Quantifying correlated truncation errors in effective field theory

J. A. Melendez [*] and R. J. Furnstahl[†]

Department of Physics, The Ohio State University, Columbus, Ohio 43210, USA

D. R. Phillips[‡]

Department of Physics and Astronomy and Institute of Nuclear and Particle Physics, Ohio University, Athens, Ohio 45701, USA;
Institut für Kernphysik, Technische Universität Darmstadt, 64289 Darmstadt, Germany;
and ExtreMe Matter Institute EMMI, GSI Helmholtzzentrum für Schwerionenforschung GmbH, 64291 Darmstadt, Germany

M. T. Pratola[§]

Department of Statistics, The Ohio State University, Columbus, Ohio 43210, USA

S. Wesolowski[‖]

Department of Mathematics and Computer Science, Salisbury University, Salisbury, Maryland 21801, USA

Effective field theories (EFTs) organize the description of complex systems into an infinite sequence of decreasing importance. Predictions are made with a finite number of terms, which induces a truncation error that is often left unquantified. We formalize the notion of EFT convergence and propose a Bayesian truncation error model for predictions that are correlated across the independent variables, e.g., energy or scattering angle. Central to our approach are Gaussian processes that encode both the naturalness and correlation structure of EFT coefficients. Our use of Gaussian processes permits efficient and accurate assessment of credible intervals, allows EFT fits to easily include correlated theory errors, and provides analytic posteriors for physical EFT-related quantities such as the expansion parameter. We demonstrate that model-checking diagnostics—applied to the case of multiple curves—are powerful tools for EFT validation. As an example, we assess a set of nucleon-nucleon scattering observables in chiral EFT. In an effort to be self-contained, appendices include thorough derivations of our statistical results. Our methods are packaged in Python code, called gsum, that is available for download on GitHub.

I. INTRODUCTION

The *power counting* of an effective field theory (EFT) mandates how to organize an infinite number of operators into a sequence of decreasing importance. The precision of EFT predictions is then governed by the uncertainty in the following quantities: (1) the fit of the parameters, or low-energy constants (LECs), of the EFT, (2) the error due to truncation of the infinite EFT series, and (3) other approximations made in prediction. For chiral EFT calculations, recent advances in precision motivate a rigorous accounting of all these theoretical uncertainties, but such an accounting is desirable for EFTs across all domains. This work will focus on the uncertainty due to truncation, but our results have consequences for how LECs should be fit [1].

In previous works, we proposed a pointwise Bayesian statistical model to estimate EFT truncation errors for predicted observables y [2,3] (see also Ref. [4]). This model formalizes the notion of convergence in y, which allows one to credibly assert which observations are consistent with theory. One can incorporate expert knowledge on the convergence pattern (inherited from the EFT power counting) into prior distributions and subsequently update these beliefs given order-by-order predictions $\{y_n\}$ [4]. If the observable is itself a function $y(x)$, e.g., a cross section at a range of energies, then Refs. [2,3] simply compute its posterior in a pointwise manner. This pointwise model is tractable, but it is flawed since it ignores information about $y(x)$ as a whole, that is, the correlations of y at nearby x.

Here we extend the pointwise model to functions $y(x)$, encoding the idea of curvewise convergence for observables via Gaussian processes (GPs). GPs are powerful tools for both regression and classification and have become popular in statistics, physics, applied mathematics, machine learning, and geostatistics [5–7]. Their popularity is due in part to their modeling flexibility and the mathematical convenience of Gaussian distributions. Despite the leap from points to curves, our algorithm remains analytic due to a convenient yet flexible

---

[*]melendez.27@osu.edu
[†]furnstahl.1@osu.edu
[‡]phillid1@ohio.edu
[§]mpratola@stat.osu.edu
[‖]scwesolowski@salisbury.edu

choice of priors. The GP parameters are interpretable from an EFT convergence standpoint and can be easily *calibrated* against known order-by-order predictions.

We develop the intuition behind our convergence model and provide a primer on GPs (Sec. II) and then explore ways to assess whether our assumptions are failing (Sec. III). To display the procedure in action, we show success and failure using toy EFT predictions (Sec. IV) and assess nucleon-nucleon–(*NN*) scattering predictions in chiral EFT (Sec. V). We then discuss future prospects and conclude the main text (Sec. VI) before providing an extended discussion on the statistical derivations relevant to the pointwise and GP models (Appendix).

II. THE MODEL

A. Setup

Predictions of experimental quantities can be inaccurate due to shortcomings of the theoretical framework and noisy measurements. This work focuses on building a model discrepancy term for EFT truncation errors. The exact manner by which specific errors propagate can be complicated, but here we assume a simple additive model of theoretical discrepancy,

$$y_{\text{exp}}(x) = y_{\text{th}}(x; \vec{a}) + \delta y_{\text{th}}(x) + \delta y_{\text{exp}}(x), \qquad (1)$$

where $x \in \mathbb{R}^d$ is the independent variable,[1] $\vec{a}$ are the fitted parameters of the theory prediction y_{th} (here the LECs of an EFT[2]), and y_{exp} are experimental measurements. The uncertainties associated with the theory and experiment are given by δy_{th} and δy_{exp}. Of course, if the exact values of δy_{th} and δy_{exp} were known, then there would be no uncertainty. Instead, δy_{th} and δy_{exp} are treated as random variables, which makes Eq. (1) true in a distributional sense. The simple relationship (1) directly impacts both the fitting protocol and error propagation to observables. For example, if one assumes that $\delta y_{\text{th}} = 0$ and δy_{exp} is normally distributed, then this leads to a least-squares log-likelihood function for $\vec{a}$ given y_{exp} [1]. Here we use general properties of EFTs to develop a Bayesian model for $\delta y_{\text{th}}(x)$ with a nontrivial correlation structure depending on x. Unless otherwise stated, we assume that the EFT has already been fit to experimental data and return to discuss the fitting procedure in Sec. II D.

Denote an observable y calculated with an nth-order EFT (N^{n}LO) as y_n. Suppose that the EFT expansion has only been computed up to kth order. Though y_k may work well as an approximation to reality, we seek a probability distribution for the full summation[3] $y \equiv y_\infty$ based on the convergence pattern

---

[1]The differential cross section, as a function of both energy and scattering angle, is an example where $d = 2$.

[2]Note that the LECs $\vec{a}$ are *not* the same as the observable coefficients c_n introduced in Eq. (3). These parameters $\vec{a}$ do not play an explicit role in our treatment of truncation errors, although they do appear via a brief discussion of EFT parameter fitting in Sec. II D.

[3]We are aware that EFTs are often *asymptotic* in nature. We seek the best possible estimate up to where the sum starts diverging. Our assumption is that the divergence occurs sufficiently beyond the kth order that the truncation model error estimate up to order ∞ is a good approximation.
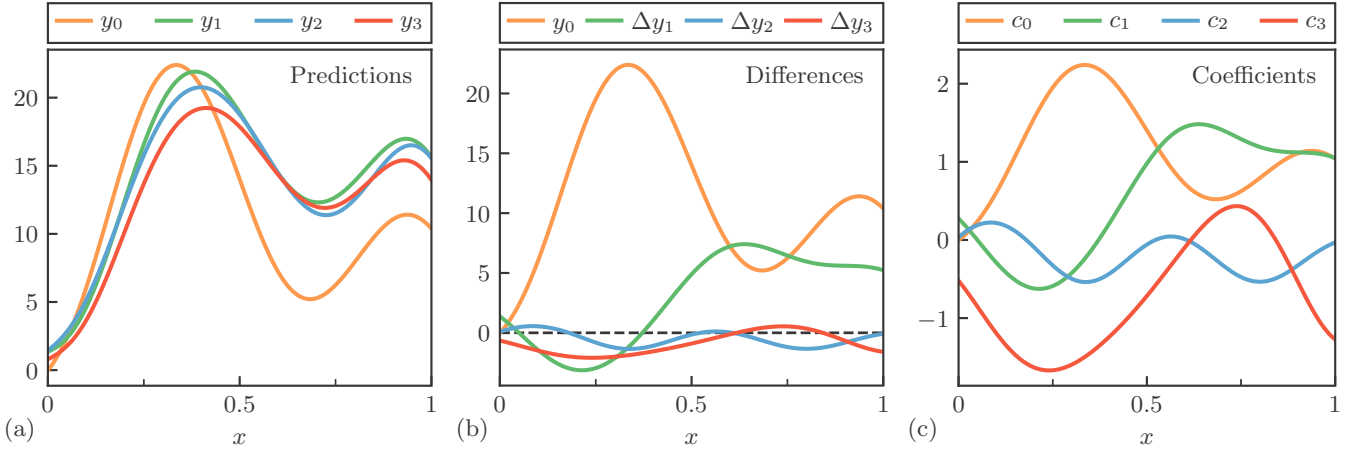
of all known predictions $y_0, y_1, \ldots, y_k$. That is, $y_{\text{th}} = y_k$ and we wish to quantify $\delta y_{\text{th}} = \delta y_k$. This hierarchy of predictions could instead be written as a leading-order calculation y_0 and a set of higher-order corrections Δy_n, with $n \in \{1, 2, \ldots, k\}$. Given this change of variables, the $n \leqslant k$-order prediction is

$$y_n(x) = y_0(x) + \Delta y_1(x) + \cdots + \Delta y_n(x), \qquad (2)$$

where each term in the series is known.

Without further knowledge of the convergence pattern in Eq. (2), it is difficult to make progress in approximating y. To proceed, we employ prior information about the construction of EFTs, that is, if the EFT is working as advertised, then each correction should be roughly suppressed by the dimensionless expansion parameter Q—in accordance with the EFT power counting. Here we assume for simplicity the following power counting: the N^{n}LO correction Δy_n is suppressed by Q^n, though other prescriptions, e.g., Q^{2n}, can be easily incorporated into this framework. That is, up to dimensionful scales, one would generally expect y_0 is $\mathcal{O}(Q^0)$ and Δy_n is $\mathcal{O}(Q^n)$.

Inspired by Eq. (2) and the power counting of the EFT, Ref. [2] proposed the following factorization:

$$y_k(x) = y_{\text{ref}}(x) \sum_{n=0}^{k} c_n(x) Q^n(x). \qquad (3)$$

Here y_{ref} is a dimensionful quantity that sets the scale of variation, while the c_n are dimensionless observable coefficients. Since all scales have been factored into y_{ref} and Q, the c_n should be *natural*, or order 1, assuming they have not been fine tuned. Combinatorial factors, if known, should also appear in Eq. (3), else the naturalness assumption may need to be modified to account for them.

The road-map from continuous observable quantities $y_0, \ldots, y_k$ to coefficients is shown visually for toy EFT predictions in Fig. 1. Predictions for the four lowest orders of the EFT are plotted in Fig. 1(a) and show a steady convergence toward the final result. Suppose that the N^{3}LO EFT is the state of the art; thus our task is to estimate the uncertainty in y_3. The change of variables described by Eq. (2) is illustrated in Fig. 1(b). Note that there is a clear hierarchy to the corrections: They tend to become smaller as order increases. Assume that y_{ref} is known, say, from dimensional analysis, and Q is known from theory arguments. The coefficients of the expansion can then be extracted using Eq. (3), as shown in Fig. 1(c). *The hierarchy has vanished, and we are left with a set of curves that appear randomly drawn from an underlying process.* Reference [3] displays c_n for a collection of *NN*-scattering observables in chiral EFT, many of which exhibit these properties.

Before moving on, we would like to correct some common misconceptions about Eq. (3). We have made no assumptions about the structure of the expansion thus far. In fact, Eq. (3) is *not* a formal expansion of y in powers of Q. Rather, we have only noted that if we assume values for y_{ref} and Q along with the factorization of Eq. (3), then there is a one-to-one correspondence between the predictions $\{y_0, \ldots, y_k\}$ and the coefficients c_n given by

$$y_0(x) \equiv y_{\text{ref}}(x) c_0(x)$$
$$\Delta y_n(x) \equiv y_{\text{ref}}(x) c_n(x) Q^n(x). \qquad (4)$$

FIG. 1. The roadmap from EFT predictions to prediction coefficients: (a) the predictions themselves as a function of a generic variable x, (b) the leading-order prediction y_0 and the increasingly suppressed order-by-order corrections Δy_n, and (c) the dimensionless coefficients c_n (using $y_{\text{ref}} = 10$ and $Q = 0.5$). One might object that $|\Delta y_3|$ is often larger than $|\Delta y_2|$ or that the coefficients in (c) do not appear sufficiently random. This is a consequence of a small sample size and seeing patterns in randomness: The coefficients c_n shown above are actual random draws from an underlying GP and were used to build parts (a) and (b). By chance c_2 is relatively small, and thus the correction $|\Delta y_2|$ is actually smaller than $|\Delta y_3|$ for this choice of Q. It pays to remember this example when considering real EFT predictions.

The coefficients are *not* the LECs of the EFT. Because we are making predictions that take the LECs as input, the c_n are potentially complicated functions of the LECs, Q, and other variables. Therefore, we do not have a general proof that the naturalness of the LECs propagates to the c_n; rather, it must be verified in each application. If some of the c_n do not conform to our assumptions, they may need to be left out of the analysis.

Our manipulations thus far have simply defined our terminology, but the arguments above lead to a physics-based uncertainty model that can be tested against reality. By a logical extension of Eq. (3), the truncation error δy_k would consist of all terms not included in the sum, i.e.,

$$\delta y_k(x) = y_{\text{ref}}(x) \sum_{n=k+1}^{\infty} c_n(x) Q^n(x). \qquad (5)$$

We can now describe the details of the EFT convergence model. By induction, we assume that the properties of the unobserved c_n for $n > k$ are the same as the c_n for $n \leqslant k$. Specifically, we assume that the c_n are independent and identically distributed (i.i.d.) random curves. Moreover, we have some prior knowledge of the curves' properties: They should be smooth and naturally sized, i.e., have a standard deviation of order 1. Data from the lower-order predictions can then be used to refine the estimates for the sizes and shapes of all subsequent c_n and, via Eq. (5), estimate the truncation error. This is shown visually in the Bayesian network of Fig. 2.

Because the coefficients are treated as random curves, the distribution from which they are drawn is known as a stochastic *process*. A stochastic process is essentially an infinite dimensional generalization of a probability distribution. Processes associate a random variable with each point in a continuous domain, hence their infinite dimensionality, and correlations between the random variables at different points

can be built in. Through the appropriate assumption on this correlation structure, one can enforce desirable qualities in random draws from the process, such as continuity of the drawn functions and their derivatives. The GP is a particularly useful type of stochastic process and will be employed throughout this paper.

B. A brief introduction to Gaussian processes

1. Definitions

GPs are a popular tool for nonparametric regression due to their flexible nature and tractable analytic forms. They are used in fields such as statistics, machine learning, and geostatistics, where GP regression is known as *kriging* [5–7]. For appropriate priors, GPs arise as the limit of a specific type of neural network [8] and are equivalent to common interpolants, e.g., splines [9]. Here we will review the most common applications of GPs: interpolation and regression. Then we discuss the calibration of the GP parameters, which is a crucial step in our truncation error model. For more in-depth discussions of GPs, see Refs. [7,9,10].

The defining quality of a GP is given in Ref. [7]:

A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution.

A GP is specified by a mean function $m(x)$ and a positive semidefinite covariance function, or kernel, $\kappa(x, x')$, where $x \in \mathbb{R}^d$. Assuming both the mean and covariance functions are known, a GP $f(x)$ is denoted by[4]

$$f(x) \sim \mathcal{GP}[m(x), \kappa(x, x')]. \qquad (6)$$

---

[4]The $z \sim \cdots$ notation is a common shorthand in statistical literature for "z is distributed as." Some authors use $\text{pr}(z) = \cdots$ as well [11]. Also "$z \mid I$" is read as "z given I."
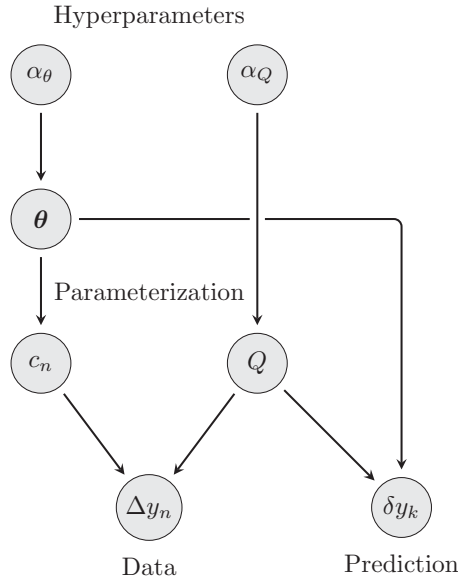
Hyperparameters

FIG. 2. A Bayesian network that demonstrates the causal connections between the random variables in our model. Order-by-order corrections Δy_n are assumed to be composed of coefficients c_n, which are draws from a GP, and factors of Q, as in Eq. (4). The observed values of Δy_n are used to update our beliefs of the GP parameters $\boldsymbol{\theta}$ and the expansion parameter Q, which themselves are random variables governed by the hyperparameters α_θ and α_Q, respectively. For example, $\eta_0, V_0, \nu_0, \tau_0^2 \in \alpha_\theta$ as assumed in Eq. (24). Once posteriors for $\boldsymbol{\theta}$ and Q have been obtained, they can be used to create a posterior for δy_k, hence quantifying the uncertainties in the predictions. Note that information only flows from the differences Δy_n to the truncation error δy_k via the GP parameters $\boldsymbol{\theta}$ and the expansion parameter Q. The reference scale y_{ref} is assumed to be known throughout and thus is not included in this diagram.

We do not operate on the infinite-dimensional object in practice; rather, the definition of GPs allows us to compute using a finite number of points. Let $\mathbf{x} = \{x_i\}_{i=1}^N$ be a set of N input points and $\mathbf{f} = \{f(x_i)\}_{i=1}^N$ be the corresponding set of function values.[5] Henceforth, we refer to coefficient values computed at particular input points as "data," because these are the points used to train the GP emulator.

