Model orthogonalization and Bayesian forecast mixing via principal component analysis

P. Giuliani [1,*] K. Godbey [1,†] V. Kejzlar [2] and W. Nazarewicz [1,3,‡]

[1] *Facility for Rare Isotope Beams, Michigan State University, East Lansing, Michigan 48824, USA*

[2] *Mathematics and Statistics Department, Skidmore College, Saratoga Springs, New York 12866, USA*

[3] *Department of Physics and Astronomy, Michigan State University, East Lansing, Michigan 48824, USA*

(Received 2 July 2024; accepted 19 August 2024; published 9 September 2024)

One can improve predictability in the unknown domain by combining forecasts of imperfect complex computational models using a Bayesian statistical machine learning framework. In many cases, however, the models used in the mixing process are similar. In addition to contaminating the model space, the existence of such similar, or even redundant, models during the multimodeling process can result in misinterpretation of results and deterioration of predictive performance. In this paper we describe a method based on the principal component analysis that eliminates model redundancy. We show that by adding model orthogonalization to the proposed Bayesian model combination framework, one can arrive at better prediction accuracy and reach excellent uncertainty quantification performance.

DOI: 10.1103/PhysRevResearch.6.033266

I. INTRODUCTION

Modeling is a crucial part of many scientific disciplines. Within the framework of the scientific method, models are designed to create postdictions about past data, describe phenomena, and make predictions about the future observations. In many cases, several alternative (and competing) models are available to describe a given physical phenomenon. These models might be based on different theoretical foundations, be calibrated to different datasets, involve different computational algorithms, and often will have a different accuracy when it comes to forecasting (i.e., postdictions or predictions).

Choosing one of the models either arbitrarily or using off-the-shelf model selection method leads to poor uncertainty quantification (UQ). To this end, combining together a set of different models is advisable [1–3]. One of the key aspects of multimodeling is the choice of individual models whose forecasts are combined and the elimination of very similar, or even redundant, models is a challenge.

The objective of this paper is to find the effective number of models in the model set and determine their relative contributions to the combined forecast obtained within the Bayesian setting. To this end, we use principal component analysis (PCA). We shall demonstrate that incorporating model preselection and model orthogonalization via PCA into the multimodel framework leads to (i) faster and scalable forecasting (only the reduced set of orthogonalized models is mixed); (ii) improved computational robustness of multimodeling; (iii) increased interpretability through elimination of similar models; and (iv) improved predictive performance as properly orthogonalized models are less prone to overfitting.

Below, we first briefly review the fundamentals of our multimodel framework. As an illustration, we apply our approach to a pedagogic example of predicting nuclear binding energies using a simple analytic model and study common modeling scenarios. Finally, we show the opportunities provided by our method for practitioners on a case study of predicting binding energies using a set of realistic models based on the nuclear density functional theory.

II. MULTIMODELING

A. Reasonable models

Let us consider a set of models $\mathcal{M}_1, \ldots, \mathcal{M}_m$ which are used to forecast observations of a physical process at locations $x_i \in \mathcal{X} \subset \mathbb{R}^n$, $i = 1, \ldots, n$, where $\mathcal{X}$ indicates the input domain of observations. As the main goal of this paper is to develop a framework for quantified extrapolations, we introduce a notion of "reasonable models," i.e., the models which (i) are well suited to provide sound quantified forecasts within the input domain $\mathcal{X}_0 \subset \mathcal{X}$ in which experimental observations exist and (ii) have sound physical/microscopic foundations enabling extrapolations and UQ outside of $\mathcal{X}_0$ into the unknown domain $\mathcal{X}^* = \mathcal{X} - \mathcal{X}_0$. Each model $\mathcal{M}_k$ is calibrated to a dataset within the input subdomain $\mathcal{X}_k \subset \mathcal{X}_0$. The condition (ii) excludes phenomenological, many-parameter formulas fitted to experimental data, which yield uncontrolled extrapolations in the unknown domain $\mathcal{X}^*$.

B. Combining forecasts

In this section, we briefly overview three basic approaches to combining forecasts of reasonable models (for a

*Contact author: giulianp@frib.msu.edu

†Contact author: godbey@frib.msu.edu

‡Contact author: witek@frib.msu.edu

comprehensive discussion, the reader is referred to Refs. [1,2]). In general, the goal of forecast combination is to use several models to predict observations $y(x)$ of a physical process at new locations $x^* \in \mathcal{X}^*$ using information from both measurements/observations $\boldsymbol{y} = [y(x_1), \ldots, y(x_n)]$ and model calculations at the input locations $x_i \in \mathcal{X}_0$.

One common approach for combining forecasts is *Bayesian model averaging* (BMA) [4–6]. Here, the resulting prediction is given by a mixture of individual models' posterior predictive distributions where the BMA model weights reflect the fit of a statistical model to data independently of the set of available models and are obtained by marginalizing over model parameters (i.e., using Bayesian evidence). However, BMA relies on theoretical assumptions which are inappropriate for approximate modeling of complex systems (e.g., one of the candidate models is a "true" model that perfectly describes the physical reality).

The *Bayesian model mixing* (BMM) framework, an extension of Bayesian stacking [2,3,7,8], implicitly assumes that while none of the models $\mathcal{M}_k$ is true, the underlying physical process is well captured by a linear combination of the models. A resulting statistical model can be written as [3,8–10]

$$y(x_i) = \sum_{k=1}^{m} \omega_k(x_i) f_k(x_i) + \sigma_i \epsilon_i, \tag{1}$$

where σ_i represents the scale of the error of the mixture model (possibly including experimental and theoretical errors), $\epsilon_i \overset{\text{iid}}{\sim} N(0,1)$, $f_1(x_i), \ldots, f_m(x_i)$ are forecasts for the datum $y(x_i)$ provided by the m theoretical models considered, and $\boldsymbol{\omega}(x_i) \equiv [\omega_1(x_i), \ldots, \omega_m(x_i)]$ are their respective *weights* which are often adjusted to fulfill the simplex constraint:

$$\omega_1, \ldots, \omega_m \geqslant 0, \ \sum_{k=1}^{m} \omega_k = 1. \tag{2}$$

The weights can, in principle, depend locally on the input domain or can be global, i.e., domain independent. The distribution of BMM model weights additionally depends on the modeling choice for $\boldsymbol{\omega}$ and the set of models considered [10]. As demonstrated in previous studies [10–12], combining models using BMM outperforms BMA in terms of both prediction accuracy and UQ. Consequently, in this paper's case studies we do not pursue the BMA strategy.

