medRxiv preprint doi: https://doi.org/10.1101/2025.01.13.25320493; this version posted January 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Evaluation of stochastic trajectory-based epidemic models using the energy score

Clara Bay[1*], Kunpeng Mu[1], Guillaume St-Onge[2,1], Matteo Chinazzi[2,1], Jessica T. Davis[1], Alessandro Vespignani[1,2],

1 Laboratory for the Modeling of Biological and Socio-technical Systems, Northeastern University, Network Science Institute, Boston, MA, USA
2 The Roux Institute, Northeastern University, Portland, ME, USA

* bay.c@northeastern.edu

Abstract

Scoring rules are critical for evaluating the predictive performance of epidemic models by quantifying how well their projections and forecasts align with observed data. In this study, we introduce the energy score as a robust performance metric for stochastic trajectory-based epidemic models. As a multivariate extension of the continuous ranked probability score (CRPS), the energy score provides a single, unified measure for time-series predictions. It evaluates both calibration and sharpness by considering the distances between individual trajectories and observed data, as well as the inter-trajectory variability. We provide an overview of how the energy score can be applied to assess both scenario projections and forecasts in this format, with a particular focus on a detailed analysis of the Scenario Modeling Hub results for the 2023-2024 influenza season. By comparing the energy score to the widely used weighted interval score (WIS), we demonstrate its utility as a powerful tool for evaluating epidemic models, especially in scenarios requiring integration of predictions across multiple target outcomes into a single, interpretable metric.

Author summary

Epidemic model predictions are often evaluated using scoring rules, such as the weighted interval score (WIS), which require outputs in interval or quantile

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2025.01.13.25320493; this version posted January 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

formats. However, epidemic models often produce outputs as collections of stochastic trajectories, which are then summarized into quantiles for evaluation. In this study, we introduce the energy score as a scoring metric specifically designed for evaluating stochastic trajectories without requiring conversion to other formats. The energy score provides a rigorous assessment by accounting for both the variability among trajectories and their alignment with observed data. Using publicly available data, we demonstrate that the energy score is a reliable and effective metric for evaluating epidemic model predictions in their native stochastic trajectory format.

Introduction

Epidemic model predictions are typically probabilistic, offering a range of potential outcomes rather than a single deterministic forecast. Epidemic forecasting and scenario modeling groups such as the CDC Flusight Forecasting Challenge, COVID-19 Forecast Hub, and Scenario Modeling Hub (SMH) have required predictions to be reported in quantile format [1–3]. Scoring rules, such as the weighted interval score (WIS), can then be applied to quantile format outputs to analyze the performance of projections with respect to observed surveillance data. Epidemic predictions however are often generated from a collection of stochastic trajectories, each one representing a single potential realization of how an epidemic might unfold [4]. For this reason, recently, there has been a growing trend toward reporting individual stochastic trajectories to collaborative hubs, as demonstrated in recent rounds of the Scenario Modeling Hub [5]. However, evaluation methods for epidemic projections have not yet been widely adapted to effectively handle this emerging format.

In this paper, we study the utility and significance of the energy score as a performance metric for evaluating epidemic model projections reported in a stochastic trajectory format, illustrated with examples from scenario modeling. The energy score has been applied across various fields, including weather forecasting [6, 7], electricity market pricing [8, 9], and wind/solar power generation [10–13]. In epidemic modeling, the energy score has been applied in a

medRxiv preprint doi: https://doi.org/10.1101/2025.01.13.25320493; this version posted January 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

limited number of cases to analyze multivariate time-series models [14–16]. Moreover, computational packages used to assess probabilistic forecasts with proper scores have implemented the energy score [17–19]. However, it has yet to gain widespread adoption as a standard metric for performance evaluation in epidemic forecasting and prediction.

In this study, we define the energy score, outline methods to adapt it for specific applications, and perform synthetic experiments to explore its properties. We then apply the energy score to evaluate the performance of models contributing to the 2023–24 Flu Scenario Modeling Hub (SMH). With its recent transition to trajectory-based submissions, the SMH provides an ideal dataset for demonstrating the utility of the energy score in a real-world context [3, 20]. While scenario projections serve a different purpose than forecasts, their evaluation often focuses on how well the projected trajectories capture the future dynamics of the epidemic [5, 21]. Our analysis shows that the energy score is a rigorous and versatile metric for assessing the performance of both individual models and ensemble projections, making it an ideal scoring rule for trajectory-based epidemic predictions.

Methods

When evaluating probabilistic predictions, it is important to use proper scoring rules for model evaluation. Proper scoring rules are evaluation measures such that a forecaster has no incentive to predict anything other than their own true belief [22, 23]. If G is the underlying generative process of the observations y, the score comparing the observed data with G will on average give the optimal score. A scoring rule is strictly proper if the generative process of the observations G gives the best score $S(G, y)$, against the observed data y (a single realization of G), and any prediction P, with score $S(P, y)$ will be greater unless $P = G$. This can be shown by

$$S(G, y) \leq S(P, y) \tag{1}$$

medRxiv preprint doi: https://doi.org/10.1101/2025.01.13.25320493; this version posted January 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

and is equal only if $P = G$, where $S(F, y)$ describes the score of a stochastic process F with respect to the true observation y, and a smaller score is considered better [6]. A score is proper but not strictly proper if the inequality holds, but is not uniquely minimized by the generative process G. This means that a prediction could give an optimal score even if it is not identical to the generative process of the observed data. The energy score is a strictly proper negatively-oriented score and it is the multivariate generalization of the continuous ranked probability score (CRPS) [6, 23].

The Energy Score for Trajectory-based Projections

The concept of energy score is derived from energy statistics and it measures distances between statistical observations to quantify differences between distributions [24, 25]. The energy score (ES(P, y)) of a multivariate distribution P, where $\mathbf{X}^{(i)}$ and $\mathbf{X}^{(j)}$ are vectors of independent random variables drawn from P, and $\mathbf{y}$ is the vector of observed values, is defined as:

$$\mathrm{ES}(P, \mathbf{y}) = \sum_{i=1}^{N} p_i ||\mathbf{X}^{(i)} - \mathbf{y}|| - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} p_i p_j ||\mathbf{X}^{(i)} - \mathbf{X}^{(j)}||, \tag{2}$$

where $|| \cdot ||$ is the Euclidean norm, p_i is the weight attributed to each individual trajectory i, and N is the total number of trajectories being analyzed [11, 23]. The weights p_i can be defined such that trajectories with a higher probability of occurrence are given more weight in the evaluation [11]. This approach could be applied when evaluating an ensemble model composed of multiple individual models or scenarios grouped together, where trajectories from certain models or scenarios are expected to perform better than others. In the following, we will assume that in projections from a single model, all trajectories should be weighted equally, as changing the weights affects the interpretation of the energy score. If we assume that all trajectories are equally weighted with weight $p_i = \frac{1}{N}$, and we

medRxiv preprint doi: https://doi.org/10.1101/2025.01.13.25320493; this version posted January 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

expand the Euclidean norm, then equation 2 can be written as:

$$\mathrm{ES}(P, \mathbf{y}) = \underbrace{\frac{1}{N}\sum_{i=1}^{N}\sqrt{\sum_{m=1}^{M}\left(x_m^{(i)} - y_m\right)^2}}_{\text{distance from surveillance data}} - \underbrace{\frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sqrt{\sum_{m=1}^{M}\left(x_m^{(i)} - x_m^{(j)}\right)^2}}_{\text{distance between trajectories}}, \quad (3)$$

where M is the number of elements in each vector (i.e. the number of time points projected), $x_m^{(i)}$ is the predicted value by trajectory i at time m, and y_m is the value of the surveillance data at time m. In this definition, the trajectory could be a time-series, where each entry is a prediction for a given date, or more generally a vector of predictions, such as different outcome targets. As we will discuss later, the score is composed of two components, the first term which compares the distance between the predicted trajectories and the observed data, and the second reflects the distances of the trajectories to each other.