Define $\mathbf{m} = m(\mathbf{x}) \in \mathbb{R}^N$ and $K = \kappa(\mathbf{x}, \mathbf{x}) \in \mathbb{R}^{N \times N}$. $\mathbf{f}$ is then distributed as the multivariate normal

$$\mathbf{f} \,|\, \mathbf{x} \sim \mathcal{N}(\mathbf{m}, K). \tag{7}$$

The fact that Eq. (6) implies Eq. (7) is the definition of a GP in a mathematical form. Often the explicit conditioning on input points $\mathbf{x}$ from $\mathbf{f} \,|\, \mathbf{x}$ is dropped for notational simplicity. The N-dimensional Gaussian distribution of Eq. (7) can then

---

[5]To specify how the GP enters in our truncation model, we must work in multiple vector spaces. The bold notation for vectors signals an element of $\mathbb{R}^N$ while we use arrows to indicate a vector in the space of EFT orders $\mathbb{R}^{n_c}$. Bold with arrows is used for elements of both spaces. The d-dimensional input space is not explicitly indicated. This notation is summarized in Table II.

easily be used to draw samples of $\mathbf{f}$ using the Cholesky decomposition of K.

The mean function is an *a priori* "best guess" of f and should capture overall trends in the data; the GP then soaks up deviations from this mean in accordance with the chosen kernel [11]. Basis function models for $m(x)$ are convenient choices due to their analytic tractability and flexibility [11]. But the effects of mean functions can also be shuffled into the kernel (see Appendix) or simply subtracted from the data in preprocessing, hence many authors set the mean to zero. The convergence model proposed here uses a constant (possibly unknown) mean μ, which we leave in the mean function slot for clarity.

One of the most popular covariance functions is the squared exponential (or radial basis function or Gaussian). It is defined by

$$\kappa(x, x'; \bar{c}, L) = \bar{c}^2 r(x, x'; L)$$
$$= \bar{c}^2 e^{-\frac{1}{2}(x-x')^\mathsf{T} L^{-1}(x-x')}, \tag{8}$$
$$\to \bar{c}^2 e^{-(x-x')^\mathsf{T}(x-x')/2\ell}, \tag{9}$$

where $r(x, x'; L)$ is the correlation function and L is a $d \times d$ diagonal matrix of correlation length parameters, allowing one to model cases where the correlation length varies by direction [7]. For simplicity our examples are in one dimension ($d = 1$), so L reduces to a single correlation length ℓ. Functions drawn from a GP with a squared exponential kernel are infinitely differentiable. The marginal variance $\bar{c}^2$ controls the width of variation about the mean, while the length scale ℓ controls the correlation between points in the input domain, see Fig. 3. Some have argued that the squared exponential is *too* smooth, and thus not applicable in many situations [12]. Thus, an alternative kernel is the Matérn, which provides an extra parameter ν and results in realizations that are $\lceil \nu \rceil - 1$ times differentiable, reproducing the squared exponential in the limit $\nu \to \infty$.

The squared exponential kernel, along with the Matérn, are *stationary*, which means that they only depend on the absolute difference in input space $|x - x'|$. Draws from a stationary process would, on average, behave similarly across the domain of interest. Stationarity is a strong assumption about the physics of the system, which can be advantageous in data-sparse regimes that could benefit from extra constraints. But its validity should be checked. We introduce diagnostics that can point to issues with the assumption of stationarity, along with the choices of mean and covariance, in Sec. III.

We assume the functional form of the mean and covariance functions (e.g., squared exponential) are known. Their parameters,[6] on the other hand, can be updated based on observed data. We denote the set of all parameters in both m and κ (e.g., $\mu, \bar{c}^2, \ell$) as $\boldsymbol{\theta}$. We are careful to write $\cdot \,|\, \boldsymbol{\theta}$ when $\boldsymbol{\theta}$ is given (i.e., *conditioned* on) and omit $\boldsymbol{\theta}$ otherwise.

---

[6]We choose to call $\mu, \bar{c}$, etc., the *parameters* of the GP to distinguish them from the hyperparameters we will introduce later. They are more regularly referred to as hyperparameters.
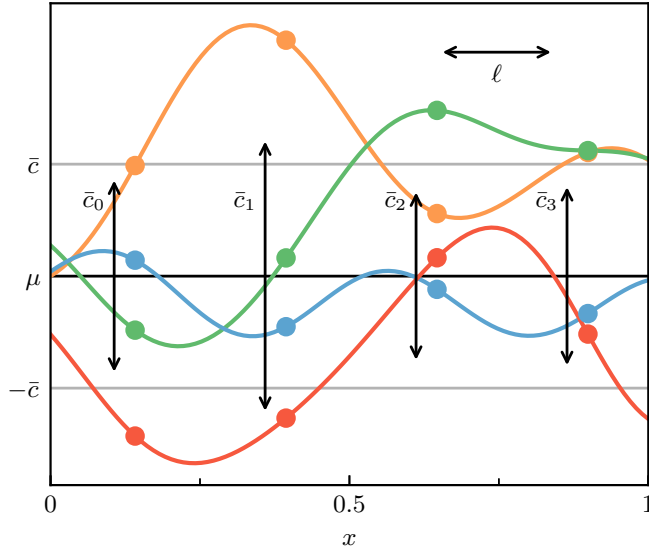
FIG. 3. The random draws from a GP used in Fig. 1(c). The GP has mean μ, marginal standard deviation $\bar{c}$, and a Gaussian covariance function with length scale ℓ, each of which are annotated on the plot. The "uncorrelated" and "fully correlated" truncation error model discussed in Ref. [1] correspond to this model with $\ell = 0$ and $\ell = \infty$, respectively, and $\mu = 0$. By using a finite set of points along each curve, shown as dots, these parameters can be estimated (or calibrated) rather than assumed. This GP model is contrasted with the pointwise model from Refs. [2,3], where data at each x_i are used to estimate a $\bar{c}_i \equiv \bar{c}(x_i)$ (and where there is no ℓ).

2. Interpolation and regression

To use GPs for interpolation or regression, we must be able to incorporate (noisy) observations into our predictions. Interpolation, in this case, simply refers to the procedure of finding a curve that passes *directly* through known data, regardless of whether the predictions are made within or outside the support of the data. Consider the partition $\mathbf{x} = [\mathbf{x}_1 \quad \mathbf{x}_2]^\mathsf{T}$ and $\mathbf{f} = f(\mathbf{x}) = [\mathbf{f}_1 \quad \mathbf{f}_2]^\mathsf{T}$ into N_1 training and N_2 test points, with corresponding partitioning of $\mathbf{m}$ and K. We would like to predict the unseen function values $\mathbf{f}_2$ given observed function values $\mathbf{f}_1$. By definition of GPs, their $N_1 + N_2$-dimensional joint distribution is Gaussian:

$$\begin{bmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \end{bmatrix} \bigg| \, \mathbf{x}, \boldsymbol{\theta} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{m}_1 \\ \mathbf{m}_2 \end{bmatrix}, \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}\right). \quad (10)$$

Then, by standard Gaussian identities [7],

$$\mathbf{f}_2 \,|\, \mathbf{x}, \mathbf{f}_1, \boldsymbol{\theta} \sim \mathcal{N}(\tilde{\mathbf{m}}_2, \tilde{K}_{22}), \quad (11)$$

where

$$\tilde{\mathbf{m}}_2 \equiv \mathbf{m}_2 + K_{21}K_{11}^{-1}(\mathbf{f}_1 - \mathbf{m}_1), \quad (12)$$

$$\tilde{K}_{22} \equiv K_{22} - K_{21}K_{11}^{-1}K_{12}. \quad (13)$$

Since we assume that K can factorize into a marginal variance $\bar{c}^2$ and a correlation matrix R, i.e., $K = \bar{c}^2 R$, this can be written as

$$\tilde{\mathbf{m}}_2 = \mathbf{m}_2 + R_{21}R_{11}^{-1}(\mathbf{f}_1 - \mathbf{m}_1), \quad (14)$$

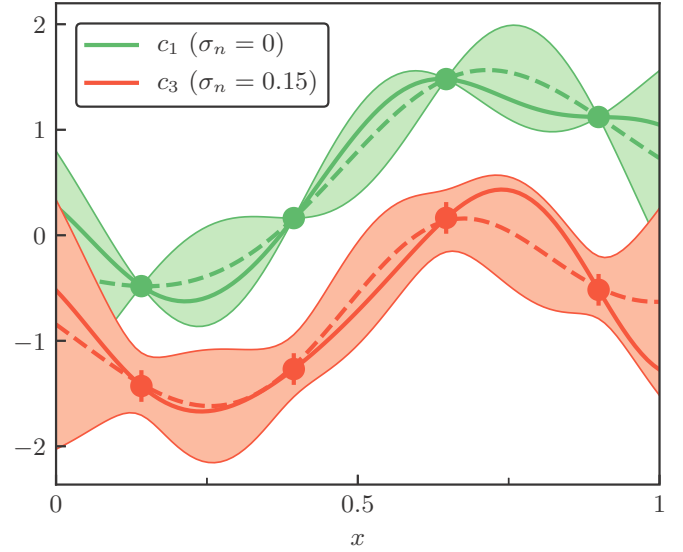$$\tilde{K}_{22} = \bar{c}^2\left(R_{22} - R_{21}R_{11}^{-1}R_{12}\right) \equiv \bar{c}^2 \tilde{R}_{22}. \quad (15)$$

FIG. 4. Examples of interpolation ($\sigma_n = 0$) and regression ($\sigma_n = 0.15$) using the GP described in Fig. 3. The true curves are shown as solid lines and the interpolants are shown as dashed lines. The bands denote 2σ marginal confidence intervals. As one might expect, there is a bubblelike structure to the interpolation error when between known points. The process will eventually return to the mean when far from the support of the data.

This work considers a hierarchy of noiseless compwuter simulations as its data, so interpolation is the primary fitting method and an example is shown in Fig. 4. Nevertheless, interpolation may require a small amount of noise called a *nugget* for the numerical stability of the matrix inversion [13–17].

To make predictions of $\mathbf{f}_2$ assuming Gaussian white noise in the observed data $\mathbf{f}_1$, one need only make the replacement $K_{11} \to K_{11} + \sigma_n^2 \mathbb{I}_{N_1}$ in Eqs. (12) and (13), leaving K_{21}, etc., unchanged. This is what we refer to as regression. The identity matrix is denoted by $\mathbb{I}$ with the rank as subscript, and here σ_n^2 is the nugget used to regularize the matrix inversion. To make predictions that also include the effects of noise, one must also make the replacement $K_{22} \to K_{22} + \sigma_n^2 \mathbb{I}_{N_2}$. An example of regression with noise is shown in Fig. 4. We include the nugget via $\bar{c}^2(R + \sigma_n^2 \mathbb{I}_N)$ to integrate $\bar{c}^2$ out analytically.

3. Calibration

Until now, we have assumed that the GP parameters $\boldsymbol{\theta}$ are known and correct, but that is rarely ever the case in practice. We often want to find the $\boldsymbol{\theta}$ that produce the best fit in interpolation/regression, or $\boldsymbol{\theta}$ may be of interest in its own right. The procedure of finding the posterior for $\boldsymbol{\theta}$ values is known as *calibration* or simply Bayesian inference for the model parameters [18].

Given data $\mathbf{f}$ at input points $\mathbf{x}$, the posterior for $\boldsymbol{\theta}$ is

$$\mathrm{pr}(\boldsymbol{\theta} \,|\, \mathbf{x}, \mathbf{f}) \propto \mathrm{pr}(\mathbf{f} \,|\, \mathbf{x}, \boldsymbol{\theta})\mathrm{pr}(\boldsymbol{\theta}). \quad (16)$$

If f is modeled as a GP, then Eq. (16) is, up to a normalization constant, Eq. (7) multiplied by the priors. The posterior can be sampled via Markov-chain Monte Carlo (MCMC) methods or optimized using gradient descent to find its maximum *a*

posteriori (MAP) value $\boldsymbol{\theta}_{\text{MAP}}$. When a single value for $\boldsymbol{\theta}$ is used as a "best guess," e.g., $\boldsymbol{\theta}_{\text{MAP}}$, this is called a *point estimate*.

For certain cases, one can find analytic posteriors for parameters in $\boldsymbol{\theta}$ by making use of *conjugate priors*, whose posteriors have the same functional form as the prior [11]. For our applications only a few features of the prior information are important; we are generally insensitive to details of the priors' functional form. Our strategy is then to make use of conjugate priors where they are flexible enough to capture distributions' relevant features and find MAP values for the remaining parameters.

The conjugate priors on the GP parameters μ and $\bar{c}^2$ are characterized by hyperparameters that quantify the mean and variance for the prior distributions of both, as well as the heaviness of the $\bar{c}^2$ distribution's tail. Our model is also easily built into existing MCMC samplers, such as PyMC3 [19], where any functional form for the priors can be used.

To make predictions unconditional on $\boldsymbol{\theta}$, which take all plausible values into account, one must compute

$$\text{pr}(\mathbf{f}_2 \mid \mathbf{x}, \mathbf{f}_1) = \int \text{pr}(\mathbf{f}_2 \mid \mathbf{x}, \mathbf{f}_1, \boldsymbol{\theta}) \text{pr}(\boldsymbol{\theta} \mid \mathbf{x}_1, \mathbf{f}_1) d\boldsymbol{\theta}. \quad (17)$$

This can be approximated by replacing the integral over $\boldsymbol{\theta}$ by evaluation of the integrand at $\boldsymbol{\theta}_{\text{MAP}}$, computed numerically with Monte Carlo methods, or evaluated analytically with the help of conjugacy. For example, Appendix shows how μ and $\bar{c}^2$ are integrated out of Eq. (17) analytically via conjugacy.

This brief introduction to GPs is enough to set the stage for our model of truncation errors.

C. The GP model of EFT convergence

Recall from Sec. II A that after making the transformation from prediction functions $\{y_n\}$ to coefficient functions $\{c_n\}$, the coefficients appeared as naturally sized random curves. Here we will specify exactly what we mean by random and discuss how we can extract meaningful insight from this assumption.

We formalize our EFT convergence assumptions via Eqs. (3) and (5), where the coefficients c_n are independent draws from a single underlying GP. We assume a constant mean function $m(x) = \mu$ for simplicity and a kernel $\kappa(x, x'; \bar{c}, \ell) = \bar{c}^2 r(x, x'; \ell)$ that need not be stationary. Thus,

$$c_n(x) \mid \boldsymbol{\theta} \overset{\text{iid}}{\sim} \mathcal{GP}[\mu, \bar{c}^2 r(x, x'; \ell)], \quad (18)$$

$$\boldsymbol{\theta} \equiv \{\mu, \bar{c}^2, \ell\}. \quad (19)$$

Figure 3 shows random draws from such a process and a comparison to the model described in Refs. [2,3].

The parametrization of Eq. (18) can be generalized as appropriate. Nonconstant mean functions can be incorporated by promoting μ to a vector of regression coefficients multiplying basis functions. For the model of the c_n's that we use here this basis would contain only one, constant, function. But if one wants to build specific structure into the mean function for y_k and δy_k, then the notion of a basis is useful for discussing the distribution of these quantities. We discuss that more general case, adopting vector and matrix notation, in Appendix. The factorization of the kernel into a marginal

variance $\bar{c}^2$ and a correlation function r that depends on a correlation length ℓ may seem restrictive. On the contrary, this kernel can incorporate a wide range of functional structures since r can be nonstationary in general and ℓ can stand in for any set of correlation parameters. Thus here we make μ, $\bar{c}^2$, and ℓ the explicit parameters of our GP model: This allows us to show how they can be calibrated or integrated out of this system.

The distribution of δy_k, and thus the full prediction $y = y_k + \delta y_k$, can be derived using the remarkable property of Gaussian random variables: they are closed under addition and matrix multiplication. Let $X \sim \mathcal{N}(\mu_1, K_1)$ and $Y \sim \mathcal{N}(\mu_2, K_2)$ be length-N independent random variables and let A, B be known $M \times N$ matrices. Then

$$AX + BY \sim \mathcal{N}(A\mu_1 + B\mu_2, AK_1A^{\mathsf{T}} + BK_2B^{\mathsf{T}}). \quad (20)$$

Equation (5), which defines the truncation error δy_k, is a geometric sum over normally distributed variables. Using Eqs. (18) and (20), one can readily find the prior

$$\delta y_k(x) \mid \boldsymbol{\theta}, Q \sim \mathcal{GP}[m_{\delta k}(x), \bar{c}^2 R_{\delta k}(x, x'; \ell)], \quad (21)$$

where

$$m_{\delta k}(x) \equiv y_{\text{ref}}(x) \frac{Q(x)^{k+1}}{1 - Q(x)} \mu \equiv b_{\delta k}(x)\mu, \quad (22)$$

$$R_{\delta k}(x, x'; \ell) \equiv y_{\text{ref}}(x) y_{\text{ref}}(x') \frac{[Q(x)Q(x')]^{k+1}}{1 - Q(x)Q(x')} r(x, x'; \ell). \quad (23)$$

We have defined the basis function $b_{\delta k}(x)$ as it will recur below. Although $R_{\delta k}$ may look like a correlation matrix, the inclusion of the y_{ref} and Q factors mean that $R_{\delta k}(x, x; \ell) \neq 1$, in general. The marginal variance of Eq. (21) is thus, in general, x dependent and equal to $\bar{c}^2 R_{\delta k}(x, x; \ell)$.

