experiments, as well as from sPHENIX, regarding questions of methodology and data reporting. We received fruitful feedback and guidance on many of the analyses we made use of.

APPENDIX G: FULL DEFAULT POSTERIOR CORRELATIONS

In this Appendix, we show, in two parts, the full 2D correlations between all 20 of the model parameters in the default calibration in Figs. 28 and 29. This is a more complete picture of the correlations shown in Fig. 7, which showed only a selection of parameters for readability and focus.

APPENDIX H: OBSERVABLE SENSITIVITY

We can use our trained emulator to efficiently probe the local sensitivity of observables to our model parameters. In Fig. 30 in this Appendix, we show the impact on a sampling of observables due to changes in a single parameter at a time. This may in principle be done in any region of parameter space, although, for these figures, we choose the locus to be the MAP parameter set. We show the dependence on the maximum of the bulk viscosity, the shadowing factor, the rapidity loss, and the remnant energy-loss factor. The plots show the emulator prediction at each parameter point and its corresponding uncertainty band in red, and the experimental measurement and its corresponding uncertainty in gray, each scaled by the value of the experimental data for that observable bin.

APPENDIX I: FULL OBSERVABLE POSTERIOR FOR INDEPENDENT SYSTEM CALIBRATIONS

In Sec. VB, we show and discuss the posterior for a selection of observables in the Au-Au and d -Au independent calibrations. In Figs. 31 and 32 in this Appendix, we show for completeness a fuller set of observable posteriors for those calibrations as plotted in Fig. 5 for the default calibration,

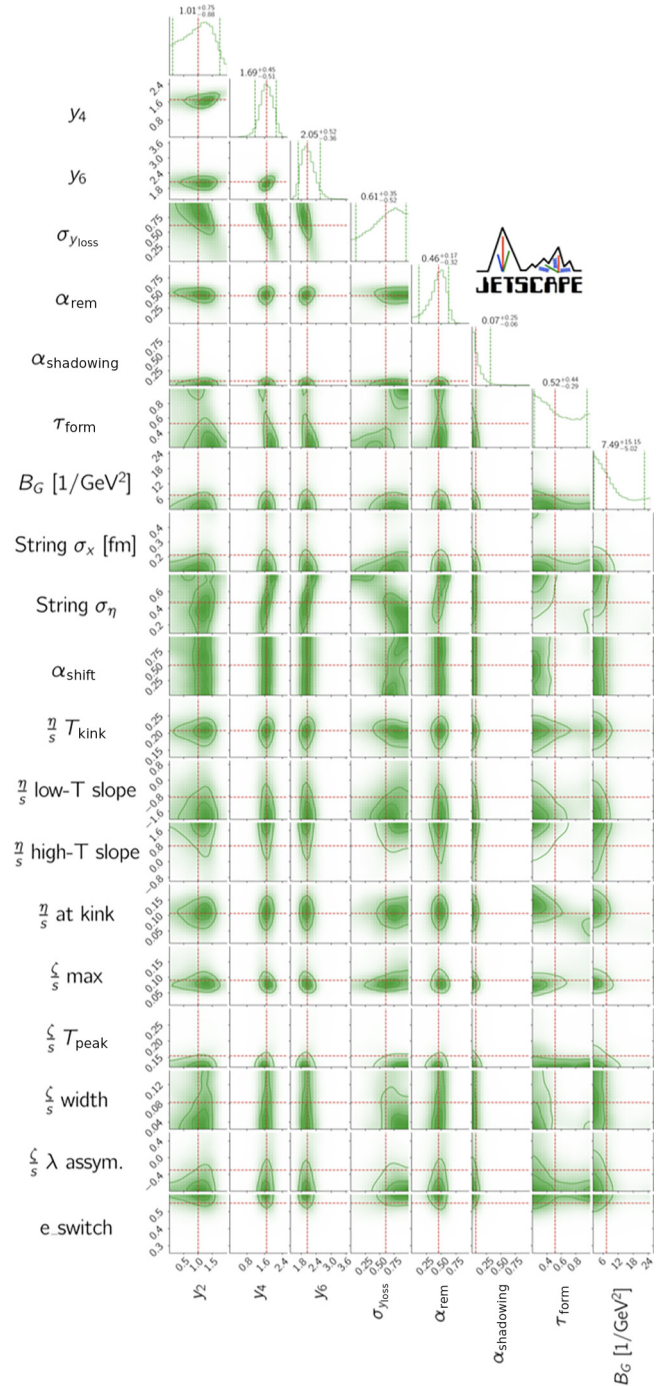


FIG. 28. A portion of the 1D and 2D marginalized parameter posteriors for the default calibration. Together with the plots in Fig. 29, this represents the full, 20-dimensional posterior of parameters calibrated in the analysis. The plots across the main diagonal show the 1D marginalized posteriors for each of the parameters. The off-diagonal plots show the 2D correlations between pairs of parameters. The median and 90% credible interval bounds are indicated by the red and green dashed lines, respectively.

again using the emulator to estimate the model predictions for one hundred thousand samples of the respective posteriors.

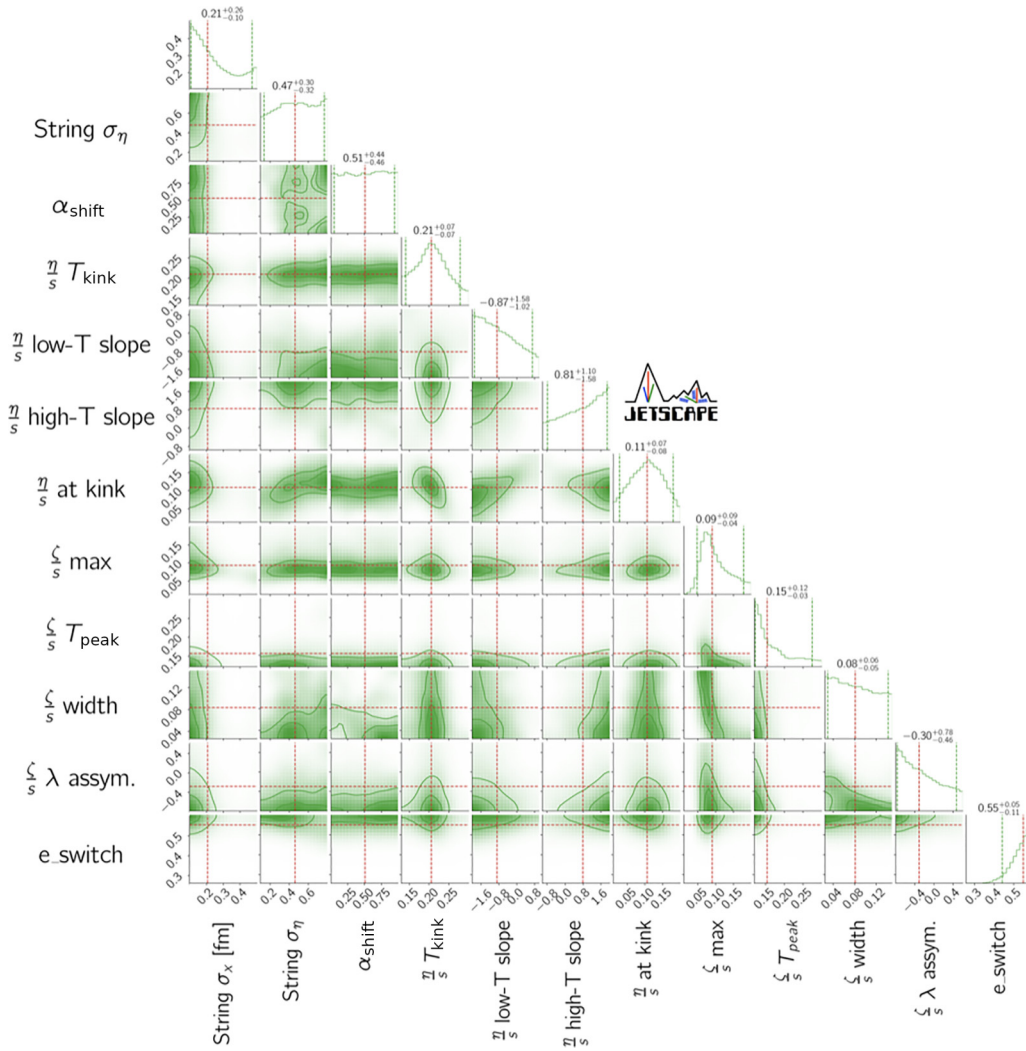


FIG. 29. The remaining portion of the 1D and 2D marginalized parameter posteriors for the default calibration. Together with the plots in Fig. 28, this represents the full, 20-dimensional posterior of parameters calibrated in the analysis. The plots across the main diagonal show the 1D marginalized posteriors for each of the parameters. The off-diagonal plots show the 2D correlations between pairs of parameters. The median and 90% credible interval bounds are indicated by the red and green dashed lines, respectively.