The energy score in Eq. 3, as well as the WIS, is an absolute measure, meaning it is strongly influenced by deviations from the signal at higher magnitudes. In other words, if a trajectory has the same relative error for two observed values of different magnitudes, the energy score will place significantly greater weight on the deviation associated with the larger magnitude. This can be a desirable feature in scenarios where errors on large data points have more serious implications. For instance, a 30% relative error in hospitalization at the onset of an epidemic might correspond to only a few weekly admissions, while the same error at the peak of the season could represent a significant underestimation of hospital bed demand. In such cases, the energy score appropriately penalizes deviations at the epidemic peak more than those at the tail (as discussed in Refs. [22, 26]).

However, in other situations, this feature may be less desirable—for example, when comparing forecast quality across states with inherently different epidemic curve magnitudes due to varying population sizes. In such cases, a relative energy score can be defined by normalizing the score by the sum of the observed time series data, thereby rescaling the score by the signal's overall size [6]. A normalized energy score can be written as:

$$\mathrm{ES}_{\mathrm{norm}} = \frac{ES(P, \mathbf{y})}{\sum_{m=1}^{M} y_m}, \quad (4)$$

medRxiv preprint doi: https://doi.org/10.1101/2025.01.13.25320493; this version posted January 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

which provides a relative measure that facilitates comparisons of normalized energy scores across different locations. This normalization still emphasizes deviations at the peak within each state but adjusts the overall score based on the sum of the signal, enabling fairer comparisons between states with different population sizes. This approach can also be extended to compare scores across different projection targets or time periods. The issue of normalization is similarly relevant in the multi-dimensional extension of the energy score, as discussed in the next section. Unless otherwise specified, we will use the energy score as defined in Eq. 3.

Multi-Dimensional Energy Score

One benefit of the energy score is that it can be adapted into a performance measure across multiple dimensions. This would allow us to evaluate a model across multiple target outcomes (i.e. cases, deaths, or hospitalizations), age groups, locations, and so on, with a single score, giving us a comprehensive understanding of a model's performance with respect to all of its predictions in high-dimensional space. With other scoring rules, this would only be possible via a summary statistic, such as a sum or average of scores for each outcome variable. To calculate this, we look at each time point as a multi-dimensional vector, where T represents the number of prediction targets we are assessing. Now, we have a matrix where the the columns describe the predictions at each time point, and the rows show the time-series predictions for each outcome. In this definition of the energy score, we are looking at the distance between matrices instead of the distance between vectors, as in the standard energy score. Therefore, we can use the Frobenius norm to measure distance, as it is the multi-dimensional extension of the Euclidean norm, and define the multi-dimensional energy score as:

$$\mathrm{ES}_{\mathrm{dim}} = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\sum_{j=1}^{T}\Phi_j^2\sum_{m=1}^{M}(A_{jm}^{(i)} - y_{jm})^2} - \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{k=1}^{N}\sqrt{\sum_{j=1}^{T}\Phi_j^2\sum_{m=1}^{M}(A_{jm}^{(i)} - A_{jm}^{(k)})^2}, \tag{5}$$

where $A^{(i)}$ is a T x M matrix of predictions for trajectory i, with T outcomes

medRxiv preprint doi: https://doi.org/10.1101/2025.01.13.25320493; this version posted January 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

and M time points, Φ_j is a normalization factor for the magnitude of the signal along each dimension/target, N is the number of trajectories reported for each outcome target, T is the number of outcomes, and M is the number of time points. Individual trajectories for each outcome dimension are constructed into these trajectory matrices, where the number of trajectories for each dimension must be the same, and the number of trajectory matrices will equal the number of trajectory vectors reported for each outcome dimension. The T x M matrix y describes the surveillance data, matching the construction of the trajectory matrices $A^{(i)}$.

The factorΦ_j^2 rescales the signal, adjusting the contribution of each forecast target T to the multi-dimensional energy score. When $\Phi_j^2 = 1$, targets with larger magnitudes dominate the score, implying that they are more relevant for assessing model performance. However, this assumption may not always be desirable. For example, if the targets are hospitalizations and deaths, the $T = 2$ energy score will be heavily influenced by the hospitalization target, which typically has a much larger magnitude than the death target. However, it may be preferable for the model to predict both targets with equal accuracy. In such cases, we can use a rescaling factor $\Phi_j = \frac{1}{\sum_{m=1}^{M} y_{jm}}$, dividing by the sum of the observation vectors for each target outcome to ensure that all outcome dimensions contribute similarly to the multi-dimensional energy score. The same principle applies when forecast targets correspond to different geographical locations. For instance, if each U.S. state is treated as a forecast target, $T = 50$, then the energy score will be dominated by states with larger populations, which typically have higher hospitalization or death counts. Applying a rescaling factor in this context ensures that performance across all states is weighted equally, regardless of population size. The choice of $\Phi_j^2 = 1$ can be adapted on a case-by-case basis, depending on the objectives of the energy score assessment and the desired balance between different forecast targets.

Finally, when grouping predictions into multi-dimensional vectors, it is crucial to pair trajectories such that those from the same simulation are used in the same matrix A. This is because the construction of the time series matrix directly influences the energy score value. In other words, the projections for each specific

medRxiv preprint doi: https://doi.org/10.1101/2025.01.13.25320493; this version posted January 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

target must originate from the same simulation trajectory. Only when the projections for each target are generated independently can different low-dimensional trajectories be randomly combined into a single higher-dimensional trajectory without compromising important correlations in the modeling output.

Comparison of scoring rules for synthetic data

Several scoring rules have been proposed in the literature to evaluate the performance of probabilistic epidemic projections. Among these, the weighted interval score (WIS) has emerged as a widely adopted standard in forecasting and scenario modeling efforts [3, 22]. The WIS is a negatively-oriented proper score applied to $(1-\alpha)\times 100\%$ prediction intervals. The score consists of three terms that describe the width or uncertainty in the prediction interval, and penalties if the surveillance data lies outside the prediction interval. It is computed at each time point with prediction P and observed value y as a weighted sum of the interval score for each $(1-\alpha)\times 100\%$ prediction interval of interest, and approximates the continuous ranked probability score (CRPS) [22]. In order to evaluate the performance of a full projection time series using the WIS, we take the average of the WIS calculated at each time point. We further discuss the WIS and CRPS in the Supporting Information (SI).

The energy score and WIS both evaluate a projection based on its calibration, or the distance between the predicted and observed values, and sharpness, which is the amount of uncertainty given by the prediction. We show in S1 Fig that the energy score and WIS are similar when evaluating a synthetic predictive distribution at a single time point. However, there are important differences between the two metrics that must be considered. Most importantly, the energy score is strictly proper for the full projection where the WIS is proper, but not strictly proper. This has implications for how these scores evaluate specific probabilistic predictions. We present proofs for the propriety of the energy score and WIS in the Supporting Information.

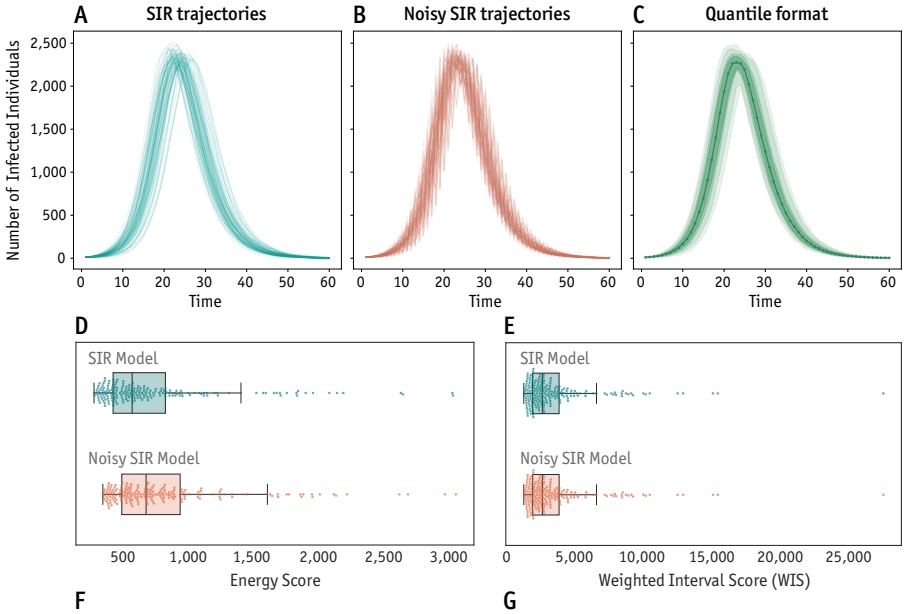
To provide a visual intuition of this difference, we generate trajectories from

