Northeastern University
Network Science Institute

Evaluating the performance of epidemic scenario projections in stochastic trajectory format with the energy score

Laboratory for the
Modeling of Biological +
Socio-technical Systems

Clara Bay
bay.c@northeastern.edu

Scoring rules

Scoring rules are used to evaluate model predictions and to assess how closely predictions agree with observed surveillance data.

Proper scoring rules are such that a forecaster has no incentive to predict anything other than their own true belief.

A **proper score** comparing the observed data with itself will always give the optimal score.

$$S(y, y) \leq S(P, y)$$

A score is **strictly proper** if the score is uniquely minimized by the observed values.

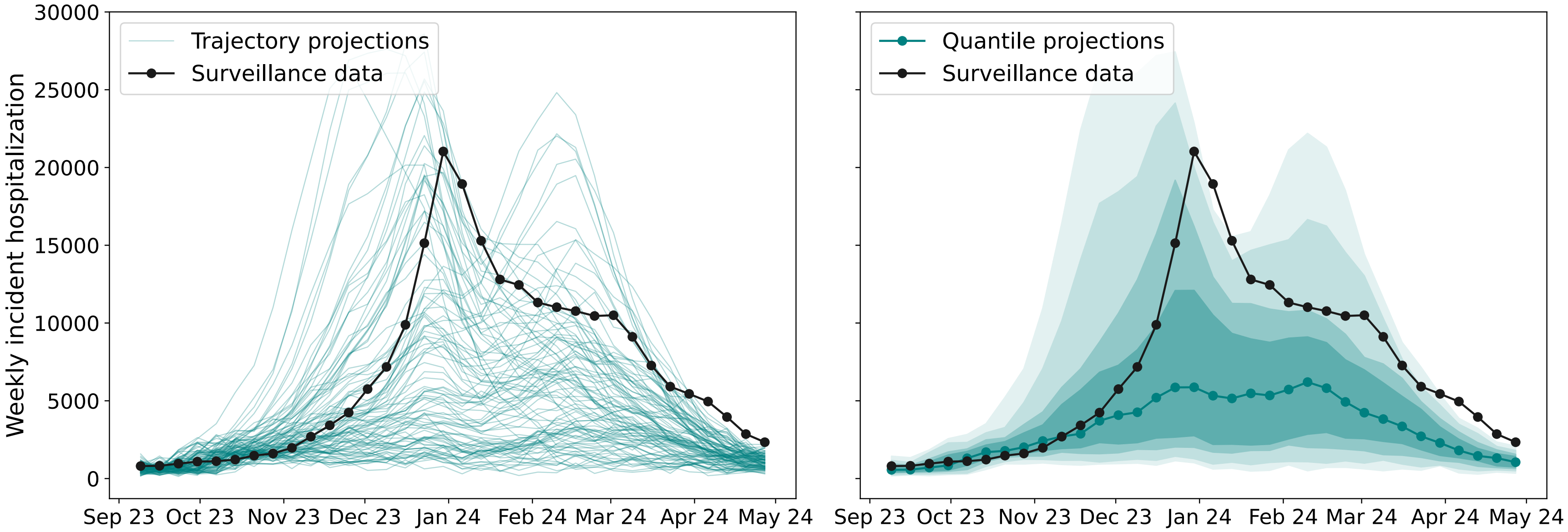
$S(P, y)$ = scoring rule, P = model prediction, y = observed data

Bracher et al. 2021. Gneiting et al. 2007.

Trajectory and Quantile Format

Historically, epidemic predictions (ex. CDC Flusight Forecasting Challenge and Scenario Modeling Hub) are reported in quantile format.

Model output is aggregated at each time point such that only the required quantiles are reported for each target outcome.