APPENDIX J: PARAMETER POSTERIOR FOR INDEPENDENT SYSTEM CALIBRATIONS

In Sec. VB, we show the posterior of the viscosity parameters for the independent Au-Au and *d*-Au calibrations in Fig. 14. Here, in Fig. 33, we show additionally the posteriors for the remaining 12 parameters in marginalized form.

APPENDIX K: PARAMETER VALUES OF FIVE POSTERIOR SAMPLES AND MAP

In Table IV, we include the values of the 20 model parameters for each of the five samples of the posterior we have simulated and used to make predictions throughout Sec. VI as well as for the MAP parameter set. To reproduce the results included in this paper, these parameters should be used in conjunction with the default parameters of the 3D Glauber and

iEBE-MUSIC framework, either through the native package or through X-SCAPE.

APPENDIX L: DEPENDENCE OF FLOW COEFFICIENTS ON THE FLOW REFERENCE

The dependence of the flow coefficients v_2 and v_3 at midrapidity as a function of the reference rapidity was shown in Fig. 18. Here, in Fig. 34, we show a version of that result scaled by the value of the coefficients using, as a benchmark, a midrapidity reference. The scaled version allows for an easier comparison of the impact of the rapidity region of the reference across the full rapidity range and between systems. We note that the apparent decrease in v_n at midrapidity with the distance of the reference from midrapidity is a convolution of multiple effects, including the decorrelation of the event planes in rapidity as well as the rapidity dependence of the magnitude of flow itself.

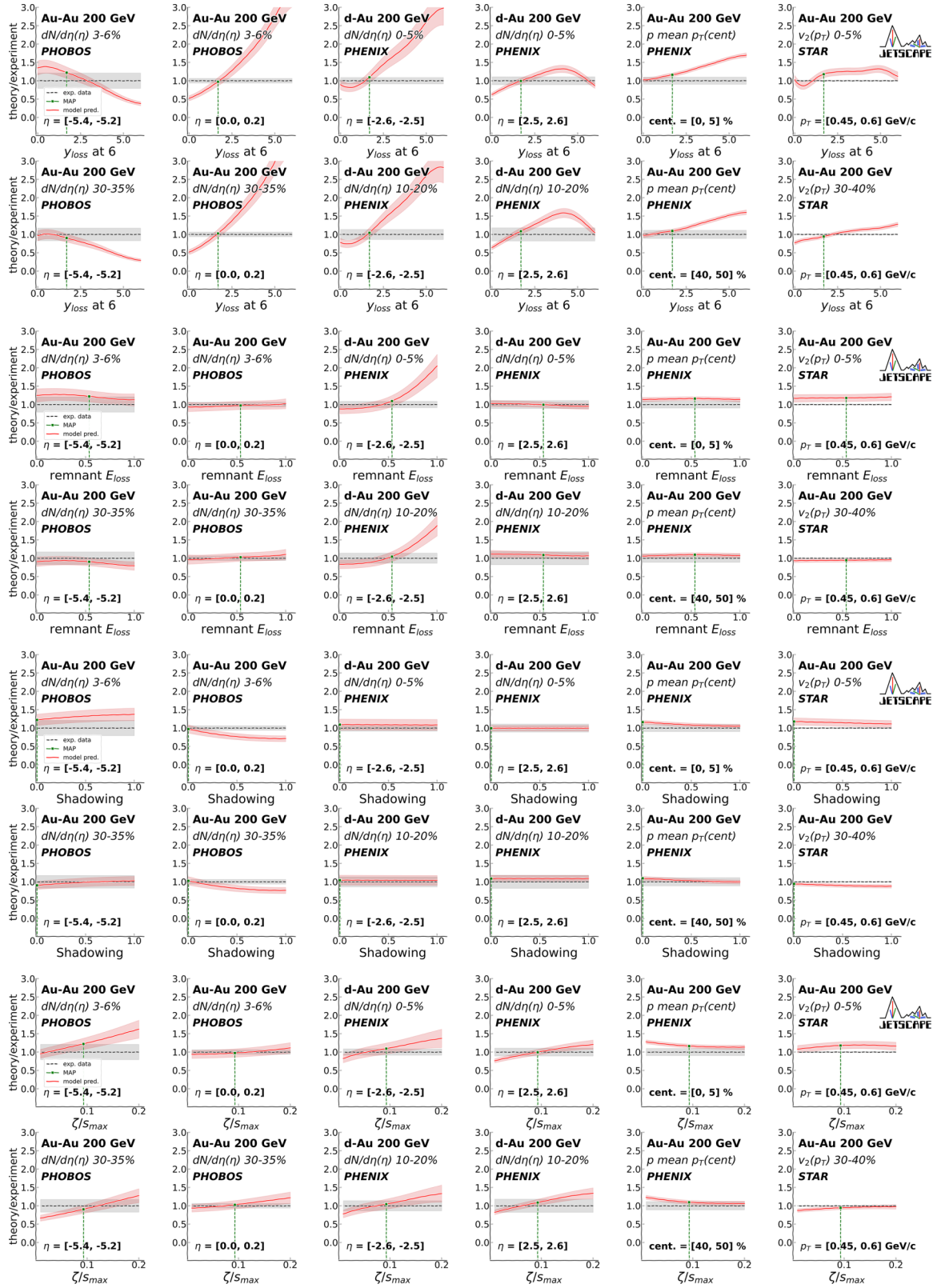


FIG. 30. The sensitivity of select observable bins to varying one of the rapidity loss parameters, the remnant energy-loss factor, the shadowing factor, and the maximum of the bulk viscosity, calculated using the emulator. The red band indicates the emulated prediction and uncertainty, the gray band the experimental measurement and uncertainty, while the dashed line marks the MAP value of the parameter. Both the emulated predictions and the experimental data are scaled by the values of the respective experimental measurements to facilitate comparison across observables.