medRxiv preprint doi: https://doi.org/10.1101/2025.01.13.25320493; this version posted January 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

two stochastic epidemic processes that have the same marginal distribution at each time step. First, we generate trajectories from a stochastic SIR model using a chain binomial process. We parameterize the SIR model such that the transmissibility is $\beta = 0.625$ and the recovery rate is $\mu = 0.25$ giving us a reproduction number $R_0 = 2.5$. For the second process, we create a noisy SIR model, where we randomly shuffle the values for the number of infectious individuals at each time point in the SIR model trajectories. Then we randomly group these values across time to produce noisy epidemic trajectories. This gives us two epidemic-like processes with the same marginal distribution at each time step, one with time-correlated, and one with time-uncorrelated stochastic trajectories. An example of the two trajectory processes are shown in Fig. 1A and B, where we also show the quantile format for these trajectories in Fig. 1C, which is identical for each trajectory process. We then score the two processes using as observation vector a single realization of the standard stochastic SIR model. Using this framework, we calculate the energy score and WIS for 200 iterations, generating 100 trajectories at 60 time points for each model at every iteration.

In Fig. 1D and E, we show the distribution of the scores for each epidemic trajectory process, and descriptive statistics of these distributions in Fig. 1F and G. From the boxplots and tables of descriptive statistics, we find that the WIS for both model processes is the same, but the energy scores are not. The energy score is able to distinguish between two processes with the same marginal distribution but differing individual behavior, where the WIS scores them identically. Moreover, the SIR model process, on average, produced better values of the energy score than the noisy model, which agrees with the knowledge that the observed values were generated from the SIR model. This is due to the energy score being strictly proper, while the WIS is proper but not strictly proper. If a prediction P has the same marginal distribution as the true underlying process G, it would give the ideal WIS score to both even if $P \neq G$; on the other hand since the energy score is strictly proper, only a prediction $P = G$ can give the ideal energy score [23].

medRxiv preprint doi: https://doi.org/10.1101/2025.01.13.25320493; this version posted January 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .



F

Energy Score	Mean	Median	25% percentile	75% percentile
SIR Model	749.6	563.8	417.3	819.8
Noisy SIR Model	829.9	672.9	484.4	933.3

G

WIS	Mean	Median	25% percentile	75% percentile
SIR Model	3521.6	2595.4	1885.5	3819.7
Noisy SIR Model	3521.6	2595.4	1885.5	3819.7

Fig 1. Synthetic experiments comparing energy score and WIS propriety and evaluation. Comparing behavior of the WIS and energy score for (A) SIR model and (B) noisy SIR epidemic trajectory process with the same marginal distribution at each time point, generating the same quantile format (C). Boxplot of the energy score (D) and WIS (E) obtained from each trajectory process given 100 trajectories each, showing the distribution for each score and epidemic trajectory process across 200 iterations. For visualization purposes, the maximum values for each score and model type is not shown. Boxplots are created such that the box shows the 25%, 50% and 75% quantiles, and the whiskers represent 1.5× interquartile range (IQR). Descriptive statistics of the (F) energy score and (G) WIS for the SIR and noisy SIR model.

Results

We illustrate the application of the energy score to epidemic scenario projections for the 2023-24 projection round of the Flu Scenario Modeling Hub. In the Flu Scenario Modeling Hub, modeling teams provide predictions about future influenza trajectories under certain assumptions about human behavior, environmental factors, or circulating strains. The Scenario Modeling Hub (SMH) has performed 5 cycles of influenza scenario projections; 3 during the 2022-23 influenza season and 1 for the 2023-24 and 2024-25 seasons in addition to 18 scenario projections for COVID-19 and 2 for RSV [27]. In previous projection

medRxiv preprint doi: https://doi.org/10.1101/2025.01.13.25320493; this version posted January 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

rounds, modeling teams were only required to report quantiles of their predictions aggregated for each week, but beginning in the 2023-24 influenza round, teams submitted 100 individual trajectories from their model output for each scenario, and location [5]. The objective of this projection round was to explore the implications of different vaccine coverage levels (high, normal, or low) and the dominant circulating strain (either A/H3N2 or A/H1N1) on the trajectory of weekly hospitalizations during the 2023-2024 flu season for US states and nationally [28]. Modeling teams reported projections for these 6 scenarios, from September 3, 2023 to June 1, 2024. Further information about the projection round and model output can be found in the Flu SMH's GitHub repository. S1 Table describes the 2023-24 SMH scenarios in further detail. In Fig. 2, we show the reported model output for the 'MOBS_NEU-GLEAM_FLU' model in the US nationally for Scenarios C and D of the 2023-24 Flu SMH round, which compares an A/H3N2 dominant season and an A/H1N1 dominant season given historically typical vaccination coverage. We see how the two example scenarios result in distinct epidemic outcomes given assumptions on the circulating influenza strain. We analyze the incident hospitalization projections from September 9, 2023 to April 27, 2024. Using this data, we evaluate the performance of scenario projections with the energy score.

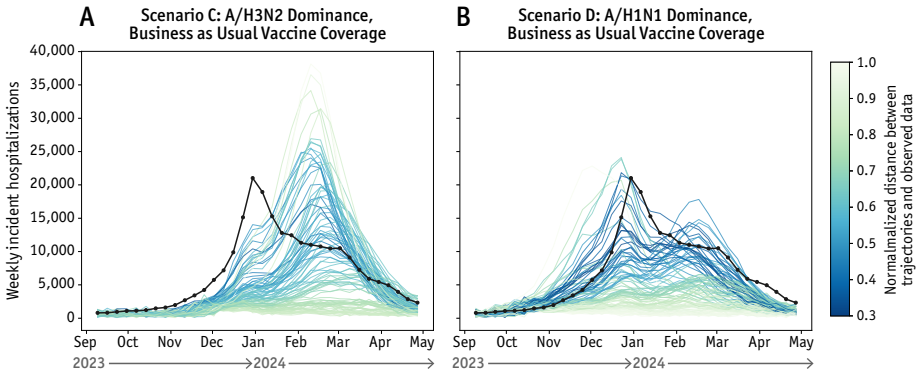


Fig 2. Epidemic predictions in the trajectory format. One hundred trajectories for incident hospitalizations for the 'MOBS_NEU-GLEAM_FLU' model for (A) scenario C and (B) scenario D in the Flu Scenario Modeling Hub 2023-24 round 1 with observed surveillance data (black) nationally in the United States. Trajectories are colored by their normalized distance from the observed data. Lighter colors represent a larger distance.

Given the widespread use of the weighted interval score (WIS) for evaluating

medRxiv preprint doi: https://doi.org/10.1101/2025.01.13.25320493; this version posted January 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

probabilistic epidemic predictions, we compare the performance assessment provided by the WIS to that of the energy score to understand how these metrics evaluate model projections. To compute the WIS, we first estimate the quantiles of the projections using the submitted trajectories for each model, scenario, and location. These quantiles are then used to calculate the WIS based on the corresponding prediction intervals. For consistency, we use the same prediction intervals employed in previous SMH projection rounds, which required quantile-based submissions.; namely the 98%, 95%, 90%, 80%,..., and 10% prediction intervals based on 23 submitted quantiles [22]. The WIS is calculated at individual time points, so we compute the average of these scores across all weeks to obtain the WIS for an entire time series. In the following, we begin our analysis by examining the energy score for the projections of a single model submitted to the Flu SMH, before extending the evaluation to multiple models.