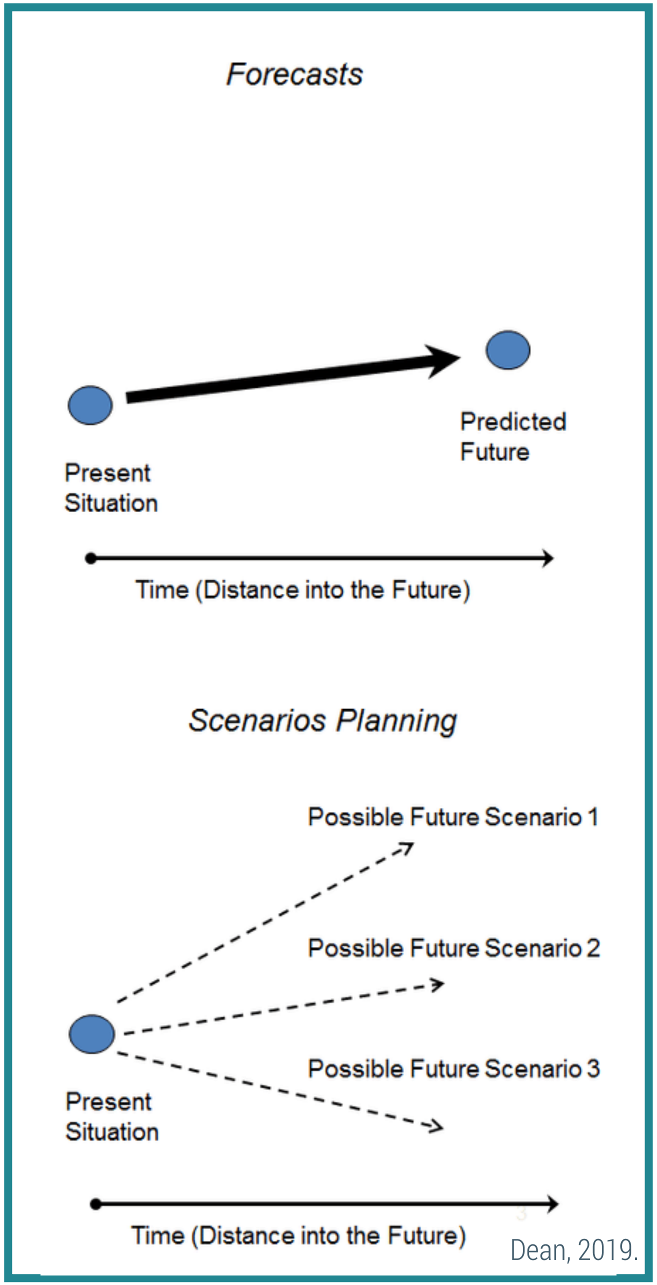


Scenario Modeling

How will an epidemic unfold long-term given assumptions on individual behavior, policies, or disease characteristics?

Different than forecasts.

Not just looking at how close predictions are to observed data.


Forecasts
Present Situation
Predicted Future
Time (Distance into the Future)
Scenarios Planning
Possible Future Scenario 1
Possible Future Scenario 2
Possible Future Scenario 3
Present Situation
Time (Distance into the Future)

Dean, 2019.

Weighted Interval Score (WIS)

The WIS has been widely used to evaluate epidemic predictions in the quantile format.

Negatively-oriented proper score that assesses forecasts by their sharpness and calibration.

Computed at each time point with prediction P and observed value y:

$$\mathrm{WIS}_{\alpha_{0:K}}(P, y) = \frac{1}{K + 0.5}\left(w_0 \,|\, y - m \,|\, + \sum_{k=1}^{K} w_k \mathrm{IS}_{\alpha_k}(P, y) \right)$$

$$\mathrm{IS}_{\alpha}(P, y) = (u_\alpha - l_\alpha) + \frac{2}{\alpha}(l_\alpha - y)\mathbf{1}(y < l_\alpha) + \frac{2}{\alpha}(y - u_\alpha)\mathbf{1}(y > u_\alpha)$$

P = model prediction, y = observed data, K = # prediction intervals, $l_\alpha(u_\alpha)$=lower (upper) bound of prediction interval, w_k = weights, $(1 - \alpha) \cdot 100$ = prediction interval, m = median of predictions

Weighted Interval Score (WIS)

The WIS has been widely used to evaluate epidemic predictions in the quantile format.

Negatively-oriented proper score that assesses forecasts by their sharpness and calibration.

Computed at each time point with prediction P and observed value y:

$$\mathrm{WIS}_{\alpha_{0:K}}(P, y) = \frac{1}{K + 0.5}\left(w_0 \,|\, y - m \,|\, + \sum_{k=1}^{K} w_k \mathrm{IS}_{\alpha_k}(P, y)\right)$$

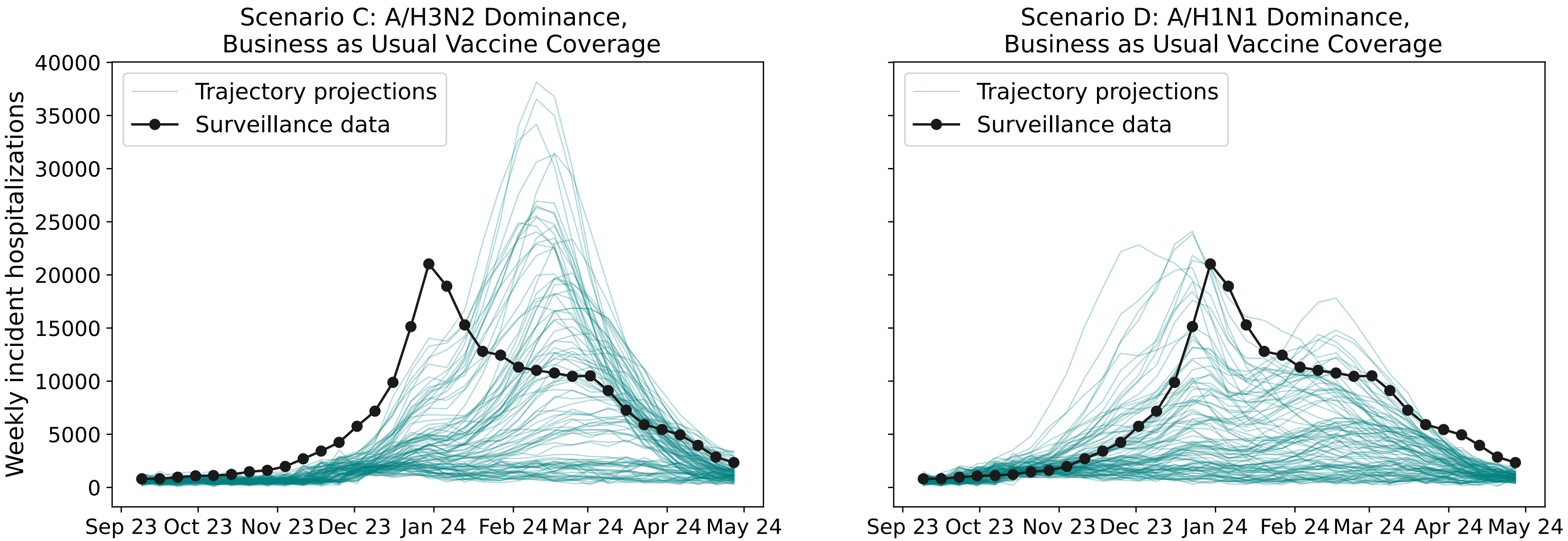
$$\mathrm{IS}_\alpha(P, y) = \underset{\substack{\downarrow \\ \text{dispersion}}}{\boxed{(u_\alpha - l_\alpha)}} + \underset{\substack{\downarrow \\ \text{penalty for} \\ \text{underprediction}}}{\boxed{\frac{2}{\alpha}(l_\alpha - y)\mathbf{1}(y < l_\alpha)}} + \underset{\substack{\downarrow \\ \text{penalty for} \\ \text{overprediction}}}{\boxed{\frac{2}{\alpha}(y - u_\alpha)\mathbf{1}(y > u_\alpha)}}$$