Equations (21)–(23) define the functional form of the truncation errors, but how should one estimate their parameters in a statistically rigorous manner? The simplest (and least rigorous) way is to assume some reasonable values for $\boldsymbol{\theta}$ and Q based on expert knowledge of the system. For example, $\mu = 0$ and $\bar{c} \approx 1$ is a good place to start if one can assume that the coefficients are naturally sized and equally likely to be positive or negative. Then ℓ and Q could be based on arguments specific to the observable and EFT under consideration. Again, y_{ref} is assumed to be known in this work, possibly based on dimensional analysis for the observable. The next best way would be to combine expert knowledge of the system with the known low-order predictions to compute point estimates of $\boldsymbol{\theta}$ and Q. (Remember, information only propagates from the order-by-order predictions to δy_k via $\boldsymbol{\theta}$ and Q, as shown in Fig. 2.) Finally, the most rigorous Bayesian method is to compute the full posterior for $\boldsymbol{\theta}$ and Q and marginalize (integrate) over all of their possible values. Depending on the specific application of this model and the needs of the user, different degrees of rigor may be warranted.

Assuming that we want to combine expert knowledge of the system with the order-by-order EFT predictions to estimate δy_k, we must start by formalizing our knowledge as Bayesian priors. We have checked the prior dependence of our pointwise model in previous works and found that truncation error estimates are fairly insensitive to their exact choice
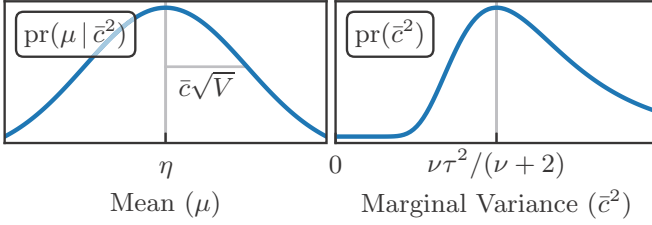
FIG. 5. Slices of the distribution $\mu, \bar{c}^2 \sim N\chi^{-2}(\eta, V, \nu, \tau)$. The left plot is a Gaussian while the right is a scaled inverse χ^2 distribution.

[2,3]. Therefore, we choose to work with conjugate priors, which yield intuitive analytic posteriors and point estimates for the GP parameters $\boldsymbol{\theta}$. We place a (scaled) normal-inverse-χ^2 prior[7] on μ and $\bar{c}^2$,

$$\mu, \bar{c}^2 \sim N\chi^{-2}(\eta_0, V_0, \nu_0, \tau_0^2). \tag{24}$$

This joint prior factorizes as $\mathrm{pr}(\mu, \bar{c}^2) = \mathrm{pr}(\mu \,|\, \bar{c}^2)\mathrm{pr}(\bar{c}^2)$, where

$$\mu \,|\, \bar{c}^2 \sim \mathcal{N}(\eta_0, \bar{c}^2 V_0), \tag{25}$$

$$\bar{c}^2 \sim \chi^{-2}(\nu_0, \tau_0^2). \tag{26}$$

The inverse χ^2 density is given by

$$f(z; \nu, \tau^2) = \frac{(\nu\tau^2/2)^{\nu/2}}{\Gamma(\nu/2)z^{1+\nu/2}} \exp\left(-\frac{\nu\tau^2}{2z}\right). \tag{27}$$

See Fig. 5 for a graphical representation of $\mathrm{pr}(\mu, \bar{c}^2)$. The hyperparameters can be interpreted as follows:

(i) *The prior mean η_0*: Our best guess for μ before seeing the data.
(ii) *The prior dispersion[8] V_0*: When combined with $\bar{c}^2$, this is our estimated uncertainty in μ. It makes sense that if the spread of the c_n is large, then we would be less certain about their mean.
(iii) *The prior degrees-of-freedom ν_0*: This quantity essentially describes our uncertainty in $\bar{c}^2$. For $\nu = 0$, Eq. (27) reduces to the scale-invariant Jeffreys prior used in Ref. [3], whereas for $\nu \to \infty$, it becomes sharply peaked.
(iv) *The prior scale τ_0^2*: For large ν, this is our best guess for $\bar{c}^2$. The mode and mean of Eq. (27) are given by $\nu\tau^2/(\nu \pm 2)$, respectively. The mean is only defined for $\nu > 2$.

There are no conjugate priors for ℓ or Q. Reasonable choices could be a strictly positive log-normal prior for ℓ and a beta prior for Q, since it is restricted to be between zero and one.

---

[7] A more common and mathematically equivalent choice is the normal-inverse-gamma prior. We find the normal-inverse-χ^2 prior more intuitive to work with in this context and thus promote its use.

[8] We have not found a standard name for this quantity in the literature. The dispersion, a generic term describing the variability or spread of a distribution, is as good as any.

Once priors have been chosen for $\boldsymbol{\theta}$ and Q, the rest of the process is algorithmic. Given order-by-order data $\vec{\mathbf{y}}_k \equiv \{\mathbf{y}_i\}_{i \leqslant k}$, one can derive analytic posteriors for all of $\boldsymbol{\theta}$ and Q. See Appendix for details on how this is done. From now on we assume that ℓ and Q are known *a priori* or point estimates are obtained from their posteriors. Analytic results for the truncation error posterior follow. Formulae in which ℓ and Q are marginalized over are more complicated and not analytic. Point estimates of ℓ and Q are often good approximations in our application, making numerical integration a needless complication. MCMC sampling can be used in cases where this approximation is not adequate.

Conjugacy ensures that the posterior for $\mu, \bar{c}^2$ has the same functional form as the prior. That is,

$$\mu, \bar{c}^2 \,|\, \vec{\mathbf{y}}_k, \ell, Q \sim N\chi^{-2}(\eta, V, \nu, \tau^2) \tag{28}$$

for updated hyperparameters η, V, ν, and τ^2 given by Eqs. (A27)–(A30). The path forward depends on whether point estimates of μ and $\bar{c}^2$ are sufficient. If so, then the mean values $\mathbb{E}[\cdot]$ of μ and $\bar{c}^2$, given by

$$\mathbb{E}[\mu \,|\, \vec{\mathbf{y}}_k, \ell, Q] = \eta, \tag{29}$$

$$\mathbb{E}[\bar{c}^2 \,|\, \vec{\mathbf{y}}_k, \ell, Q] = \frac{\nu\tau^2}{\nu - 2}, \tag{30}$$

can be used as point estimates in Eq. (21). It is interesting to note that these estimates are Bayesian analogs to standard frequentist estimators for the mean and variance. If one integrates over all possible values of μ and $\bar{c}^2$, then the posterior is a *Student-t process* (TP) [20],

$$\delta y_k(x) \,|\, \vec{\mathbf{y}}_k, \ell, Q$$
$$\sim \mathcal{TP}_\nu\{b_{\delta k}(x)\eta, \tau^2[R_{\delta k}(x, x'; \ell) + b_{\delta k}(x)V b_{\delta k}(x')]\}. \tag{31}$$

Like a GP, any finite collection of random variables from a TP has a joint Student-t distribution. Equation (31) is centered at our best guess for $b_{\delta k}(x)\mu$ but also includes the uncertainty in μ via $b_{\delta k}(x)V b_{\delta k}(x')$ and the uncertainty in $\bar{c}^2$ via the heavy tails of the t distribution (which are heavier for smaller ν). The $\kappa(x, x')$ in $\mathcal{TP}_\nu[m(x), \kappa(x, x')]$ is not the covariance function in our notation but instead a scale function. The covariance between the response at x and x' for a TP is given by $\nu\kappa(x, x')/(\nu - 2)$ and is only defined for $\nu > 2$. Given this fact, note the relationship between the point estimate in Eq. (30) and the covariance in Eq. (31).

D. Application types

We have shown how to arrive at a physically motivated distribution for the truncation error, which we will now use as a springboard to discuss how our EFT convergence model is applicable to two situations that can arise when fitting and predicting.

(1) For certain physical systems, predictions may be expensive to compute. This can affect the fitting of the LECs, where predictions must be made repeatedly to find a best fit (or posterior), but can also affect predictions with fixed LECs for particularly expensive systems. For example, fitting chiral EFT beyond the two-body sector becomes computationally expensive.
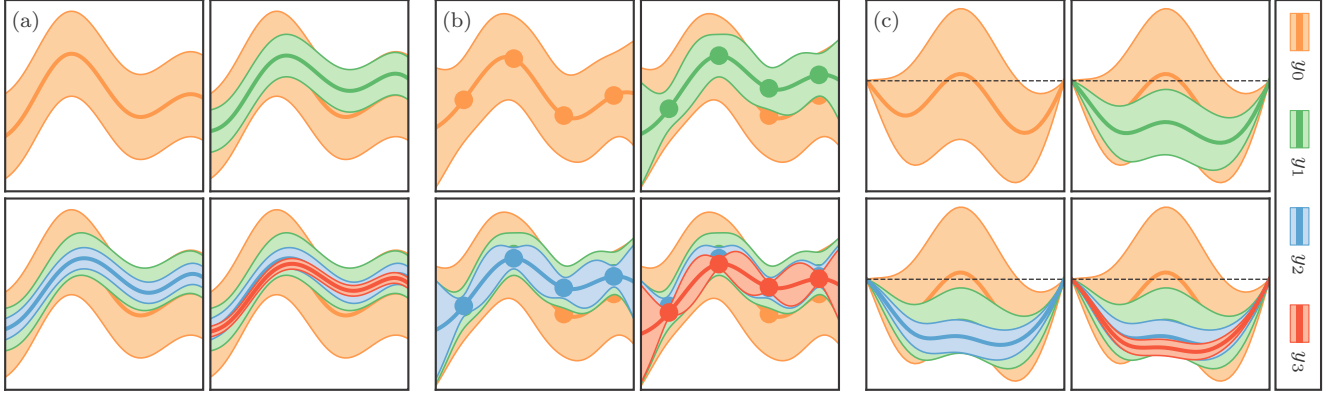
FIG. 6. Toy predictions of $y(x) = y_n(x) + \delta y_n(x)$ vs x for $n \leqslant 3$. Shaded regions denote 2σ bands, whose colors correspond to orders described in the legend. True values for $\boldsymbol{\theta}$ and Q are assumed throughout. (a) A computationally inexpensive system with truncation error uncertainty. Higher-order predictions are generally contained within the lower-order bands. (b) A computationally expensive system that combines both truncation error and interpolation error quickly and analytically. Three data points from (a) were used to fit $y_n(x)$ at each order, with zero numerical noise assumed. Solid lines are the interpolants, which are close to the $y_n(x)$ from (a) to within their bubblelike interpolation errors. Note that uncertainty in $y_3(x)$ is almost purely due to interpolation. (c) An inexpensive system with the corrections $c_n(x)$ constrained to be zero at the endpoints. This imposes a constraint on $\delta y_n(x)$ as well. As expected, the truncation error smoothly vanishes when approaching the endpoints.

(2) Certain observables may have symmetry constraints, which provides information about contributions from higher orders in the EFT. For example, certain spin observables in the *NN* sector of chiral EFT must vanish at $\theta = 0°$ or $180°$. Thus, the distribution for the truncation error should reflect this fact.

Here we discuss how our correlated convergence model can solve both of these problems and how to use this model when fitting an EFT. We assume that point estimates from Eqs. (29) and (30) are used so that everything remains Gaussian and leave the TP case for Appendix.

1. Inexpensive predictions

If predictions are inexpensive, then $y_k(x)$ can be computed at all x of interest. In this case, we do not need to interpolate $y_k(x)$ and can instead focus on estimating the truncation error, i.e., estimating $\boldsymbol{\theta}$ and Q. In this data-rich environment, we must be careful when learning the point estimates for these quantities, since too much data within one correlation length ℓ can cause ill-conditioned matrices and thus poor point estimates [13–17]. We suggest simply choosing a representative subset of data for learning $\boldsymbol{\theta}$ and Q and using a small nugget to regularize the matrix inversion. Then δy_k can be produced at *all* points of interest without issue using Eq. (21). The distribution for the full observable $y = y_k + \delta y_k$ has mean and covariance function given by

$$m_{\text{th}}(x) = y_k(x) + m_{\delta k}(x), \tag{32}$$

$$\Sigma_{\text{th}}(x, x'; \ell) = \bar{c}^2 R_{\delta k}(x, x'; \ell). \tag{33}$$

An example of the posteriors for $y(x)$ assuming different max orders k is given in Fig. 6(a).

2. Expensive predictions

If predictions are expensive, then one may want an estimate of $y_k(x)$ at some x that was not explicitly computed with the EFT. Thus, we now have both interpolation error and truncation error to worry about. Fortunately, GPs are *made* for interpolation and are very fast. In this case, we need a distribution for $y_k(x)$ as opposed to $\delta y_k(x)$. Following Eqs. (3), (18), and (20), one can write

$$y_k(x) \,|\, \boldsymbol{\theta}, Q \sim \mathcal{GP}[m_k(x), \bar{c}^2 R_k(x, x'; \ell)], \tag{34}$$

where

$$m_k(x) \equiv y_{\text{ref}}(x) \frac{1 - Q(x)^{k+1}}{1 - Q(x)} \mu \equiv b_k(x)\mu, \tag{35}$$

$$R_k(x, x'; \ell) \equiv y_{\text{ref}}(x) y_{\text{ref}}(x') r(x, x'; \ell) \frac{1 - [Q(x)Q(x')]^{k+1}}{1 - Q(x)Q(x')}. \tag{36}$$

Now it is a simple matter to estimate $\boldsymbol{\theta}$ and Q as before and then apply Eqs. (11)–(13) to Eq. (34). Because the conditional distribution of Eq. (11) is *still Gaussian*, one can again use the fact that GPs are closed under addition to compute $y = y_k + \delta y_k$, which has mean and covariance functions

$$m_{\text{th}}(x) = \tilde{m}_k(x) + m_{\delta k}(x), \tag{37}$$

$$\Sigma_{\text{th}}(x, x'; \ell) = \bar{c}^2 [\tilde{R}_k(x, x'; \ell) + R_{\delta k}(x, x'; \ell)]. \tag{38}$$

[Remember that the tildes above m and R refer to the conditioning shown in Eqs. (11)–(13).] See Fig. 6(b) for an example when only a few points are computed at each order and used to estimate the full prediction with combined interpolation errors and truncation errors.

3. Symmetry constraints

Above we discussed how to constrain $y_k(x)$ based on known data, but sometimes we could constrain $\delta y_k(x)$ based

on, e.g., symmetry arguments. That is, we may know that higher-order coefficients are zero at a particular x. In this case, we need only constrain $\delta y_k(x)$ via Eqs. (11)–(13) according to our knowledge of the system. This could apply regardless of whether $y_k(x)$ is expensive or inexpensive. Again, since Eq. (11) is still Gaussian, $y = y_k + \delta y_k$ is a GP. One need only make the replacements $m_{\delta k} \to \tilde{m}_{\delta k}$ and $R_{\delta k} \to \tilde{R}_{\delta k}$ in Eqs. (32)–(33) and (37)–(38). Figure 6(c) shows an example of how information about symmetries can be applied to an inexpensive system.

4. EFT fitting

Our model permits—for the first time—a rigorous inclusion of all higher-order terms and their correlations within the fitting procedure. Interpolation of expensive predictions fits seamlessly into this framework, which allows more experimental data to be used than might otherwise be possible. The posterior for the LECs $\vec{a}_k$ of the kth-order EFT, given data for different observables $\mathcal{D} = \{\mathbf{y}_{exp}\}_i$, is given by Bayes' theorem

$$\mathrm{pr}(\vec{a}_k \,|\, \mathcal{D}, I) \propto \mathrm{pr}(\mathcal{D} \,|\, \vec{a}_k, I)\mathrm{pr}(\vec{a}_k) = \mathrm{pr}(\vec{a}_k) \prod_i \mathrm{pr}(\mathcal{D}_i \,|\, \vec{a}_k, I_i),$$

(39)

where I_i is the prior information for the ith observable, and we have assumed that different observables are conditionally independent of one another given $\vec{a}_k$.

Consider one specific observable y. Prior information about y could include $\boldsymbol{\theta}$ and Q (from previous lower-order fits of the EFT, for example) and also its experimental uncertainty Σ_{exp}. The values of $\vec{a}_k$, $\boldsymbol{\theta}$, and Q are sufficient to construct the distribution for $y(x; \vec{a}_k) = y_k(x; \vec{a}_k) + \delta y_k(x)$, which is a GP with mean function $m_{th}(x; \vec{a}_k, \boldsymbol{\theta}, Q)$ and covariance function $\Sigma_{th}(x, x'; \boldsymbol{\theta}, Q)$. The exact form of these functions is dependent on whether the predictions are interpolated and whether the truncation error is constrained, as discussed above. Once the distribution for $y(x; \vec{a}_k)$ is determined, then the likelihood for this observable is given by

$$\mathrm{pr}(\mathcal{D}_i \,|\, \vec{a}_k, I_i) \propto e^{-\frac{1}{2}(\mathbf{y}_{exp} - \mathbf{m}_{th})^{\mathsf{T}}(\Sigma_{th} + \Sigma_{exp})^{-1}(\mathbf{y}_{exp} - \mathbf{m}_{th})}.$$

(40)

This is an extension of the likelihood proposed in Ref. [1], whose "uncorrelated" and "fully correlated" models correspond to $\ell = 0$ and $\ell = \infty$ limits. Applying Eq. (40) to data with finite positive ℓ is a subject for future work.

III. MODEL CHECKING

Though the application of our model to any given data set is simple enough in practice, one must validate that the modeling assumptions made in Sec. II are appropriate for the system under consideration. If this model is inappropriate, then the estimates of $\boldsymbol{\theta}$ and Q, along with the truncation error distributions, could be biased or even meaningless. Conversely, if one assumes our model describes how an EFT *should* behave, then it would be of great interest to know when a given EFT is failing. Thus a well-chosen set of model-checking diagnostics are critical to test our assumptions against the data at hand.

To test our assumptions against reality, we must first enumerate them:

(1) The coefficients c_n are i.i.d. realizations of a GP.
(2) The c_n GP's mean and covariance functions—and their parameters $\boldsymbol{\theta}$—have been correctly identified.
(3) From the knowledge of a few coefficients, we can create a statistically meaningful distribution for the truncation error.