Single Model Evaluation

We focus the first part of our analysis on the 'MOBS_NEU-GLEAM_FLU' model submitted to the Flu Scenario Modeling Hub, which is a multi-scale, age-structured, stochastic metapopulation epidemic model that uses global flight and commuting data to simulate the spread of an infectious disease [29, 30]. In this section, we examine the energy score for this single model across time periods, locations, and scenarios.

As noted earlier, the energy score is influenced by the absolute values defining a trajectory. This dependency complicates the comparison of scores across locations, time periods, or target outcomes with differing magnitudes. In Fig. 3A, we compare the energy score for the 'MOBS_NEU-GLEAM_FLU' model influenza hospitalization projections for each scenario and location to the sum of the corresponding surveillance time series. We find that the non-normalized energy score is strongly correlated to the size of the signal. However, Fig. 3B shows that the normalized energy score depends much less on the magnitude of the surveillance data, with an R^2 of 0.018. This means that scores can be compared with the normalized energy score, and demonstrates its nature as a relative score.

medRxiv preprint doi: https://doi.org/10.1101/2025.01.13.25320493; this version posted January 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

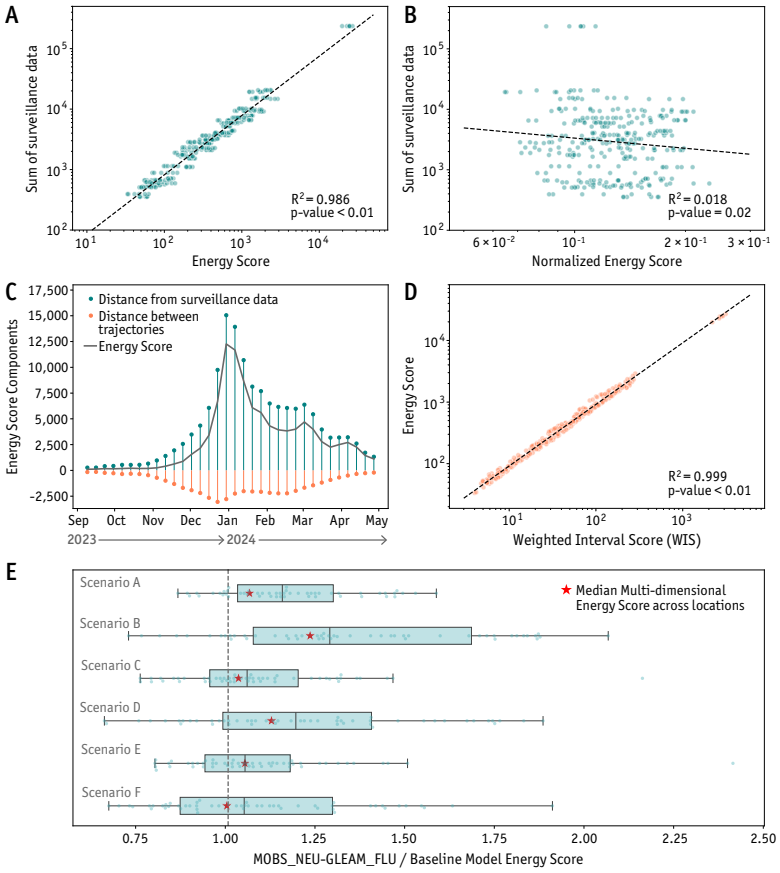


Fig 3. Evaluation of a single model with the energy score. Comparison of the dependence of the (A) energy score and (B) normalized energy score on the sum of the surveillance data of each scenario and location for the model projections. (C) Decomposition of the energy score terms at each week for the Scenario D projections, where the blue represents the term describing the distance of the trajectories from observed data, and the orange describes the term representing the distance between all pairs of trajectories. The gray line shows the full energy score calculated at each week. (D) Relationship between the WIS and energy score. Each point describes the WIS and energy score for a location and scenario. The WIS was found by estimating the quantiles estimated from the trajectories. (E) Boxplot of energy score ratio across 52 locations for the 'MOBS_NEU-GLEAM_FLU' model compared to a 4-week-ahead naive baseline model for each scenario. Vertical dashed line shows where the 'MOBS_NEU-GLEAM_FLU' and baseline model have the same scores, where ratios below one describe when the model performs better than the baseline. The overlaid red stars show the median of the multi-dimensional energy score ratio across locations for 50 iterations of randomizing the trajectory pairings. Boxplots are created such that the box shows the 25%, 50% and 75% quantiles, and the whiskers represent 1.5 × IQR. Scatter plots (A,B,D) show a dashed line of the linear regression fit, with the R^2 and two-sided p-value describing the fit. All results are shown for the 'MOBS_NEU-GLEAM_FLU' model 2023-24 Flu SMH incident hospitalization projections.

medRxiv preprint doi: https://doi.org/10.1101/2025.01.13.25320493; this version posted January 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

While the energy score is defined for a multivariate time series, we analyze its behavior at individual time points to study how the energy score changes throughout an outbreak. In Fig. 3C, we show the energy score at each week of the flu season for the 'MOBS_NEU-GLEAM_FLU' model, scoring incident hospitalization projections for Scenario D (A/H1N1 Dominance, Business as usual vaccination coverage) nationally in the U.S., and we decompose these scores to look at the contribution of individual terms. The energy score is made up of a positive term describing the distance between the trajectories and surveillance data, and the pairwise distance between all trajectories subtracted from this value. This negative term is why the energy score value is less than the component describing the distance from surveillance data. In this example, the energy score is largely composed of the distance between each trajectory and the observed data. The score increases at time points near the peak as expected because the energy score is an absolute metric depending on the magnitude of the surveillance signal, which means it will typically give higher weights near the peak of an epidemic curve. We show in S1 Fig that the WIS follows similar patterns when evaluated at each week of the influenza season.

Since the WIS is a commonly used score for evaluating epidemic projections, we compare the performance of the 'MOBS_NEU-GLEAM_FLU' model using the energy score and WIS. If we calculate the energy score and WIS for the predictions made for each location and scenario for incident hospitalizations in the 2023-24 Flu Scenario Modeling Hub round, we find a strong correlation between the two scores, shown in Fig. 3D. This highlights that a prediction that performs well under the WIS is likely to also be scored well by the energy score.

Using the energy score, we are also able to compare performance across scenarios and locations. In Fig. 3E, we evaluate the performance for all scenarios, looking at the distribution of scores for predictions at each location. To evaluate model performance, we create a naive baseline forecast for 4-week-ahead projections, to use as a reference point against which we compare the performance of long-term scenario projections [31]. While, alternative baseline methods can be devised for scenarios projections, this approach has been used previously to assess the performance of several rounds of the SMH COVID-19 projections [3]. We

medRxiv preprint doi: https://doi.org/10.1101/2025.01.13.25320493; this version posted January 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

report the details of the generation of the naive baseline model and its trajectories in the Supporting Information. From the baseline model, we create an energy score ratio where we divide the energy score for each model, location, and scenario by the energy score of the baseline model for each location. A ratio less than 1 describes when the scenario model performs better than the naive baseline model. It is important to note that scenario projections are not necessarily expected to outperform the baseline model. This is because the baseline model is continuously updated with new data each week to produce 4-week-ahead predictions, whereas scenario projections are made several months in advance without incorporating recent surveillance data. In our analysis, we find that most scenarios for the 'MOBS_NEU-GLEAM_FLU' model perform similar to or slightly worse than the naive baseline model. In S2 Table, we show descriptive statistics for the energy score ratio 'MOBS_NEU-GLEAM_FLU' model for each scenario.