Bracher et al. 2021.

MOBS LAB

Trajectory Format

How do we evaluate model performance using individual stochastic trajectories that are given as output from an epidemic model?



Energy Score

Based on the concepts of energy statistics and energy distance, which describe the similarity of distributions by measuring the distance between statistical objects.

Strictly proper, negatively-oriented score that is a multivariate generalization of the continuous ranked probability score (CRPS).

Assesses forecasts based on their sharpness and calibration.

$$\mathrm{ES}(P, y) = E_P||X^{(i)} - y|| - \frac{1}{2}E_P||X^{(i)} - X^{(j)}||$$

P = model prediction, y = observed data, $X^{(i)}$ = trajectory drawn from P, $||\cdot||$ = Euclidean norm, E_P = expected value

Gneiting et al. 2007. Székely et al. 2013. Székely et al. 2017.

Energy Score

$$\mathrm{ES}(P, y) = E_P||X^{(i)} - y|| - \frac{1}{2}E_P||X^{(i)} - X^{(j)}||$$

Assume all trajectories are equally weighted and expand norm:

$$\mathrm{ES}(P, y) = \underbrace{\boxed{\frac{1}{N}\sum_{i=1}^{N}\sqrt{\sum_{t=1}^{M}\left(x_t^{(i)} - y_t\right)^2}}}_{\text{calibration}} - \underbrace{\boxed{\frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sqrt{\sum_{t=1}^{M}\left(x_t^{(i)} - x_t^{(j)}\right)^2}}}_{\text{sharpness}}$$

P = model prediction, y = observed data, N = # trajectories, M = length of trajectory (# weeks predicted)

Normalized Energy Score

Like the WIS, the energy score is an absolute measure, so we must normalize it if we want to compare scores across locations, outcome targets, time periods, etc.

Use the absolute percentage error instead of the typical absolute error to create a relative measure.

$$\mathrm{ES}_{\mathrm{norm}}(P, y) = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\sum_{t=1}^{M}\left(\frac{x_t^{(i)} - y_t}{y_t}\right)^2} - \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sqrt{\sum_{t=1}^{M}\left(\frac{x_t^{(i)} - x_t^{(j)}}{y_t}\right)^2}$$

Energy Score for Multiple Targets

Can be used as a performance measure across multiple target outcomes (ex. cases/deaths/hospitalizations, age groups, etc.).

Gives a comprehensive understanding of a model's performance with respect to all of its predictions instead of a sum or average.

Have a matrix of predictions instead of a vector.

Trajectories from different targets must be paired.

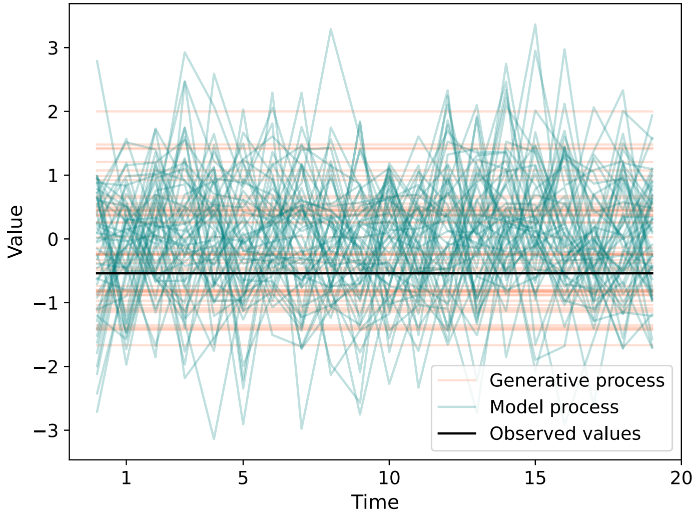
$$\mathrm{ES}_{\mathrm{mt}} = \frac{1}{N}\sum_{i=1}^{N}\left|\left|\frac{A^{(i)}-y}{y}\right|\right|_F - \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|\left|\frac{A^{(i)}-A^{(j)}}{y}\right|\right|_F$$