The *Bayesian model combination* (BMC) strategy aims to find a combined forecast that outperforms individual forecasts by hoping that systematic deficiencies of different models will compensate. Here, the focus is on the overall performance rather than the relation of the models $\mathcal{M}_k$ to the true model. In BMC, one assumes the mixture model in the form [1,13]

$$y(x_i) = \sum_{k=1}^{m} c_k(x_i) f_k(x_i) + f_0 + \sigma_i \epsilon_i, \tag{3}$$

where $\boldsymbol{c} = [c_k]$ are model *amplitudes* and f_0 is an optional constant term. If $f_0 = 0$, BMC is reduced to the unrestricted combined forecast whose amplitudes $\boldsymbol{c}$ can be determined by unrestricted chi-square minimization or by constructing a Bayesian posterior distribution given the data. In general, the amplitudes of BMC do not have to be positive [11,13]. Some applications of BMC [14] impose the simplex constraint (2); this results in a worsened performance [13].

C. Model similarity and redundancy

In many cases, physics models may have a similar mathematical foundation but their parameters are calibrated using different methodologies. It is also possible that models are in fact identical in spite of their different formulation. Consider, e.g., (i) a model given by a polynomial of order n (Taylor expansion) and (ii) a model given by a Legendre multipole expansion of order n. Both models are manifestly identical, if calibrated to the same dataset. This extreme situation is referred to as *model redundancy* [15].

In addition to "polluting" the model space $[\mathcal{M}_k]$, the existence of redundant or similar models during the multimodeling process can result in difficulties with obtaining reliable inferences and hence misinterpretation of results and deterioration of predictive performance. The standard application of BMA will particularly suffer from this situation given that each model weight is calculated independently of the set of available models, allowing for overemphasis of model classes with repeated representations. Consequently, adding model preselection and orthogonalization to model combination pipelines is important and, as we show in the following, relatively straightforward.

III. MODEL ORTHOGONALIZATION

PCA [16] and singular value decomposition (SVD) [17] are two related methods that have become essential tools for data compression, signal processing, data visualization, feature selection, and dimensionality reduction across science and engineering [18]. In the past [19], PCA has been specifically applied to model orthogonalization; see also other PCA applications to combining forecasts [20,21], including a recent application to nuclear mass models [22].

For the purpose of this paper, we will use PCA to identify the first p principal components, or directions of maximal variability, across a set of m theoretical mass models ($p \leqslant m$). We first consider forecasts of the m different models: $f_k(x_i)$ ($i = 1, \ldots, n$), where n is the number of model results. For our specific application, these forecasts consist of the $n \approx 600$ computed nuclear binding energies of different even-even nuclei characterized by the number of protons Z and neutrons N, i.e., the domain of interest is defined by $x_i = (Z_i, N_i)$; see Ref. [23]. We arrange these model results into a matrix $\boldsymbol{X^0} = (X^0_{i,k})$, where a fixed column represents the forecast of a single model f_k across all n measurements, while a fixed row represents the predictions of all m models on a fixed mass of nucleus x_i. From this matrix, we construct the centered matrix $\boldsymbol{X^c} \equiv (X^c_{i,k})$,

$$X^c_{i,k} = X^0_{i,k} - \phi_0(x_i), \tag{4}$$

by subtracting the average $\phi_0(x)$ of all the models (columns):

$$\phi_0(x_i) = \frac{1}{m} \sum_{k=1}^{m} X^0_{i,k} = \frac{1}{m} \sum_{k=1}^{m} f_k(x_i). \tag{5}$$

The vector $\phi_0(\boldsymbol{x})$, where $\boldsymbol{x} = [x_i]$ denotes the list of inputs, represents the average forecast of all models and in principle

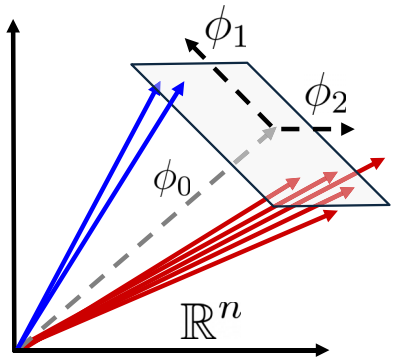

FIG. 1. Schematic representation of the PCA approach for the model combination. Here, two model classes consist of two and five models, respectively, and are represented as vectors in a space $\mathbb{R}^n$. This collection of seven models is approximated in the affine space (gray rectangle) spanned by the constant ϕ_0 term (dashed light-gray arrow) and the two principal components ϕ_1 and ϕ_2 (dashed black arrows).

contains the main features that a reasonable physical model should have. The deviations from this average, contained in the matrix $\boldsymbol{X}^{\boldsymbol{c}}$, serve to characterize individual models. The matrix $\boldsymbol{X}^{\boldsymbol{c}}$ can be expressed in the singular value decomposition form:

$$\begin{aligned} \boldsymbol{X}^{\boldsymbol{c}}_{n\times m} &= \boldsymbol{U}_{n\times n}\ \boldsymbol{S}_{n\times m}\ \boldsymbol{V}^{T}_{m\times m} \\ &\approx\ \hat{\boldsymbol{X}}^{\boldsymbol{c}}{}_{n\times p} = \hat{\boldsymbol{U}}_{n\times p}\ \hat{\boldsymbol{S}}_{p\times p}\ \hat{\boldsymbol{V}}^{T}_{p\times m}. \end{aligned} \tag{6}$$

In Eq. (6), the reduced-dimension matrix $\hat{\boldsymbol{X}}^{\boldsymbol{c}}$ optimally [18,24] approximates the original matrix $\boldsymbol{X}^{\boldsymbol{c}}$ by keeping only the first $p \leqslant m$ singular values s_j of $\boldsymbol{S}$. The total number p of components kept can be chosen in several ways (see for example the partial sum criterion [16]), and in this paper we treat it as a hyperparameter that is selected by analyzing the performance of the overall model across a validation set.

Once the truncation is done, the retained p principal components are obtained by the columns of $\hat{\boldsymbol{U}}$. We label these components as $\phi_j(x)$ (with $j = 1, \ldots, p$). As illustrated in in Fig. 1, the forecast of any of the original models can be approximated by a linear combination of this smaller set of p orthogonal components $\phi_j(\boldsymbol{x})$ identified by the SVD algorithm, plus the original average forecast:

$$f_k(\boldsymbol{x}) \approx \phi_0(\boldsymbol{x}) + \sum_{j=1}^{p} v_j^{(k)} \phi_j(\boldsymbol{x}), \quad \text{for } k = 1, \ldots, m. \tag{7}$$

The coefficients $v_j^{(k)}$ can be obtained by multiplying the kth column of $\hat{\boldsymbol{V}}^T$ by the respective singular values of $\hat{\boldsymbol{S}}$. This reduction presents several advantages for BMM and BMC that we will discuss in the following.