We also utilize the multi-dimensional energy score to compute a single comprehensive score for each scenario, providing an overall assessment of the 'MOBS_NEU-GLEAM_FLU' model's performance across all U.S. states. The multi-dimensional energy score requires trajectories to be paired consistently across all dimensions, including across states. However, the SMH does not explicitly provide the pairing of trajectories across different locations. For the 'MOBS_NEU-GLEAM_FLU' model, trajectories in different states are originated from independent calibrations, allowing us to randomize the trajectory identifiers and conduct a sensitivity analysis to evaluate how different pairings influence the multi-dimensional energy score, repeating this process over 50 iterations. Additionally, we calculate an energy score ratio, as previously described, to compare these values against the multi-dimensional energy score of the 4-week-ahead naive baseline model. In Fig. 3E, the red stars show the median multi-dimensional energy score ratio across these iterations for each scenario. In S5 Table, we show that the uncertainty around the multi-dimensional energy score for randomizing the trajectory pairings is very small. Additionally, we find that the distribution of the energy score ratio across locations in the boxplots and the multi-dimensional energy score rank the scenarios similarly, with both medians following the same pattern. This shows that the multi-dimensional energy score is

medRxiv preprint doi: https://doi.org/10.1101/2025.01.13.25320493; this version posted January 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

evaluating the model similarly to the standard energy score, but in an aggregated manner. In addition, we examine the impact of removing the normalization factor from the multi-dimensional energy score expression in order to keep emphasis on locations with larger signals in S2 Fig.

Multi-Model Analysis

In this section, we focus our analysis on the performance of all models submitted to this Flu SMH projection round that provided incident hospitalization projections for multiple locations. This selection resulted in six individual models being included in the analysis, with four models excluded from the study because they only reported data for one location. Ensemble models generated by the SMH were not considered, as they are reported only in a quantile format. To illustrate the utility of the energy score in comparing performance across multiple models, we analyze model rankings and performance across different locations. In Fig. 4 we show how the energy score varies across models, scenarios, and locations. We calculate the energy score ratio in comparison with the 4-week-ahead naive baseline model [31] described previously for each scenario, location, and model, where a ratio below one represents a case where the scenario model performed better than the naive baseline model. We find that there are models that consistently perform better than others, and that some models perform very similar to or slightly better than the baseline model. S3 Table presents descriptive statistics of this data for all scenarios and models by summarizing the energy score ratio for all models and scenarios examined. This demonstrates the ability of the energy score to compare models and differentiate between better-performing models. This tells us about relative model performance within each scenario.

We also investigate how model rankings based on the energy score compare to those derived from the WIS. For each scenario and location, we rank the models from lowest to highest score using both the energy score and WIS. To assess the agreement between these rankings, we calculate Kendall's τ rank correlation across all models. The Kendall's τ correlation coefficient compares rankings by counting the number of pairs of objects that are in an incorrect order, divided by

medRxiv preprint doi: https://doi.org/10.1101/2025.01.13.25320493; this version posted January 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

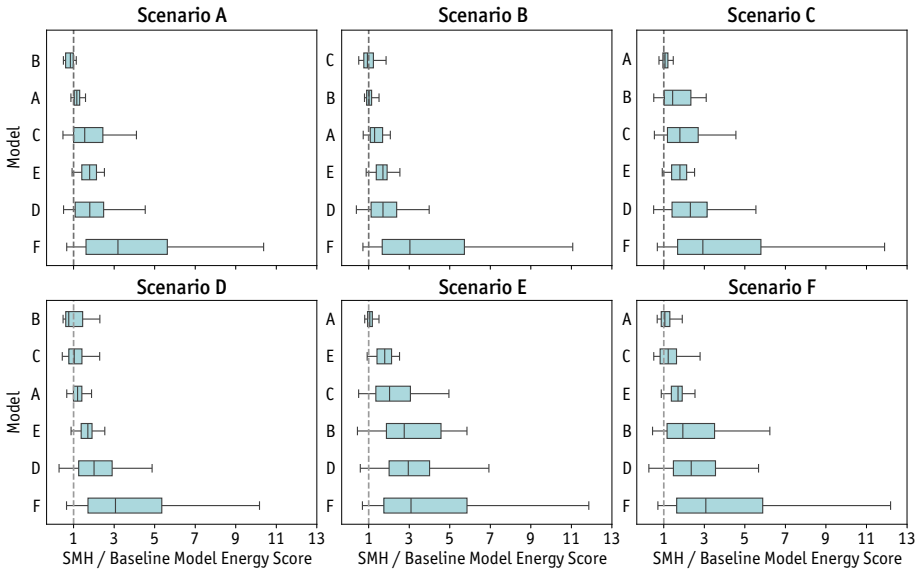


Fig 4. Energy score ranking across models. Boxplots showing the distribution of the ratio of the energy score for each model, location, and scenario in Flu Scenario Modeling Hub divided by the energy score for the 4 week ahead naive baseline flu model at each location. Each boxplot shows the energy score ratio distribution over the number of locations the corresponding model reported (shown in SI). Vertical dashed line shows where the ensemble and baseline model has equal energy scores such that ratios below one describe when the SMH model performs better than the baseline. Models are ordered by median energy score ratio within each scenario. Boxplots are such that the box shows the 25%, 50% and 75% quantiles, and the whiskers represent 1.5 × IQR. Note that Model A represents the 'MOBS_NEU-GLEAM_FLU' model, which can be compared to results in Fig. 3.

the total possible pairs [32]. A rank correlation of 1 means that the rankings of the energy score and WIS are identical. In S1 Fig, we show a histogram of the Kendall's τ rank correlation between the energy score and WIS for all models and scenarios. The mean rank correlation coefficient between the energy score and WIS is 0.87. This shows that the energy score and WIS rank models quite similarly, but there can be cases where the rankings are not identical. More precisely, we find that for 49.7% of projections, the energy score and WIS have a Kendall's τ correlation coefficient of 1.

Trajectory-Based Ensemble

In multi-model analysis, it is common practice to create an ensemble model by combining predictions from the multiple individual models. [21, 31]. In the

medRxiv preprint doi: https://doi.org/10.1101/2025.01.13.25320493; this version posted January 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

ensemble, the information and uncertainty from multiple models are aggregated together to provide a "consensus" projection of future possibilities aggregating the different assumptions and methodologies of individual models [26]. The performance of ensemble models has been shown to generate improved future predictions that characterize uncertainty better than individual models [21, 26, 33, 34]. The SMH includes three ensemble models as part of its analysis. However, these ensembles are built by first converting each model's trajectories into a quantile format before applying the ensemble methods. Here, we propose an alternative approach for generating an ensemble model that directly utilizes the trajectories reported by each modeling team. In this method, we simply bundle all trajectories from each model into the definition of a single ensemble, assigning equal weight to each trajectory. This approach avoids the need to summarize stochastic model outputs into quantile format, leveraging the raw data provided by the models instead. In S3 Fig, we compare this ensemble of trajectories method with the three ensemble models reported by the SMH.

In Fig. 5, we explore the energy score across locations using the trajectory-based ensemble. We include all models reported to the 2023-24 Flu SMH that submit projections of incident hospitalizations for any number of the 50 U.S. states, Washington D.C., or nationally. Fig. 5A shows the ratio of the energy score of the trajectory-based ensemble model divided by the energy score of the 4-week-ahead naive baseline model at each location (U.S. states and the District of Columbia) for Scenario D. Locations in blue show where the ensemble model performed better than the baseline model (ratio < 1), where those in orange show where the naive baseline model had better performance (ratio > 1). We observe heterogeneous performance across the United States, with the ensemble model outperforming the naive baseline in 24 locations, but falling short in 27. These results indicate areas where model performance could be improved for future predictions. Notably, in this example, the ensemble tends to perform better in the Southern U.S. states.