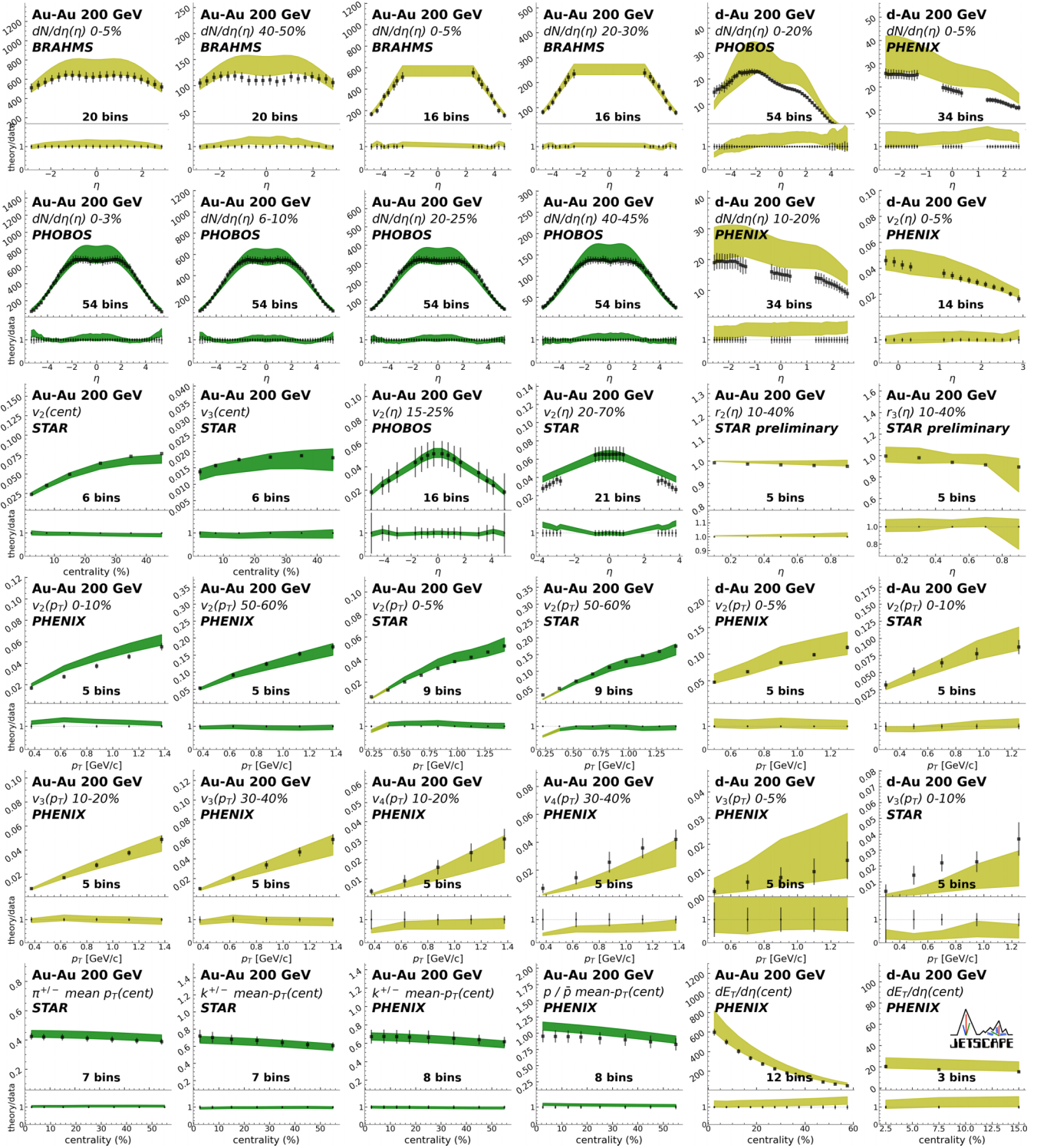


FIG. 31. The observable posterior for the data included in the Au-Au-only calibration, plotted in green, as well as predictions for some not included, plotted in olive. The solid markers are the experimental data, while the bands show a 90% interval of the posterior calculated using the emulator. Each plot is labeled with the number of data bins corresponding to each observable in a given plot.

APPENDIX M: SAME-RAPIDITY CORRELATIONS

In this Appendix, we show the dependence of v_n on η if the reference rapidity region coincides with the region of interest. This is sometimes the case in experimental measurements—

particularly in some older measurements where only a single detector, confined to a limited rapidity region, was used to obtain v_n . In some measurements from STAR, for example, the observable was measured at midrapidity while enforcing an η -separation between individual pairs of particles within

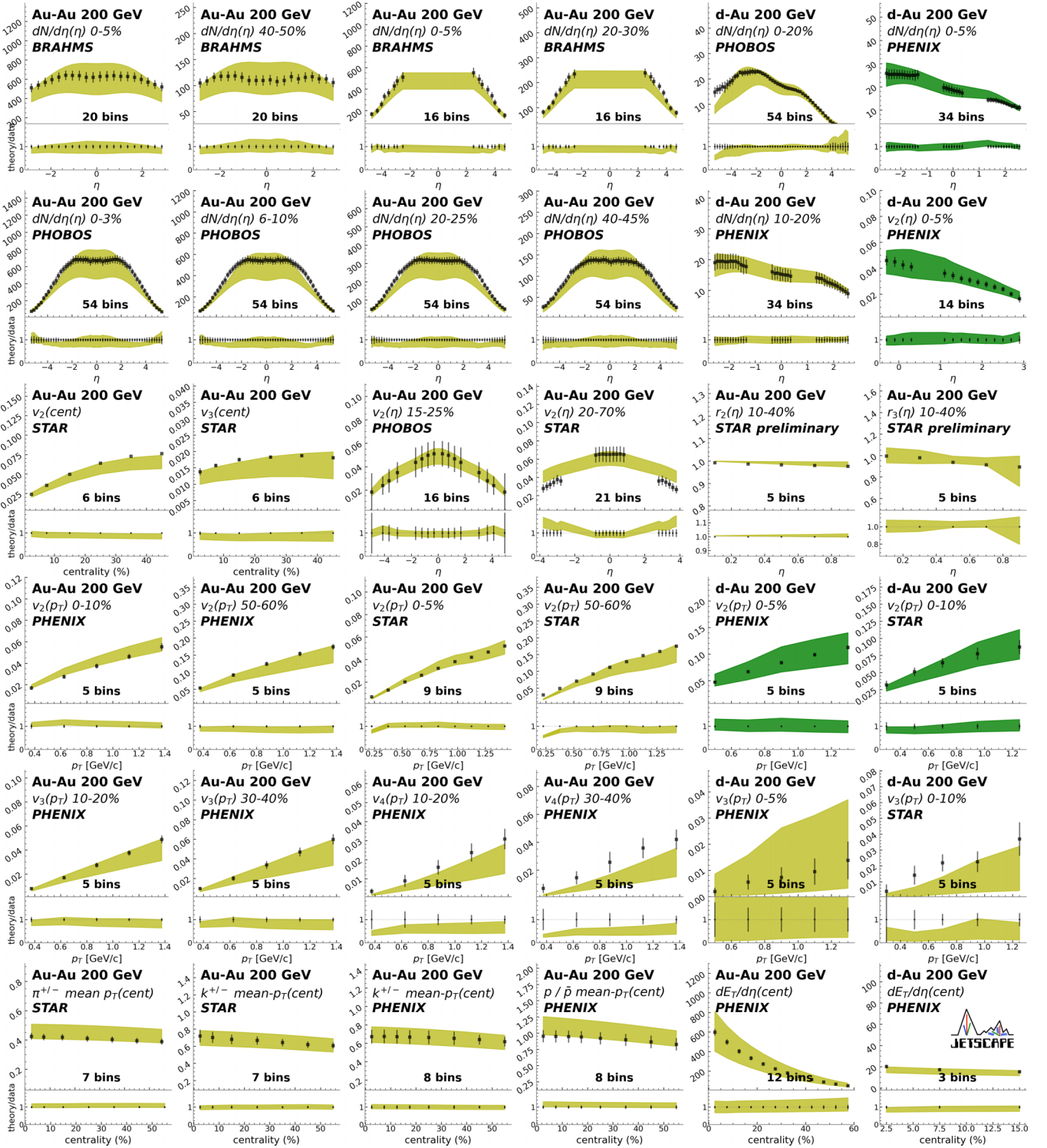


FIG. 32. The observable posterior for the data included in the d -Au-only calibration, plotted in green, as well as predictions for some not included, plotted in olive. The solid markers are the experimental data, while the bands show a 90% interval of the posterior calculated using the emulator. Each plot is labeled with the number of data bins corresponding to each observable in a given plot.

the midrapidity region covered by the central time projection chamber (TPC). If such measurements are compared with—as they are in this paper—care must be taken to explicitly account for the potential autocorrelations that may arise from correlating particles from the same rapidity region. In Fig. 35,

we show our calculations of $v_n(\eta)$ when $\eta_{\text{ref}} = \eta$ (in bins of width 0.5 in η) for one sample of the default posterior. To highlight the potential impact of autocorrelations if not accounted for properly, we show the calculations including those contributions, as well as after subtraction. Overall, the v_n

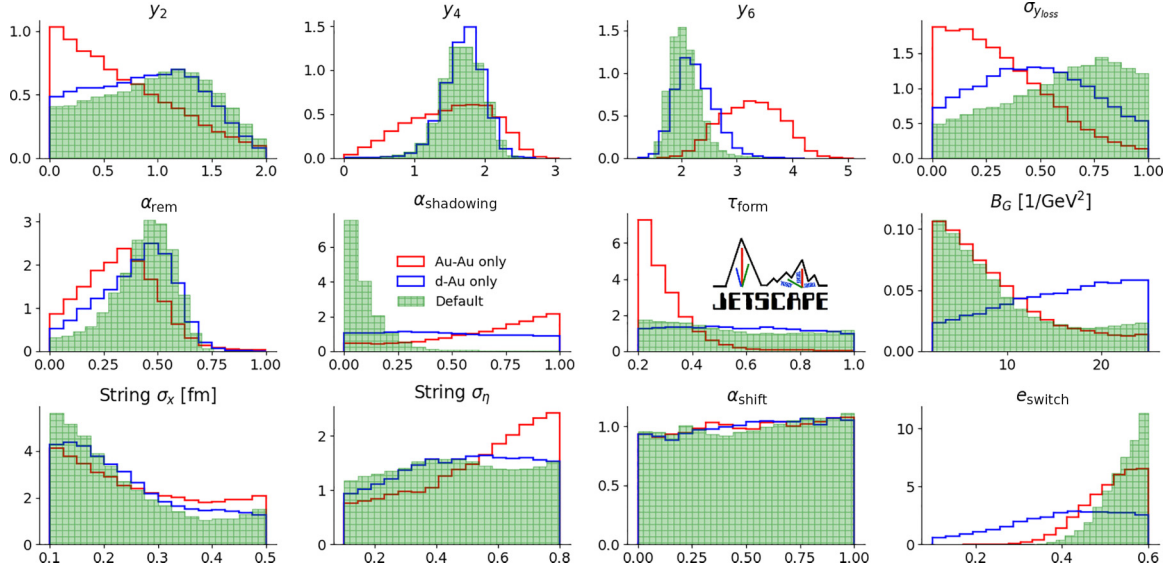


FIG. 33. The posteriors for the twelve model parameters (excluding the viscosity parameters), for three calibrations: the default, in hatched green, the Au-Au only calibration, in red, and the d -Au-only calibration in blue.