Model-checking diagnostics can assess whether these assumptions are violated at a statistically significant level. The question of how to diagnose issues with GP fitting and prediction has been addressed thoroughly in Ref. [21]. However, our situation is novel compared to that of the everyday GP practitioner: (1) they typically have points along one curve, we have multiple i.i.d. curves; (2) their main focus is often the accuracy of the GP regressor's predictions, our focus is mainly on the parameters $\boldsymbol{\theta}$ themselves; and (3) our model for the c_n leads to a truncation error distribution which itself can be tested. Nevertheless, the same techniques discussed in Ref. [21] are directly applicable here.

Throughout the discussion below, we assume that the parameters for a GP have either been fixed *a priori*, or calibrated/marginalized via training data $\mathbf{f}_{train}$. The resulting GP (or TP) is then to be evaluated against test (or validation) data $\mathbf{f}_{val}$. One should take care that the training set $\mathbf{f}_{train}$ does not overlap with any points in $\mathbf{f}_{val}$. We have greater-than-average freedom with how we partition $\mathbf{f}_{train}$ and $\mathbf{f}_{val}$ because our data contain multiple i.i.d. curves c_n that could be tested against one another and because our truncation error model can be tested against reality. For now, assume that $\mathbf{f}_{train}$ contains data on *one* curve and was used to construct an *emulator* via Eqs. (11)–(13) to predict $\mathbf{f}_{val}$ from that same curve. We will return to the details on how one might choose different emulators, or how to partition the training/test sets differently, in Sec. III D. Below, we use the generic notation of $\mathbf{m}$ and K for the mean and covariance of the emulator (with ν degrees of freedom in the TP case).

A. Mahalanobis distance

A common way to measure the loss, or incorrectness, of a prediction is through the sum of squared residuals. This metric as a summary assumes that the errors at different input points x are completely uncorrelated; the individual weighted residuals at each point are simply summed up. The Mahalanobis distance is the multivariate analog of this idea that takes into account these correlations. That is,

$$\mathrm{D}_{MD}^2(\mathbf{f}_{val}) = (\mathbf{f}_{val} - \mathbf{m})^{\mathsf{T}} K^{-1} (\mathbf{f}_{val} - \mathbf{m}).$$

(41)

Large D_{MD}^2 implies that the validation data does not match the emulator. But what defines "large" and "small" in this context? A necessary component of any diagnostic is a *reference distribution*. A reference distribution sets the scale of variation and defines what normal and abnormal diagnostic values look like.

Suppose that we have validation data at M locations. Then, if the emulator is Gaussian, then the reference distribution for $\mathrm{pr}(\mathrm{D}_{MD}^2)$ is a χ^2 distribution with M degrees of freedom. If the emulator is a TP with ν degrees of freedom, then the

reference distribution is an F distribution $F_{M,\nu}$ that is scaled by $s = (\nu - 2)M/\nu$ [21]. (That is, D_{MD}^2/s follows a standard $F_{M,\nu}$ distribution.) The reference distributions in either the GP or TP cases do not depend on the specific form of the means and covariance functions: only the number of validation data and, if applicable, the degrees of freedom. Both the χ^2 and the F distributions are available in standard statistical libraries and permit simple analytic evaluations of credible intervals.

But, in general, the distributions of such diagnostics may not have a simple form. A surefire way to generate a distribution in any case is to simply sample from the emulator and compute the diagnostic for each $\mathbf{f}$ in the sample. Then the validation data $\mathbf{f}_{val}$ could then be rejected at a 68% or 95% level, for example, based on the empirical reference distribution.

B. Pivoted Cholesky decomposition

The MD is great as a one-number summary, but more information can be gained by considering decompositions of this quantity. Let G be defined by $K = GG^\mathsf{T}$. Then

$$\mathbf{D}_G = G^{-1}(\mathbf{f}_{val} - \mathbf{m}) \qquad (42)$$

is a variance decomposition of $D_{MD}^2 = \mathbf{D}_G^\mathsf{T}\mathbf{D}_G$. There is freedom in the exact definition of this diagnostic because G is not unique. The *pivoted* Cholesky decomposition of K leads to a diagnostic ($\mathbf{D}_{PC}$) that pinpoints the data contributing to a failing D_{MD}^2 and shows useful patterns when plotted graphically vs index [21]. A misestimated variance leads to unusually sized $\mathbf{D}_{PC}$ across all indices, whereas an incorrect estimation of the length scale leads to failing $\mathbf{D}_{PC}$ at large index.

The reference for each component of $\mathbf{D}_G$ is distributed as a standard Gaussian (for a GP emulator) since the points have been scaled and decorrelated by G. For the TP emulator, the reference is a Student-t distribution with ν degrees of freedom. The desired credible interval (here 2σ) can be computed accordingly.

C. Credible interval diagnostic

The credible interval diagnostic tests the accuracy of the emulator. An emulator is accurate if a $100\alpha\%$ credible interval (CI) contains approximately $100\alpha\%$ of the validation data. This diagnostic could be particularly useful when comparing experimental data $\mathbf{y}_{exp}$ to the full prediction with truncation uncertainty, since it captures whether the error behaves as advertised. Following Ref. [21], start by building $CI_i(\alpha)$, the $100\alpha\%$ marginal CI at the point x_i, for each of the M corresponding $f_i \in \mathbf{f}_{val}$. Then compute the accuracy of the credible intervals:

$$D_{CI}(\alpha, \mathbf{f}_{val}) = \frac{1}{M} \sum_{i=1}^{M} \mathbf{1}[f_i \in CI_i(\alpha)], \qquad (43)$$

where $\mathbf{1}$ is an indicator function, i.e., it equates to 1 if its argument is true and equates to 0 otherwise. Note that although the CI may be informed by training data at many points, the D_{CI} is computed point by point.

This diagnostic is the analog to the "consistency plots" shown in Ref. [3], which used the pointwise truncation model. It was shown that in this pointwise case, where the data are assumed uncorrelated with one another, the reference distribution is a simple binomial. Since the data here are curves, the reference is more complicated. Here we compute the reference distribution by simulation, that is, we sample the emulator and compute D_{CI} for each sample. The creation of a reference in this manner takes into account correlations by widening the acceptable region of accuracy, though the diagnostic is still inherently a point-by-point summary. Interestingly, we have found that the simulated reference is approximately a binomial with an effective number of data $M_{eff} = \Delta x/\ell$, where Δx is the difference between the minimum and maximum values of x.

D. Honorable mentions

We have tested many more diagnostics than are shown graphically in this work, such as the following:

(a) Variogram: A common tool used to investigate correlation structure by looking at the variability of pairs of points across the domain [22,23]. We have found that it is too sensitive to random fluctuations for our application.

(b) KL divergence: Another one-number summary for emulators. We did not find that it gave any new information beyond D_{MD}^2.

(c) Other variance decompositions: In addition to the $\mathbf{D}_{PC}$, we considered the standard Cholesky decomposition and the eigendecomposition. The Cholesky decomposition has the benefit that each index corresponds to a single validation datum, but it does not order the indices in a visually intuitive manner for diagnosing problems. The eigendecomposition, on the other hand, displays similar qualities as the pivoted Cholesky decomposition when plotted vs index, but its indices do not correspond to individual validation points. Thus, we choose to display only the pivoted Cholesky here, which enjoys the best of both worlds.

The above diagnostics are still available in the gsum [24] package despite not appearing here.

E. How to choose an emulator

We have described the most common approach for diagnosing GP emulators: $\mathbf{f}_{train}$ and $\mathbf{f}_{val}$ are function values from one curve evaluated at mutually exclusive sets of input points x. Now we discuss alternatives to this approach based on the questions our diagnostics are designed to address. This discussion is split into two parts: (1) evaluating coefficients c_n and (2) evaluating predictions y_n.

By testing the coefficients we can address the question of whether the observed convergence pattern matches our assumptions. Because they are assumed to be i.i.d., the parameters of their process can be learned from training data along each c_n simultaneously or along a subset of the curves. Then an emulator for some c_n could be used to predict test

data between the interpolating points, regardless of whether those points along c_n were used to estimate $\boldsymbol{\theta}$.

A problem occurs when using interpolating processes if the training data are too close [13–17]. The bubblelike interpolation uncertainty becomes very small and can approach the size of the nugget used for numerical stability. In this case the diagnostics become highly sensitive to the nugget regulator, a clear sign of diagnostics gone awry. One way to avoid this issue is to spread the training points out relative to the length scale of the system, but this solution comes with a price. The estimates of $\boldsymbol{\theta}$ and the accuracy of the interpolants will suffer.

But the issue of numerical instability can be mitigated by avoiding interpolating processes altogether. Our goal is to test whether the c_n are drawn from one underlying process, so why not test the coefficients against this process rather than interpolating? The marginal variance of the underlying process should always be large with respect to the nugget size, which sidesteps the problem of comparable sizes. Because the emulator is not forced to interpolate the training data (only $\boldsymbol{\theta}$ is tuned), it may be less crucial to exclude training data from the validation set and would allow for more data to be used in the analysis. Diagnosing against the underlying process is thus the method of choice in this work.

Now we outline the diagnosis procedure for y_n. The predictions y_n allow us to test whether our truncation (and possibly interpolation) model behaves as advertised. The data used to estimate the parameters $\boldsymbol{\theta}$ could include all EFT predictions up to the highest order available, with the caveat that data that are too closely spaced could cause numerical instability in their estimates. Then an emulator for the full prediction can be constructed as discussed in Sec. II D, and the validation data could then consist of experimental data $\mathbf{y}_{\mathrm{exp}}$. If testing the truncation error, there is no issue computing diagnostics with the same x as was used in the training set.

Modulo the possible numerical issues discussed above, poorly behaving diagnostics are a sign that something is wrong with the EFT or the convergence model assumptions. To diagnose such problems requires a working knowledge of what success and failure look like. Hence, it will be more illuminating to see the diagnostics in action rather than to further discuss theory. After the following illustrations it should be clear that statistics is a powerful tool to diagnose physics and EFT failure modes.

IV. TOY APPLICATION

Figure 1 is a summary of how one takes the theoretical predictions of an EFT and extracts coefficients, which then inform the truncation error model as described in Sec. II. This section takes the predictions from Fig. 1(a) as given and provides an example workflow for extracting physical insights from the data. By working with toy predictions, we are able to assess the accuracy of both the truncation error bands and the extracted GP parameters $\boldsymbol{\theta}$. We test the distribution of the coefficients in Fig. 1(c) against our GP model, and then compute the truncation error distributions for each y_n in Fig. 1(a), followed by an assessment against a to-all-orders EFT prediction. Throughout, we show visually

how the diagnostics from Sec. III behave and how to know when they are failing.

Let us assume that mapping from predictions y_n to coefficients c_n is well understood, that is, the values of y_{ref} and Q are known. (We will return later to relax this assumption.) Then we are left with two projects: estimating the process from whence the coefficients were drawn and subsequently using that process to compute the truncation error distribution, all the while validating assumptions along the way. The first step is to outline the priors on the GP parameters $\boldsymbol{\theta}$. Here we take $\eta_0 = 0$ and $V_0 = 0$ (i.e., the mean of the coefficients is known to be zero), which corresponds to the belief that the coefficients are just as likely to be positive as negative and matches the distribution from which they are drawn. The prior on the variance $\bar{c}^2$ is given $\nu_0 = 1$ and $\tau_0^2 = 1$, which is fairly uninformative with a mode at $1/3$ and a very heavy tail. Finally, an improper uniform prior on the interval $(0, \infty)$ is taken for the length scale ℓ.

Once priors have been assigned to $\boldsymbol{\theta}$, the next step is to update our beliefs using the coefficients as data. We choose five points along each of the four c_n with $\Delta x \approx 0.25$, shown as dots in Fig. 7(a), and use them to compute $\mathbb{E}[\bar{c}^2]$ and ℓ_{MAP} [Eqs. (30) and (A54)]. This optimization procedure results in estimates that are remarkably close to the true parameter values, see Table I. Plugging these estimates into Eq. (18) yields an estimate of the *underlying* process that generated the c_n, whose mean and marginal variance are plotted on Fig. 7(a) as black solid and dashed lines. The interpolating processes can then be computed for each c_n [Eqs. (11)–(13)], whose means and marginal variances are given by colored dashed lines and filled areas in Fig. 7(a).

We know that the estimated process is quite close to the generating process, but we have no such guarantee when considering real EFT predictions. This is where the diagnostics proposed in Sec. III come into play. We choose to compare the c_n validation data, whose locations are indicated by minor ticks in Fig. 7(a), to the underlying GP, as opposed to each interpolating process. Thus, our diagnostics will indicate whether these data points could have feasibly been drawn from the GP that we have learned from the training points. The results of the diagnostics are shown in Figs. 7(b) and 7(c). As expected, the diagnostics behave quite well because the emulator matches the generating process. The D_{MD}^2 does return a small value for c_2, reflecting our intuition that c_2 looks abnormally small compared to the others. The pivoted Cholesky decomposition vector, plotted vs its index, shows an essentially random variation of points, meaning that the variation of the c_n about the mean are distributed

TABLE I. Parameter estimates for the emulators in Fig. 7(a) using a nugget of $\sigma_n^2 = 10^{-10}$.

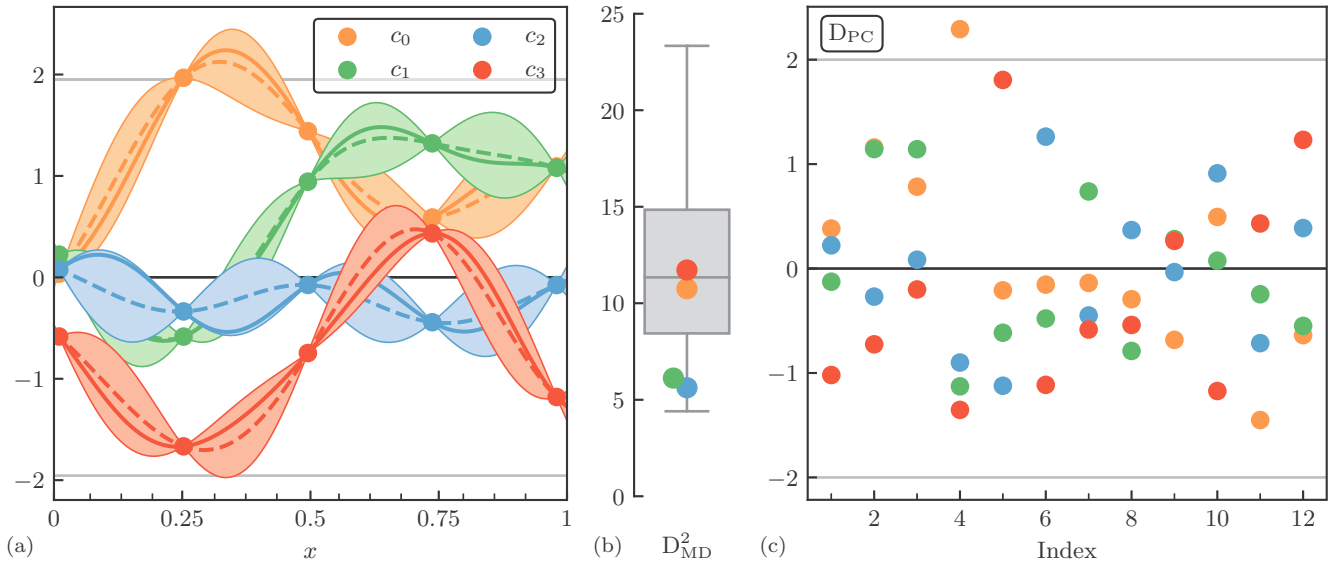| $\boldsymbol{\theta}$	Prior	Estimate	True
μ	$\mathcal{N}(0, 0)$	0	0
$\bar{c}^2$	$\chi^{-2}(1, 1)$	0.98^2	1
ℓ	$U(0, \infty)$	0.20	0.2
Q	$U(0, 1)$	0.48	0.5

FIG. 7. Toy coefficients and their corresponding diagnostics computed against the underlying process. Colors match across each subplot. (a) The coefficients from Fig. 1 and their corresponding GP emulators. The mean and 2σ bands of the calibrated underlying process are shown as black and gray lines, respectively. The colored dashed lines and corresponding bands are the interpolants and 2σ bands found by fitting this underlying GP to each c_n individually with Eqs. (11)–(13). The training data are denoted by dots, whereas the validation data locations are indicated by the minor ticks. (b) The Mahalanobis distance computed against the *underlying* (not interpolating) process. The interior line, box end caps, and whiskers on the box plot show the median, 50% credible intervals, and 95% credible intervals, respectively. The blue D_{MD}^2 for c_2 is the smallest, which shows that the diagnostic reflects our intuition as discussed in Fig. 1. Overall, the D_{MD}^2 are reasonably sized. (c) The pivoted Cholesky diagnostic $\mathbf{D}_{PC}$ vs index, with gray lines that represent its 2σ error bands. The points seem to be distributed as expected: approximately distributed as a standard Gaussian, with no real pattern vs index.

approximately as expected. Again, c_2 is rather small, given that the vertical variation should follow a standard normal, but the others show a mixture of large and small values of $\mathbf{D}_{PC}$. Such diagnostics show that the coefficients from the computed EFT orders conform to the assumptions of our convergence model.

The computed EFT orders are only half of the story; we desire truncation bands that take into account all not yet computed orders. By inserting point estimates of θ and Q into Eq. (21) [or by using its marginalized analog, Eq. (31)], posteriors for the full EFT summation are just as simple as the posterior for the c_n. The truncation error bands for each y_n, using the point estimates of θ in Table I, are shown in Fig. 8(a). As an example, a higher-order prediction y_{20} is taken as the true value of the prediction. Given the truncation error processes and the validation data of y_{20}, we can again compute diagnostics to determine whether the truncation errors behave as expected. We show the credible interval diagnostic in Fig. 8(b), which assesses the accuracy of the truncation error credible intervals. On average, the error estimates contain the true value of y in the proportion dictated by the credible interval. This is the sign of a healthy EFT convergence pattern.

