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

Clara Bay[1], Guillaume St-Onge[1], Jessica Davis[1], Matteo Chinazzi[1,2], and Alessandro Vespignani[1]

[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

Introduction

Scenario modeling plays a pivotal role in guiding decisions during pandemics and epidemics. Scenario models are aimed at exploring possible futures of an epidemic, which introduces additional complexities in their evaluation [1]. Typically, projections statistically aggregate the individual stochastic trajectories derived by sampling the parameter space. Evaluation metrics tailored for quantile interval projections, such as the weighted interval score (WIS) have found widespread use in infectious disease performance assessment [2].

We employ the energy score to evaluate the performance of scenario model projections reported in a trajectory format. The energy score has been applied to a variety of fields to analyze weather [3], electricity market price [4], and wind power generation [5], but it is not an extensively used metric for performance analysis in epidemic forecasting and prediction. Here, we describe the utility and significance of the energy score for epidemic prediction, with examples from scenario modeling.

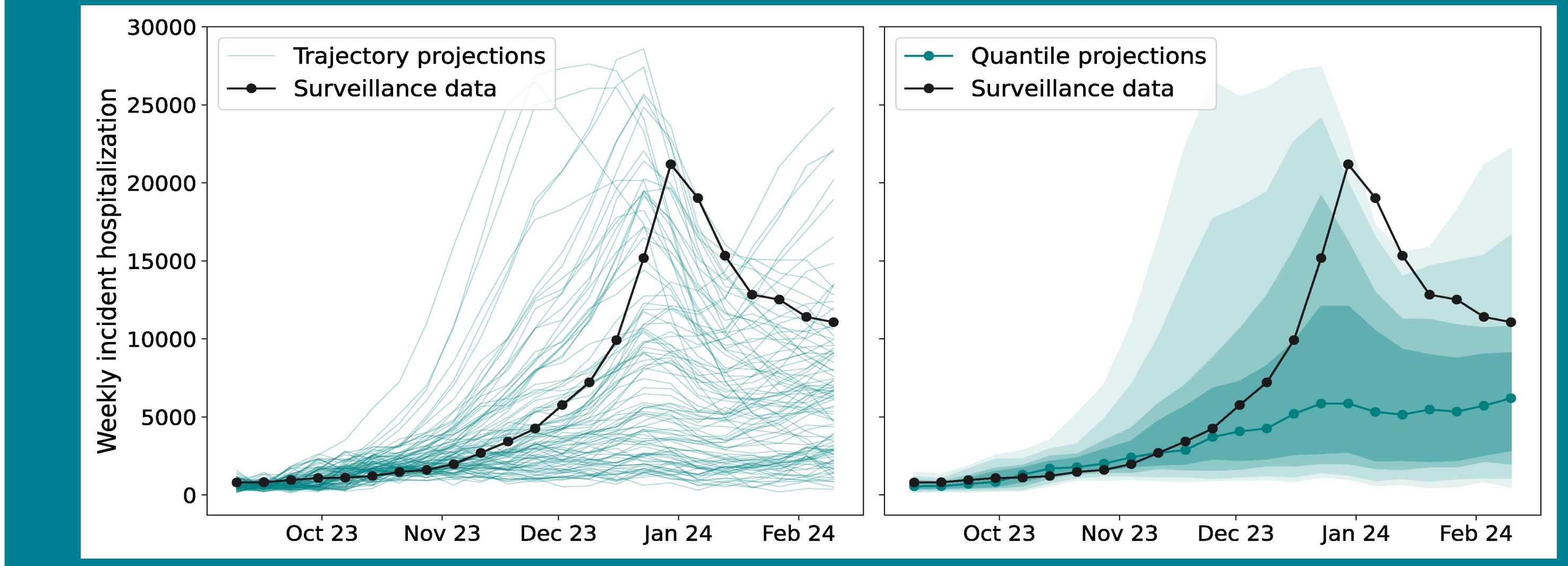

Figure 1: Comparison of epidemic predictions in the trajectory (left) versus quantile (right) format. (left) One hundred trajectories for a single model for scenario F in the Flu Scenario Modeling Hub 2023-24 round 1 (blue) with observed surveillance data (black) nationally in the United States. (right) The corresponding quantiles showing the median (blue line), 50%, 70%, 90%, and 98% prediction intervals (shaded) with the observed data (black).

