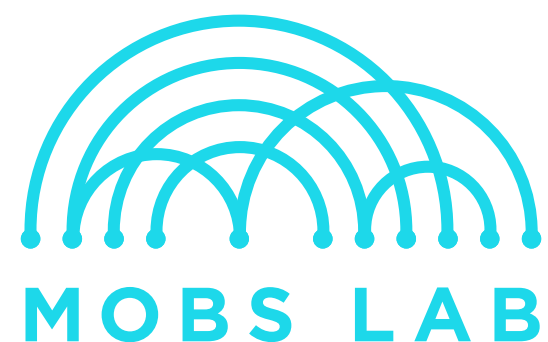




# Energy Score

Evaluating the performance of epidemic projections in stochastic trajectory format



Laboratory for the  
Modeling of Biological +  
Socio-technical Systems

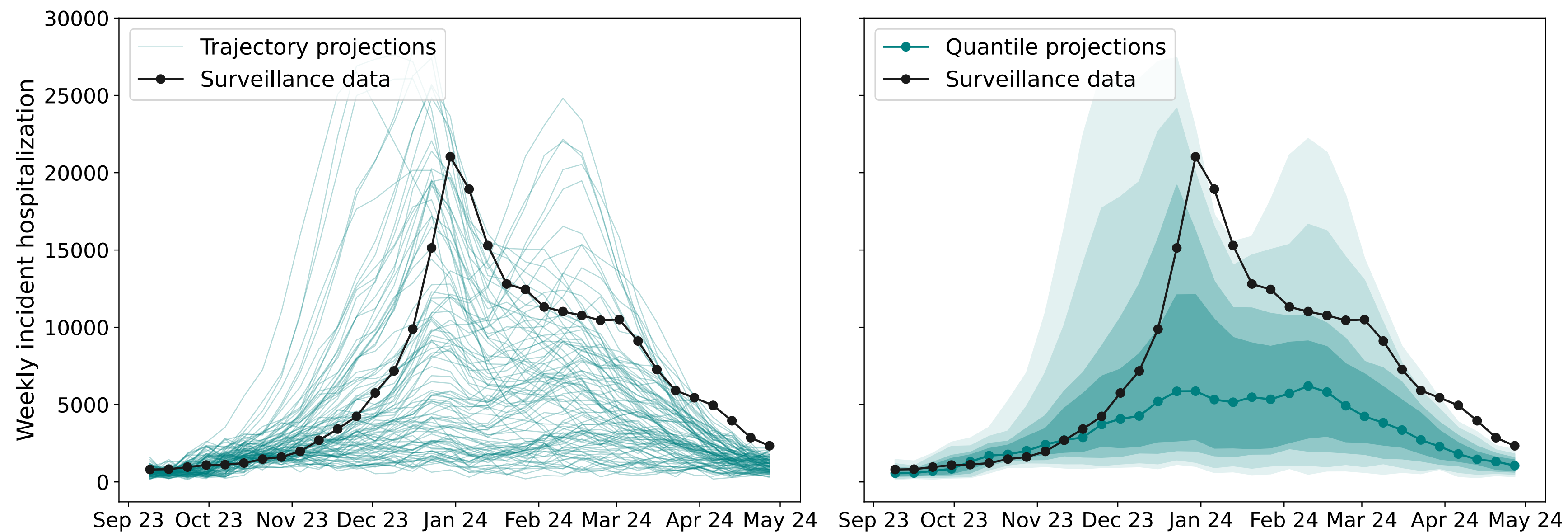
Clara Bay

[bay.c@northeastern.edu](mailto:bay.c@northeastern.edu)

# Epidemic Model Formats

Historically, epidemic predictions are reported in quantile format.

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



# 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

# 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). Gives one score for a group of time series.

Assesses forecasts based on their calibration and sharpness. Assume all trajectories are equally weighted.

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

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

# Normalized Energy Score

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

Divide energy score by the sum of surveillance data.

Keeps focus on periods with large magnitude.

$$ES_{\text{norm}} = \frac{ES(P, y)}{\sum_{m=1}^M y_m}$$



# Multi-Dimensional Energy Score


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

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



Trajectories      Targets      Weeks      Normalize by sum of surveillance data

# Weighted Interval Score (WIS)

Commonly used proper score for evaluating probabilistic predictions in quantile format.

Evaluates based on calibration and sharpness.