Thus far we have assumed for simplicity that the pathway from order-by-order predictions to naturally sized coefficients, via Eq. (4), was known in advance. But first there is some work to do to even make this conversion: Assumptions must be made for y_{ref} and Q. As we have stated previously, y_{ref} can be derived based on dimensional analysis for the ob-

servable at hand, and Q can be based on a separation-of-scales argument specific to the EFT in use. But this can be easier said than done. We have found that one of the most informative pathways to physics insight is graphically exploring a data set. Plotting coefficients generated from different assumptions on y_{ref} and Q can show very different patterns, some of which better conform to the assumption that the c_n are identically distributed. Empirical exploration of coefficients' behavior can often lead to useful insights regarding the parameters that enter the EFT convergence model. For example, the assignment $y_{ref} = 1$ for spin observables in Ref. [3] was first hit on through such exploration. This process should be iterative, with scale arguments from the EFT reinforcing or critiquing such empirical observations. The diagnostics we have presented here lend statistical rigor to this process.

An important tool for exploring the EFT convergence is the model evidence [25] $\mathrm{pr}(\vec{\mathbf{y}}_k \mid \ell, Q, y_{ref})$ (where dependence on y_{ref} is shown explicitly here). Up to prior factors and an unimportant normalization constant, the model evidence is equivalent to the posterior for ℓ, Q, and y_{ref}. Suppose that ℓ and Q are both scalar quantities with uniform priors and y_{ref} is known. Then the posterior $\mathrm{pr}(\ell, Q \mid \vec{\mathbf{y}}_k)$ can be computed analytically, as is shown in Fig. 9 and summarized in Table I. Figure 9 is a pathway to learning the convergence pattern, and other EFT quantities, through data. In fact, even if Q is a function of x and other parameters, such as the EFT breakdown scale Λ_b, then the posterior for these parameters is still analytic and can be optimized to find a MAP value [see Eq. (A55)].
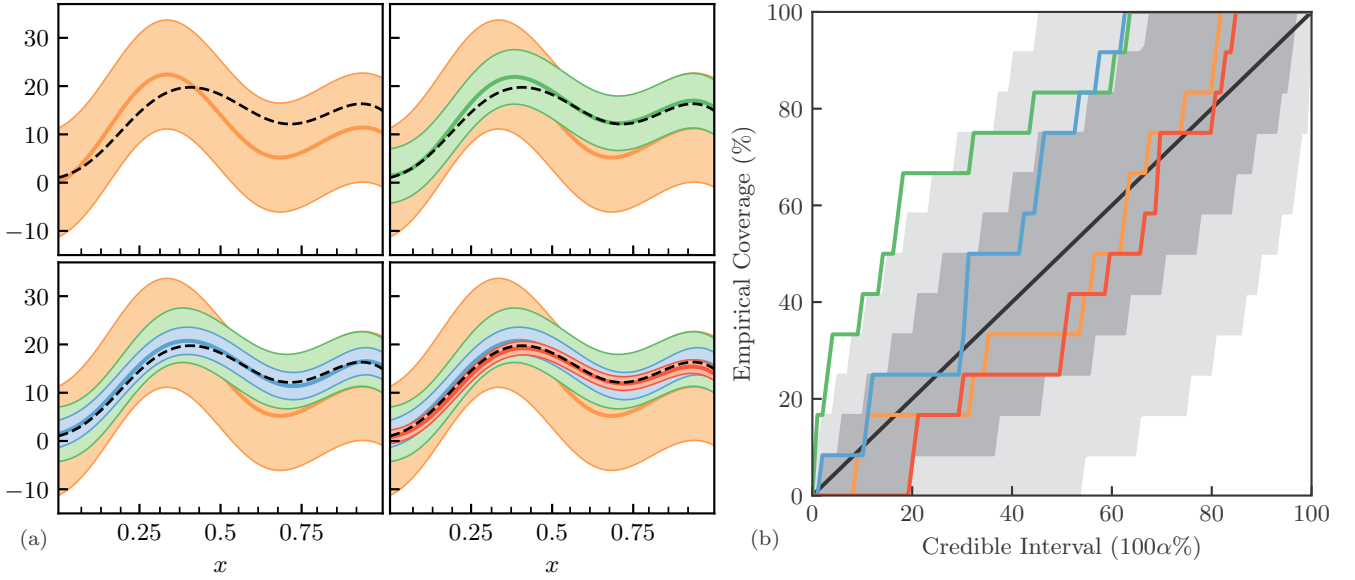
FIG. 8. (a) Toy predictions from Fig. 1 and their corresponding 2σ truncation error estimates generated from the underlying process in Fig. 7. The black dashed line is the observable computed up to order $n = 20$, which represents the true value of the observable and is used as validation data. Minor ticks denote the validation data locations used in the diagnostic. (b) The credible interval diagnostic for the truncation error bands. The stepwise nature of the lines arises from the finite number of points used in the diagnostic. Dark and light bands represent 1σ and 2σ credible intervals. By chance, y_1 is very close to the true curve, which results in a D_{CI} that is too accurate at small α.

But choosing Q may not be as simple as finding the best Λ_b, rather, one may need to decide between a discrete set of functional forms for $Q(x)$. In this case, differing $Q(x)$ can be thought of as models to choose between. Moreover, one could question the choice, or functional form, of y_{ref}. The
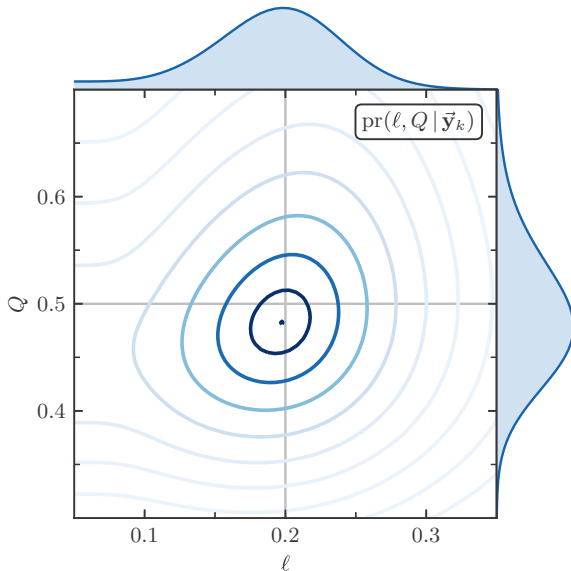

FIG. 9. The joint and marginal posteriors for ℓ and Q given data in Fig. 7(a), with a uniform prior $\text{pr}(\ell, Q) \propto 1$. The contours increment in approximately half-standard-deviation intervals, with a point denoting the MAP value. The gray lines represent the true values of ℓ and Q. The training data are spaced with $\Delta x \approx 0.25$, meaning that the ℓ posterior cannot entirely discount the possibility of small ℓ without prior information.

model evidence is useful as a diagnostic tool in this case as well. (ℓ could be integrated out numerically, if desired.) In this case, one can compute Bayes' factors, i.e., ratios of the model evidence for different choices (Q, y_{ref}), to determine the choice statistically favored by the EFT convergence pattern [25,26].

As an example of what can occur when Q is misestimated, we now choose $Q = 0.3$ (opposed to the true $Q = 0.5$) and repeat the analysis of Fig. 7. The results of this analysis are shown in Fig. 10. Since Q is underestimated, then the c_n grow with n, as dictated by Eq. (4). Thus c_3 is large compared to the others and causes the $\bar{c}$ estimate to increase. In general, the marginal variance is biased upward by large outliers. This is reflected by the Mahalanobis distance only capturing c_3, but the other coefficients are abnormally small. The assumption that the curves are identically distributed has broken down. The pivoted Cholesky decomposition in Fig. 10(c) provides further evidence of this breakdown; the lower-order components are too small compared to the reference distribution. Moreover, there is a trend of smaller errors near the right side of the chart: evidence that the length scale has been misestimated. Indeed, the length scale found by optimizing the likelihood is $\ell = 0.17$, small enough compared to the true value of $\ell = 0.2$ to show up in this diagnostic.

V. APPLICATION TO *NN* SCATTERING WITH CHIRAL EFT POTENTIALS

Now that we have introduced both the truncation error model and its diagnostics, we dedicate this section to a real-world example in low-energy nuclear physics. In this regime, chiral EFT is a popular tool used to develop a
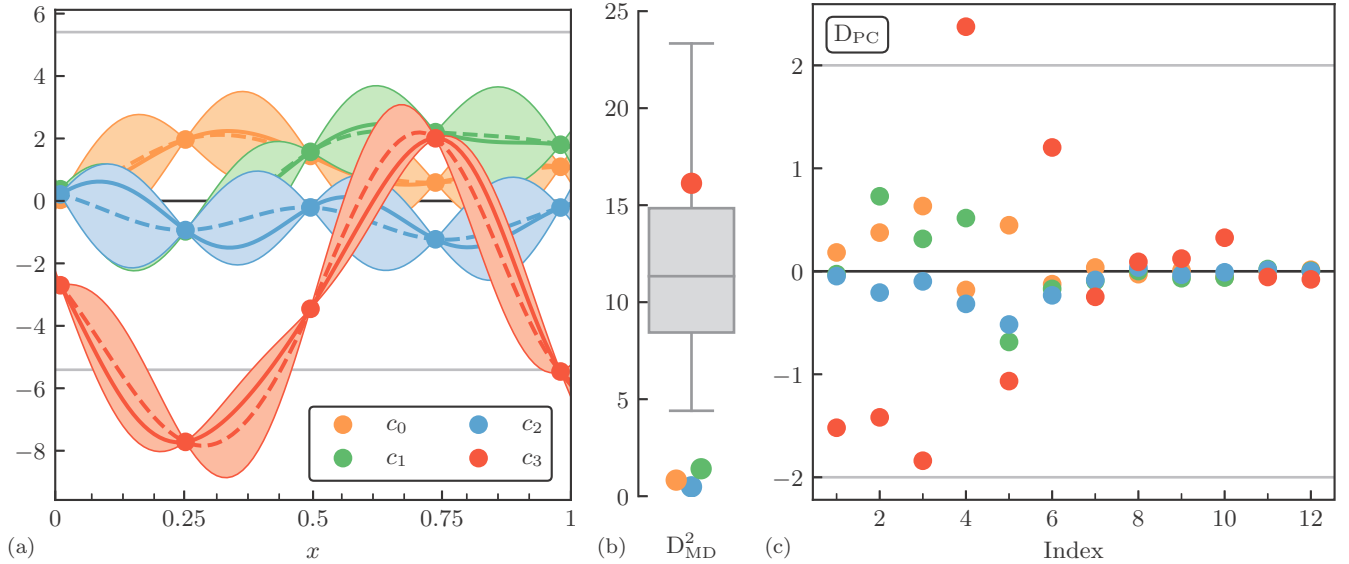
FIG. 10. Toy coefficients extracted using $Q = 0.3$ (opposed to the true $Q = 0.5$) and their corresponding diagnostics. In (a), the training data are denoted by dots, whereas the validation data locations are indicated by the minor ticks. See Fig. 7 for the remaining figure notation. The coefficients no longer look particularly i.i.d. given the large size of c_3. This biases the estimates of $\bar{c}^2$ and ℓ and results in the failure modes of the Mahalanobis distance and its pivoted Cholesky distribution. The funnel-like behavior of the pivoted Cholesky when moving from low to high index signals a misspecification of the length scale, which indeed is biased downward due to the incorrectly chosen Q.

systematically improvable description of nuclear observables. The most popular version of chiral EFT is the original proposed in Weinberg's seminal works, where a nuclear potential is written down and resummed using the Lippmann-Schwinger (LS) equation [27,28]. This resummation obscures the power counting that was manifest in the potential. Nevertheless, for certain regulators, the convergence of nucleon-nucleon scattering observables appears consistent with this power counting [2,3].

But the promising results in Refs. [2,3] were obtained using the pointwise convergence model. Though it permitted a statistical extraction of the EFT breakdown scale Λ_b, strong conclusions about Λ_b could not be made due to the lack of correlations inherent in the pointwise approach. Here we readdress a subset of these analyses using our GP approach. A more thorough treatment of this and other systems is a topic for future work.

We consider three neutron-proton (np) scattering observables: the differential cross section $\sigma(\theta)$ as a function of center-of-mass angle θ, the total cross section σ_{tot} as a function of laboratory energy E_{lab}, and the spin observable A as a function of θ. Each of the observables are computed using the semilocal potential of Epelbaum, Krebs, and Meißner (EKM) [29,30]. We find that, for NN data in chiral EFT, the coefficients look very smooth and center around a mean of zero, see Ref. [3]. Thus, we model the correlation structure using the squared exponential kernel, Eq. (9), and choose a mean function of zero.

Once the coefficients' mean and covariance functions have been specified, choices must be made for y_{ref} and Q. In Ref. [3], $y_{\text{ref}} = y_0$ was used for the total and differential cross section, while it was argued that $y_{\text{ref}} = 1$ is the appropriate scale for spin observables. Here we argue for a slight modifi-

cation: We choose $y_{\text{ref}} = y_5$ for the differential cross section since the leading-order prediction y_0 gets dangerously close to 0 in some regions, which causes unnatural peaks in the coefficient curves.

The expansion parameter Q is a ratio of low- to high-energy scales. The high-energy scale is the breakdown scale of the EFT Λ_b, whose value is obscured by the LS equation. For np scattering in chiral EFT, the low-energy scales include quantities such as the beam kinetic energy in the laboratory frame E_{lab} and the pion mass m_π. In the center-of-momentum frame, one can rewrite the laboratory energy as the relative momentum given the proton mass M_p and neutron mass M_n,

$$p_{\text{rel}}^2 = \frac{E_{\text{lab}} M_p^2 (E_{\text{lab}} + 2M_n)}{(M_p + M_n)^2 + 2M_p E_{\text{lab}}}. \tag{44}$$

With p_{rel} a function of E_{lab}, the expansion parameter is defined as

$$Q(E_{\text{lab}}) = \frac{f(p_{\text{rel}}, m_\pi)}{\Lambda_b}, \tag{45}$$

where $f(p_{\text{rel}}, m_\pi)$ is some mapping of p_{rel}, m_π. We take $\Lambda_b = 600$ MeV here and return to estimating its value in future work.

Again, due to the resummation performed by the LS equation, the exact mapping $f(p_{\text{rel}}, m_\pi)$ is not known. One would expect that $\lim_{p_{\text{rel}} \gg m_\pi} f(p_{\text{rel}}, m_\pi) = p_{\text{rel}}$ and $\lim_{p_{\text{rel}} \ll m_\pi} f(p_{\text{rel}}, m_\pi) = m_\pi$, but many functional forms could satisfy these constraints. In prior work, we have assumed the mapping $f(p_{\text{rel}}, m_\pi) \rightarrow \max(p_{\text{rel}}, m_\pi)$ or a smoothed max function

$$f(p_{\text{rel}}, m_\pi) \rightarrow \frac{p_{\text{rel}}^n + m_\pi^n}{p_{\text{rel}}^{n-1} + m_\pi^{n-1}} \tag{46}$$
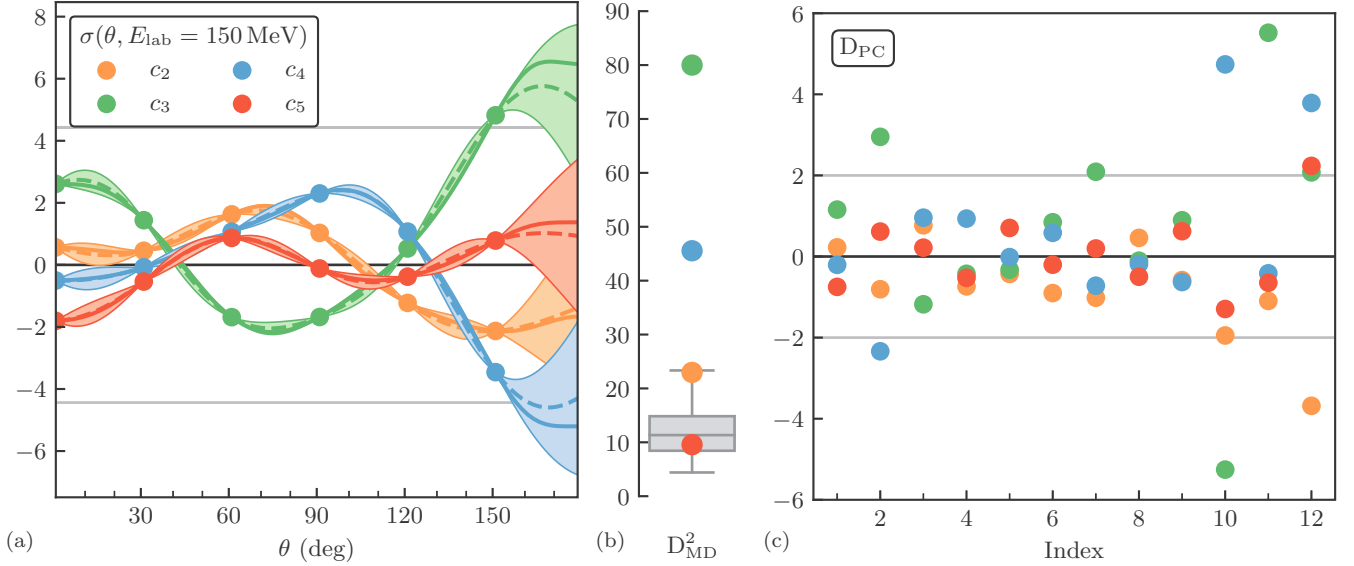
FIG. 11. Differential cross-section coefficients using the EKM $R = 0.9$ fm potential evaluated at $E_{\text{lab}} = 150$ MeV. In (a), the training data are denoted by dots, whereas the validation data locations are indicated by the minor ticks. See Fig. 7 for the remaining figure notation. The problems with the diagnostics for c_3 and c_4 are driven primarily by the backward angles, which are unusually large. Outlier values at large index in the pivoted Cholesky plot point to problems with nonstationarity.

with $n = 8$ [3]. Here we choose the smooth max function, Eq. (46).