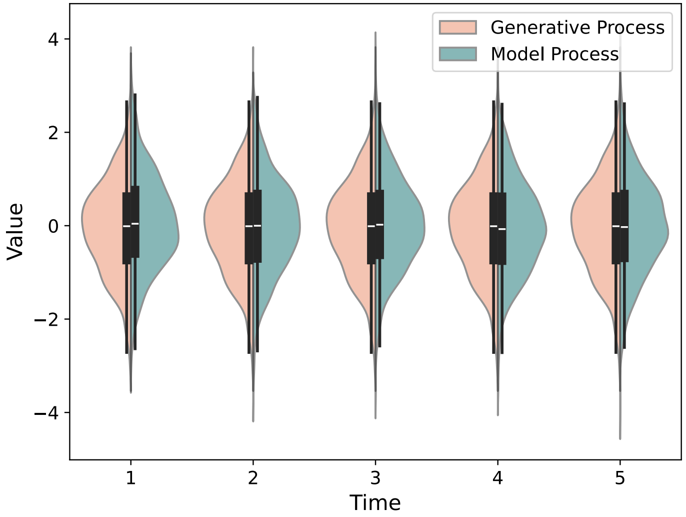
A= matrix of trajectories for each target, y = observed data, N = # trajectories, $||\cdot||_F$ = Frobenius norm

Energy Score vs. WIS

Energy score and WIS both evaluate model projections based on their calibration and sharpness.

Energy score is strictly proper. WIS is proper, but **not** strictly proper.

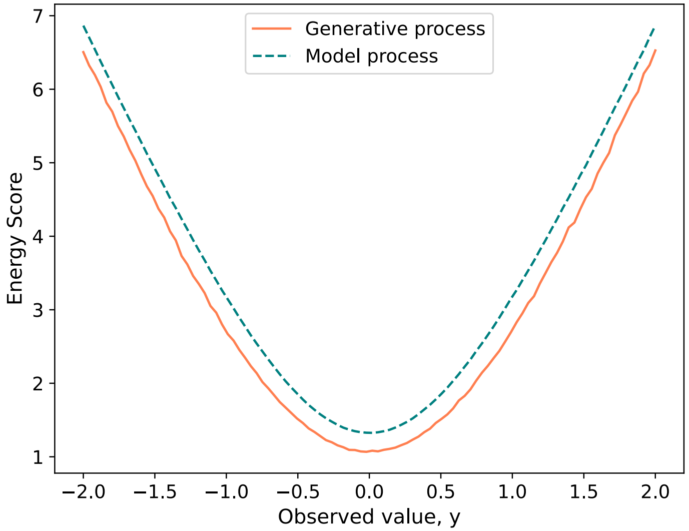


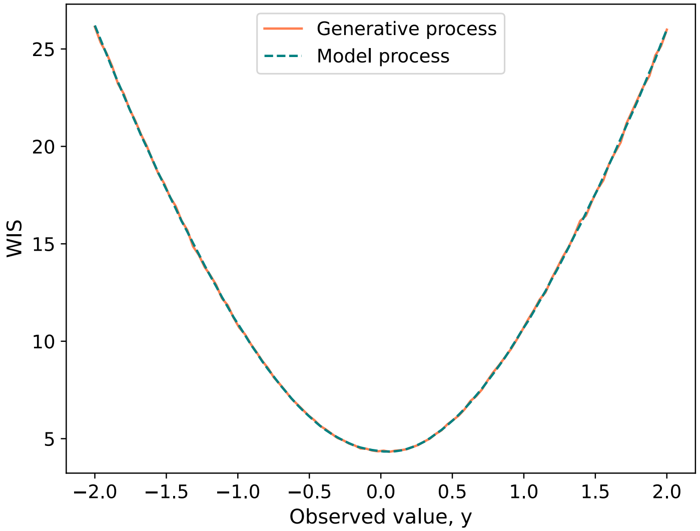


Energy Score vs. WIS

Energy score and WIS both evaluate model projections based on their calibration and sharpness.