The Energy Score

The energy score is a negatively-oriented proper score that is a multivariate generalization of the continuous ranked probability score (CRPS) [3,6]. The idea for the energy score is based on the concepts of energy statistics, which are functions of distances between statistical observations, and energy distance, which can describe the similarity of distributions [7,8]. The energy score accounts for both calibration and sharpness by comparing both the distance between individual trajectories and the observed data as well as the distance between all pairs of trajectories.

For a multivariate distribution *P*, where **X**$^{(i)}$ and **X**$^{(i)}$ are vectors of independent random variables drawn from *P*, and **y** is the vector of true (observed) values, the energy score is defined as:

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

If we assume that all trajectories are equally weighted with weight 1/N for N trajectories, we can expand the Euclidean norm to rewrite the energy score as:

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

where M is the number of elements in each trajectory, $x_t^{(i)}$ is the predicted value specified by a single model trajectory *i* at time *t*, and y_t is the value of the surveillance data at time *t*.

The energy score is an absolute measure, so if we want to compare scores across locations, or time periods with dissimilar surveillance magnitudes, we perform standardization by using the absolute percentage error instead of the typical absolute error in our energy score calculation.

The energy score and WIS evaluate models similarly.

We illustrate the application of the energy score to epidemic scenario modeling projections for 2023-24 round 1 of the Flu Scenario Modeling Hub (SMH). The implications of vaccine coverage levels and the dominant circulating strain was analyzed in this round through 6 scenarios, with 10 modeling teams reporting projections from September 3, 2023 to June 1, 2024 [9].