Let us begin by testing the convergence pattern of the EFT for the total and differential cross sections. To do so, we follow the path set out in Sec. IV, that is, we extract the observable coefficients using the physically motivated assumptions of y_{ref} and Q described above and evaluate their features using the diagnostics of Sec. III. The analysis of the differential cross section evaluated at $E_{\text{lab}} = 150$ MeV is given in Fig. 11. The coefficients look beautifully i.i.d. and stationary until θ approaches backward angles, at which point c_3 and c_4 grow in size. This observation is reflected in the diagnostics computed at the validation points. It is clear that both c_3 and c_4 are outliers by the Mahalanobis distance, whereas c_2 and c_5 are appropriately sized. The pivoted Cholesky provides some insight into the problem: Unusually sized values at high index point to problems in the correlation structure or nonstationarity. And it turns out that these outliers at large index are exactly the validation points at large θ, as can be verified by removing all points greater than $\theta = 125°$ from the diagnostic. If the backward angles are excluded, then both diagnostics show excellent agreement with our model assumptions.

Now we turn to the analysis of the total cross section, which was extensively studied in Refs. [2,3] for differently regulated potentials. The coefficients and their diagnostics are shown in Fig. 12. The choice of test and validation points merit mention here. The highly nonstationary behavior at low energy is not representative of the high-energy data and would bias the analysis if a stationary process were used. This is a standard artifact that appears in our tested NN observables when plotted against laboratory energy [3]. Hence these points were omitted from this preliminary analysis. Additionally,

the training and test points must be spaced out—due to the very large length scale in this system—otherwise numerical instabilities appear. Given these caveats, the analysis of the total cross section nevertheless show interesting patterns. The diagnostics show that c_5 is a clear outlier, even when restricted to high-energy data. The pivoted Cholesky points again to the onset of nonstationarity or a misestimation of the length scale.

How does chiral EFT fare when subjected to these diagnostics? Both the total and differential cross sections show behavior that is consistent with our convergence model in some regimes yet inconsistent in others. In this case, there are plausible physical explanations for the diagnostics of unusual size. Likely explanations for the failures of the differential and total cross sections are a misspecification of Q at backward angles and low energies, respectively. This problem with the choice of Q will then result in incorrect truncation error estimates. It may be that the correct expansion parameter takes into account the momentum transfer for angle-dependent observables like the differential cross section. Meanwhile, the nonstationarity appearing at low energies occurs in the regime where $p_{\text{rel}} \approx m_\pi$, pointing to the possibility that the crossover region of Q was not parameterized correctly. Barring the wrong choice of Q, these diagnostic failure modes could simply point to the possibility that the EFT convergence pattern fails in these regimes according to the definition of convergence proposed in this work.

We have shown how observable coefficients can be analyzed in real EFT predictions; truncation errors follow as in Sec. IV. We illustrate this for the spin observable A, which is constrained to be zero at $\theta = 0°$ if the (small) magnetic-moment interaction is neglected, as it is here. Therefore contributions from all orders of the EFT are zero as well.
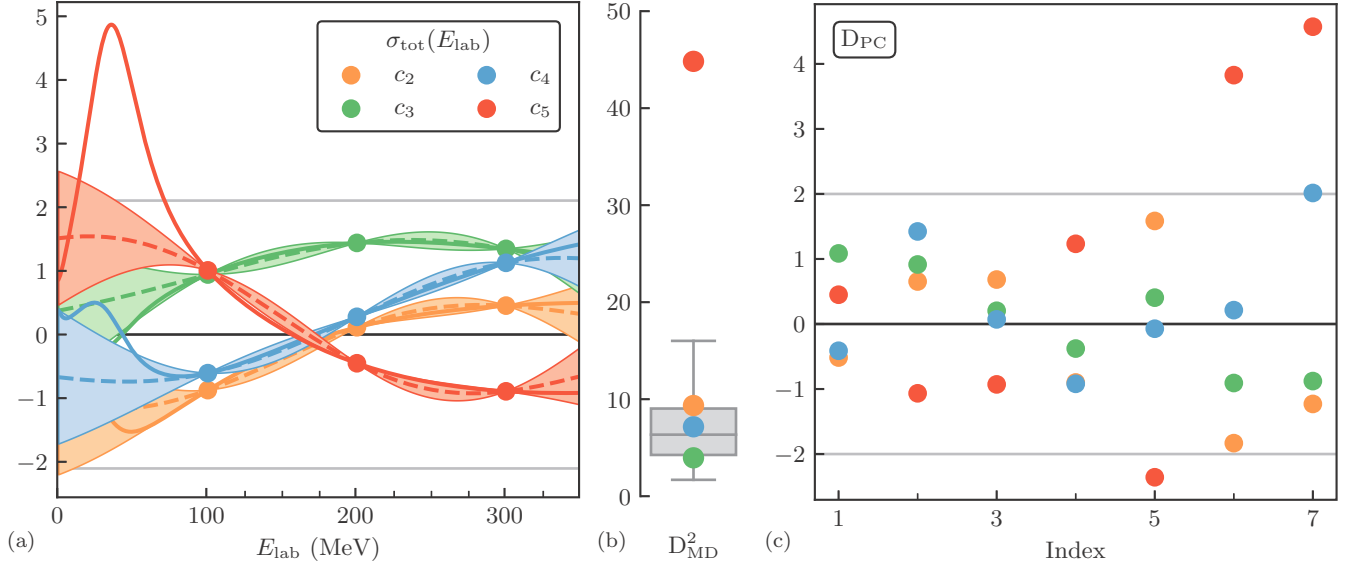
FIG. 12. (a) Total cross-section coefficients using the EKM $R = 0.9$ potential, as shown in Ref. [3]. The training data are denoted by dots, whereas the validation data locations are indicated by the minor ticks. See Fig. 7 for the remaining figure notation. (b) The D_{MD}^2 clearly shows a problem with c_5, whereas the others are behaving well. (c) When decomposed as $\mathbf{D}_{PC}$, the issue with c_5 appears at large index, indicating potential problems with nonstationarity. If low E_{lab} validation data were included, then the problem becomes much more visible.

We can accommodate the constraint using the formalism described in Sec. II D and shown in Fig. 6(c). Two sets of EFT truncation errors are proposed in Fig. 13, without and with the constraint. The bands in Fig. 13(b) are unrealistically large at $\theta = 0°$ given the information we know about this system. On the other hand, the bands in Fig. 13(c) vanish at $\theta = 0°$ as expected. This example uses $E_{lab} = 96$ MeV, but the point is general. The truncation errors are smoothly connected between the constrained and unconstrained regions in a manner harmonious with the way that lower-order coefficients approach zero. That is, the extent of the domain affected by this constraint is dependent on the length scale ℓ extracted from the known coefficients.
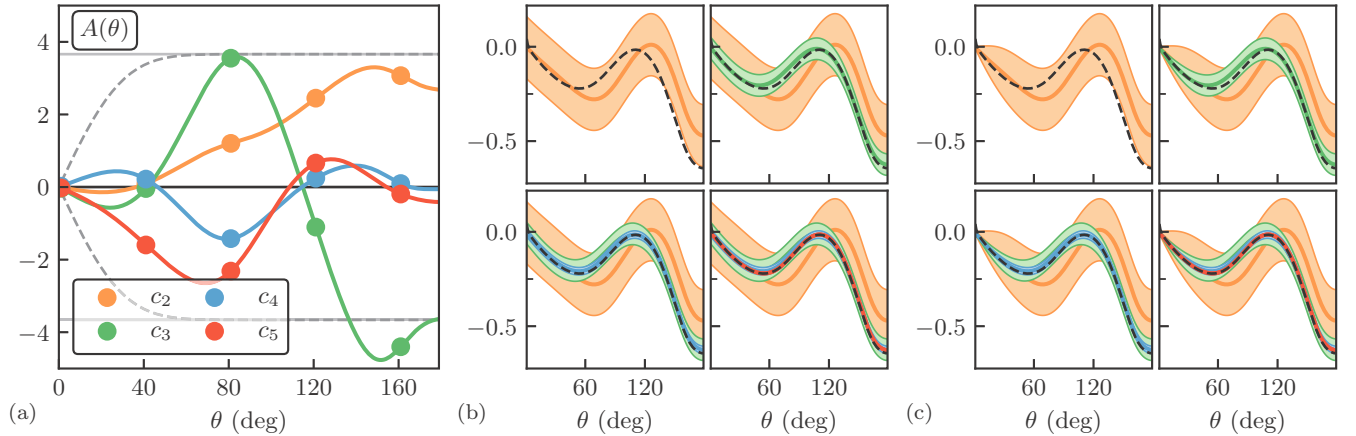

FIG. 13. The spin observable A, calculated at $E_{lab} = 96$ MeV using the EKM $R = 0.9$ fm potential, with the magnetic-moment interaction neglected for simplicity. A is then constrained to be zero at $\theta = 0°$. (a) The observable coefficients and the points used to learn $\bar{c}$ and ℓ of their underlying distribution. One can choose to estimate the truncation error using the coefficients' underlying distribution, shown as a solid gray line, or the underlying distribution conditioned to be 0 at $\theta = 0°$, shown as a darker gray dashed line. The conditioned distribution more closely follows the pattern of the coefficients and will more accurately describe the higher-order coefficients due to the symmetry of the system. (b) The EFT predictions and their 2σ truncation errors using the underlying unconstrained coefficient distribution. The truncation error does not vanish at $\theta = 0°$. (c) The EFT predictions and their 2σ truncation errors using the constrained coefficient distribution. Now the truncation error does vanish at $\theta = 0°$, as expected by symmetry arguments.

VI. SUMMARY AND OUTLOOK

Order-by-order predictions of a well-behaved EFT should converge regularly toward the all-orders value. We have formalized this idea into a falsifiable EFT convergence model. Rigorous estimates of the truncation error—induced by stopping the EFT expansion at a finite order—are now made possible, as are novel techniques for estimating EFT-related quantities, such as the expansion parameter Q, breakdown scale Λ_b, and LECs. We believe that formalizing and testing convergence is a crucial step when using an EFT, otherwise failure may go unrecognized.

This work answers several open questions faced by EFT practitioners that were left unresolved by our earlier work [1–3].

(1) *EFTs make predictions $y_n(x)$ that may be correlated in x, which in turn induces correlations in the truncation errors; how should that be taken into account when making predictions and fitting?* With the EFT power counting as our guide, we propose a change of variables that allows us to isolate the correlated quantities: observable coefficients $c_n(x)$ (see Fig. 1). The truncation error distribution follows by modeling the c_n as draws from a GP and summing all higher-order contributions. Originally, we had treated predictions at each kinematic value in a pointwise manner [2,3] and later modeled correlations in two extreme limits, which correspond to zero and infinite GP length scale ℓ [1]. Now ℓ can be estimated using data across multiple curves. This GP model could be adapted to correlations between discrete predictions as well.

(2) *Is this Bayesian formalism computationally demanding?* Bayesian updating of our convergence model parameters $\boldsymbol{\theta}$ is made simple and analytic by the use of conjugate priors. Our self-contained treatment shows how prior information on naturalness and correlations is updated using *all* available data and derives a novel analytic posterior for the EFT expansion parameter Q (see Fig. 9 and Appendix).

(3) *What about computationally expensive systems?* If it is too expensive to compute an EFT prediction at all points x or orders n, then $\boldsymbol{\theta}$ and Q, which control the truncation error distribution, can still be estimated from the available predictions without problem. (Even for inexpensive systems, if some low-order corrections are known to be unrepresentative of those at higher orders, then one would not want to use these in estimating the truncation error; for example, c_1 vanishes by symmetry in chiral EFT.) In these cases, the posteriors of $\boldsymbol{\theta}$ and Q are simply more influenced by the priors. After estimating $\boldsymbol{\theta}$ and Q, our convergence model then seamlessly combines truncation estimates with fast interpolation formulas to make predictions at all desired x (see Fig. 6 and Sec. II D, which also shows how to incorporate symmetry constraints).

(4) *How do these results affect EFT fitting?* We derived a novel likelihood function [Eq. (40)] that enables the EFT fit to account for all higher-order terms and their correlations in x. This extends to fitting with interpolation as well, allowing more experimental data to be used than might otherwise be computationally feasible. Because it is Gaussian, this likelihood can easily be incorporated into existing fitting procedures.

(5) *How is this falsifiable?* We provide a menu of model checking diagnostics that enable anyone to assess whether an EFT is working as advertised. We emphasize that these diagnostics require reference distributions (whose functional forms are given in Sec. III E) to assign statistical significance to their output. Claims about EFT convergence or truncation errors can then be falsified at any desired level of significance. We demonstrate the usefulness of these diagnostics on a toy example in Sec. IV, which exemplifies the general workflow of testing an EFT. Section V does the same for selected *NN*-scattering observables in chiral EFT. The success and failure modes within chiral EFT point to new research directions to pursue. This ability to probe for failure using statistical diagnostics and visualization means that our framework is not just a way to establish theoretical error bars but provides powerful tools for identifying limitations in our statistical model *and* in the physical understanding encoded in an EFT.

The stage is now set for a wide range of applications within chiral EFT; some of the paths in progress are as follows:

(i) A reevaluation of the EFTs critiqued with the pointwise convergence model in Ref. [3] in light of the GP model. Additionally, new variants of chiral EFT, which may show promising convergence properties in the *NN* sector and fewer regulator artifacts [31,32], should be included in the analysis.

(ii) A systematic study of EFT convergence beyond *NN* observables to assess its validity. This includes few-body, nuclear matter, and Compton scattering observables.

(iii) An exploration of the effects of our proposed likelihood [Eq. (40)] on the LECs' posteriors and in particular the effect on three-body LECs.

The models and tools we have developed apply generally to other EFTs and perturbative theories in general, so we anticipate many other applications. The pointwise and curve-wise models and associated model checking diagnostics are implemented in the `gsum` [24] Python package.

The `gsum` Python package is and can be freely downloaded and used [24]. This includes a notebook that creates all figures shown here, among other examples.

APPENDIX: DERIVATIONS

The mathematical details of the model described in Sec. II are shown here. Due to the choice of conjugate priors on μ, $\bar{c}^2$, almost all posteriors of interest can be provided in closed form.

Before getting into the fine details of the GP model discussed in this work, we believe it instructive to begin with a short aside. We have considered a simpler pointwise model in previous works [2,3], which is shown graphically in Fig. 3. To our knowledge, no one has yet worked out analytic expressions for the case of conjugate priors, whose derivations mirror the correlated case. Hence we will provide the results below for completeness. Again, conjugacy allows for efficient computation of truncation errors, parameter posteriors, and evidences using standard statistical libraries. We promote the use of these conjugate priors over the options laid out in previous works.

See Table II for a summary of the vector notation. Footnote 4 gives a brief overview of the statistics notation.

1. Pointwise model

Consider a set of EFT predictions $\{y_n\}$, each of which is a number rather than a function. In this case, the quantity of interest y_n could itself be a scalar or we could be considering a functional quantity $y_n(x)$ at a specific input point x in its domain. Here we wish to predict the truncation error independently of the function values at any surrounding input points. This procedure could be repeated at all points in the domain to produce error bands, as was done in Ref. [3]. This is what we call the *pointwise* model.

If the expansion parameter Q and reference scale $y_{\rm ref}$ are assumed to be known, then the data $\vec{y}_k$ can be converted to coefficients $\vec{c}_k$,

$$\vec{y}_k \equiv [y_0 \; y_1 \; \cdots \; y_k]^{\mathsf{T}} \Rightarrow \vec{c}_k \equiv [c_0 \; c_1 \; \cdots \; c_k]^{\mathsf{T}}. \quad (A1)$$

Here we assume that the coefficients c_n are normal given $\bar{c}^2$ and that $\bar{c}^2$ has a scaled inverse-χ^2 prior [11],

$$c_n \,|\, \bar{c}^2 \overset{\mathrm{iid}}{\sim} \mathcal{N}(0, \bar{c}^2), \quad (A2)$$

$$\bar{c}^2 \sim \chi^{-2}(\nu_0, \tau_0^2). \quad (A3)$$

TABLE II. Notation for vectors in different spaces.

Notation	Example	Description
—	$x \in \mathbb{R}^d$	Input space (suppressed)
Bold	$\mathbf{c}_n \in \mathbb{R}^N$	A set using inputs $\{x_i\}_{i=1}^N$
Arrow	$\vec{c}_k \in \mathbb{R}^{n_c}$	Orders $c_0, c_1, \ldots, c_k$
Bold italics	$\boldsymbol{\theta}$	GP parameters $\{\mu, \bar{c}, \ell\}$
—	μ, η, V	Mean basis space
Composition	$\vec{\mathbf{c}}_k \in \mathbb{R}^{n_c} \times \mathbb{R}^N$	Combination of above
	$\mathbf{x} \in \mathbb{R}^N \times \mathbb{R}^d$	

Equation (A3) has qualitatively similar properties to priors used in past works [2,3]: It is strictly positive, can penalize $\bar{c}^2$ for being or large or too small, and reduces to a scale-invariant prior for $\nu_0 = 0$. Unlike past priors on $\bar{c}^2$, the scaled inverse-χ^2 permits intuitive analytic posteriors for δy_k and Q for *all* choices of ν_0 and τ_0^2.