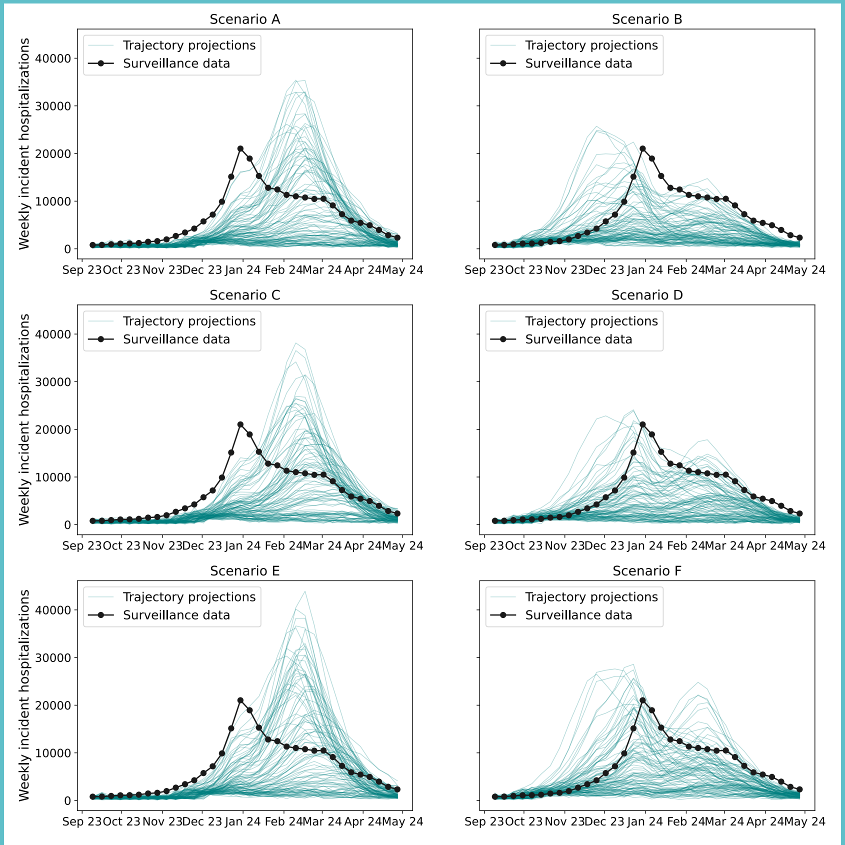
Energy score is strictly proper. WIS is proper, but **not** strictly proper.





The energy score is able to distinguish between the two stochastic trajectory processes, meaning it is strictly proper.

Case Study



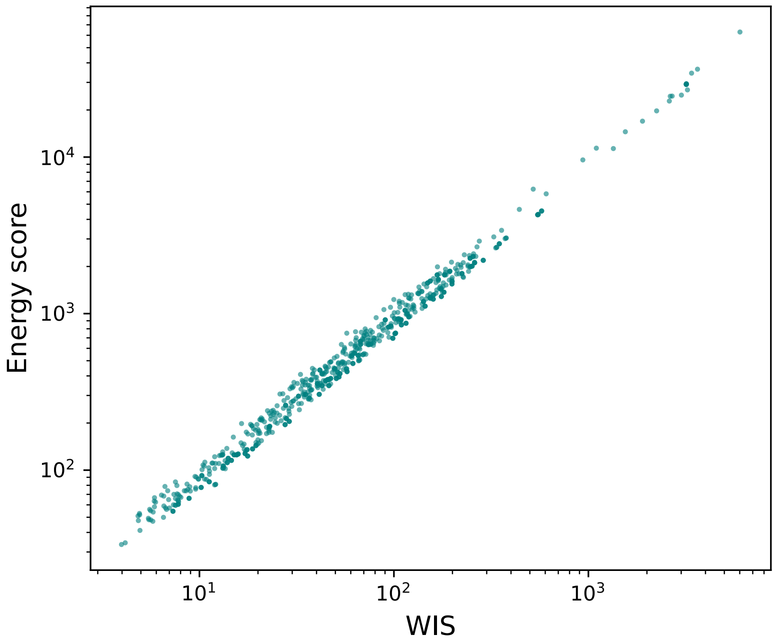
MOBS_NEU-GLEAM_FLU model

- Flu Scenario Modeling Hub 2023-24 Projection Round
- Modeling teams are now required to submit 100 trajectories.
- 6 scenarios:
    - Vaccine coverage levels (high, normal, low).
    - Dominant circulating strain (A/H3N2, A/H1N1).
- September 3, 2023 - June 1, 2024
- Hospitalization predictions.