A. Global model mixing and model combination with principal components

Conveniently, we can construct a combined model $f^\dagger$ by mixing (globally) the identified principal components instead of the original forecasts:

$$f^\dagger(\boldsymbol{x}; \boldsymbol{b}) = \phi_0(\boldsymbol{x}) + \sum_{j=1}^{p} b_j \phi_j(\boldsymbol{x}), \tag{8}$$

where b_j corresponds to the global weight of the jth principal component ϕ_j. Note that since each principal component ϕ_j is itself a linear combination of the original forecasts f_k, we can express Eq. (8) in a BMM-like form,

$$f^\dagger(\boldsymbol{x}; \boldsymbol{b}) = \sum_{k=1}^{m} \omega_k(\boldsymbol{b}) f_k(\boldsymbol{x}), \tag{9}$$

where the m weights ω_k depend on the p latent variables $\boldsymbol{b} = (b_1, \ldots, b_p)$ as degrees of freedom:

$$\boldsymbol{\omega}(\boldsymbol{b}) = \frac{1}{m}\mathbf{1} + \left(\hat{\boldsymbol{V}}_{m\times p}\hat{\boldsymbol{S}}^{-1}_{p\times p}\right)\boldsymbol{b}, \tag{10}$$

where $\mathbf{1}$ is the all-ones vector of dimension m with all elements equal to 1.

By construction, the sum of the weights ω_k in Eq. (9) adds up to 1, i.e., it satisfies the second part of the simplex constraint (2). Indeed, since ϕ_0 is the average over all models, it contributes with $m \times \frac{1}{m} = 1$ to the total sum of the weights, while every principal component ϕ_j is itself a linear combination of the columns of the matrix $\boldsymbol{X}^{\boldsymbol{c}}$, each of which adds net zero total sum of model weights (see Fig. 1). We note, however, that while the weights ω_k fulfill the second simplex constraint, they do not have to be positive.

Given the available data, the weights $\boldsymbol{b}$ can be jointly estimated [25], with an assumed combined model error scale σ, within a Bayesian framework:

$$p(\boldsymbol{b}, \sigma\,|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{b}, \sigma)p(\boldsymbol{b}, \sigma). \tag{11}$$

Here, $p(\boldsymbol{y}|\boldsymbol{b}, \sigma)$ is the data likelihood function with the standard Gaussian-noise assumption as in Eq. (1), namely,

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{b}, \sigma) &\propto \sigma^{-n} \exp\left(-\frac{1}{2}\chi^2\right), \\ \chi^2 &= \sum_{i=1}^{n} \frac{[f^\dagger(x_i; \boldsymbol{b}) - y(x_i)]^2}{\sigma^2}, \end{aligned} \tag{12}$$

and $p(\boldsymbol{b}, \sigma)$ is the joint prior distribution of the mixing weights and the common error scale σ.

The assumption that the deviations between our model predictions $f^\dagger(x_i; \boldsymbol{b})$ and the observations $y(x_i)$ follow independent Gaussian distributions with the same noise scale ($\sigma_i \equiv \sigma$) is a common choice for calibrating nuclear models with extremely precise data such as nuclear masses [26–28]. This assumption can be easily modified for different applications of the framework if needed.

A reasonable choice for $p(\boldsymbol{b}, \sigma)$ is multivariate Gaussian prior distribution for $\boldsymbol{b}$, informed by the empirical distribution of the original weights $v_j^{(k)}$ when reproducing the models in (7), and a Gamma prior distribution for the precision $1/\sigma^2$ parametrized by a shape parameter ν_0 and a scale parameter σ_0 (see Ref. [29]). We create a weakly informed prior by selecting $\nu_0 = 10$ and $\sigma_0 = 2$ MeV for the first case study, and $\nu_0 = 10$ and $\sigma_0 = 0.5$ MeV for the realistic case. These choices do not appreciably impact the obtained posterior distributions.

Notably, our procedure allows for an efficient approximation of the posterior distribution (11) using the standard Gibbs sampler, because the multivariate Gaussian and Gamma priors $p(\boldsymbol{b}, \sigma)$ together with the likelihood (12) form the classical normal-inverse-gamma semiconjugate model (see chapter 9 of Ref. [29]).

This prior choice effectively resembles BMC in the context of the combined model (9), because the weights $\omega_k(\boldsymbol{b})$ are unrestricted in the sense that they can be positive and negative. If one wishes to impose a further constraint on the model weights, such as the simplex constraint (2), the original multivariate Gaussian prior distribution for $\boldsymbol{b}$ can be easily modified so that it is zero whenever the constraint is not satisfied.

The combined model (8), referred to as BMC + PCA in the following, parametrizes an approximated manifold of the currently developed theoretical models through the mixing weights $\boldsymbol{b}$, as illustrated by the gray affine subspace in Fig. 1. In the Bayesian context, this combined model, equipped with a posterior probability distribution for the weights (11), has the appealing statistical interpretation of representing an overarching distribution of plausible models that can explain the observed data. Within this framework, each original forecast f_k could be interpreted as a random—not necessarily independent—draw from this distribution, and through Eq. (11) we aim to create a more informed quantified prediction for new observables. Furthermore, performing the mixing on the principal components instead of the original forecasts has the advantage of both filtering out parametric directions that could be associated with noise instead of important features, and preventing the combined-model parameters $\boldsymbol{b}$ from becoming ill conditioned in the presence of similar or redundant forecasts f_k. We demonstrate these features in the next section.

IV. RESULTS

A. Case study I: Redundant and similar models

To test the proposed global BMC framework (8), we first consider nuclear binding energy forecasts generated by several variants of the analytic liquid drop model [30,31]. Within this seven-parameter model, the binding energy of a nucleus, $\mathcal{E}(N, Z)$, is given by [32]

$$\mathcal{E}(N, Z; \boldsymbol{a}) = a_{\rm vol}A + a_{\rm surf}A^{2/3} + a_{\rm curv}A^{1/3} + a_{\rm sym}I^2A + a_{\rm ssym}I^2A^{2/3} + a^{(2)}_{\rm sym}I^4A + a_{\rm Coul}\frac{Z^2}{A^{1/3}}, \quad (13)$$

where $A = N + Z$ is the mass number and $I = (N - Z)/A$ is the isospin excess. The parameters $\boldsymbol{a} = [a_{\rm vol}, a_{\rm surf}, a_{\rm curv}, a_{\rm sym}, a_{\rm ssym}, a_{\rm sym}, a_{\rm Coul}]$ have a well-defined physical meaning; they represent volume, surface, curvature, symmetry, surface-symmetry, second-order-symmetry, and Coulomb terms respectively.