values show similar η dependence as when the regions of reference and interest are separated, with perhaps a slightly more peaked shape around midrapidity. We can see that the impact from autocorrelations is relatively small for large values of the v_n and where the particle multiplicity is largest (notably at midrapidity). The impact grows as we move to large absolute rapidities for both of those reasons and becomes significant around $|\eta| = 4$, and dominant quickly beyond that. We note also the smaller impact on the d -Au v_n compared with Au-Au, chiefly attributable to the larger values of the observable for central events in that system.

We also show in this Appendix a cross-check of the impact of self-correlations due specifically to the reference correlator in the denominator of the scalar product definition. We define the scalar product as

$$v_n = \frac{\langle Q_A \cdot Q_B^* \rangle}{\langle M_A \rangle \sqrt{\langle Q_B \cdot Q_B^* \rangle}},$$

where we only invoke two “subevents” separated in rapidity, namely, A and B, corresponding respectively to the regions of interest and reference, and where M_A stands for the multiplicity

TABLE IV. All 20 model parameters for each of the five posterior samples used to make high-statistics posterior predictions throughout this analysis, as well as the MAP parameter set and the prior range for each parameter.

Parameters	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	MAP	Prior
y_2	0.659	1.054	1.820	0.441	0.274	1.610	[0,2]
y_4	1.357	1.684	1.936	1.812	1.032	1.685	$[y_2, 4]$
y_6	1.801	1.977	2.078	2.054	2.048	1.685	$[y_4, 6]$
$\sigma_{y_{\text{loss}}}$	0.990	0.420	0.354	0.774	0.912	0.682	[0,1]
α_{rem}	0.326	0.434	0.232	0.625	0.554	0.536	[0,1]
$\alpha_{\text{shadowing}}$	0.104	0.034	0.032	0.122	0.095	0.001	[0,1]
$B_G [\text{GeV}^{-2}]$	24.020	6.020	21.860	4.440	21.220	3.960	[2,25]
$\sigma_x [\text{fm}]$	0.105	0.100	0.287	0.210	0.107	0.207	[0.1,0.5]
σ_η	0.294	0.473	0.393	0.566	0.284	0.458	[0.1,0.8]
α_{shift}	0.927	0.779	0.075	0.510	0.011	0.477	[0,1]
$\tau_{\text{form}} [\text{fm}]$	0.822	0.453	0.219	0.557	0.952	0.424	[0.2,1]
$(\eta/s) T_{\text{kink}} [\text{GeV}]$	0.277	0.227	0.229	0.260	0.200	0.206	[0.13,0.3]
$m_{\text{low}} [\text{GeV}^{-1}]$	0.838	-0.766	-1.705	-1.396	-0.779	-1.999	$[-2,1]$
$m_{\text{high}} [\text{GeV}^{-1}]$	0.500	1.351	1.880	1.218	1.211	1.999	$[-1,2]$
$(\eta/s)_{\text{kink}}$	0.123	0.145	0.041	0.063	0.036	0.108	[0.01,0.2]
$(\zeta/s)_{\text{max}}$	0.147	0.103	0.055	0.090	0.134	0.092	[0.01,0.2]
$(\zeta/s) T_{\text{max}} [\text{GeV}]$	0.130	0.153	0.232	0.169	0.152	0.180	[0.12,0.3]
$w_\zeta [\text{GeV}]$	0.068	0.081	0.127	0.026	0.055	0.067	[0.025,0.15]
λ_ζ	-0.358	-0.644	-0.334	0.004	-0.673	-0.799	$[-0.8,0.6]$
$e_{\text{switch}} [\text{GeV}/\text{fm}^3]$	0.591	0.527	0.476	0.423	0.588	0.520	[0.1,0.6]

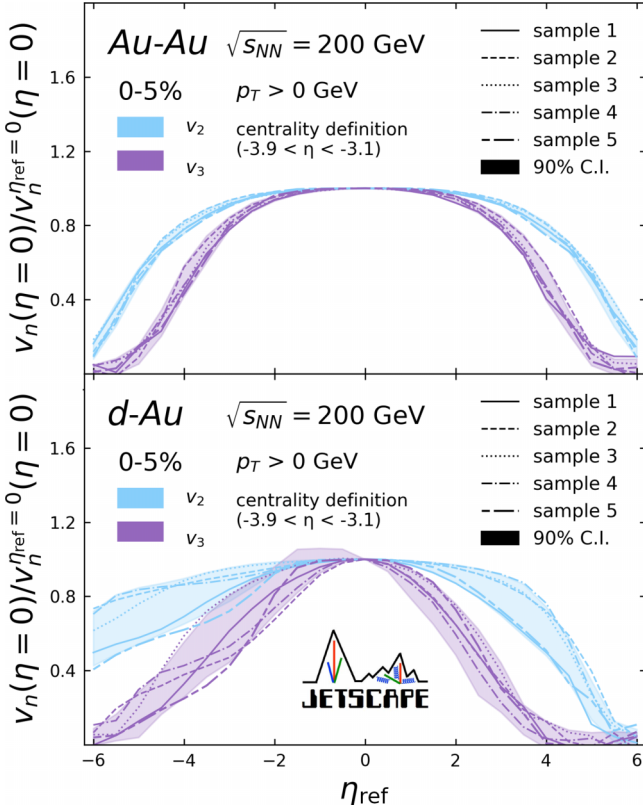


FIG. 34. The calibrated model calculations for flow coefficients in Au-Au and d -Au at midrapidity as a function of the reference rapidity region, scaled by the value of the coefficients using a midrapidity reference. The calculations are plotted for five individual parameter samples of the posterior simulated for ten thousand events each. In addition, a 90% credible interval band, estimated from ten such high-statistics random samples, is shown.

ity in subevent A and the averages are taken over all events in a centrality bin.

In defining the anisotropy coefficient this way, we make the simplification, used throughout the analysis, of correlating particles from the same subevent to compute the reference correlator $Q_B \cdot Q_B^*$. Thus, among the methodological differences with various flow measurements that we treat as negligible in this analysis, we also include the contribution of autocorrelations to this denominator. In Fig. 36 we show that the contribution from such autocorrelations is indeed negligible for both Au-Au and (especially) d -Au, even for small values of the v_n . The largest impact, as shown in this figure, is on $v_3(\eta)$, an observable which does not in fact feature in our analysis in η -differential form and for which we in any case have relatively large emulation uncertainties. Lastly, since the region of interest and reference do not in general overlap in the analysis, we do not have an autocorrelation contribution coming from the numerator correlator.