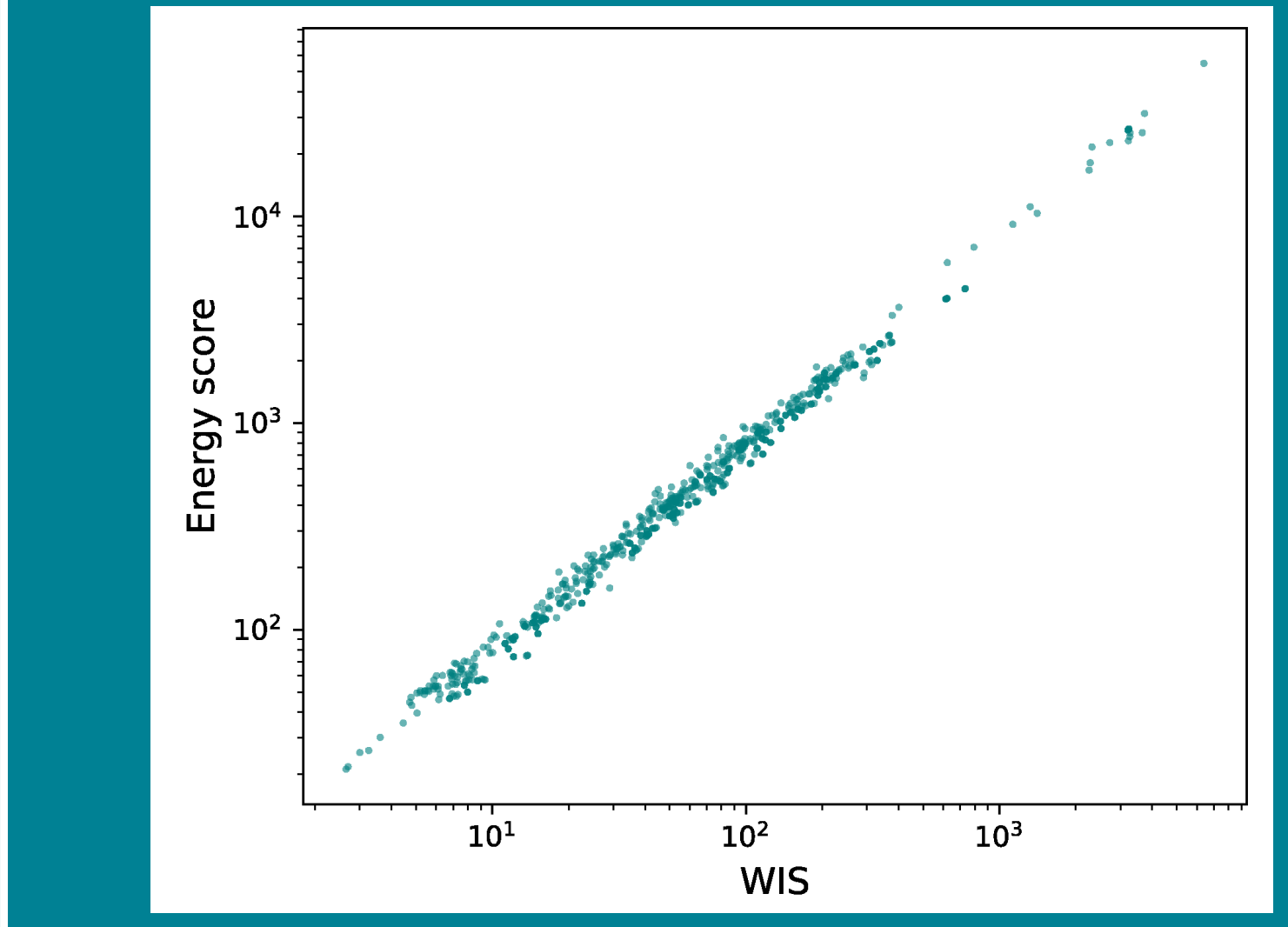

Figure 2: Relationship between the weighted interval score and interval score for models in 2023-24 Round 1 of the Flu Scenario Modeling Hub that reported both trajectories and quantiles.

In Fig. 2, we show that the energy score has a strong correlation with WIS. Projections that are scored well by the WIS are likely to be scored well by the energy score.

We rank the performance of each individual model at each location and scenario in Fig. 3 where models with a higher rank have better performance. By comparing the energy score and WIS, we find that while there are differences between the rankings of both methods, there are similarities in how the energy score and WIS evaluate model performance.