We extend this to show the distribution in the energy score ratio compared to the 4-week-ahead naive baseline model across locations for each scenario in Fig. 5B. This allows us to assess the overall performance of the trajectory-based

medRxiv preprint doi: https://doi.org/10.1101/2025.01.13.25320493; this version posted January 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

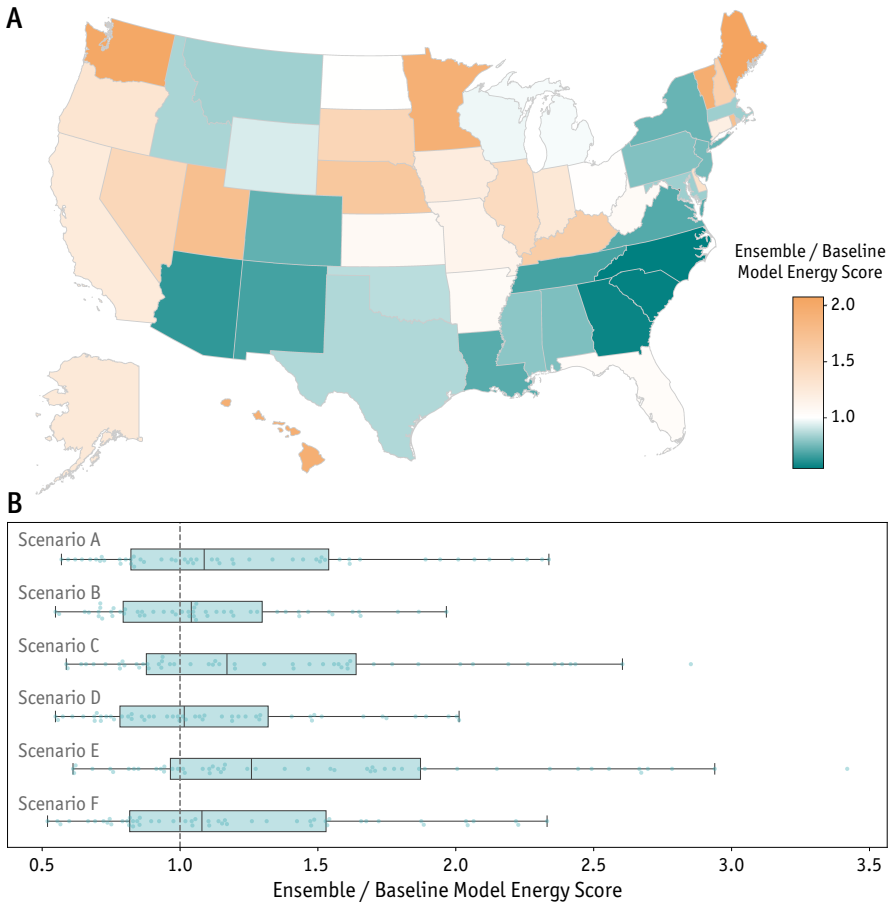


Fig 5. Energy score for the trajectory-based ensemble model. (A) Map of the US showing the energy score ratio (trajecory-based Ensemble / baseline model energy score) for each US state of the 2023-24 Flu Scenario Modeling Hub round for Scenario D. Ratios below one (blue) show where the energy score for the trajectory-based ensemble model for a state was better than the naive baseline, where the ratios above one (orange) show where the baseline model had better performance. (B) Boxplots of the energy score ratio across 52 locations for all scenarios compared to a 4-week-ahead naive baseline model. Vertical dashed line shows where the ensemble and baseline model has equal performance, where ratios below one describe when the ensemble model performs better than the baseline. Boxplots are such that the box shows the 25%, 50% and 75% quantiles, and the whiskers represent 1.5 × IQR.

ensemble model for each scenario, and compare the scenarios to each other. This analysis illustrates that scenarios B, D, and F are the best-performing scenarios for the ensemble model, which corresponds to the scenarios describing H1N1 as the dominant circulating strain. This is in agreement with the observed strain dynamics in the United States for the 2023-24 influenza season [35]. We show summary statistics of this data in S4 Table.

medRxiv preprint doi: https://doi.org/10.1101/2025.01.13.25320493; this version posted January 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

It is worth noting that we do not calculate the multi-dimensional energy score for the ensemble model due to the assumptions required about trajectory independence and pairing—information that is unavailable for the other modeling teams within the SMH.

This ensemble of trajectories method could be easily translated to a different weighting process where models or even specific trajectories are weighted differently as data is gathered [4]. This allows for a flexible framework for ensemble creation and performance analysis using only single trajectories generated by individual models without the need to summarize and aggregate model output using quantiles.

Discussion

In this paper, we discuss the benefits and application of the energy score to evaluate probabilistic epidemic predictions given in a trajectory format. It is important to highlight that the energy score is not a new method, and it has been used in the evaluation of probabilistic forecasts in many fields [6]. Using multiple scoring rules to assess predictions is shown to be useful [36], and the energy score adds a strictly proper performance measure to the toolbox of probabilistic prediction analysis methods. The energy score is a natural way of assessing epidemic model output given in trajectory format, as it does not rely on summarizing stochastic model output into quantiles. Moreover, the energy score acknowledges the stochastic nature of epidemic forecasting. It follows the full path of each individual trajectory over a time series, and is a multivariate score that assesses a probabilistic time series with one score value. We show how the strictly proper nature of the energy score is able to differentiate between predictions with the same marginal distribution but different generating processes. We introduce the multi-dimensional energy score in order to gain a comprehensive understanding of a model's performance across all of its predictions.

One limitation of the energy score is the increasing computational cost of the energy score as the number of trajectories increases. The second term in the energy score expression requires the calculation of the pairwise distance between

medRxiv preprint doi: https://doi.org/10.1101/2025.01.13.25320493; this version posted January 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

all trajectories. This can become computationally challenging when the number of trajectories is large. We show in S4 Fig that randomly sampling even a small percent of trajectories gives good agreement with the energy score found through using the full set of stochastic trajectories.

The information contained within epidemic projections in quantile format versus trajectory format has both advantages and disadvantages. Trajectories allow for greater flexibility in weighting and analysis of time-series features, where the quantile format does not retain information on variability within potential epidemic outcomes [4]. The energy score acknowledges the individual behavior of single trajectories, where scores evaluated using a quantile format evaluate the descriptive statistics at each week. This can suppress important epidemiological information and obscure the true uncertainty of different epidemic outcomes [4, 37, 38]. Typically, predictions in a quantile format highlight the most likely outcomes, rather than the worst-case scenario, which is critically important to public health decision-makers [39]. In this paper, we show how the individual behavior of a stochastic model realization can change the energy score, even if descriptive statistics are the same. This is important to consider when choosing the best evaluation method to apply to any given research question.

The discrimination ability of a scoring rule illustrates the differences in scores generated by forecasts of differing quality. For example, a scoring rule has low (high) discrimination ability if forecasts with a very different quality result in the same (different) scoring value [40]. A proper scoring rule can still have poor discrimination ability. It has been shown that the energy score lacks discrimination ability between forecasts with different correlation structures, but discriminates well between predictions with different means or variances [14, 40, 41]. Many of these works are aimed at the capability of the energy score to correctly identify the true underlying distribution driving the dynamics. While this is important to consider when employing the energy score for the performance analysis of epidemic forecasts, we believe that it should not limit our use of this tool. The main goal in evaluating epidemic projections and forecasts is not the specification of the true data-generating distribution, but the identification of projections that are closest to reality.

medRxiv preprint doi: https://doi.org/10.1101/2025.01.13.25320493; this version posted January 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Our results consider equally weighted trajectories, but the energy score definition allows for the weight of each trajectory to be individually defined. This could lead to developing scoring strategies that allows changing trajectory weights based on past performance, or a particular outcome of interest. This adjustment may alter the intuition behind the energy score, but it could be useful for generating ensembles or calibrating models as we gather more information about the progression of an epidemic. While we use the Flu Scenario Modeling Hub as an example of how the energy score can be utilized, the energy score can be extended beyond scenario projections to any epidemic forecast or prediction that produces stochastic realizations. As trajectories become more common in epidemic model reporting, the energy score should be considered in model evaluation as a way to utilize the information contained within individual trajectories. We do not believe that this should discourage the use of other scoring rules, but we illustrate the energy score as a robust performance metric if trajectories are available.

Author contributions

Conceptualization: CB, KM, GS, JTD, AV; Investigation: CB, KM, GS, MC, JTD, AV; Methodology: CB, KM, GS, JTD, AV; Software: CB; Supervision: JTD, AV; Writing–original draft: CB, AV; Analyze the results, writing–review & editing the paper: CB, KM, GS, MC, JTD, AV.

Declaration of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data and Code Availability

For surveillance data, we use the National Healthcare Safety Network (NHSN) Weekly Hospital Respiratory Data [42] for confirmed influenza hospital admissions.

medRxiv preprint doi: https://doi.org/10.1101/2025.01.13.25320493; this version posted January 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

We use a formatted version of this data provided by the CDC FluSight Forecast Hub [43]. Scenario model projections are made available by the Flu Scenario Modeling Hub [28] and in their GitHub repository. Our code for this project is publicly available on Zenodo [44].