APPENDIX N: A NOTE ON MAKING LHC PREDICTIONS

The current analysis constrains our 3D hydrodynamic model with data collected at the RHIC top energy. While the

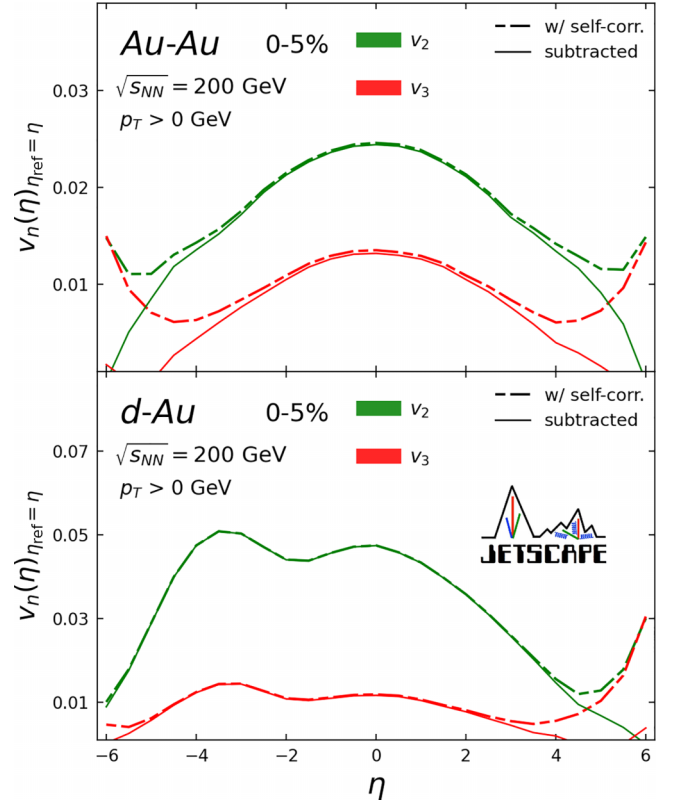


FIG. 35. The calibrated model calculations for $v_2(\eta)$ and $v_3(\eta)$ in Au-Au (top) and d -Au (bottom) for the case where the reference rapidity region coincides with the rapidity bin of interest, for all η , in bins of 0.5 units. The calculations are shown for sample 3 of the default posterior distribution for 0%–5% most central collisions in both systems. The dashed lines represent the calculation with the contribution due to autocorrelations among particles included (labeled “w/ self-corr.”), while the solid lines represent the calculation with the contribution explicitly subtracted (labeled “subtracted”).

sensitivity of the rapidity loss parameters varies with beam energy (and additional constraints must thus be obtained at the LHC energy), one might reasonably assume that the remaining parameters have no explicit dependence on the beam energy. Under this assumption, then, we can make predictions for LHC observables starting from our current posterior.

While, as we have done throughout this paper, it is best to make model predictions using a sampling of the posterior (in order to show a realistic quantification of the uncertainty on those predictions), this may not be practical if an additional fitting procedure is required. Thus, given this is the case for the rapidity loss when making LHC predictions, it may be best to use only the MAP parameter set as a starting point. We lay out here the parameter set resulting from such a fit, which includes not only a y_{10} parameter for the rapidity loss at $y_{\text{beam}} = 10$ (with $y_{\text{LHC}} = 8.58$ for 5.02 TeV center-of-mass energy), but additionally a refitted e_{switch} . The latter parameter is not of strong significance for predicting most LHC observables. However, given the exclusion of proton yield data in our current calibration, this observable can be well-described at the LHC (as well as at RHIC) by our model only if the

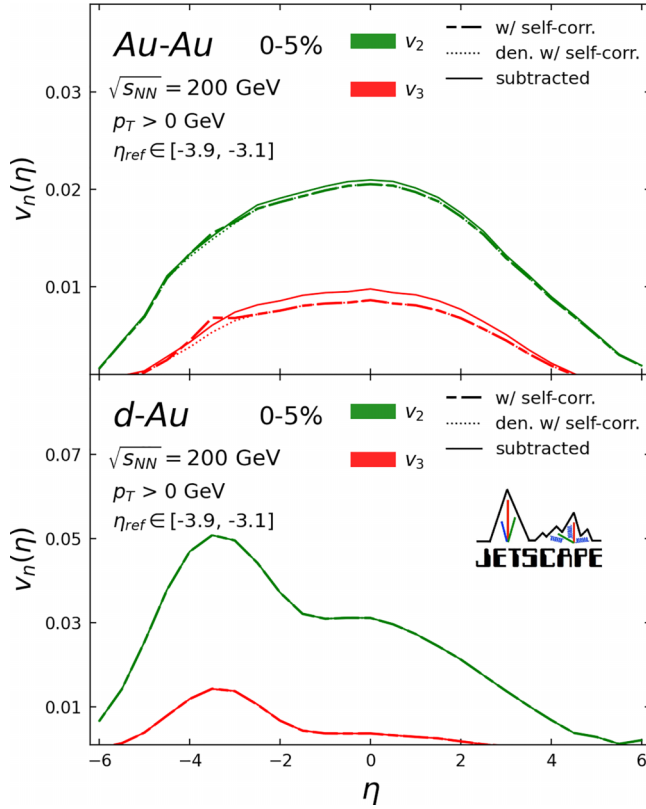


FIG. 36. The calibrated model calculations for $v_2(\eta)$ and $v_3(\eta)$ in Au-Au (top) and d-Au (bottom) using the generic PHENIX Beam-Beam Counter reference region ($[-3.9, -3.1]$), for all η in bins of 0.5 units. The calculations are shown for sample 3 of the default posterior distribution for 0%–5% most central collisions in both systems. The dashed lines represent the calculation with the contribution due to autocorrelations included (labeled “w/ self-corr.”), while the dotted lines show the observables having subtracted only the numerator autocorrelations (which are nonzero only for η in the Beam-Beam Counter region) labeled as “den. w/ self-corr.” The solid lines (labeled “subtracted”) meanwhile represent the fully subtracted calculations, additionally removing autocorrelations originating from the denominator correlator. Note that the subtraction from the denominator naturally leads to a larger v_n . In the d-Au (bottom) panel all three curves essentially overlap. In the Au-Au (top) panel, where the effect is more significant, the dashed and dotted lines overlap everywhere except in the η -region that coincides with the reference, because the Q vectors in the numerator correlator overlap in only that region.

switching energy density is refit to a lower value. Previous studies found that the switching point is indeed especially sensitive to identified particle observables [7, 10].

The parameter set resulting from this LHC fit to Pb-Pb data is shown in Table V and, again, features only two changes from the MAP quoted in Table IV.

While this LHC tune is not a direct prediction from our posterior, it is a fit directly informed by our calibration’s best fit to RHIC top energy data. For small LHC systems such as p-Pb, O-O, or Ne-Ne, the same (or a similar) fitting procedure may be used to make predictions. The model may be run through the iEBE-MUSIC [22] or X-SCAPE frameworks [67].

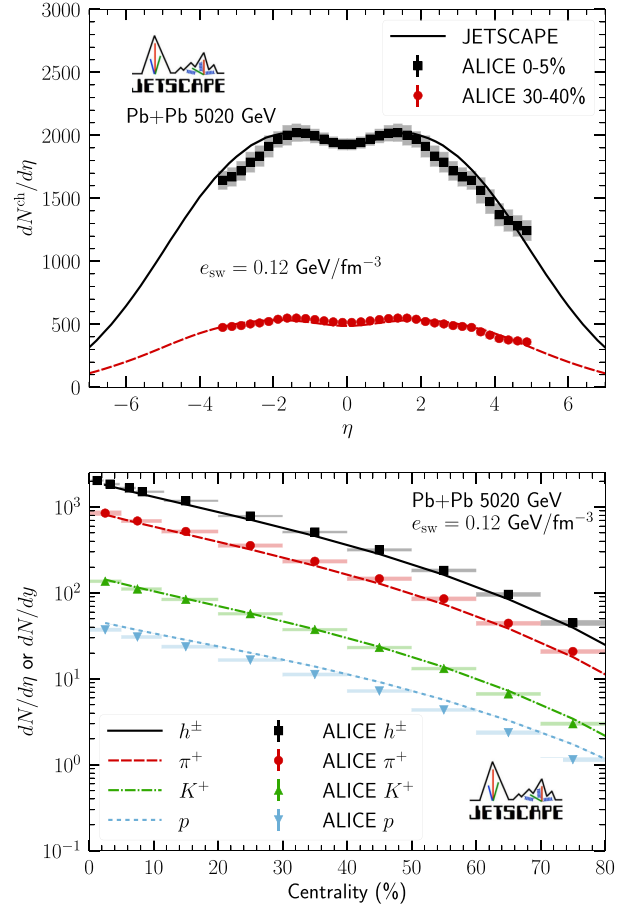


FIG. 37. The model calculations of the multiplicity as a function of η and the integrated charged and identified hadrons for Pb-Pb at 5.02 TeV for the retuned parameter set obtained from our calibration’s MAP compared with experimental measurements from ALICE.

A comparison of LHC predictions using this tune and the experimental data from ALICE for the Pb-Pb $dN/d\eta$ and integrated charged and identified particle yields may be seen in Fig. 37.