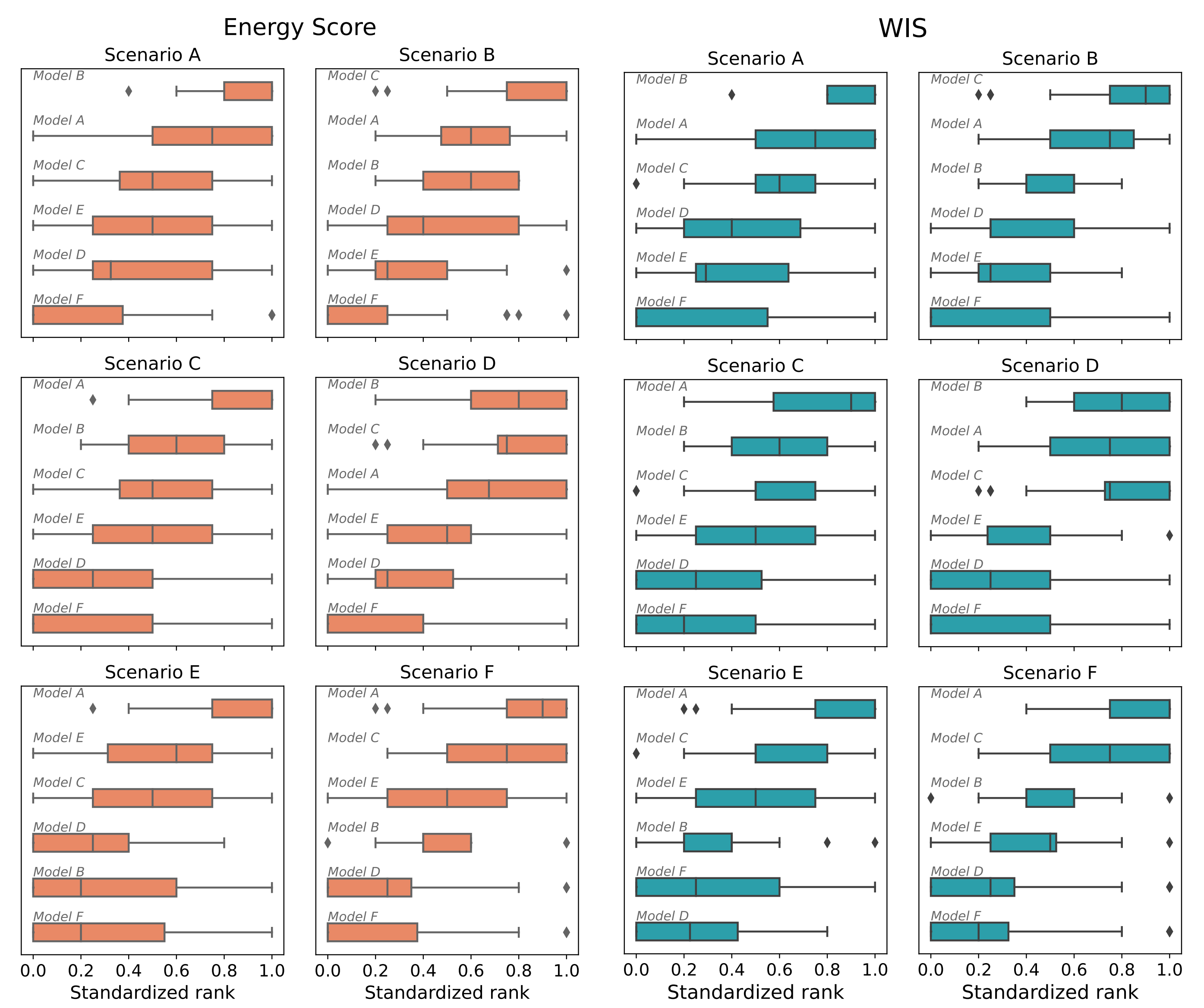

Figure 3: Distribution of the standard rank for each individual model in 2023-24 Round 1 of the Flu Scenario Modeling Hub for the (left) energy score and (right) WIS calculated by deriving quantiles from the reported trajectories. The standardized rank is calculated such that a higher rank corresponds to a better-performing model.

The energy score is strictly proper.

For two models with the same marginal distribution but different generative processes, the WIS will calculate the same score, but the energy score will not.

The energy score rewards being bold as opposed to conservative in predictions. This incentivizes honest forecasts because only a prediction identical to the truth will have the best score [6].

Trajectory Ensemble Method

We propose an alternative method of generating an ensemble model that utilizes the trajectories reported by each modeling team as opposed to transforming them into quantiles. To do this, we simply bundle all the trajectories for each model together and assign equal weight to each one. This group of trajectories is the resulting ensemble that we call the trajectory ensemble. In Fig. 4, we compare the trajectory ensemble with the 3 ensemble models reported by the SMH, and we find that while it does not exactly match any of the SMH models, it most closely resembles the level of uncertainty given by the Ensemble_LOP_untrimmed model.