Computed at each time point with prediction  $\mathbf{P}$  and observed value  $\mathbf{y}$ . Take the average to get a score for a full time series.

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

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

↓  
dispersion

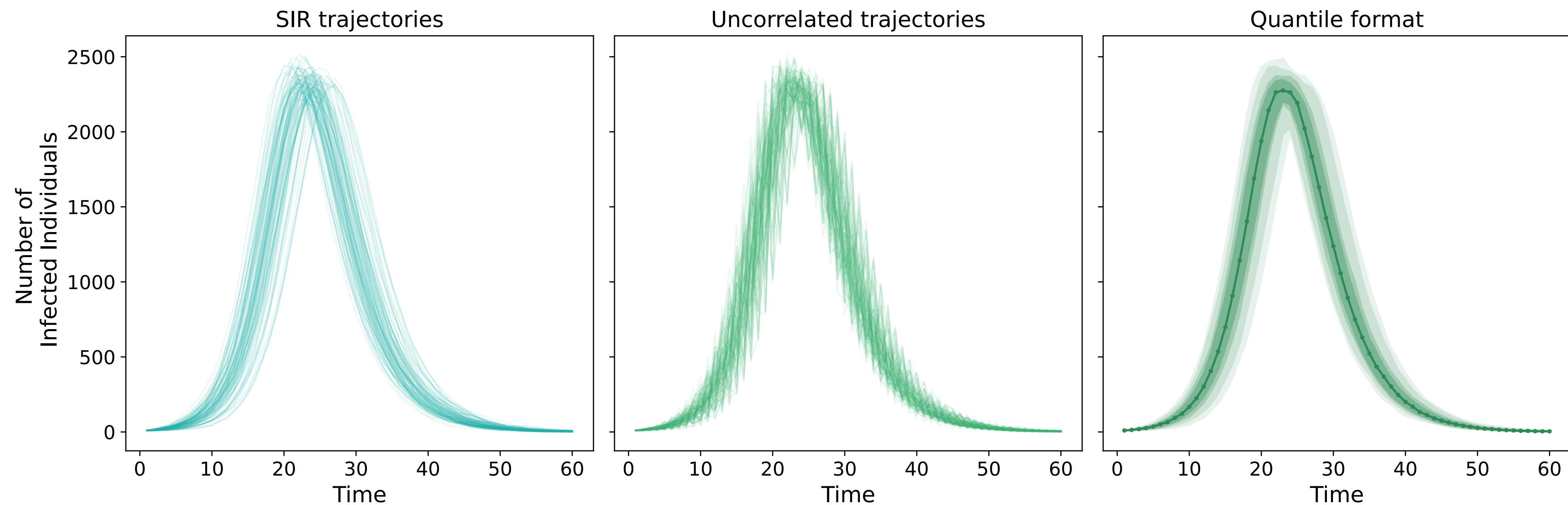
↓  
penalty for  
underprediction

↓  
penalty for  
overprediction

# Properness of Energy Score and 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.

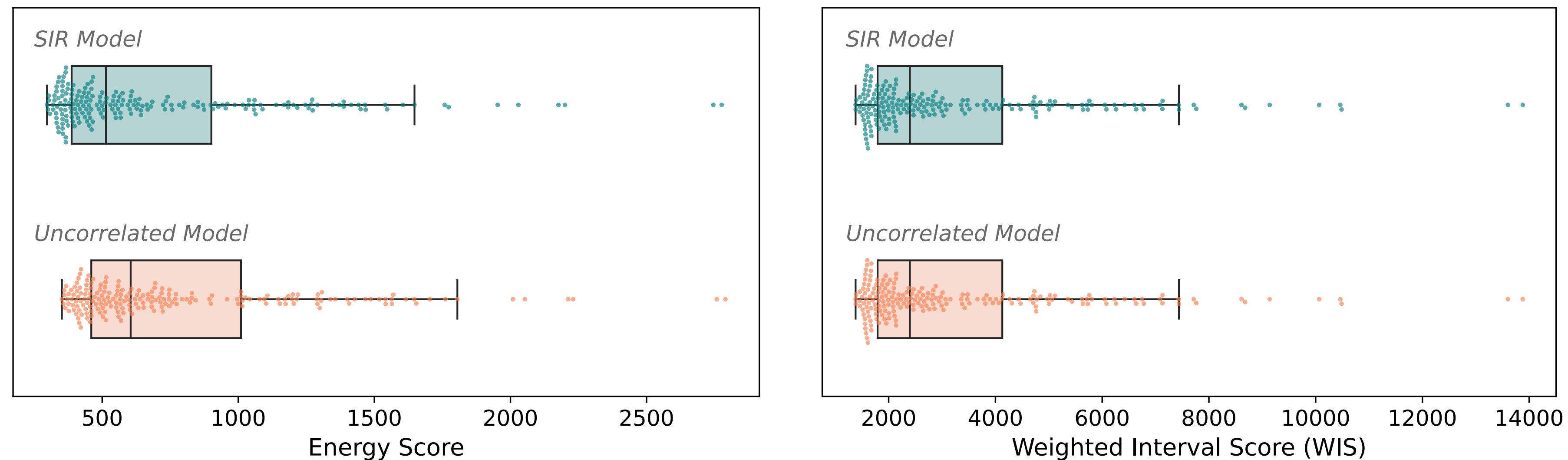




# Properness of Energy Score and WIS

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.