TABLE V. All 21 model parameters for the retuned MAP parameter set to fit 5.02 TeV Pb-Pb LHC data.

Parameters	MAP	Parameters	MAP
y_2	1.610	τ_{form} [fm]	0.424
y_4	1.685	$(\eta/s) T_{\text{kink}}$ [GeV]	0.206
y_6	1.685	m_{low} [GeV $^{-1}$]	−1.999
y_{10}	2.800	m_{high} [GeV $^{-1}$]	1.999
σ_{yloss}	0.682	$(\eta/s)_{\text{kink}}$	0.108
α_{rem}	0.536	$(\zeta/s)_{\text{max}}$	0.092
$\alpha_{\text{shadowing}}$	0.001	$(\zeta/s) T_{\text{max}}$ [GeV]	0.180
B_G [GeV $^{-2}$]	3.960	w_ζ [GeV]	0.067
σ_x [fm]	0.207	λ_ζ	−0.799
σ_η	0.458	e_{switch} [GeV/fm 3]	0.120
α_{shift} [fm]	0.477		

- [1] I. Arsene *et al.* (BRAHMS Collaboration), Quark gluon plasma and color glass condensate at RHIC? The perspective from the BRAHMS experiment, *Nucl. Phys. A* **757**, 1 (2005).
- [2] B. B. Back *et al.* (PHOBOS Collaboration), The PHOBOS perspective on discoveries at RHIC, *Nucl. Phys. A* **757**, 28 (2005).
- [3] J. Adams *et al.* (STAR Collaboration), Experimental and theoretical challenges in the search for the quark gluon plasma: The STAR Collaboration's critical assessment of the evidence from RHIC collisions, *Nucl. Phys. A* **757**, 102 (2005).
- [4] K. Adcox *et al.* (PHENIX Collaboration), Formation of dense partonic matter in relativistic nucleus-nucleus collisions at RHIC: Experimental evaluation by the PHENIX collaboration, *Nucl. Phys. A* **757**, 184 (2005).
- [5] B. Müller, J. Schukraft, and B. Wyslouch, First results from Pb+Pb collisions at the LHC, *Annu. Rev. Nucl. Part. Sci.* **62**, 361 (2012).
- [6] 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).
- [7] J. Bernhard, J. Moreland, and S. Bass, Bayesian estimation of the specific shear and bulk viscosity of quark-gluon plasma, *Nat. Phys.* **15**, 1113 (2019).
- [8] W. Ke, J. S. Moreland, J. E. Bernhard, and S. A. Bass, Constraints on rapidity-dependent initial conditions from charged particle pseudorapidity densities and two-particle correlations, *Phys. Rev. C* **96**, 044912 (2017).
- [9] J. Auvinen, J. E. Bernhard, S. A. Bass, and I. Karpenko, Investigating the collision energy dependence of η/s in the beam energy scan at the BNL Relativistic Heavy Ion Collider using Bayesian statistics, *Phys. Rev. C* **97**, 044905 (2018).
- [10] D. Everett *et al.* (JETSCAPE Collaboration), Multisystem Bayesian constraints on the transport coefficients of QCD matter, *Phys. Rev. C* **103**, 054904 (2021).
- [11] 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).
- [12] G. Nijs, W. van der Schee, U. Gürsoy, and R. Snellings, Bayesian analysis of heavy ion collisions with the heavy ion computational framework Trajectum, *Phys. Rev. C* **103**, 054909 (2021).
- [13] G. Nijs, W. van der Schee, U. Gürsoy, and R. Snellings, Transverse momentum differential global analysis of heavy-ion collisions, *Phys. Rev. Lett.* **126**, 202301 (2021).
- [14] G. Nijs and W. van der Schee, Hadronic nucleus-nucleus cross section and the nucleon size, *Phys. Rev. Lett.* **129**, 232301 (2022).
- [15] 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*.
- [16] M. Virta, J. Parkkila, and D. J. Kim, Enhancing Bayesian parameter estimation by adapting to multiple energy scales in heavy-ion collisions at RHIC and at the LHC, *Phys. Rev. C* **111**, 044903 (2025).
- [17] J. E. Parkkila, A. Onnerstad, and D. J. Kim, Bayesian estimation of the specific shear and bulk viscosity of the quark-gluon plasma with additional flow harmonic observables, *Phys. Rev. C* **104**, 054904 (2021).
- [18] Z. Yang and L.-W. Chen, Bayesian inference of the specific shear and bulk viscosities of the quark-gluon plasma at crossover from ϕ and Ω observables, *Phys. Rev. C* **107**, 064910 (2023).
- [19] M. R. Heffernan, C. Gale, S. Jeon, and J.-F. Paquet, Early-times Yang-Mills dynamics and the characterization of strongly interacting matter with statistical learning, *Phys. Rev. Lett.* **132**, 252301 (2024).
- [20] M. R. Heffernan, C. Gale, S. Jeon, and J.-F. Paquet, Bayesian quantification of strongly interacting matter with color glass condensate initial conditions, *Phys. Rev. C* **109**, 065207 (2024).
- [21] S. A. Jahan, H. Roch, and C. Shen, Bayesian analysis of (3+1)D relativistic nuclear dynamics with the RHIC beam energy scan data, *Phys. Rev. C* **110**, 054905 (2024).
- [22] C. Shen, B. Schenke, and W. Zhao, Viscosities of the baryon-rich quark-gluon plasma from beam energy scan data, *Phys. Rev. Lett.* **132**, 072301 (2024).
- [23] C. Shen, B. Schenke, and W. Zhao, The effects of pseudorapidity-dependent observables on (3+1)D Bayesian inference of relativistic heavy-ion collisions, *EPJ Web Conf.* **296**, 14001 (2024).
- [24] N. Götz, I. Karpenko, and H. Elfner, Bayesian analysis of a (3+1)D hybrid approach with initial conditions from hadronic transport, *Phys. Rev. C* **112**, 014910 (2025).
- [25] A. Mankolli, 3D multi-system Bayesian calibration with energy conservation to study rapidity-dependent dynamics of nuclear collisions, *EPJ Web Conf.* **296**, 05010 (2024).
- [26] M. Heffernan, Quantification of the quark-gluon plasma with statistical learning, Ph.D. thesis, McGill University, 2023.
- [27] D. S. Everett, Quantifying the quark gluon plasma, Ph.D. thesis, Ohio State University, 2021.
- [28] J. Parkkila, Quantifying the transport properties of quark-gluon plasma through measurement of higher harmonic flow and their non-linear response, Ph.D. thesis, Jyväskylä University, 2021.
- [29] J.-F. Paquet, Applications of emulation and Bayesian methods in heavy-ion physics, *J. Phys. G* **51**, 103001 (2024).
- [30] V. Khachatryan *et al.* (CMS Collaboration), Observation of long-range near-side angular correlations in proton-proton collisions at the LHC, *J. High Energy Phys.* **09** (2010) 091.
- [31] G. Aad *et al.* (ATLAS Collaboration), Observation of long-range elliptic azimuthal anisotropies in $\sqrt{s} = 13$ and 2.76 TeV pp collisions with the ATLAS detector, *Phys. Rev. Lett.* **09** (2016) 172301.
- [32] S. Chatrchyan *et al.* (CMS Collaboration), Observation of long-range near-side angular correlations in proton-lead collisions at the LHC, *Phys. Lett. B* **718**, 795 (2013).
- [33] G. Aad *et al.* (ATLAS Collaboration), Observation of associated near-side and away-side long-range correlations in $\sqrt{s_{NN}} = 5.02$ TeV proton-lead collisions with the ATLAS detector, *Phys. Rev. Lett.* **110**, 182302 (2013).
- [34] B. Abelev *et al.* (ALICE Collaboration), Long-range angular correlations on the near and away side in p -Pb collisions at $\sqrt{s_{NN}} = 5.02$ TeV, *Phys. Lett. B* **719**, 29 (2013).
- [35] A. Adare *et al.* (PHENIX Collaboration), Quadrupole anisotropy in dihadron azimuthal correlations in central $d + Au$ collisions at $\sqrt{s_{NN}} = 200$ GeV, *Phys. Rev. Lett.* **111**, 212301 (2013).