To test various common modeling scenarios, we create four model classes of the form

$$y^{(t)}_{\rm th}(N, Z) = \mathcal{E}(N, Z; \boldsymbol{a}^{(t)}), \quad \text{for } t \in \{\text{P, G, I, B}\}, \quad (14)$$

where the model-class index P, G, I, and B stands for "perfect," "good," "intermediate," and "bad" models, respectively. These labels reflect how close these models are to the reference model that generates the synthetic data. Each model class has parameters centered around parameters $\boldsymbol{a}^{(t)}$ defined in Table I. For some scenarios, to obtain nondegenerate forecasts, the parameters are shifted by a small random amount $\delta\boldsymbol{a}^{(t)}$, which is a Gaussian with a width of 2‰ of $\boldsymbol{a}^{(t)}$. In some cases, we also add a Gaussian noise term with the width $\sigma_{\rm noise} = 1$ MeV. These two sources of error, the shift in the parameters and the overall Gaussian noise, simulate a situation in which models within the same class yield predictions that deviate both in a coherent ($\delta\boldsymbol{a}^{(t)}$) and uncorrelated ($\sigma_{\rm noise}$) way. [The spread of model predictions due to these sources of noise is shown in Fig. 2(a).] The reference forecast $y_{\rm true}(N, Z)$ consists of $n = 629$ binding energies of even-even nuclei with $8 \leqslant Z \leqslant 102$ computed with the SkO parametrization with the noise term with $\sigma_{\rm noise}$ added.

TABLE I. Four model classes used in this paper and their respective parameters as defined in Table I of Ref. [32]. The models belonging to the class "Bad" are based on the NL1 parametrization with the terms $\{a_{\rm sym}, a_{\rm ssym}, a^{(2)}_{\rm sym}\}$ set to zero; we label this parametrization as NL1*. Three scenarios (S1–S3) considered are shown in columns 3–5 that list the number $N_{\rm rep,k}$ of repeated (redundant) models belonging to class k; the number of principal components p kept; and the individual model weights ω_k (9) obtained by maximizing the likelihood (12) in the case without noise (both $\delta\boldsymbol{a}^{(t)}$ and σ_0 are set to zero). Since no noise was added, every repeated model within the same class has identical weight. For each scenario, $\sum \omega_k N_{\rm rep,k} = 1$. The weights from S1 are correctly redistributed among the repetitions in S2, i.e., $(\omega_k N_{\rm rep,k})_{S1} = (\omega_k N_{\rm rep,k})_{S2}$ for each k. The perfect model is selected in S3.

		S1		S2		S3	
		$p = 2$		$p = 2$		$p = 3$	
Model class	Parameter center $\boldsymbol{a}^{(t)}$	$N_{\rm rep}$	ω_k	$N_{\rm rep}$	ω_k	$N_{\rm rep}$	ω_k
Perfect	SkO	0		0		1	1.000
Good	SLy4	1	0.710	3	0.237	3	0.000
Intermediate	NL1	1	0.309	5	0.062	5	0.000
Bad	NL1*	1	−0.019	10	−0.002	10	0.000

The perfect model is selected in S3.

We study three scenarios S1–S3, which are described in Table I and illustrated in Fig. 2 for scenario S3. In these scenarios, we use different numbers of models in each class with the objective of demonstrating that the proposed algorithm works as intended. The singular values of the SVD (6) can give an initial estimate of the expected number of effective components, as shown Fig. 2(b). The projections of each model on the identified principal components ϕ_k can visually help identify model classes, as is also shown in the inset of Fig. 2(a).

For the remaining part of this section, we use the BMC + PCA model (8). The synthetic data are separated into three groups: training dataset ($\mathcal{X}^{\rm tr}_0$) with 300 data points, validation dataset ($\mathcal{X}^{\rm va}_0$) with 71 data points, and testing dataset ($\mathcal{X}^{\rm te}_0$) with 258 data points, as is often done in machine learning applications [18], including studies focusing on nuclear mass models [10,33–35]. The specific way the three sets for our paper were chosen [see Fig. 3(a)] reflects that one of the main objectives of our model forecast combination lies in model

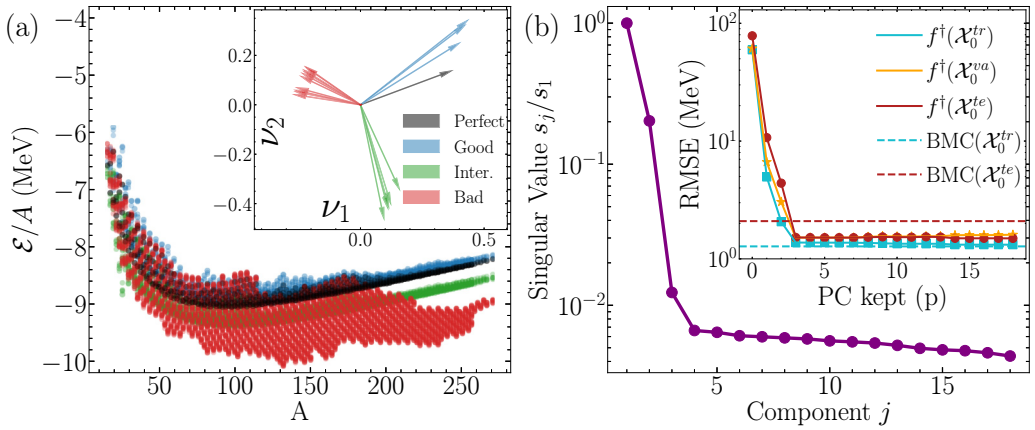

FIG. 2. Illustration of S3 of Table I. (a) Forecasts of the binding energy per nucleon produced by 19 different models: one perfect model (black), three good models (blue), five intermediate models (green), and ten bad models (red). The spread of the results comes from the noise terms added. The inset shows the projection $\nu_j^{(k)}$ defined in Eq. (7) for each of the 19 models onto the first two principal components, clearly identifying the existence of three model classes, with the perfect model and three good models being nearly aligned. (b) Decay of the singular values s_j. The inset shows the evolution of the RMSE (15) for the training (cyan blue squares), validation (yellow stars), and testing (dark red circles) datasets as the number of principal components kept in the expansion (8) is increased (zero corresponds to ϕ_0). The BMC + PCA results are marked by solid lines. The dashed lines show the RMSE obtained when combining all 19 models without projecting on principal components (pure BMC), which shows signs of overfitting: lower RMSE, training dataset; higher RMSE, testing set.