Analogously to Eq. (21), one can derive from Eqs. (5), (20), and (A2) that

$$\delta y_k \,|\, \bar{c}^2, Q \sim \mathcal{N}\left[0, y_{\rm ref}^2 \frac{Q^{2(k+1)}}{1 - Q^2} \bar{c}^2\right]. \quad (A4)$$

The above equations outline our prior beliefs about the coefficients, their widths, and how they combine to form a discrepancy term for y_k. If $\bar{c}$ was precisely known from *a priori* arguments or otherwise, then Eq. (A4) is sufficient to estimate the truncation error. Since this is not the case in general, we would like to update these priors based on order-by-order predictions.

Our belief about an unseen coefficient c_n is updated via

$$\mathrm{pr}(c_n \,|\, \vec{c}_k) = \int_0^\infty \mathrm{pr}(c_n \,|\, \bar{c}^2)\mathrm{pr}(\bar{c}^2 \,|\, \vec{c}_k)d\bar{c}^2. \quad (A5)$$

By marginalizing in information about $\bar{c}^2$, it is clear that the effect of data on c_n flows through $\bar{c}^2$. To compute the *posterior predictive distribution*, Eq. (A5), we must first understand how the observed coefficients $\vec{c}_k$ impact our understanding of $\bar{c}^2$. This is where the utility of conjugate priors is realized. The posterior of $\bar{c}^2$ has the same functional form as the prior—an inverse χ^2 distribution—but with ν and τ^2 that are updated by the data. This property simplifies our analysis greatly: Computing the posterior $\bar{c}^2 \,|\, \vec{c}_k$ is reduced to determining ν and τ^2. From Bayes' theorem and the fact that the $\vec{c}_k$ are marginally independent given $\bar{c}^2$,

$$\mathrm{pr}(\bar{c}^2 \,|\, \vec{c}_k) \propto \mathrm{pr}(\vec{c}_k \,|\, \bar{c}^2)\mathrm{pr}(\bar{c}^2) = \mathrm{pr}(\bar{c}^2)\prod_n \mathrm{pr}(c_n \,|\, \bar{c}^2)$$

$$\propto \frac{1}{\bar{c}^{\nu_0+n_c+2}} \exp\left[-\frac{1}{2\bar{c}^2}\left(\nu_0\tau_0^2 + \vec{c}_k^2\right)\right]. \quad (A6)$$

Above we have defined n_c to be the number of coefficients in $\vec{c}_k$ that are useful for induction (which is not always $k+1$, see Ref. [2]). On comparison with Eq. (27), we can read off

$$\nu = \nu_0 + n_c, \quad (A7)$$

$$\nu\tau^2 = \nu_0\tau_0^2 + \vec{c}_k^2. \quad (A8)$$

Therefore, we have shown that

$$\bar{c}^2 \,|\, \vec{c}_k \sim \chi^{-2}(\nu, \tau^2), \quad (A9)$$

with ν and τ^2 defined by Eqs. (A7) and (A8).

Now that the posterior for $\bar{c}^2$ has been established as Eq. (A9), we can return our attention to Eq. (A5). This integral evaluates to

$$\mathrm{pr}(c_n \,|\, \vec{c}_k) = \frac{1}{\sqrt{\pi\nu\tau^2}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{c_n^2}{\nu\tau^2}\right)^{-\frac{\nu+1}{2}}. \quad (A10)$$

Equation (A10) is the pdf for the Student-t distribution with degrees of freedom ν, mean 0, and scale τ, i.e.,

$$c_n \,|\, \vec{c}_k \sim t_\nu(0, \tau^2). \quad (A11)$$

In contrast to the notation of the Gaussian distribution, the variance of the Student-t is given by $\nu\tau^2/(\nu-2)$ and is only defined when $\nu > 2$.

Let us pause and ponder the implications of what we have derived. Equations (A2), (A7)–(A9), and (A11) fit together within one cohesive picture and relate to common estimators employed in frequentist statistics.

   (i) The variance of $\bar{c}^2$ decreases as the degrees of freedom ν increases. This uncertainty in $\bar{c}^2$ maps directly to the heaviness of the tails in the t distribution, which is also controlled by ν.

   (ii) A strict prior ($\nu_0 \gg 1$) or a lot of data ($n_c \gg 1$) both lead to $\mathrm{pr}(\bar{c}^2 \,|\, \vec{c}_k)$ being sharply peaked at $\bar{c}^2 \approx \tau^2$. This informative distribution for $\bar{c}^2$ corresponds to a $\mathrm{pr}(c_n \,|\, \vec{c}_k)$ that approximates a Gaussian with variance τ^2.

   (iii) Given a lot of data, then $\bar{c}^2 \approx \tau^2 \approx \vec{c}_k^2/n_c$, which is a frequentist estimator known as the *sample variance*. In this case $\mathrm{pr}(c_n \,|\, \vec{c}_k)$ will approximate a Gaussian with $\vec{c}_k^2/n_c$ as its variance.

   (iv) Rather than integrating over $\bar{c}^2$, we could use its mean value as a point estimate for the variance in Eq. (A2). The mean of Eq. (A9) is given by $\nu\tau^2/(\nu-2)$, which is exactly the variance of the t distribution in Eq. (A11).

By computing or assuming values for ν and τ^2, we are able to relax the assumption of a known $\bar{c}^2$ in Eq. (A2) by giving c_n a distribution with tails and a scale compatible with our understanding of $\bar{c}^2$. A Bayesian perspective allows us to incorporate prior beliefs about naturalness and update them with finite sets of data in a manner that is harmonious with the limiting cases assumed in frequentist statistics.

A better understanding of unseen observable coefficients c_n is useful as a means to better understand the truncation error δy_k. The steps in deriving Eq. (A11) lead analogously to the posterior

$$\delta y_k \,|\, \vec{y}_k, Q \sim t_\nu\left[0, y_{\mathrm{ref}}^2 \frac{Q^{2(k+1)}}{1-Q^2}\tau^2\right], \tag{A12}$$

where ν and τ^2 are computed via Eqs. (A7) and (A8). The distribution for the full prediction $y = y_k + \delta y_k$ is thus

$$y \,|\, \vec{y}_k, Q \sim t_\nu\left[y_k, y_{\mathrm{ref}}^2 \frac{Q^{2(k+1)}}{1-Q^2}\tau^2\right]. \tag{A13}$$

Equation (A13) is a novel result from this work and is the most straightforward method to compute posteriors of y in the pointwise model. Statistical packages such as SciPy have built-in t distributions which can easily calculate credible intervals for Eq. (A13).

Until now, we have assumed that the expansion parameter Q is known, but in practice this may not be the case. In fact, Q may even be a function of x in certain circumstances, which we assume here for generality. Fortunately, conjugacy permits the analytic computation of the (un-normalized) posterior

$$\mathrm{pr}(\mathbf{Q} \,|\, \vec{y}_k) \propto \mathrm{pr}(\vec{y}_k \,|\, \mathbf{Q})\mathrm{pr}(\mathbf{Q}). \tag{A14}$$

The right-hand side is (up to a simple integral over Q) the model evidence, which is a useful quantity for model comparison, and the likelihood $\mathrm{pr}(\vec{y}_k \,|\, \mathbf{Q})$ is exactly calculable. The pointwise model assumes that each of the $\vec{y}_k(x_i)$ are independent and each have their own $\bar{c}_i$, so that $\mathrm{pr}(\vec{y}_k \,|\, \mathbf{Q}) = \prod_i \mathrm{pr}[\vec{y}_k(x_i) \,|\, Q(x_i)]$. We consider each x_i individually and drop x_i for convenience. Note that since $y_0 = y_{\mathrm{ref}}c_0$ and $y_n = y_{n-1} + y_{\mathrm{ref}}c_n Q^n$, we can make a change of variables

$$\mathrm{pr}(\vec{y}_k \,|\, Q) = \frac{\mathrm{pr}(\vec{c}_k)}{\prod_n |y_{\mathrm{ref}}Q^n|}, \tag{A15}$$

where $c_0 = y_0/y_{\mathrm{ref}}$ and $c_n = \Delta y_n/y_{\mathrm{ref}}Q^n$ implicitly depend on Q. Thus we need only compute the joint distribution $\mathrm{pr}(\vec{c}_k)$. Although the distribution of one coefficient c_n follows a t distribution Eq. (A11), the same is not true of the joint distribution of multiple coefficients $\vec{c}_k$. Nevertheless, we can find an analytic form for their distribution using the clever manipulation of normalization constants. Let primes denote un-normalized distributions. Then both

$$\mathrm{pr}(\bar{c}^2 \,|\, \vec{c}_k) = \frac{(\nu_0\tau_0^2/2)^{\nu_0/2}}{\mathrm{pr}(\vec{c}_k)\Gamma(\nu_0/2)}\,\chi^{-2'}(\nu_0, \tau_0^2)\prod_n \frac{1}{\sqrt{2\pi}}\mathcal{N}'(0, \bar{c}^2), \tag{A16}$$

from Bayes' theorem, and

$$\mathrm{pr}(\bar{c}^2 \,|\, \vec{c}_k) = \frac{(\nu\tau^2/2)^{\nu/2}}{\Gamma(\nu/2)}\,\chi^{-2'}(\nu, \tau^2), \tag{A17}$$

from Eq. (A9), are true statements. Since $\mathrm{pr}(\vec{c}_k)$ does not depend on $\bar{c}^2$, equating the normalization coefficients yields the desired quantity

$$\mathrm{pr}(\vec{c}_k) = \frac{\Gamma(\nu/2)}{\Gamma(\nu_0/2)}\sqrt{\frac{1}{(2\pi)^{n_c}}\frac{(\nu_0\tau_0^2/2)^{\nu_0}}{(\nu\tau^2/2)^{\nu}}}. \tag{A18}$$

Through Eqs. (A15) and (A18) we have completely specified $\vec{y}_k \,|\, Q$, and $\vec{\mathbf{y}}_k \,|\, \mathbf{Q}$ is simply the product of these quantities.

Finally, we can write an un-normalized expression for the expansion parameter posterior by dropping terms independent of Q and multiplying $\vec{y}_k(x_i) \,|\, Q(x_i)$ at each x_i

$$\mathrm{pr}(\mathbf{Q} \,|\, \vec{\mathbf{y}}_k) \propto \mathrm{pr}(\mathbf{Q})\prod_i\left[\tau_i^{\nu_i}\prod_n Q_i^n\right]^{-1}. \tag{A19}$$

Rather than dealing with the multidimensional object $\mathbf{Q}$, one can instead parametrize Q as a low-energy scale $f(x)$ and a high-energy scale Λ_b, known as the breakdown scale of the EFT. These are related to Q via $Q(x) = f(x)/\Lambda_b$. The posterior for Λ_b, assuming $f(x)$ is known, has a simple relationship to the posterior for Q:

$$\mathrm{pr}(\Lambda_b \,|\, \vec{\mathbf{y}}_k, \mathbf{f}) \propto \mathrm{pr}(\vec{\mathbf{y}}_k \,|\, \mathbf{f}, \Lambda_b)\mathrm{pr}(\Lambda_b) = \mathrm{pr}(\vec{\mathbf{y}}_k \,|\, \mathbf{Q})\mathrm{pr}(\Lambda_b)$$

$$\propto \mathrm{pr}(\Lambda_b)\prod_i\left[\tau_i^{\nu_i}\prod_n Q_i^n\right]^{-1}. \tag{A20}$$

This is a more general version of the posterior given in Ref. [3], where the restriction $\nu_0 = 0$ was required. Here, thanks to conjugacy, ν_0 and τ_0^2 are free to be chosen by the

modeler. The extension to further parameters of Q requires only a modification of the prior $pr(\Lambda_b)$ in Eq. (A20).

2. Gaussian process model

Here we will expand on the template provided by subsection 1 of Appendix and derive analogous results for the GP truncation error model introduced in Sec. II. The results in this Appendix follow exactly from the analysis of the Bayesian linear model (see Ref. [33], chaps. 9 and 10) but are reproduced here for completeness. As a reminder, we have assumed that the c_n are GPs with the following form:

$$c_n(x) \,|\, \theta \stackrel{\mathrm{iid}}{\sim} \mathcal{GP}[\mu, \bar{c}^2 r(x, x'; \ell)], \tag{18'}$$

and placed a normal-inverse-χ^2 prior on $\mu, \bar{c}^2$,

$$\mu, \bar{c}^2 \sim \mathrm{N}\chi^{-2}\left(\eta_0, V_0, \nu_0, \tau_0^2\right). \tag{24'}$$

We now generalize the parametrization as promised in Sec. II C by introducing a length p vector of functions $b(x)$ that multiplies a length p vector of regression coefficients μ to allow for a non-constant mean function as in, e.g., Ref. [21]. That is, we make the replacement $\mu \to b^\mathsf{T}(x)\mu$ in Eq. (18). Both η_0 and V_0, previously scalars, are promoted to a length p vector and a $p \times p$ matrix, respectively. For the examples of EFT truncations shown in this work, we use $b(x) = 1$.

As before, we would like to condition on data to gain a better understanding of the observable coefficients and ultimately the discrepancy $\delta y_k(x)$. Again, assume that Q and y_{ref} are known so that our data are N vectors of coefficients $\vec{c}_k$ (each of length n_c) at the chosen input points. The combined $(n_c \times N)$-shaped observations are denoted $\vec{\mathbf{c}}_k$ (see Table II). The posterior predictive distribution for a new curve $c_n(x)$ is then

$$pr(c_n(x) \,|\, \vec{\mathbf{c}}_k) = \int pr[c_n(x) \,|\, \theta] pr(\theta \,|\, \vec{\mathbf{c}}_k) d\theta, \tag{A21}$$

where $\theta = \{\mu, \bar{c}^2, \ell\}$. As before, the data influence $c_n(x)$ by updating our beliefs about θ. Because of the conjugate prior on μ and $\bar{c}^2$, they can be updated and integrated out of Eq. (A21) analytically, but there is no conjugate prior for ℓ. For now, assume that ℓ is known; we will return to finding its posterior later. With a posterior for ℓ, one could use its MAP value as a point estimate in the following equations or marginalize over it numerically.

To make progress on Eq. (A21) requires deriving $pr(\mu, \bar{c}^2 \,|\, \vec{\mathbf{c}}_k, \ell)$. Again, since we know that this posterior has the same functional form as the prior, our task is to determine η, V, ν, τ^2. As usual, Bayes' theorem gets the ball rolling:

$$pr(\mu, \bar{c}^2 \,|\, \vec{\mathbf{c}}_k, \ell, Q) \propto pr(\vec{\mathbf{c}}_k \,|\, \theta, Q) pr(\mu, \bar{c}^2)$$

$$\propto pr(\mu, \bar{c}^2) \prod_n pr(\mathbf{c}_n \,|\, \theta). \tag{A22}$$

Next, by inserting Eqs. (18) and (24),

$$pr(\mu, \bar{c}^2 \,|\, \vec{\mathbf{c}}_k, \ell, Q)$$

$$\propto \bar{c}^{-(\nu_0 + Nn_c + 3)} \exp \frac{-1}{2\bar{c}^2}\left[\nu_0\tau_0^2 + (\mu - \eta_0)^\mathsf{T} V_0^{-1}(\mu - \eta_0)\right]$$

$$\times \exp \frac{-1}{2\bar{c}^2} \sum_n (\mathbf{c}_n - B\mu)^\mathsf{T} R_\ell^{-1}(\mathbf{c}_n - B\mu). \tag{A23}$$

where $R_\ell = r(\mathbf{x}, \mathbf{x}; \ell)$ and $B = b^\mathsf{T}(\mathbf{x})$ are $N \times N$ and $N \times p$ matrices, respectively. If $b(x) = 1$, as is assumed for the examples here, then $B = \mathbf{1}$, a length N vector of ones.

So far we can read off the updated value of $\nu = \nu_0 + Nn_c$ but must inspect the exponent further for the others. Define $\langle\vec{\mathbf{c}}_k\rangle$ as an $N \times 1$ vector that is the average over the n_c orders,

$$\langle\vec{\mathbf{c}}_k\rangle \equiv \frac{1}{n_c} \sum_n \mathbf{c}_n. \tag{A24}$$

The sum over the quadratic form can be rearranged as

$$\sum_n (\mathbf{c}_n - B\mu)^\mathsf{T} R_\ell^{-1}(\mathbf{c}_n - B\mu)$$

$$= n_c(\langle\vec{\mathbf{c}}_k\rangle - B\mu)^\mathsf{T} R_\ell^{-1}(\langle\vec{\mathbf{c}}_k\rangle - B\mu)$$

$$+ \sum_n (\mathbf{c}_n - \langle\vec{\mathbf{c}}_k\rangle)^\mathsf{T} R_\ell^{-1}(\mathbf{c}_n - \langle\vec{\mathbf{c}}_k\rangle), \tag{A25}$$

so that we can complete the square for μ in the exponent of Eq. (A23). The result is

$$\frac{-1}{2\bar{c}^2}[\nu\tau^2 + (\mu - \eta)^\mathsf{T} V^{-1}(\mu - \eta)], \tag{A26}$$

where

$$\eta \equiv V\left(V_0^{-1}\eta_0 + n_c B^\mathsf{T} R_\ell^{-1}\langle\vec{\mathbf{c}}_k\rangle\right), \tag{A27}$$

$$V \equiv \left(V_0^{-1} + n_c B^\mathsf{T} R_\ell^{-1} B\right)^{-1}, \tag{A28}$$

$$\nu \equiv \nu_0 + Nn_c, \tag{A29}$$

$$\nu\tau^2 \equiv \nu_0\tau_0^2 + \eta_0^\mathsf{T} V_0^{-1}\eta_0 + \sum_n \mathbf{c}_n^\mathsf{T} R_\ell^{-1}\mathbf{c}_n - \eta^\mathsf{T} V^{-1}\eta. \tag{A30}$$

After some algebra, one can rewrite Eq. (A30) in an intuitive way [33]:

$$\nu\tau^2 = \nu_0\tau_0^2 + n_c s^2$$

$$+ (\langle\vec{\mathbf{c}}_k\rangle - B\eta_0)^\mathsf{T}\left(\frac{R_\ell}{n_c} + BV_0B^\mathsf{T}\right)^{-1}(\langle\vec{\mathbf{c}}_k\rangle - B\eta_0), \tag{A31}$$

which updates the scale τ^2 by combining prior information with two other terms: one due to the discrepancy between the prior mean and sample mean and the other due to the sample variance s^2 for *correlated* observations,

$$s^2 \equiv \frac{1}{n_c} \sum_n (\mathbf{c}_n - \langle\vec{\mathbf{c}}_k\rangle)^\mathsf{T} R_\ell^{-1}(\mathbf{c}_n - \langle\vec{\mathbf{c}}_k\rangle). \tag{A32}$$

When $n_c \gg 1$, then $\tau^2 \approx s^2/N$. If one only wanted point estimates so that the c_n or δy_k posteriors remain Gaussian, then $\mu \approx \eta$ and $\bar{c}^2 \approx \nu\tau^2/(\nu - 2)$ provide approximate Bayesian estimators [see Eqs. (29) and (30)].