Acknowledgments

We acknowledge support from HHS/CDC 5U01IP0001137, and the cooperative agreement CDC-RFA-FT-23-0069 from the CDC's Center for Forecasting and Outbreak Analytics. The findings and conclusions in this study are those of the authors and do not necessarily represent the official position of the funding agencies, the CDC, or the U.S. Department of Health and Human Services of the United States.

References

1. Mathis SM, Webber AE, León TM, Murray EL, Sun M, White LA, et al. Evaluation of FluSight influenza forecasting in the 2021-22 and 2022-23 seasons with a new target laboratory-confirmed influenza hospitalizations. medRxiv. 2023;doi:10.1101/2023.12.08.23299726.

2. Cramer EY, Huang Y, Wang Y, Ray EL, Cornell M, Bracher J, et al. The United States COVID-19 Forecast Hub dataset. Scientific Data. 2022;9(1):462. doi:10.1038/s41597-022-01517-w.

3. Howerton E, Contamin L, Mullany LC, Qin M, Reich NG, Bents S, et al. Evaluation of the US COVID-19 Scenario Modeling Hub for informing pandemic response under uncertainty. Nature Communications. 2023;14(1):7260. doi:10.1038/s41467-023-42680-x.

4. Sherratt K, Srivastava A, Ainslie K, Singh DE, Cublier A, Marinescu MC, et al. Characterising information gains and losses when collecting multiple epidemic model outputs. Epidemics. 2024;47:100765. doi:https://doi.org/10.1016/j.epidem.2024.100765.

medRxiv preprint doi: https://doi.org/10.1101/2025.01.13.25320493; this version posted January 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

5. Loo SL, Howerton E, Contamin L, Smith CP, Borchering RK, Mullany LC, et al. The US COVID-19 and Influenza Scenario Modeling Hubs: Delivering long-term projections to guide policy. Epidemics. 2024;46:100738. doi:https://doi.org/10.1016/j.epidem.2023.100738.

6. Gneiting T, Stanberry LI, Grimit EP, Held L, Johnson NA. Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. TEST. 2008;17(2):211–235. doi:10.1007/s11749-008-0114-x.

7. Kapoor A, Negi A, Marshall L, Chandra R. Cyclone trajectory and intensity prediction with uncertainty quantification using variational recurrent neural networks. Environmental Modelling & Software. 2023;162:105654. doi:https://doi.org/10.1016/j.envsoft.2023.105654.

8. Cramer E, Witthaut D, Mitsos A, Dahmen M. Multivariate probabilistic forecasting of intraday electricity prices using normalizing flows. Applied Energy. 2023;346:121370. doi:https://doi.org/10.1016/j.apenergy.2023.121370.

9. Grothe O, Kächele F, Krüger F. From point forecasts to multivariate probabilistic forecasts: The Schaake shuffle for day-ahead electricity price forecasting. Energy Economics. 2023;120:106602. doi:https://doi.org/10.1016/j.eneco.2023.106602.

10. Pinson P, Girard R. Evaluating the quality of scenarios of short-term wind power generation. Applied Energy. 2012;96:12–20. doi:https://doi.org/10.1016/j.apenergy.2011.11.004.

11. Staid A, Watson JP, Wets RJB, Woodruff DL. Generating short-term probabilistic wind power scenarios via nonparametric forecast error density estimators. Wind Energy. 2017;20(12):1911–1925. doi:https://doi.org/10.1002/we.2129.

12. Zhang Y, Wang J, Wang X. Review on probabilistic forecasting of wind power generation. Renewable and Sustainable Energy Reviews. 2014;32:255–270. doi:https://doi.org/10.1016/j.rser.2014.01.033.

medRxiv preprint doi: https://doi.org/10.1101/2025.01.13.25320493; this version posted January 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

13. van der Meer D. A benchmark for multivariate probabilistic solar irradiance forecasts. Solar Energy. 2021;225:286–296. doi:https://doi.org/10.1016/j.solener.2021.07.010.

14. Held L, Meyer S, Bracher J. Probabilistic forecasting in infectious disease epidemiology: the 13th Armitage lecture. Statistics in Medicine. 2017;36(22):3443–3460. doi:https://doi.org/10.1002/sim.7363.

15. Engebretsen S, Diz-Lois Palomares A, Rø G, Kristoffersen AB, Lindstrøm JC, Engø-Monsen K, et al. A real-time regional model for COVID-19: Probabilistic situational awareness and forecasting. PLOS Computational Biology. 2023;19(1):1–26. doi:10.1371/journal.pcbi.1010860.

16. Bonacina F, Boëlle PY, Colizza V, Lopez O, Thomas M, Poletto C. Characterization and forecast of global influenza (sub)type dynamics. medRxiv. 2024;doi:10.1101/2024.08.01.24311336.

17. Jordan A, Krüger F, Lerch S. Evaluating Probabilistic Forecasts with scoringRules. Journal of Statistical Software. 2019;90(12):1–37. doi:10.18637/jss.v090.i12.

18. Allen S. Weighted scoringRules: Emphasizing Particular Outcomes When Evaluating Probabilistic Forecasts. Journal of Statistical Software. 2024;110(8):1–26. doi:10.18637/jss.v110.i08.

19. Zanetta F, Allen S. Scoringrules: a python library for probabilistic forecast evaluation; 2024. Available from: `https://github.com/frazane/scoringrules`.

20. Runge MC, Shea K, Howerton E, Yan K, Hochheiser H, Rosenstrom E, et al. Scenario design for infectious disease projections: Integrating concepts from decision analysis and experimental design. Epidemics. 2024;47:100775. doi:https://doi.org/10.1016/j.epidem.2024.100775.

21. Reich NG, et al. Collaborative Hubs: Making the Most of Predictive Epidemic Modeling. Am J Public Health. 2022;112(6):839–842. doi:10.2105/AJPH.2022.306831.

medRxiv preprint doi: https://doi.org/10.1101/2025.01.13.25320493; this version posted January 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

22. Bracher J, Ray EL, Gneiting T, Reich NG. Evaluating epidemic forecasts in an interval format. PLOS Computational Biology. 2021;17(2):1–15. doi:10.1371/journal.pcbi.1008618.

23. Gneiting T, Raftery AE. Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association. 2007;102(477):359–378. doi:10.1198/016214506000001437.

24. Székely GJ, Rizzo ML. The Energy of Data. Annual Review of Statistics and Its Application. 2017;4(1):447–479. doi:10.1146/annurev-statistics-060116-054026.

25. Székely GJ, Rizzo ML. Energy statistics: A class of statistics based on distances. Journal of Statistical Planning and Inference. 2013;143(8):1249–1272. doi:https://doi.org/10.1016/j.jspi.2013.03.018.

26. Cramer EY, et al. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. Proc Natl Acad Sci USA. 2022;119(15):e2113561119. doi:10.1073/pnas.2113561119.

27. Scenario Modeling Hub. Scenario Modeling Hub; 2024. `https://scenariomodelinghub.org/`.

28. Scenario Modeling Hub. Flu Scenario Modeling Hub; 2024. `https://fluscenariomodelinghub.org/`.

29. Balcan D, Gonçalves B, Hu H, Ramasco JJ, Colizza V, Vespignani A. Modeling the spatial spread of infectious diseases: The GLobal Epidemic and Mobility computational model. Journal of Computational Science. 2010;1(3):132–145. doi:https://doi.org/10.1016/j.jocs.2010.07.002.

30. Chinazzi M, Davis JT, y Piontti AP, Mu K, Gozzi N, Ajelli M, et al. A multiscale modeling framework for Scenario Modeling: Characterizing the heterogeneity of the COVID-19 epidemic in the US. Epidemics. 2024;47:100757. doi:https://doi.org/10.1016/j.epidem.2024.100757.