- [36] J. L. Nagle and W. A. Zajc, Small system collectivity in relativistic hadronic and nuclear collisions, *Annu. Rev. Nucl. Part. Sci.* **68**, 211 (2018).
- [37] U. W. Heinz and J. S. Moreland, Hydrodynamic flow in small systems or: “How the heck is it possible that a system emitting only a dozen particles can be described by fluid dynamics?”, *J. Phys.: Conf. Ser.* **1271**, 012018 (2019).
- [38] C. Aidala *et al.* (PHENIX Collaboration), Creation of quark–gluon plasma droplets with three distinct geometries, *Nat. Phys.* **15**, 214 (2019).
- [39] M. I. Abdulhamid *et al.* (STAR Collaboration), Measurement of flow coefficients in high-multiplicity $p + \text{Au}$, $d + \text{Au}$, and $^3\text{He} + \text{Au}$ collisions at $\sqrt{s_{\text{NN}}} = 200 \text{ GeV}$, *Phys. Rev. C* **110**, 064902 (2024).
- [40] M. Nie (STAR Collaboration), Energy dependence of longitudinal flow decorrelation from STAR, *Nucl. Phys. A* **1005**, 121783 (2021).
- [41] J. Jia, S. Huang, C. Zhang, and S. Bhatta, Sources of longitudinal flow decorrelations in high-energy nuclear collisions, [arXiv:2408.15006](https://arxiv.org/abs/2408.15006).
- [42] S. A. Bass and A. Dumitru, Dynamics of hot bulk QCD matter: From the quark gluon plasma to hadronic freezeout, *Phys. Rev. C* **61**, 064909 (2000).
- [43] D. Teaney, J. Lauret, and E. V. Shuryak, A hydrodynamic description of heavy ion collisions at the SPS and RHIC, [arXiv:nucl-th/0110037](https://arxiv.org/abs/nucl-th/0110037).
- [44] U. Heinz and R. Snellings, Collective flow and viscosity in relativistic heavy-ion collisions, *Annu. Rev. Nucl. Part. Sci.* **63**, 123 (2013).
- [45] C. Gale, S. Jeon, and B. Schenke, Hydrodynamic modeling of heavy-ion collisions, *Int. J. Mod. Phys. A* **28**, 1340011 (2013).
- [46] R. Derradi de Souza, T. Koide, and T. Kodama, Hydrodynamic approaches in relativistic heavy ion reactions, *Prog. Part. Nucl. Phys.* **86**, 35 (2016).
- [47] J. D. Bjorken, Highly relativistic nucleus-nucleus collisions: The central rapidity region, *Phys. Rev. D* **27**, 140 (1983).
- [48] C. Shen and B. Schenke, Dynamical initial state model for relativistic heavy-ion collisions, *Phys. Rev. C* **97**, 024907 (2018).
- [49] C. Shen and B. Schenke, Longitudinal dynamics and particle production in relativistic nuclear collisions, *Phys. Rev. C* **105**, 064905 (2022).
- [50] S. Ryu, B. Schenke, C. Shen, and W. Zhao, The role of longitudinal decorrelations for measurements of anisotropic flow in small collision systems, *EPJ Web Conf.* **296**, 15001 (2024).
- [51] B. Schenke, S. Jeon, and C. Gale, (3 + 1)D hydrodynamic simulation of relativistic heavy-ion collisions, *Phys. Rev. C* **82**, 014903 (2010).
- [52] B. Schenke, S. Jeon, and C. Gale, Elliptic and triangular flow in event-by-event (3 + 1)D viscous hydrodynamics, *Phys. Rev. Lett.* **106**, 042301 (2011).
- [53] J.-F. Paquet, C. Shen, G. S. Denicol, M. Luzum, B. Schenke, S. Jeon, and C. Gale, Production of photons in relativistic heavy-ion collisions, *Phys. Rev. C* **93**, 044906 (2016).
- [54] S. A. Bass *et al.*, Microscopic models for ultrarelativistic heavy ion collisions, *Prog. Part. Nucl. Phys.* **41**, 255 (1998).
- [55] M. Bleicher *et al.*, Relativistic hadron-hadron collisions in the ultra-relativistic quantum molecular dynamics model, *J. Phys. G* **25**, 1859 (1999).
- [56] P. M. Chesler and L. G. Yaffe, Holography and off-center collisions of localized shock waves, *J. High Energy Phys.* **10** (2015) 070.
- [57] W. Busza, K. Rajagopal, and W. van der Schee, Heavy ion collisions: The big picture, and the big questions, *Annu. Rev. Nucl. Part. Sci.* **68**, 339 (2018).
- [58] S. Schlichting and D. Teaney, The first fm/c of heavy-ion collisions, *Annu. Rev. Nucl. Part. Sci.* **69**, 447 (2019).
- [59] J. Berges, M. P. Heller, A. Mazeliauskas, and R. Venugopalan, QCD thermalization: *Ab initio* approaches and interdisciplinary connections, *Rev. Mod. Phys.* **93**, 035003 (2021).
- [60] H. Song, S. A. Bass, U. Heinz, T. Hirano, and C. Shen, 200 A GeV Au + Au collisions serve a nearly perfect quark-gluon liquid, *Phys. Rev. Lett.* **106**, 192301 (2011); **109**, 139904(E) (2012).
- [61] J. S. Moreland, J. E. Bernhard, and S. A. Bass, Alternative ansatz to wounded nucleon and binary collision scaling in high-energy nuclear collisions, *Phys. Rev. C* **92**, 011901 (2015).
- [62] M. L. Miller, K. Reygers, S. J. Sanders, and P. Steinberg, Glauber modeling in high energy nuclear collisions, *Annu. Rev. Nucl. Part. Sci.* **57**, 205 (2007).
- [63] C. Loizides, Glauber modeling of high-energy nuclear collisions at the subnucleon level, *Phys. Rev. C* **94**, 024914 (2016).
- [64] B. Alver, M. Baker, C. Loizides, and P. Steinberg, The PHOBOS Glauber Monte Carlo, [arXiv:0805.4411](https://arxiv.org/abs/0805.4411).
- [65] C. Loizides, J. Nagle, and P. Steinberg, Improved version of the PHOBOS Glauber Monte Carlo, *SoftwareX* **1-2**, 13 (2015).
- [66] C. Shen and S. Alzhrani, Collision-geometry-based 3D initial condition for relativistic heavy-ion collisions, *Phys. Rev. C* **102**, 014909 (2020).
- [67] I. Soudi *et al.* (JETSCAPE Collaboration), Soft-hard framework with exact four-momentum conservation for small systems, *Phys. Rev. C* **112**, 014905 (2025).
- [68] K. J. Eskola, H. Paukkunen, and C. A. Salgado, EPS09: A new generation of NLO and LO nuclear parton distribution functions, *J. High Energy Phys.* **04** (2009) 065.
- [69] A. Bazavov *et al.* (HotQCD Collaboration), Equation of state in (2 + 1)-flavor QCD, *Phys. Rev. D* **90**, 094503 (2014).
- [70] F. Cooper and G. Frye, Comment on the single particle distribution in the hydrodynamic and statistical thermodynamic models of multiparticle production, *Phys. Rev. D* **10**, 186 (1974).
- [71] C. Shen, Z. Qiu, H. Song, J. Bernhard, S. Bass, and U. Heinz, The iEBE-VISHNU code package for relativistic heavy-ion collisions, *Comput. Phys. Commun.* **199**, 61 (2016).
- [72] H. Grad, On the kinetic theory of rarefied gases, *Commun. Pure Appl. Math.* **2**, 331 (1949).
- [73] W. Israel, Nonstationary irreversible thermodynamics: A causal relativistic theory, *Ann. Phys. (NY)* **100**, 310 (1976).
- [74] W. Israel and J. M. Stewart, Transient relativistic thermodynamics and kinetic theory, *Ann. Phys. (NY)* **118**, 341 (1979).
- [75] A. Monnai and T. Hirano, Effects of bulk viscosity at freeze-out, *Phys. Rev. C* **80**, 054906 (2009).
- [76] G. S. Denicol, H. Niemi, E. Molnar, and D. H. Rischke, Derivation of transient relativistic fluid dynamics from the Boltzmann equation, *Phys. Rev. D* **85**, 114047 (2012); **91**, 039902(E) (2015).
- [77] D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman, EMCEE: The MCMC Hammer, *Publ. Astron. Soc. Pac.* **125**, 306 (2013).
- [78] B. Alver *et al.* (PHOBOS Collaboration), Phobos results on charged particle multiplicity and pseudorapidity distributions