Returning to Eq. (A21), we must integrate out μ and $\bar{c}^2$ (again, assuming ℓ is fixed). Though the integral over μ is doable, there is neat trick to see the answer right away: Rewrite the GP as a mean μ (whose distribution is Gaussian) plus a zero-mean GP, that is,

$$c_n(x) = b^\mathsf{T}(x)\mu + \epsilon_1(x), \quad \epsilon_1(x) \sim \mathcal{GP}[0, \bar{c}^2 r(x, x'; \ell)], \tag{A33}$$

$$\mu = \eta + \epsilon_2, \qquad \epsilon_2 \sim \mathcal{N}(0, \bar{c}^2 V). \tag{A34}$$

Now substitute Eq. (A34) into Eq. (A33) and use Gaussian sum rules to arrive at

$$c_n(x) \,|\, \vec{\mathbf{c}}_k, \bar{c}^2, \ell \sim \mathcal{GP}\{b^\mathsf{T}(x)\eta, \bar{c}^2[r(x, x'; \ell) + b^\mathsf{T}(x)V b(x')]\}, \tag{A35}$$

which has marginalized out μ for us. The integral over $\bar{c}^2$, follows from the multivariate analog of Eq. (A11), that is, integrating Eq. (A35) over $\bar{c}^2$ yields a Student-t process $\mathcal{TP}_\nu[m(x), \kappa(x, x')]$ [20]

$$c_n(x) \,|\, \vec{\mathbf{c}}_k, \ell \sim \mathcal{TP}_\nu\{b^\mathsf{T}(x)\eta, \tau^2[r(x, x'; \ell) + b^\mathsf{T}(x)V b(x')]\}. \tag{A36}$$

Just like GPs, TPs are collections of random variables, any finite number of which have a joint t distribution. Specifically, the density of an N-dimensional multivariate t distribution is

$$\mathrm{pr}(\mathbf{y} \,|\, \nu, \mathbf{m}, K) = \frac{|K|^{-1/2}}{(\nu\pi)^{N/2}} \frac{\Gamma(\frac{\nu+N}{2})}{\Gamma(\nu/2)}$$
$$\times \left[1 + \frac{(\mathbf{y}-\mathbf{m})^\mathsf{T} K^{-1}(\mathbf{y}-\mathbf{m})}{\nu}\right]^{-\frac{\nu+N}{2}}. \tag{A37}$$

With this notation, $\kappa(x, x')$ is not the covariance function of $\mathcal{TP}_\nu[m(x), \kappa(x, x')]$, and, similarly, K is not the covariance matrix in Eq. (A37). Rather, the covariance is given by $\nu\kappa(x, x')/(\nu - 2)$ and is only defined for $\nu > 2$. This differs from the notation in, say, Refs. [20,21]. TPs, like GPs, can be used in regression, etc., but can better handle outliers in the data. Here, TPs only show up in our convergence model due to our choice of priors but are convenient objects to work with nevertheless.

For completeness, we provide the conditional distribution for Student-t processes, which is very similar to the GP version in Sec. II B 2. Suppose $\mathbf{y}_1$ and $\mathbf{y}_2$ are length n_1 and n_2 vectors, respectively, and

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} \sim t_\nu\left(\begin{bmatrix} \mathbf{m}_1 \\ \mathbf{m}_2 \end{bmatrix}, \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}\right). \tag{A38}$$

Then

$$\mathbf{y}_2 \,|\, \mathbf{y}_1 \sim t_{\nu+n_1}\left(\tilde{\mathbf{m}}_2, \frac{\nu + d_1}{\nu + n_1}\tilde{K}_{22}\right), \tag{A39}$$

where $\tilde{\mathbf{m}}_2$ and $\tilde{K}_{22}$ are defined as in Eqs. (12) and (13), and $d_1 = (\mathbf{y}_1 - \mathbf{m}_1)^\mathsf{T} K_{11}^{-1}(\mathbf{y}_1 - \mathbf{m}_1)$. One could derive this by starting with $\mathbf{y} \,|\, \sigma^2 \sim \mathcal{N}(\mathbf{m}, \sigma^2 K)$ and $\sigma^2 \sim \chi^{-2}(\nu, 1)$, using the Gaussian conditional rules described in Sec. II B 2 and then marginalizing over σ^2 at the end.

It is important to note that Eq. (A39) is *not* how we compute conditionals for, e.g., Eq. (A36). This is because conditioning and marginalizing over the mean do not commute. Consider the case where $\mathbf{f} = [\mathbf{f}_1 \quad \mathbf{f}_2]^\mathsf{T}$ is normally distributed just as c_n. Then

$$\mathrm{pr}(\mathbf{f}_2 \,|\, \mathbf{f}_1) = \int \mathrm{pr}(\mathbf{f}_2 \,|\, \mathbf{f}_1, \mu, \bar{c}^2)\mathrm{pr}(\mu, \bar{c}^2 \,|\, \mathbf{f}_1)d\mu d\bar{c}^2, \tag{A40}$$

from which it is clear that one must first follow the conditional rules of Eqs. (11)–(13) and update the posterior of μ and $\bar{c}^2$

before integrating them out. Thus, following the same logic as above,

$$\mathrm{pr}(\mathbf{f}_2 \,|\, \mathbf{f}_1) = t_\nu\left[R_{21}R_{11}^{-1}\mathbf{f}_1 + \tilde{B}_2\eta, \tau^2\left(\tilde{R}_{22} + \tilde{B}_2 V \tilde{B}_2^\mathsf{T}\right)\right], \tag{A41}$$

where, remembering that $B = \mathbf{1}$ (a vector of ones) is the basis matrix for our coefficients,

$$\tilde{B}_2 \equiv B_2 - R_{21}R_{11}^{-1}B_1. \tag{A42}$$

Of course, there could be other i.i.d. data beyond just $\mathbf{f}_1$ that could be used to compute $\mathrm{pr}(\mu, \bar{c}^2 \,|\, \mathbf{f}_1)$, which would serve only to adjust η, V, ν, and τ^2.

The above analysis is sufficient to finish our derivation for full uncertainty quantification of y_k, though the details will vary depending on the application as described in Sec. II D. In all cases, the derivation involves simply computing m_{th} and Σ_{th} from Sec. II D and integrating out μ and $\bar{c}^2$ by following Eq. (A36) or Eq. (A41). Below we provide the mean and scale function for $\mathcal{TP}_\nu[m_{\mathrm{th}}(x), \Sigma_{\mathrm{th}}(x, x')]$ in each case.

3. Inexpensive predictions

Let primes denote means with $\mu \to \eta$, e.g.,

$$m_{\delta k}'(x) \equiv b_{\delta k}^\mathsf{T}(x)\eta = y_{\mathrm{ref}}(x)\frac{Q(x)^{k+1}}{1 - Q(x)}\eta. \tag{A43}$$

Then

$$m_{\mathrm{th}}(x) = y_k(x) + m_{\delta k}'(x), \tag{A44}$$
$$\Sigma_{\mathrm{th}}(x, x'; \ell) = \tau^2\left[R_{\delta k}(x, x'; \ell) + b_{\delta k}^\mathsf{T}(x)V b_{\delta k}(x')\right]. \tag{A45}$$

4. Expensive predictions

Define $\tilde{b}_i(x)$ for $i \in \{k, \delta k\}$ as the nonvectorized version of Eq. (A42), using the appropriate $b_i(x)$ and R_i. Then

$$m_{\mathrm{th}}(x) = \tilde{m}_k'(x) + m_{\delta k}'(x), \tag{A46}$$
$$\Sigma_{\mathrm{th}}(x, x'; \ell) = \tau^2\{\tilde{R}_k(x, x'; \ell) + R_{\delta k}(x, x'; \ell)$$
$$+ [\tilde{b}_k(x) + b_{\delta k}(x)]^\mathsf{T} V[\tilde{b}_k(x') + b_{\delta k}(x')]\}. \tag{A47}$$

5. Constraints

One only needs to make the replacements $m_{\delta k}' \to \tilde{m}_{\delta k}'$, $b_{\delta k} \to \tilde{b}_{\delta k}$, and $R_{\delta k} \to \tilde{R}_{\delta k}$ in the above equations.

In the final part of this Appendix, we derive the formulas for the posterior $\mathrm{pr}(\ell, \mathbf{Q} \,|\, \vec{\mathbf{y}}_k) \propto \mathrm{pr}(\vec{\mathbf{y}}_k \,|\, \ell, \mathbf{Q})\mathrm{pr}(\ell, \mathbf{Q})$. The quantity $\mathrm{pr}(\vec{\mathbf{y}}_k \,|\, \ell, \mathbf{Q})$ is (almost) the marginal likelihood (or, evidence) and can be computed exactly. The first step is to make a change of variables from the data $\vec{\mathbf{y}}_k$ to coefficients

$$\mathrm{pr}(\vec{\mathbf{y}}_k \,|\, \ell, \mathbf{Q}) = \frac{\mathrm{pr}(\vec{\mathbf{c}}_k \,|\, \ell)}{\prod_{n,i} |y_{\mathrm{ref}}(x_i)Q^n(x_i)|}. \tag{A48}$$

To find $\mathrm{pr}(\vec{\mathbf{c}}_k \,|\, \ell)$, we will make use of the same normalizing constant trick as in the uncorrelated case. Again, let primes denote un-normalized distributions and Z_{dist} be their

normalizing constants. By Bayes' theorem,

$$\text{pr}(\mu, \bar{c}^2 \,|\, \vec{\mathbf{c}}_k, \ell) = \frac{Z_{\chi,0}^{-1} Z_{\mathcal{N}}^{-n_c}}{\text{pr}(\vec{\mathbf{c}}_k \,|\, \ell)} \, \text{N}\chi^{-2'}\left(\eta_0, V_0, \nu_0, \tau_0^2\right)$$

$$\times \prod_n \mathcal{N}'(\mu, \bar{c}^2 R_\ell) \quad (A49)$$

and by conjugacy

$$\text{pr}(\mu, \bar{c}^2 \,|\, \vec{\mathbf{c}}_k, \ell) = \frac{1}{Z_\chi} \, \text{N}\chi^{-2'}(\eta, V, \nu, \tau^2). \quad (A50)$$

Given that

$$Z_{\chi,0}^{-1} = \frac{(\nu_0 \tau_0/2)^{\nu_0/2}}{\sqrt{|2\pi V_0|}\Gamma(\nu_0/2)}, \quad Z_{\mathcal{N}}^{-1} = \frac{1}{\sqrt{|2\pi R_\ell|}} \quad (A51)$$

with a similar form for Z_χ, then

$$\text{pr}(\vec{\mathbf{c}}_k \,|\, \ell) = \frac{\Gamma(\nu/2)}{\Gamma(\nu_0/2)} \sqrt{\frac{|V|/|V_0|}{|2\pi R_\ell|^{n_c}}} \frac{(\nu_0 \tau_0^2/2)^{\nu_0}}{(\nu \tau^2/2)^{\nu}}. \quad (A52)$$

The joint posterior $\text{pr}(\ell, \mathbf{Q} \,|\, \vec{\mathbf{y}}_k)$ follows from Eqs. (A48) and (A52) and a choice of prior $\text{pr}(\ell, \mathbf{Q})$. It follows that the unnormalized marginal posteriors for $\mathbf{Q}$ and ℓ are

$$\text{pr}(\mathbf{Q} \,|\, \vec{\mathbf{y}}_k, \ell) \propto \frac{\text{pr}(\mathbf{Q})}{\tau^\nu \prod_{n,i} |Q^n(x_i)|}, \quad (A53)$$

$$\text{pr}(\ell \,|\, \vec{\mathbf{y}}_k, \mathbf{Q}) \propto \frac{\text{pr}(\ell)}{\tau^\nu} \sqrt{\frac{|V|}{|R_\ell|^{n_c}}}. \quad (A54)$$

The posterior for the breakdown scale Λ_b, where $Q(x; \Lambda_b) = f(x)/\Lambda_b$, follows from the same logic as in subsection 1 of Appendix and is given by

$$\text{pr}(\Lambda_b \,|\, \vec{\mathbf{y}}_k, \ell, \mathbf{f}) \propto \frac{\text{pr}(\Lambda_b)}{\tau^\nu \prod_{n,i} |Q^n(x_i)|}. \quad (A55)$$

MAP values can then be found numerically and used as point estimates in the analytic equations derived in this Appendix.

[1] S. Wesolowski, R. J. Furnstahl, J. A. Melendez, and D. R. Phillips, J. Phys. G: Nucl. Part. Phys. **46**, 045102 (2019).

[2] R. J. Furnstahl, N. Klco, D. R. Phillips, and S. Wesolowski, Phys. Rev. C **92**, 024005 (2015).

[3] J. A. Melendez, S. Wesolowski, and R. J. Furnstahl, Phys. Rev. C **96**, 024003 (2017).

[4] M. Cacciari and N. Houdeau, J. High Energy Phys. 09 (2011) 039.

[5] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn, Stat. Sci. **4**, 409 (1989).

[6] N. Cressie, Terra Nova **4**, 613 (1992).

[7] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning* (The MIT Press, Cambridge, MA, 2006).

[8] R. M. Neal, in *Bayesian Learning for Neural Networks*, edited by P. Bickel, P. Diggle, S. Fienberg, K. Krickeberg, I. Olkin, N. Wermuth, and S. Zeger, Lecture Notes in Statistics, Vol. 118 (Springer, New York, 1996).

[9] D. J. C. MacKay, in *Neural Networks and Machine Learning*, NATO ASI Series, Vol. 168, edited by C. M. Bishop (Springer, New York, 1998), pp. 133–166.

[10] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms* (Cambridge University Press, Cambridge, UK, 2003).

[11] A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin, *Bayesian Data Analysis*, 3rd ed. (CRC Press, Boca Raton, FL, 2013).

[12] M. L. Stein, *Interpolation of Spatial Data: Some Theory for Kriging*, Springer Series in Statistics (Springer-Verlag, New York, 1999).

[13] R. Ababou, A. C. Bagtzoglou, and E. F. Wood, Math. Geol. **26**, 99 (1994).

[14] I. Andrianakis and P. G. Challenor, Comput. Stat. Data Anal. **56**, 4215 (2012).

[15] H. Mohammadi, R. L. Riche, N. Durrande, E. Touboul, and X. Bay, arXiv:1602.00853 [math, stat].

[16] A. Pepelyshev, in *mODa 9 Advances in Model-Oriented Design and Analysis*, edited by A. Giovagnoli, A. C. Atkinson, B. Torsney, and C. May (Physica-Verlag HD, Heidelberg, 2010), pp. 149–156, arXiv:1005.4385.

[17] P. Ranjan, R. Haynes, and R. Karsten, Technometrics **53**, 366 (2011).

[18] T. J. Santner, B. J. Williams, and W. I. Notz, *The Design and Analysis of Computer Experiments*, 2nd ed., Springer Series in Statistics (Springer-Verlag, New York, 2018).

[19] J. Salvatier, T. V. Wiecki, and C. Fonnesbeck, PeerJ Comput. Sci. **2**, e55 (2016).

[20] A. Shah, A. Wilson, and Z. Ghahramani, in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, Vol. 33, edited by S. Kaski and J. Corander (PMLR, Reykjavik, Iceland, 2014), pp. 877–885.

[21] L. S. Bastos and A. O'Hagan, Technometrics **51**, 425 (2009).

[22] N. Cressie, J. Int. Assoc. Math. Geol. **17**, 563 (1985).

[23] A. W. Bowman and R. M. Crujeiras, Comput. Stat. Data Anal. **66**, 19 (2013).

[24] J. Melendez, "gsum", Python package version 0.2, available for download from Github, github.com/buqeye/gsum.

[25] S. Wesolowski, N. Klco, R. J. Furnstahl, D. R. Phillips, and A. Thapaliya, J. Phys. G: Nucl. Part. Phys. **43**, 074001 (2016).

[26] R. Trotta, Contemp. Phys. **49**, 71 (2008).

[27] S. Weinberg, Appl. Phys. Lett. B **251**, 288 (1990).

[28] S. Weinberg, Nucl. Phys. B **363**, 3 (1991).

[29] E. Epelbaum, H. Krebs, and U.-G. Meißner, Phys. Rev. Lett. **115**, 122301 (2015).

[30] E. Epelbaum, H. Krebs, and U.-G. Meißner, Eur. Phys. J. A **51**, 53 (2015).

[31] P. Reinert, H. Krebs, and E. Epelbaum, Eur. Phys. J. A **54**, 86 (2018).

[32] D. R. Entem, R. Machleidt, and Y. Nosyk, Phys. Rev. C **96**, 024004 (2017).

[33] A. O'Hagan, *Bayesian Inference*, Kendall's Advanced Theory of Statistics (Edward Arnold, London, 1994).