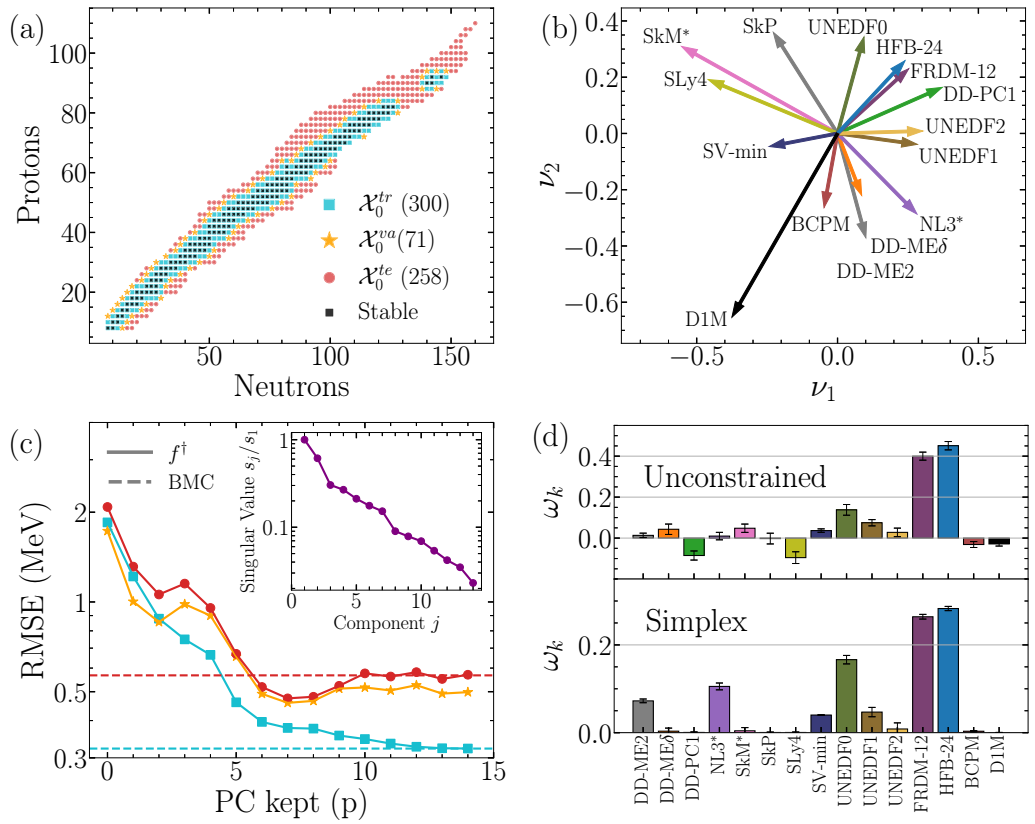

FIG. 3. Case study II results. (a) Training (squares), validation (stars), and testing (circles) datasets of binding energies of 629 even-even nuclei used in this paper. The stable isotopes are marked by small black squares. (b) Projections ν_1 and ν_2 of 15 realistic models of the nuclear binding energy into the first two principal components. This representation allows us to visualize intermodel relationships. (c) Similar to Fig. 2(b) but for the realistic mass models. The colors and symbols follow the same convention as in panel (a), with solid lines representing the BMC + PCA model of Eq. (8) and dashed lines representing the BMC of Eq. (3) with $f_0 = 0$. (d) Distribution of the weights ω_k for the individual models in the expansion (9) in the unconstrained (top) and simplex-constrained (bottom) settings [see Eq. (2)]. The vertical error bars represent a 95% region obtained from the sampled posterior.

TABLE II. Model performance quantified by the RMSE (in MeV). Columns 2, 3, 4, and 5 show the RMSE across the training, validation, testing, and full dataset, respectively. The two bottom rows show the performance of the combined BMC+PCA model $f^\dagger$, in the unconstrained and simplex variant. These RMSE were calculated by averaging the model predictions from the visited posteriors in Eq. (11), and then using Eq. (15).

Model	RMSE $\mathcal{X}_0^{\text{tr}}$	RMSE $\mathcal{X}_0^{\text{va}}$	RMSE $\mathcal{X}_0^{\text{te}}$	RMSE $\mathcal{X}_0$
DD-ME2 [39]	2.47	2.48	2.25	2.38
DD-MEδ [40]	2.46	2.19	2.18	2.32
DD-PC1 [41]	1.94	1.77	2.19	2.03
NL3* [42]	2.18	2.15	3.59	2.84
SkM* [43]	4.91	6.34	9.75	7.42
SkP [44]	3.06	3.50	4.41	3.72
SLy4 [45]	4.53	4.84	6.27	5.34
SV-min [46]	2.97	2.99	3.98	3.42
UNEDF0 [47]	1.47	2.02	1.51	1.56
UNEDF1 [48]	1.82	1.83	2.06	1.92
UNEDF2 [49]	1.83	1.66	2.04	1.90
FRDM-12 [38]	0.62	0.62	0.65	0.63
HFB-24 [37]	0.52	0.51	0.52	0.52
BCPM [50]	2.57	2.34	2.44	2.49
D1M [51]	5.02	4.91	5.63	5.27
$f^\dagger(p=7)$	0.38	0.46	0.47	0.43
$f^\dagger(\text{simplex}, p=7)$	0.74	0.69	0.71	0.72

extrapolation into the region in which experimental information does not exist. By dividing the sets in this way, we provide a more stringent test of how the combined model's performance will evolve as we go further away from the region where data currently exist.

To select the number of principal components kept p, we study how it impacts the root mean squared error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[f^\dagger(x_i;\boldsymbol{b}_0) - y_{\text{true}}(x_i)]^2}. \tag{15}$$

The parameters $\boldsymbol{b}_0$ are chosen to maximize the likelihood function (12) on the training dataset $\mathcal{X}_0^{\text{tr}}$. The RMSE is then computed across all the sets and the number of components p is chosen based on the performance on the validation set. The inset in Fig. 2(b) shows RMSE(p) for S3. The RMSE computed for the three datasets saturates after $p = 3$, as expected. Indeed, there are only four distinct model classes in S3. The training RMSE is always lower, since the parameters $\boldsymbol{b}_0$ are fitted to it, and the validation RMSE correctly serves as a proxy for the expected RMSE in the extrapolated testing dataset.

B. Case study II: Realistic nuclear mass models

We now turn to a set of realistic models of nuclear binding energy. For this paper, we have chosen 15 realistic computational models that represent a few classes of theoretical frameworks that see broad use. The models are specified in Table II, with their respective RMSE for each of the three datasets. As the model orthogonalization and mixing strategies only require precomputed data across a range of nuclei, we pull forecasts directly from published datasets. The specific mass models chosen and their parameters are those from the MassExplorer database [36] and HFB-24 [37] and FRDM-12 [38] mass tables.