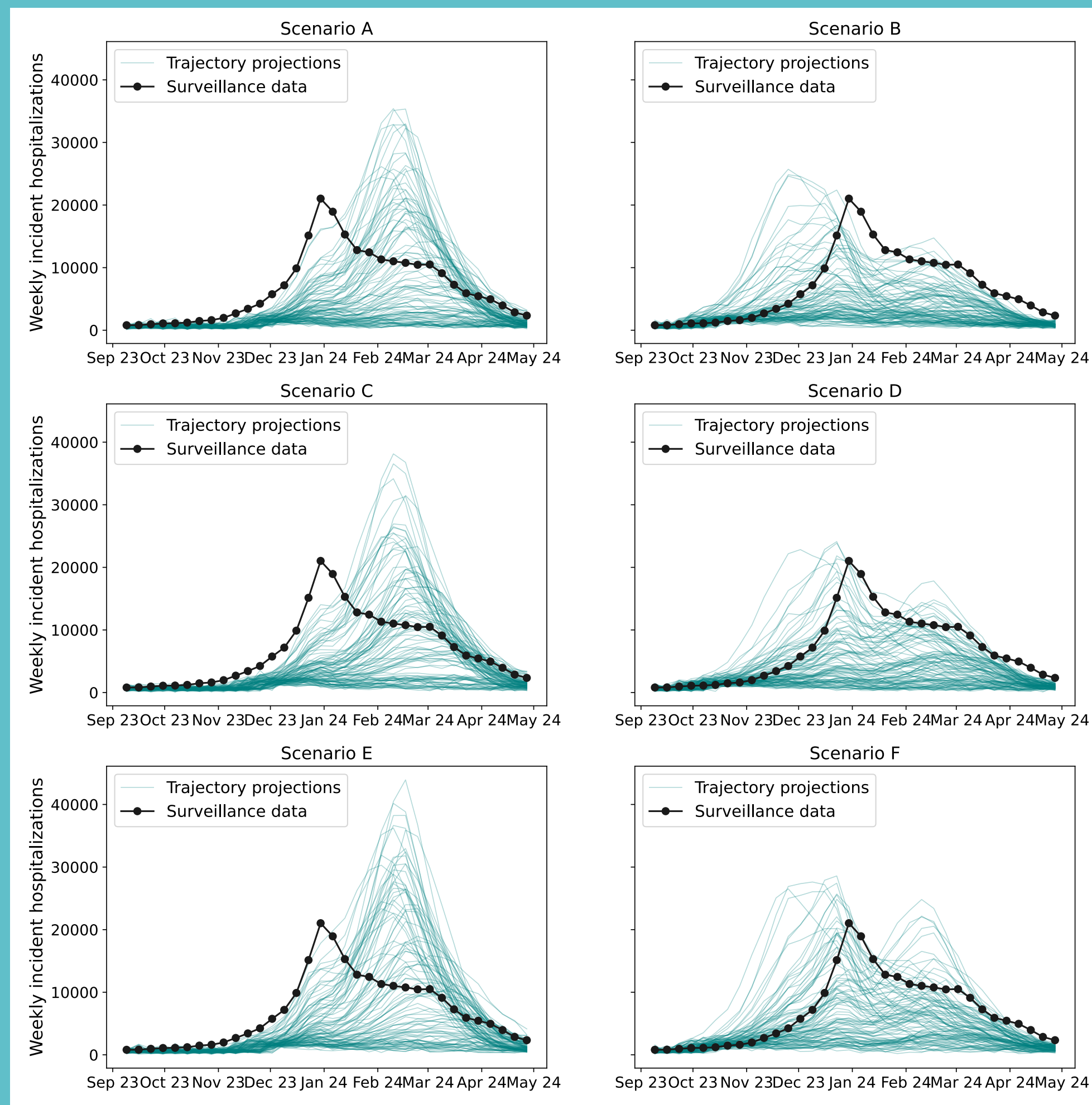


The energy score is able to distinguish between the two stochastic trajectory processes, meaning it is strictly proper.

# Case Study

## *2023-24 Flu Scenario Modeling Hub projections*

- 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.