Score Comparison

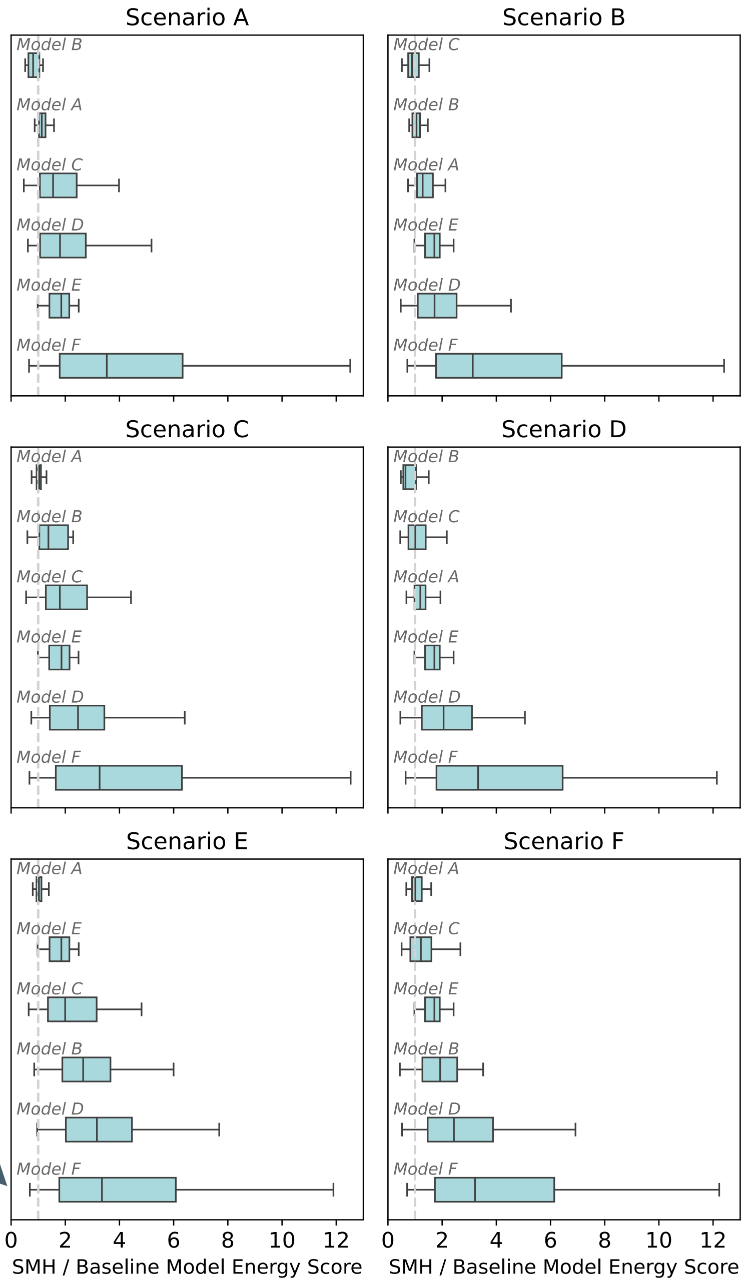
High correlation between the energy score and WIS.

A prediction that performs well under the WIS is likely to also be scored well by the energy score.



Model Ranking

The energy score can be used to compare performance across multiple models.