The domains of the experimental datasets used are shown in Fig. 3(a). As in the case study I, we divide the 629 data points into training, validation, and testing sets. We perform the SVD on the forecasts produced by the set of 15 models restricted to the training set and show their projections v_1 and v_2 on the first two principal components in Fig. 3(b). The nearly exponential decay of the singular values (6) is shown in the inset of Fig. 3(c). Note that the SVD is performed after the centering procedure, thus the principal components in Fig. 3(b) should be interpreted as the variability about the average in the space of predictions. While it is tempting to draw conclusions on model similarity from these projections, it can only be said that the forecasts themselves are similar if two vectors are close. The models HFB-24 and FRDM-12, for instance, share little in common when it comes to the underlying form and theoretical assumptions, though their predictive capability is very similar due to their parametric expressivity. The UNEDF series of interactions (UNEDF0, UNEDF1, UNEDF2) is another interesting case study in that they all have a very similar functional form, but are based on different calibrations. This difference directly manifests in the projections where UNEDF0 is nearly orthogonal to the other two models in the same family despite their close functional relationship. Additional information could likely be gleaned from a more targeted study of certain nuclei, but the global dataset of nuclear binding energies does not immediately reveal model specific physical insights.

We now proceed to create the combined BMC + PCA model $f^\dagger$ as specified in Eq. (8). To this end, we analyze the RMSE performance by maximizing the likelihood (12) of the combined model as we vary the number of principal components kept. Based on the result shown in Fig. 3(c), we retain $p = 7$ principal components for the combined model for the rest of this analysis. Indeed, as the number of components is increased beyond 7, the validation and test errors grow, suggesting overfitting. The dashed lines show the RMSE obtained when combining all 15 models without projecting on principal components, which also shows signs of overfitting: a lower RMSE for the training dataset and a higher RMSE for the test dataset. It is to be noted that even though the singular values shown in the inset of Fig. 3(c) decay nearly exponentially, they do not experience a rapid drop as in Fig. 2(b). In this case, the RMSE(p) behavior for the validation set provides a good metric to determine the number of principal components that are needed to optimally model the experimental values outside of the training set.

Figure 3(d) shows the weights ω_k of each model for both the unconstrained and simplex-constrained case. The vertical error bars show the 95% credible intervals obtained from sampling the weights using Eq. (11). In the unconstrained case, several models dominate the combination, though with significant cancellation of the model amplitudes. In the case of the simplex-constrained combination, we see a similar behavior for the models with the largest weights, FRDM-12 and HFB-24, that make up roughly 50% of the combined model. The starkest difference is in the effective nullification of many

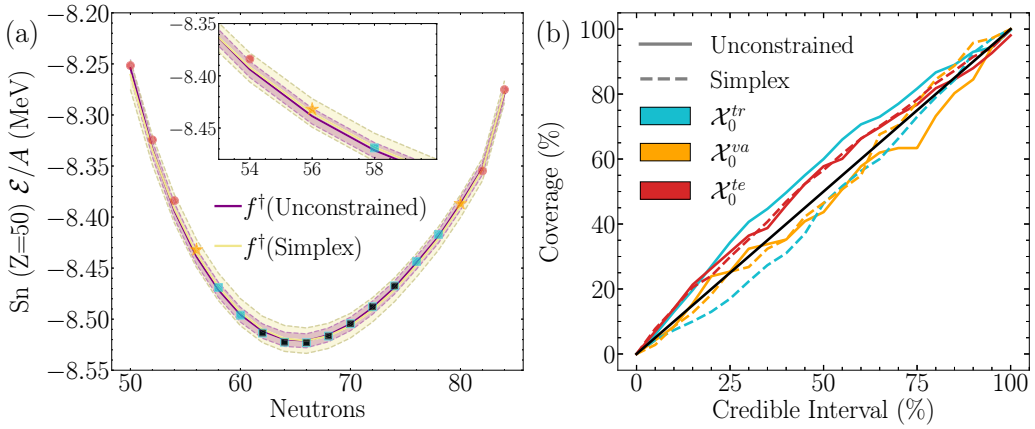

FIG. 4. (a) Predictive posterior distribution for the binding energy per nucleon of the Sn isotopes. The mean prediction and 95% credible interval of the unconstrained combined model is shown in purple, while the simplex-constrained (simplex) combined model is shown in khaki. The inset shows the detail of the plot for $N = 54$, 56, and 58. (b) ECP for unconstrained and simplex-constrained variants for training (blue), validation (yellow), and testing (red) datasets. The diagonal black line shows a reference of what a perfect statistical coverage would entail, with points above it being conservative, and those below being overconfident.

of the other models across the model space, leaving only a few active (in our case seven) in the combination. The specifics of weight distributions naturally depend on the number of principal components that are retained, though the general behavior seems to be consistent across different active subspaces. From the distributions of the weights (11), we compute the predictions of the combined model $f^\dagger$ with quantified uncertainties across the entire dataset. Figure 4(a) shows the predictions and 95% credible intervals for the Sn ($Z = 50$) isotopic chain for the unconstrained and simplex-constrained variants. One feature to note is that the simplex-constrained model has a systematically broader credible interval in its forecasts both inside and outside the training region, yet both models seem to cover the experimental data well within their credible bands. Since the constrained model can only reproduce a subset of possible combinations, we expect its performance in terms of RMSE to be less expressive in both interpolation and extrapolation, at the gain of less sensitivity to overfitting. Indeed, as can be seen in the last two rows of Table II, the RMSE scores for the constrained model are significantly worse than for the unconstrained model—as well as the two best performing models FRDM-12 and HFB-24. Yet, they remain steady in the transition from training to validation and testing, while for the unconstrained approach the testing RMSE increases by about 20% in comparison to the training RMSE. That the unconstrained model $f^\dagger$ outperforms each individual model in terms of RMSE is not surprising—an unconstrained combination of PCs fitted to a given dataset will outperform what each individual model can do alone, as has been shown for nuclear mass models in Ref. [22]. Embedding the PCA-based model combination within a full Bayesian framework and the classification of data into training, validation, and test sets then allows for more reliable forecasts with quantified uncertainties when extrapolating beyond experimentally known masses.

Indeed, by analyzing the empirical coverage of our calibrated model, we can quantitatively assess signs of overfitting. Figure 4(b) shows the empirical coverage probability (ECP) [52,53] for both the unconstrained and simplex-constrained models across the three datasets considered. The fact that the empirical curves all lie close to the diagonal reference line gives us confidence that the combined predictions are neither being overconfident (too small credible intervals) or overconservative (too big credible intervals). This is particularly reassuring for the test set, in which considerable extrapolations have been made [see Fig. 3(a)]. When considering just one data type (here nuclear binding energies), the risk for overfitting is low, yet if one wishes to consider model performance on quantities not in the original dataset, the risk for overfitting can be substantially increased.