- in Au + Au, Cu + Cu, $d + Au$, and $p + p$ collisions at ultra-relativistic energies, *Phys. Rev. C* **83**, 024913 (2011).
- [79] B. B. Back *et al.* (PHOBOS Collaboration), Centrality and pseudorapidity dependence of elliptic flow for charged hadrons in Au + Au collisions at $\sqrt{s_{NN}} = 200$ GeV, *Phys. Rev. C* **72**, 051901 (2005).
- [80] J. Adams *et al.* (STAR Collaboration), Azimuthal anisotropy in Au + Au collisions at $\sqrt{s_{NN}} = 200$ GeV, *Phys. Rev. C* **72**, 014904 (2005).
- [81] A. Adare *et al.* (PHENIX Collaboration), Measurements of higher-order flow harmonics in Au + Au collisions at $\sqrt{s_{NN}} = 200$ GeV, *Phys. Rev. Lett.* **107**, 252301 (2011).
- [82] L. Adamczyk *et al.* (STAR Collaboration), Third harmonic flow of charged particles in Au + Au collisions at $\sqrt{s_{NN}} = 200$ GeV, *Phys. Rev. C* **88**, 014904 (2013).
- [83] B. I. Abelev, *et al.* (STAR Collaboration), Systematic measurements of identified particle spectra in $p + p$, $d + Au$ and Au + Au collisions from STAR, *Phys. Rev. C* **79**, 034909 (2009).
- [84] S. S. Adler, *et al.* (PHENIX Collaboration), Identified charged particle spectra and yields in Au + Au collisions at $\sqrt{s_{NN}} = 200$ GeV, *Phys. Rev. C* **69**, 034909 (2004).
- [85] A. Adare *et al.* (PHENIX Collaboration), Pseudorapidity dependence of particle production and elliptic flow in asymmetric nuclear collisions of $p + Al$, $p + Au$, $d + Au$, and $^3He + Au$ at $\sqrt{s_{NN}} = 200$ GeV, *Phys. Rev. Lett.* **121**, 222301 (2018).
- [86] M. I. Abdulhamid, *et al.* (STAR Collaboration), Measurements of the elliptic and triangular azimuthal anisotropies in central $^3He + Au$, $d + Au$ and $p + Au$ collisions at $\sqrt{s_{NN}} = 200$ GeV, *Phys. Rev. Lett.* **130**, 242301 (2023).
- [87] I. G. Bearden *et al.* (BRAHMS Collaboration), Pseudorapidity distributions of charged particles from Au+Au collisions at the maximum RHIC energy, *Phys. Rev. Lett.* **88**, 202301 (2002).
- [88] S. S. Adler, *et al.* (PHENIX Collaboration), Transverse-energy distributions at midrapidity in $p+p$, $d+Au$, and Au+Au collisions at $\sqrt{s_{NN}} = 62.4$ –200 GeV and implications for particle-production models, *Phys. Rev. C* **89**, 044905 (2014).
- [89] B. B. Back *et al.* (PHOBOS Collaboration), Scaling of charged particle production in $d + Au$ collisions at $\sqrt{s_{NN}} = 200$ GeV, *Phys. Rev. C* **72**, 031901 (2005).
- [90] M. Luzum and J.-Y. Ollitrault, Eliminating experimental bias in anisotropic-flow measurements of high-energy nuclear collisions, *Phys. Rev. C* **87**, 044907 (2013).
- [91] P. Božek, W. Broniowski, and J. Moreira, Torqued fireballs in relativistic heavy-ion collisions, *Phys. Rev. C* **83**, 034911 (2011).
- [92] J. Jia and P. Huo, Forward-backward eccentricity and participant-plane angle fluctuations and their influences on longitudinal dynamics of collective flow, *Phys. Rev. C* **90**, 034915 (2014).
- [93] J. Jia, P. Huo, G. Ma, and M. Nie, Observables for longitudinal flow correlations in heavy-ion collisions, *J. Phys. G* **44**, 075106 (2017).
- [94] W. Zhao, S. Ryu, C. Shen, and B. Schenke, 3D structure of anisotropic flow in small collision systems at energies available at the BNL Relativistic Heavy Ion Collider, *Phys. Rev. C* **107**, 014904 (2023).
- [95] V. Khachatryan *et al.* (CMS Collaboration), Evidence for transverse momentum and pseudorapidity dependent event plane fluctuations in PbPb and pPb collisions, *Phys. Rev. C* **92**, 034911 (2015).
- [96] M. Aaboud *et al.* (ATLAS Collaboration), Measurement of longitudinal flow decorrelations in Pb+Pb collisions at $\sqrt{s_{NN}} = 2.76$ and 5.02 TeV with the ATLAS detector, *Eur. Phys. J. C* **78**, 142 (2018).
- [97] U. A. Acharya *et al.* (PHENIX Collaboration), Kinematic dependence of azimuthal anisotropies in $p + Au$, $d + Au$, and $^3He + Au$ at $\sqrt{s_{NN}} = 200$ GeV, *Phys. Rev. C* **105**, 024901 (2022).
- [98] N. J. Abdulameer *et al.* (PHENIX Collaboration), Measurements of second-harmonic Fourier coefficients from azimuthal anisotropies in $p + p$, $p + Au$, $d + Au$, and $^3He + Au$ collisions at $\sqrt{s_{NN}} = 200$ GeV, *Phys. Rev. C* **107**, 024907 (2023).
- [99] K. Li, S. Mak, J.-F. Paquet, and S. A. Bass, Additive multi-index Gaussian process modeling, with application to multi-physics surrogate modeling of the quark-gluon plasma, *J. Am. Stat. Assoc.* **121**, 1 (2025).
- [100] Y. Ji, H. S. Yuchi, D. Soeder, J.-F. Paquet, S. A. Bass, V. R. Joseph, C. F. J. Wu, and S. Mak, Conglomerate multi-fidelity Gaussian process modeling, with application to heavy-ion collisions, *SIAM/ASA J. Uncertainty Quantif.* **12**, 473 (2024).
- [101] Y. Ji, S. Mak, D. Soeder, J.-F. Paquet, and S. A. Bass, A graphical multi-fidelity Gaussian process model, with application to emulation of heavy-ion collisions, *Technometrics* **66**, 267 (2024).
- [102] S. A. Jahan, H. Roch, and C. Shen, Bayesian model selection and uncertainty propagation for beam energy scan heavy-ion collisions, *Phys. Rev. C* **113**, 024919 (2026).
- [103] A. Mankolli, Parameter posterior chains from 3D Bayesian calibration with RHIC data, Zenodo, 2026, [10.5281/zenodo.18602345](https://doi.org/10.5281/zenodo.18602345) (2026).
- [104] T. J. Boerner, S. Deems, T. R. Furlani, S. L. Knuth, and J. Towns, ACCESS: Advancing Innovation: NSF's advanced cyberinfrastructure coordination ecosystem: Services & support, in *Practice and Experience in Advanced Research Computing 2023: Computing for the Common Good*, PEARC '23 (Association for Computing Machinery, New York, 2023), pp. 173–176.
- [105] R. Pordes, D. Petravick, B. Kramer, D. Olson, M. Livny, A. Roy, P. Avery, K. Blackburn, T. Wenaus, F. Würthwein, I. Foster, R. Gardner, M. Wilde, A. Blatecky, J. McGee, and R. Quick, The open science grid, in *J. Phys. Conf. Ser.* **78**, 012057 (2007).
- [106] I. Sfiligoi, D. C. Bradley, B. Holzman, P. Mhashikar, S. Padhi, and F. Würthwein, The pilot way to grid resources using glideinwms, in *2009 WRI World Congress on Computer Science and Information Engineering*, 2 (IEEE, Los Angeles, CA, 2009), Vol. 2, pp. 428–432.
- [107] OSG, Osgpool (2006), doi:[10.21231/906P-4D78](https://doi.org/10.21231/906P-4D78).
- [108] OSG, Open science data federation (2015), doi:[10.21231/OKVZ-VE57](https://doi.org/10.21231/OKVZ-VE57).
- [109] J. E. Bernhard, Bayesian parameter estimation for relativistic Heavy-Ion Collisions, Ph.D. thesis, Duke University, 2018.
- [110] R. A. Soltz, D. A. Hangal, and A. Angerami, Simple model to investigate jet quenching and correlated errors for centrality-dependent nuclear modification factors in relativistic heavy-ion collisions, *Phys. Rev. C* **111**, 034911 (2025).