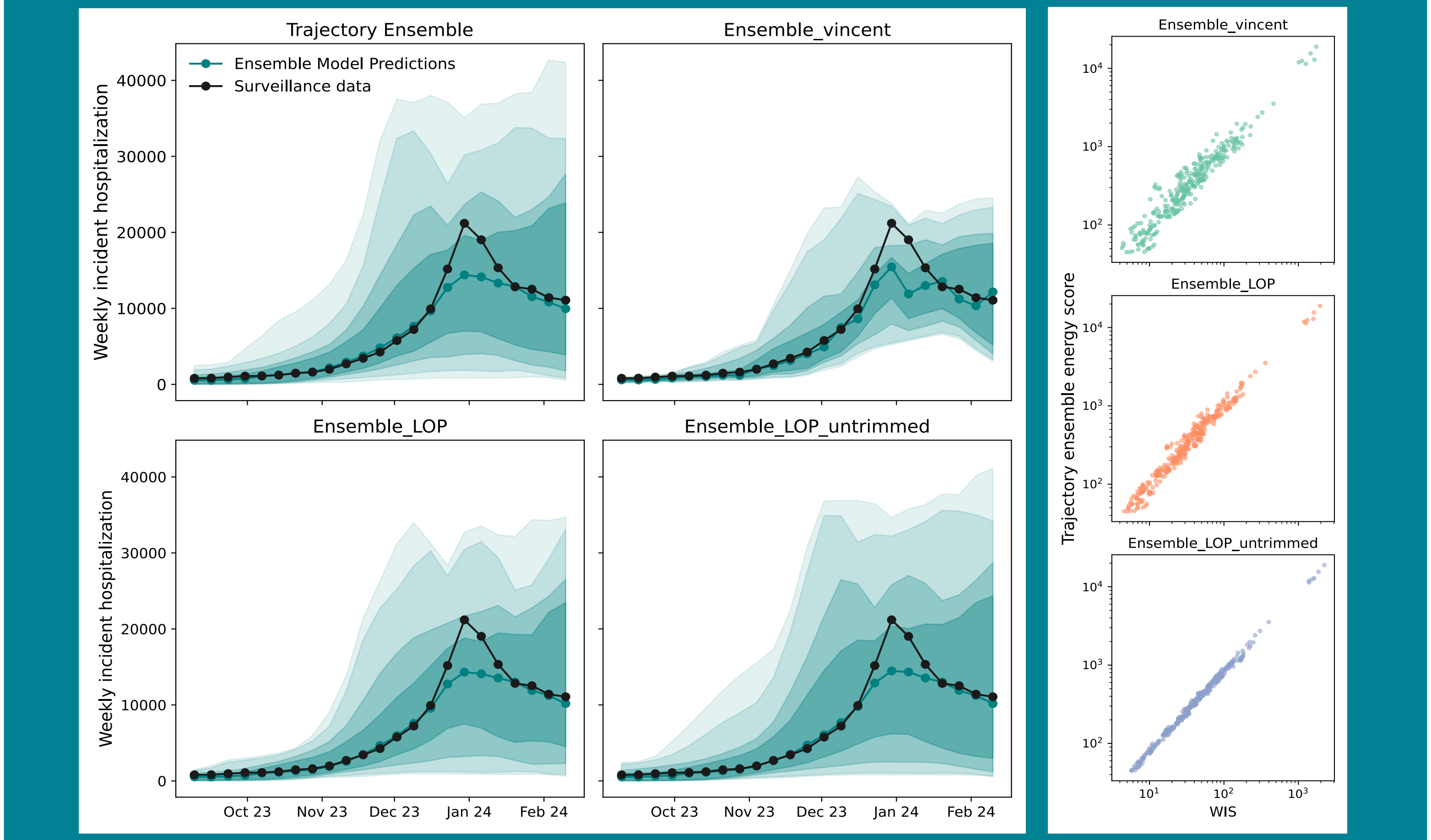

Figure 4: (left panel) Quantile model projections for the trajectory ensemble method and the three ensemble models reported by the SMH for scenario F in the Flu Scenario Modeling Hub 2023-24 round 1 (blue) with observed surveillance data (black) nationally in the United States. (right panel) Relationship between the WIS scores of the three SMH-reported ensemble models and the energy score of the trajectory ensemble.

Discussion

The energy score is a measure that uses the output of stochastic epidemic model simulations to evaluate model performance without transforming or summarizing the results into quantile format.

We show the utility of the energy score in scenario modeling performance analysis and how it evaluates models similarly. A benefit of this method is that the energy score is strictly proper, but computational cost is a potential limitation due to the pairwise comparison between all trajectories. However, we find that sampling trajectories provides similar results.

We highlight how the energy score enables a streamlined ensemble process, permitting the weighting of different models and even specific trajectories as more data is accumulated.

References

[1] Howerton, E., Contamin, L., Mullany, L.C. et al. Evaluation of the US COVID-19 Scenario Modeling Hub for informing pandemic response under uncertainty. Nat Commun 14, 7260 (2023). https://doi.org/10.1038/s41467-023-42680-x. **[2]** Bracher, J., Ray, E.L., Gneiting, T., Reich, N.G., 2021. Evaluating epidemic forecasts in an interval format. PLOS Computational Biology 17, 1–15. URL: https://doi.org/10.1371/journal.pcbi.1008618, doi:10.1371/journal.pcbi.1008618. **[3]** Gneiting, T., Stanberry, L.I., Grimit, E.P., Held, L., Johnson, N.A., 2008. Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. TEST 17, 211–235. doi:10.1007/s11749-008-0114-x. **[4]** Cramer, E., Witthaut, D., Mitsos, A., Dahmen, M., 2023. Multivariate probabilistic forecasting of intraday electricity prices using normalizing flows. Applied Energy 346, 121370. doi:https://doi.org/10.1016/j.apenergy.2023.121370. **[5]** Pinson, P., Girard, R., 2012. Evaluating the quality of scenarios of short-term wind power generation. Applied Energy 96, 12–20. doi:https://doi.org/10.1016/j.apenergy.2011.11.004. **[6]** Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102, 359–378. doi:10.1198/016214506000001437. **[7]** Szekely, G.J., Rizzo, M.L., 2017. The energy of data. Annual Review of Statistics and Its Application 4, 447–479. doi:10.1146/annurev-statistics-060116-054026. **[8]** Szekely, G.J., Rizzo, M.L., 2013. Energy statistics: A class of statistics based on distances. Journal of Statistical Planning and Inference 143, 1249–1272. doi:https://doi.org/10.1016/j.jspi.2013.03.018. **[9]** Scenario Modeling Hub, 2024. Flu Scenario Modeling Hub. https://https://fluscenariomodelinghub.org//.

MOBS LAB
Laboratory for the
Modeling of Biological +
Socio-technical Systems

netsi
Northeastern University
Network Science Institute