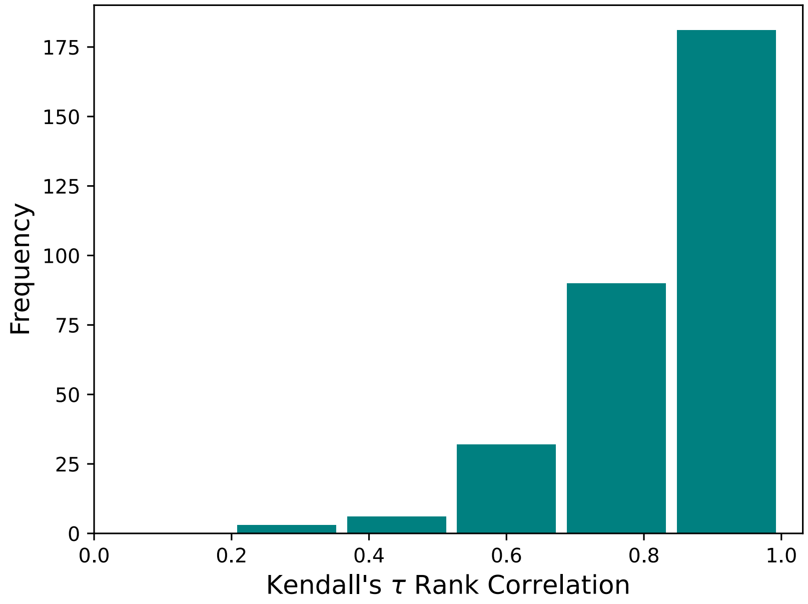


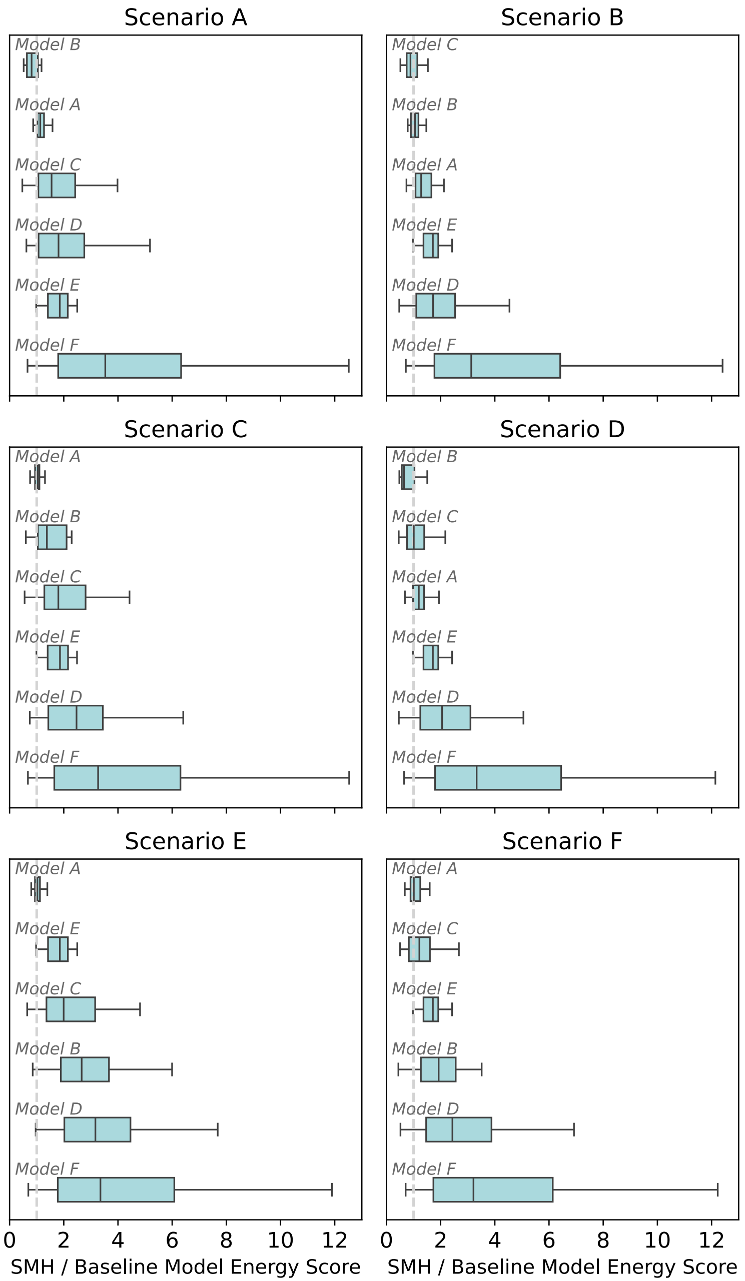
Distribution of energy score ratio across all locations (US states) for each model and scenario.

Model Ranking

Compare with the model ranking of the WIS through Kendall's τ rank correlation.

Energy score and WIS rank models very similarly, but not identically.

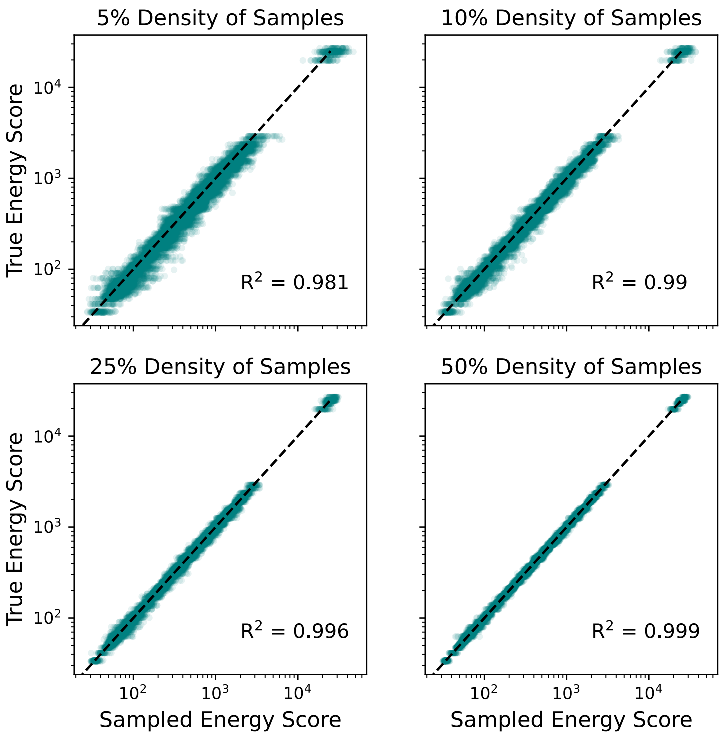


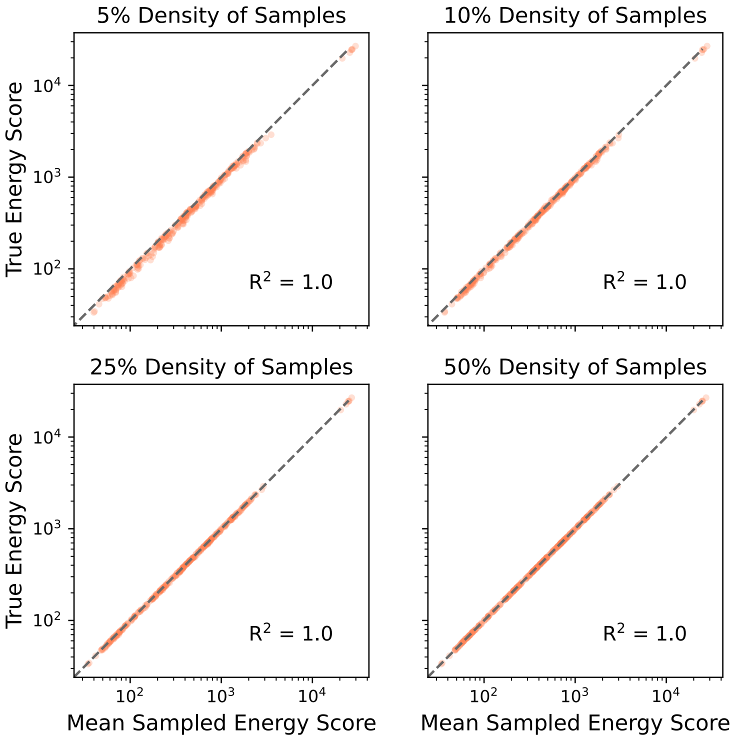


Energy Score Sampling

Limitation: pairwise comparison between all pairs of trajectories in the energy score is computationally challenging for large numbers of trajectories.