V. CONCLUSIONS AND OUTLOOK

In this paper, we propose, implement, and benchmark a Bayesian model combination framework accompanied by model orthogonalization using principal component analysis. We discuss the features of the proposed BMC + PCA method by applying it to global models of nuclear binding energy. Following the tests based on the analytic liquid drop model, we carry out realistic BMC + PCA calculations of nuclear binding energies using 15 global computational nuclear models. We demonstrate that the BMC + PCA framework performs excellently in terms of prediction accuracy and uncertainty quantification. While we have focused on the nuclear physics use case in this paper, the method itself is completely general and can be applied broadly where model forecasts are utilized. It is also easy to implement and results in an interpretable combination, simplifying the application to problems that span disciplines. Furthermore, the computational scheme is robust against model repetitions and it does not favor one model class when multiple copies are added. The BMC + PCA technique can be reduced to BMM + PCA by imposing the simplex condition. In this case, only the several best performing models remain in the combination, and we recover an interpretation of the model combination that can be compared to other traditional BMA and BMM approaches where the *positive* model weights are determined by the data. For this simplex-constrained version both the root mean squared error (Table II) as well as its uncertainty bands

(Fig. 4) are bigger than the unconstrained version, yet the simplex-constrained approach shows signs of less overfitting in terms of extrapolation, with a performance that remains stable across the three sets [Fig. 3(a)].

In addition to producing optimal multimodel forecasts with robust uncertainties, BMC + PCA is also capable of identifying model collinearities and redundancies, a functionality that stands to benefit other applications that aim to combine the wisdom of multiple nuclear models [54–56]. Furthermore, the framework is also able to perform model selection if the exact model happens to be present in the set of models. The computational efficiency of the combined model also positions it well for wide distribution on web-based platforms, such as the Bayesian Mass Explorer project [57]. While live evaluation of most of the individual models is impossible due to their inherent numerical complexity, the combined model can be evaluated, with uncertainties, on the fly. The model combination procedure itself is also efficient, meaning interested users can provide their own datasets and update the resultant BMC + PCA model.

Future developments will include a local extension of BMC + PCA by assuming domain-dependent weights, i.e., $b_j \to b_j(\boldsymbol{x})$; see Ref. [10]. This enhancement will help construct model combinations that emphasize the local performance of models in certain regions, a typical scenario in physics modeling across multiple scales. To aid adoption of the method, it is currently planned to implement the procedure into the open source model mixing software TAWERET [58]. We will also consider the extension to heterogeneous forecasts by considering data of several classes (e.g., other nuclear data like binding energies and charge radii).

ACKNOWLEDGMENTS

We thank L. Neufcourt for valuable suggestions during the early stages of the project. We also wish to thank the BMM working group of the BAND collaboration and Edgard Bonilla for useful discussions. This material is based upon work supported by the US Department of Energy, Office of Science, Office of Nuclear Physics under Grants No. DE-SC0023688 and No. DOE-DE-SC0013365, and by the National Science Foundation under Grant No. 2004601 (CSSI program, BAND collaboration).

[1] M. Höge, A. Guthke, and W. Nowak, The hydrologist's guide to Bayesian model selection, averaging and combination, J. Hydrol. **572**, 96 (2019).

[2] Y. Yao, A. Vehtari, D. Simpson, and A. Gelman, Using stacking to average Bayesian predictive distributions (with discussion), Bayesian Anal. **13**, 917 (2018).

[3] D. R. Phillips, R. Furnstahl, U. Heinz, T. Maiti, W. Nazarewicz, F. Nunes, M. Plumlee, S. Pratt, M. Pratola, F. Viens, and S. M. Wild, Get on the BAND wagon: A Bayesian framework for quantifying model uncertainties in nuclear dynamics, J. Phys. G **48**, 072001 (2021).

[4] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky, Bayesian model averaging: A tutorial, Statist. Sci. **14**, 382 (1999).

[5] L. Wasserman, Bayesian model selection and model averaging, J. Math. Psychol. **44**, 92 (2000).

[6] T. Fragoso, W. Bertoli, and F. Louzada, Bayesian model averaging: A systematic review and conceptual classification, Int. Stat. Rev. **86**, 1 (2018).

[7] T. Le and B. Clarke, A Bayes interpretation of stacking for $\mathcal{M}$-complete and $\mathcal{M}$-open settings, Bayesian Anal. **12**, 807 (2017).

[8] Y. Yao, G. Pirš, A. Vehtari, and A. Gelman, Bayesian hierarchical stacking: Some models are (somewhere) useful, Bayesian Anal. **17**, 1043 (2022).

[9] C. Fernández and P. J. Green, Modelling spatially correlated data via mixtures: A Bayesian approach, J. R. Stat. Soc. B **64**, 805 (2002).

[10] V. Kejzlar, L. Neufcourt, and W. Nazarewicz, Local Bayesian Dirichlet mixing of imperfect models, Sci. Rep. **13**, 19600 (2023).

[11] J. M. Bates and C. W. J. Granger, The combination of forecasts, J. Oper. Res. Soc. **20**, 451 (1969).

[12] T. P. Minka, Bayesian model averaging is not model combination, https://api.semanticscholar.org/CorpusID:116598428 (2002).

[13] C. W. J. Granger and R. Ramanathan, Improved methods of combining forecasts, J. Forecast. **3**, 197 (1984).

[14] K. Monteith, J. L. Carroll, K. Seppi, and T. Martinez, Turning Bayesian model averaging into Bayesian model combination, in *The 2011 International Joint Conference on Neural Networks* (IEEE, San Jose, CA, 2011), pp. 2657–2663.

[15] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference A Practical Information-Theoretic Approach* (Springer, New York, 2002).

[16] I. Jolliffe, *Principal Component Analysis*, Springer Series in Statistics (Springer, New York, 2002).

[17] A. Blum, J. Hopcroft, and R. Kannan, *Foundations of Data Science* (Cambridge University Press, New York, 2020).

[18] S. L. Brunton and J. N. Kutz, *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control* (Cambridge University Press, New York, 2019).

[19] M. Clyde, H. Desimone, and G. Parmigiani, Prediction via orthogonalized model mixing, J. Am. Stat. Assoc. **91**, 1197 (1996).

[20] P. Poncela, J. Rodríguez, R. Sánchez-Mangas, and E. Senra, Forecast combination through dimension reduction techniques, Int. J. Forecast. **27**, 224 (2011).

[21] J. H. Stock and M. W. Watson, Generalized shrinkage methods for forecasting using many predictors, J. Bus. Econ. Stat. **30**, 481 (2012).

[22] X.-H. Wu and P. Zhao, Principal components of nuclear mass models, Sci. China Phys. Mech. Astron. **67**, 272011 (2024).

[23] L. Neufcourt, Y. Cao, W. Nazarewicz, and F. Viens, Bayesian approach to model-based extrapolation of nuclear observables, Phys. Rev. C **98**, 034318 (2018).

[24] C. Eckart and G. Young, The approximation of one matrix by another of lower rank, Psychometrika **1**, 211 (1936).