31. Ray EL, Brooks LC, Bien J, Biggerstaff M, Bosse NI, Bracher J, et al. Comparing trained and untrained probabilistic ensemble forecasts of

medRxiv preprint doi: https://doi.org/10.1101/2025.01.13.25320493; this version posted January 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

COVID-19 cases and deaths in the United States. International Journal of Forecasting. 2023;39(3):1366–1383. doi:https://doi.org/10.1016/j.ijforecast.2022.06.005.

32. Kendall MG. A New Measure of Rank Correlation. Biometrika. 1938;30(1-2):81–93. doi:10.1093/biomet/30.1-2.81.

33. Bates JM, Granger CWJ. The Combination of Forecasts. J Oper Res Soc. 1969;20(4):451–468. doi:10.1057/jors.1969.103.

34. McGowan CJ, et al. Collaborative efforts to forecast seasonal influenza in the United States, 2015–2016. Sci Rep. 2019;9(1):683. doi:10.1038/s41598-018-36361-9.

35. Centers for Disease Control and Prevention. Influenza Activity in the United States during the 2023–2024 Season and Composition of the 2024–2025 Influenza Vaccine; 2024. Available from: `https://www.cdc.gov/flu/whats-new/flu-summary-2023-2024.html`.

36. Pic R, Dombry C, Naveau P, Taillardat M. Proper Scoring Rules for Multivariate Probabilistic Forecasts based on Aggregation and Transformation; 2024. Available from: `https://arxiv.org/abs/2407.00650`.

37. Juul JL, Græsbøll K, Christiansen LE, Lehmann S. Fixed-time descriptive statistics underestimate extremes of epidemic curve ensembles. Nature Physics. 2021;17(1):5–8. doi:10.1038/s41567-020-01121-y.

38. McCabe R, Kont MD, Schmit N, Whittaker C, Løchen A, Walker PGT, et al. Communicating uncertainty in epidemic models. Epidemics. 2021;37:100520. doi:10.1016/j.epidem.2021.100520.

39. Wilke CO, Bergstrom CT. Predicting an epidemic trajectory is difficult. Proceedings of the National Academy of Sciences of the United States of America. 2020;117(46):28549–28551. doi:10.1073/pnas.2020200117.

40. Pinson P, Tastu J. Discrimination ability of the Energy score. No. 15 in DTU Compute Technical Report-2013. Technical University of Denmark;

medRxiv preprint doi: https://doi.org/10.1101/2025.01.13.25320493; this version posted January 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

2013. Available from: https://orbit.dtu.dk/en/publications/discrimination-ability-of-the-energy-score.

41. Alexander C, Coulon M, Han Y, Meng X. Evaluating the discrimination ability of proper multi-variate scoring rules. Annals of Operations Research. 2022;doi:https://doi.org/10.1007/s10479-022-04611-9.

42. Centers for Disease Control and Prevention. Weekly Hospital Respiratory Data (HRD) Metrics by Jurisdiction, National Healthcare Safety Network (NHSN) (Preliminary); 2025. https://data.cdc.gov/Public-Health-Surveillance/Weekly-Hospital-Respiratory-Data-HRD-Metrics-by-Ju/mpgq-jmmr/about_data.

43. FluSight Forecast Hub. Target Data; 2025. https://github.com/cdcepi/FluSight-forecast-hub/tree/main/target-data.

44. Bay C. clarabay/energy-score; 2025. Available from: https://doi.org/10.5281/zenodo.14623543.

Supporting information

S2 Fig. Multi-dimensional energy score. Boxplot of energy score ratio across 52 locations for the 'MOBS_NEU-GLEAM_FLU' model compared to a 4-week-ahead naive baseline model for each scenario. Vertical dashed line shows where the 'MOBS_NEU-GLEAM_FLU' and baseline model have the same scores, where ratios below one describe when the model performs better than the baseline. The overlaid red stars show the median of the multi-dimensional energy score ratio across locations for 50 iterations of randomizing the trajectory pairings. The blue stars show the multi-dimensional energy score without the use of a normalization factor. Boxplots are created such that the box shows the 25%, 50% and 75% quantiles, and the whiskers represent 1.5 × IQR.

S1 Fig. WIS analysis. (A) Decomposition of the WIS terms at each week for the 'MOBS_NEU-GLEAM_FLU' model 2023-24 Flu SMH incident hospitalization

medRxiv preprint doi: https://doi.org/10.1101/2025.01.13.25320493; this version posted January 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

projections for Scenario D, where blue show the contribution of the dispersion term, red describes the penalty for underprediction, and orange shows the penalty for overprediction. The gray line shows the full score calculated at each week. (B) Comparison of the energy score (gray) and WIS (dashed orange) calculated at one time point as a function of an observed value y given an underlying predictive distribution (blue) with 100 samples drawn from the predictive distribution to calculate the scores. The predictive distribution is a negative binomial distribution with a mean of 60 and variance of 31. (C) Histogram of the Kendall's τ rank correlation between the model rankings for the energy score and WIS for each scenario and location for the 2023-24 Flu SMH incident hospitalization projections. A Kendall's τ rank correlation of 1 means that the energy score and WIS rank models identically.

S3 Fig. Ensemble model comparison. Relationship between the WIS of the ensemble models reported by the SMH: Ensemble_vincent (left), Ensemble_LOP (middle), and Ensemble_LOP_trimmed (right) and the WIS of the ensemble of trajectories method. Scatter plots show a dashed line of the linear regression fit, with the R^2 and two-sided p-value describing the correlation.

S4 Fig. Sampled energy score. (A) Sampled energy score compared to the true energy score for different densities of sampled trajectories (5, 10, 25, and 50 %) for the 'MOBS_NEU-GLEAM_FLU' model. Dots show a single sampled energy score compared to its corresponding true value for a given location and scenario. (B) Sampled energy score averaged across 50 iterations compared to the true energy score for different densities of sampled trajectories (5, 10, 25, and 50 %) for the 'MOBS_NEU-GLEAM_FLU' model. Dots show the average sampled energy score compared to its corresponding true value for a given location and scenario. Black dotted line shows the y=x line where the sampled and true energy scores are equivalent.

S1 Table 2023-24 Flu Scenario Modeling Hub round. Description of the scenarios used by each modeling team in the 2023-24 flu SMH projection round. VE describes assumptions surrounding vaccine effectiveness (VE).

medRxiv preprint doi: https://doi.org/10.1101/2025.01.13.25320493; this version posted January 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

S5 Table Descriptive statistics of the multi-dimensional energy score for the 'MOBS_NEU-GLEAM_FLU' model. Table showing the mean, standard deviation, minimum, maximum, and range for the multi-dimensional energy score across locations for the 'MOBS_NEU-GLEAM_FLU' model influenza hospitalization predictions for each scenario of the 2023-24 SMH round, given 50 iterations of randomly pairing trajectories.

S2 Table Descriptive statistics of the energy score ratio for the 'MOBS_NEU-GLEAM_FLU' model projections. Table showing the number of locations predicted, minimum, 5% quantile, median, 95% quantile and maximum for energy score ratios across locations for influenza hospitalization predictions for the 'MOBS_NEU-GLEAM_FLU' model of the 2023-24 SMH round across locations for each scenario. The energy score ratio is calculated by dividing the energy score of the SMH model at each scenario and location by the energy score of a 4-week-ahead naive baseline model at each location.

S3 Table Descriptive statistics of the energy score ratio for the all analyzed models and scenarios. Table showing the number of locations predicted, minimum, 5% quantile, median, 95% quantile and maximum for energy score ratios across locations for influenza hospitalization predictions for each scenario and model of the 2023-24 SMH round. The energy score ratio is calculated by dividing the energy score of the SMH model at each scenario and location by the energy score of a 4-week-ahead naive baseline model at each location.

S4 Table Descriptive statistics of the energy score ratio for the ensemble model projections. Table showing the number of locations predicted, minimum, 5% quantile, median, 95% quantile and maximum for energy score ratios across locations for influenza hospitalization predictions for the ensemble of trajectories model across locations for each scenario. The energy score ratio is calculated by dividing the energy score of the ensemble model at each scenario and location by the energy score of a 4-week-ahead naive baseline model at each location.