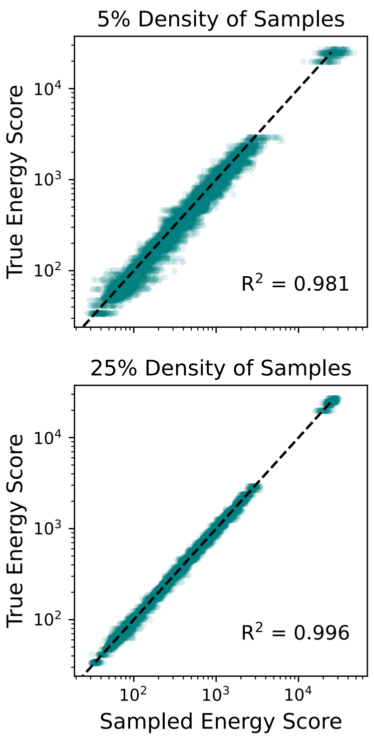
Randomly sample *n* trajectories, calculate the energy score, repeat for 50 iterations and compare with score found by using all trajectories.

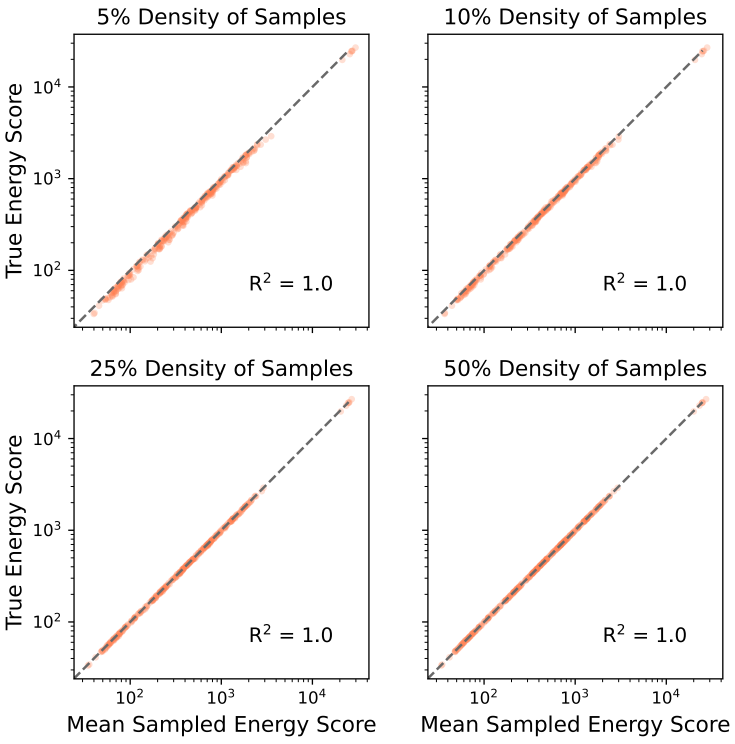




Energy Score Sampling

By taking a small sample of trajectories, we can find a close estimate of the true energy score.





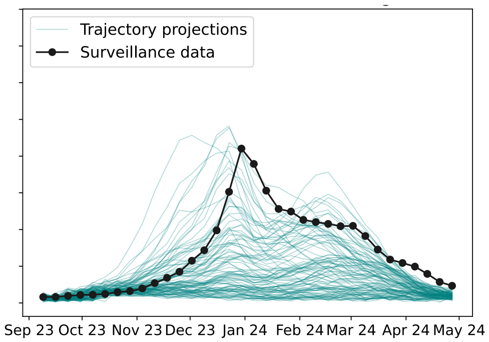
Conclusion

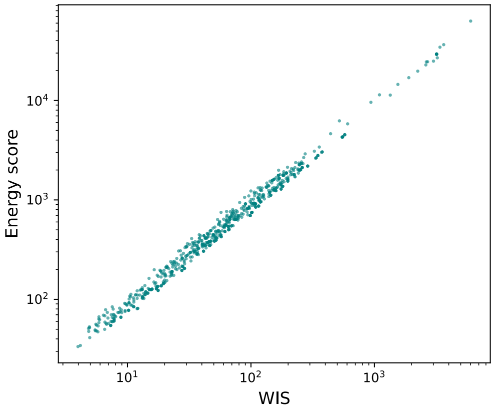
The energy score is a strictly proper scoring rule that evaluates model predictions in a stochastic trajectory format.

Analyzes model projections similarly to the WIS.

Can be extended to changing weights across trajectories based on performance.

Leads to new ensemble techniques that use the energy score to choose which trajectories to include in an ensemble model.


Trajectory projections
Surveillance data
Sep 23
Oct 23
Nov 23
Dec 23
Jan 24
Feb 24
Mar 24
Apr 24
May 24


Energy score
10^2
10^3
10^4
10^1
10^2
10^3
WIS