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

[26] V. Kejzlar, L. Neufcourt, W. Nazarewicz, and P.-G. Reinhard, Statistical aspects of nuclear mass models, J. Phys. G **47**, 094001 (2020).

[27] J. D. McDonnell, N. Schunck, D. Higdon, J. Sarich, S. M. Wild, and W. Nazarewicz, Uncertainty quantification for nuclear density functional theory and information content of new measurements, Phys. Rev. Lett. **114**, 122501 (2015).

[28] P. Giuliani, K. Godbey, E. Bonilla, F. Viens, and J. Piekarewicz, Bayes goes fast: Uncertainty quantification for a covariant energy density functional emulated by the reduced basis method, Front. Phys. **10**, 1054524 (2023).

[29] P. D. Hoff, *A First Course in Bayesian Statistical Methods*, Springer Texts in Statistics Vol. 580 (Springer, New York, 2009).

[30] C. F. v. Weizsäcker, Zur theorie der kernmassen, Z. Phys. **96**, 431 (1935).

[31] P. Ring and P. Schuck, *The Nuclear Many-Body Problem* (Springer-Verlag, Berlin, 1980).

[32] P.-G. Reinhard, M. Bender, W. Nazarewicz, and T. Vertse, From finite nuclei to the nuclear liquid drop: Leptodermous expansion based on self-consistent mean-field theory, Phys. Rev. C **73**, 014309 (2006).

[33] L. Neufcourt, Y. Cao, W. Nazarewicz, E. Olsen, and F. Viens, Neutron drip line in the Ca region from Bayesian model averaging, Phys. Rev. Lett. **122**, 062502 (2019).

[34] L. Neufcourt, Y. Cao, S. A. Giuliani, W. Nazarewicz, E. Olsen, and O. B. Tarasov, Quantified limits of the nuclear landscape, Phys. Rev. C **101**, 044307 (2020).

[35] Y. Saito, I. Dillmann, R. Krücken, M. R. Mumpower, and R. Surman, Uncertainty quantification of mass models using ensemble Bayesian model averaging, Phys. Rev. C **109**, 054301 (2024).

[36] Mass Explorer, http://massexplorer.frib.msu.edu (2020).

[37] S. Goriely, N. Chamel, and J. M. Pearson, HFB-24 mass formula, http://www.astro.ulb.ac.be/bruslib/nucdata/hfb24-dat (2020).

[38] P. Möller, A. Sierk, T. Ichikawa, and H. Sagawa, Nuclear ground-state masses and deformations: FRDM(2012), At. Data Nucl. Data Tables **109–110**, 1 (2016).

[39] G. A. Lalazissis, T. Nikšić, D. Vretenar, and P. Ring, New relativistic mean-field interaction with density-dependent meson-nucleon couplings, Phys. Rev. C **71**, 024312 (2005).

[40] X. Roca-Maza, X. Vinas, M. Centelles, P. Ring, and P. Schuck, Relativistic mean-field interaction with density-dependent meson-nucleon vertices based on microscopical calculations, Phys. Rev. C **84**, 054309 (2011).

[41] T. Nikšić, D. Vretenar, and P. Ring, Relativistic nuclear energy density functionals: Adjusting parameters to binding energies, Phys. Rev. C **78**, 034318 (2008).

[42] G. Lalazissis, S. Karatzikos, R. Fossion, D. P. Arteaga, A. Afanasjev, and P. Ring, The effective force NL3 revisited, Phys. Lett. B **671**, 36 (2009).

[43] J. Bartel, P. Quentin, M. Brack, C. Guet, and H.-B. Håkansson, Towards a better parametrisation of Skyrme-like effective forces: A critical study of the SkM force, Nucl. Phys. A **386**, 79 (1982).

[44] J. Dobaczewski, H. Flocard, and J. Treiner, Hartree-Fock-Bogolyubov description of nuclei near the neutron-drip line, Nucl. Phys. A **422**, 103 (1984).

[45] E. Chabanat, P. Bonche, P. Haensel, J. Meyer, and R. Schaeffer, New Skyrme effective forces for supernovae and neutron rich nuclei, Phys. Scr. **T56**, 231 (1995).

[46] P. Klüpfel, P.-G. Reinhard, T. J. Bürvenich, and J. A. Maruhn, Variations on a theme by Skyrme: A systematic study of adjustments of model parameters, Phys. Rev. C **79**, 034310 (2009).

[47] M. Kortelainen, T. Lesinski, J. Moré, W. Nazarewicz, J. Sarich, N. Schunck, M. V. Stoitsov, and S. Wild, Nuclear energy density optimization, Phys. Rev. C **82**, 024313 (2010).

[48] M. Kortelainen, J. McDonnell, W. Nazarewicz, P.-G. Reinhard, J. Sarich, N. Schunck, M. V. Stoitsov, and S. M. Wild, Nuclear energy density optimization: Large deformations, Phys. Rev. C **85**, 024304 (2012).

[49] M. Kortelainen, J. McDonnell, W. Nazarewicz, E. Olsen, P.-G. Reinhard, J. Sarich, N. Schunck, S. M. Wild, D. Davesne, J. Erler, and A. Pastore, Nuclear energy density optimization: Shell structure, Phys. Rev. C **89**, 054314 (2014).

[50] M. Baldo, L. M. Robledo, P. Schuck, and X. Viñas, New Kohn-Sham density functional based on microscopic nuclear and neutron matter equations of state, Phys. Rev. C **87**, 064305 (2013).

[51] S. Goriely, S. Hilaire, M. Girod, and S. Péru, First Gogny-Hartree-Fock-Bogoliubov nuclear mass model, Phys. Rev. Lett. **102**, 242501 (2009).

[52] T. Gneiting, F. Balabdaoui, and A. E. Raftery, Probabilistic forecasts, calibration and sharpness, J. Roy. Stat. Soc. Ser. B: Stat. Methodol. **69**, 243 (2007).

[53] T. Gneiting and A. E. Raftery, Strictly proper scoring rules, prediction, and estimation, J. Am. Stat. Assoc. **102**, 359 (2007).

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

[55] M. Qiu, B.-J. Cai, L.-W. Chen, C.-X. Yuan, and Z. Zhang, Bayesian model averaging for nuclear symmetry energy from effective proton-neutron chemical potential difference of neutron-rich nuclei, Phys. Lett. B **849**, 138435 (2024).

[56] V. Cirigliano *et al.*, Towards precise and accurate calculations of neutrinoless double-beta decay, J. Phys. G **49**, 120502 (2022).

[57] K. Godbey, L. Buskirk, and P. Giuliani, BMEX The Bayesian Mass Explorer; https://bmex.dev/.

[58] K. Ingles, D. Liyanage, A. C. Semposki, and J. C. Yannotty, Taweret A Python package for Bayesian model mixing, J. Open Source Softw. **9**, 6175 (2024).