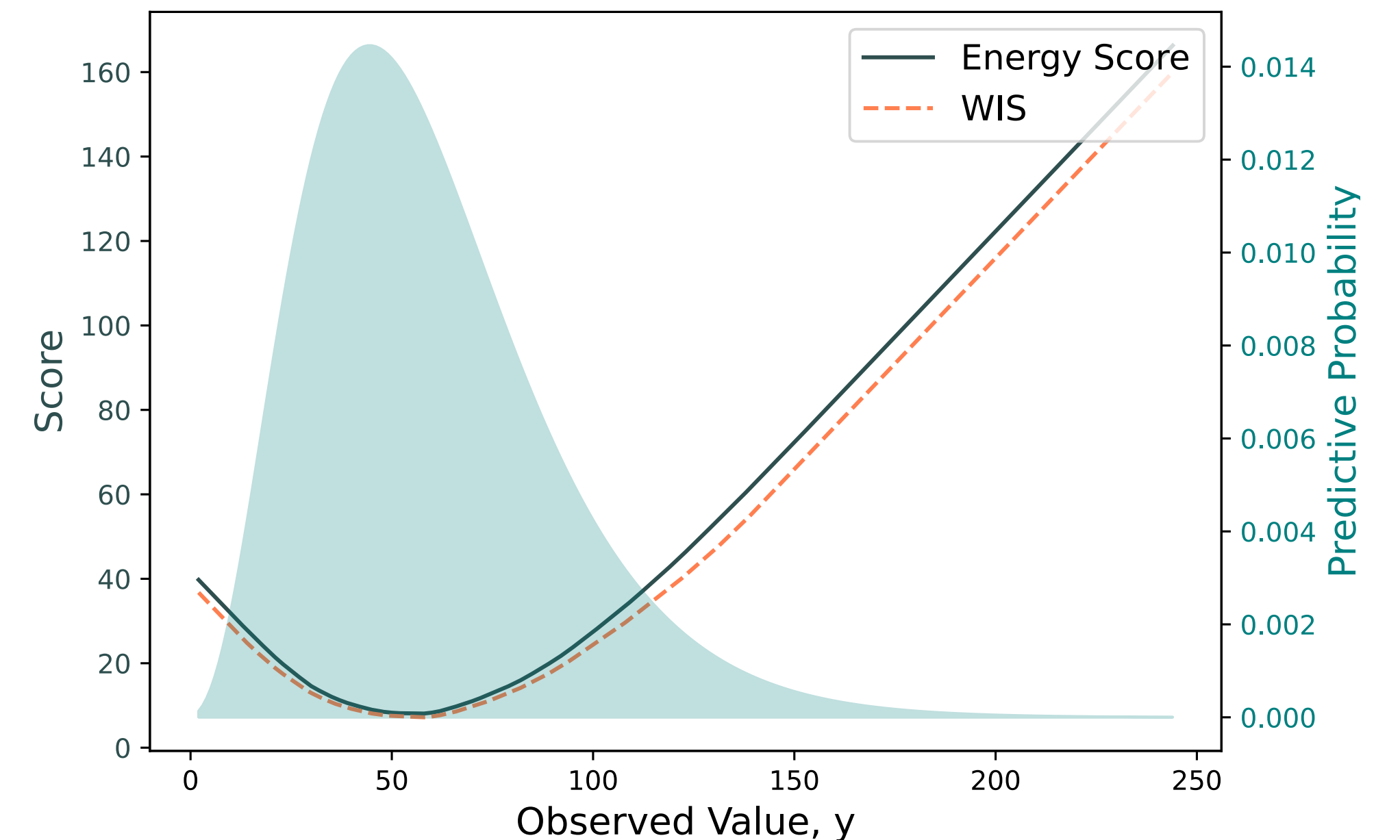
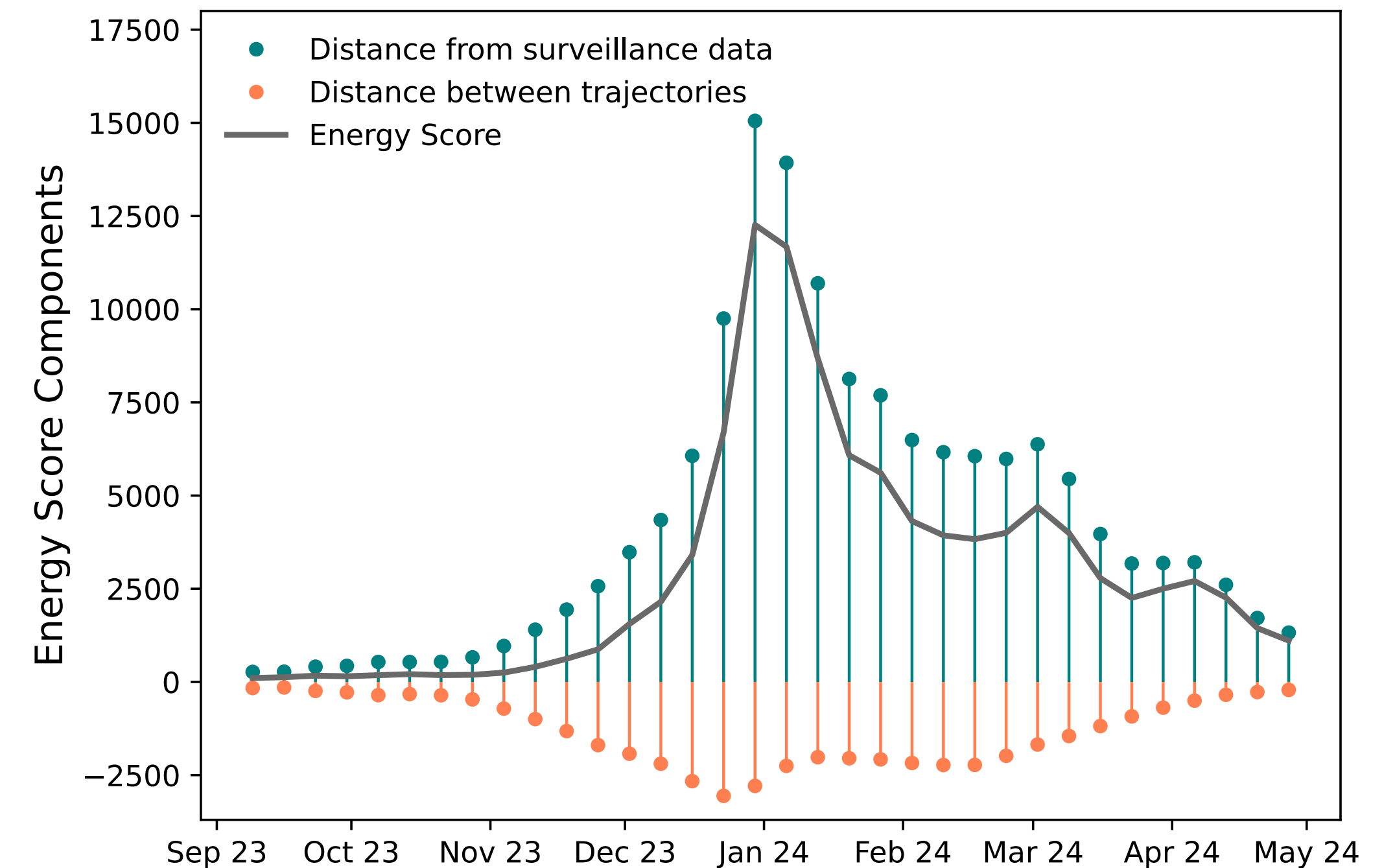


# Energy score at a single time point

Calculate energy score at each week of flu season. Trajectories are single points.

Energy score weights periods with high activity more heavily.

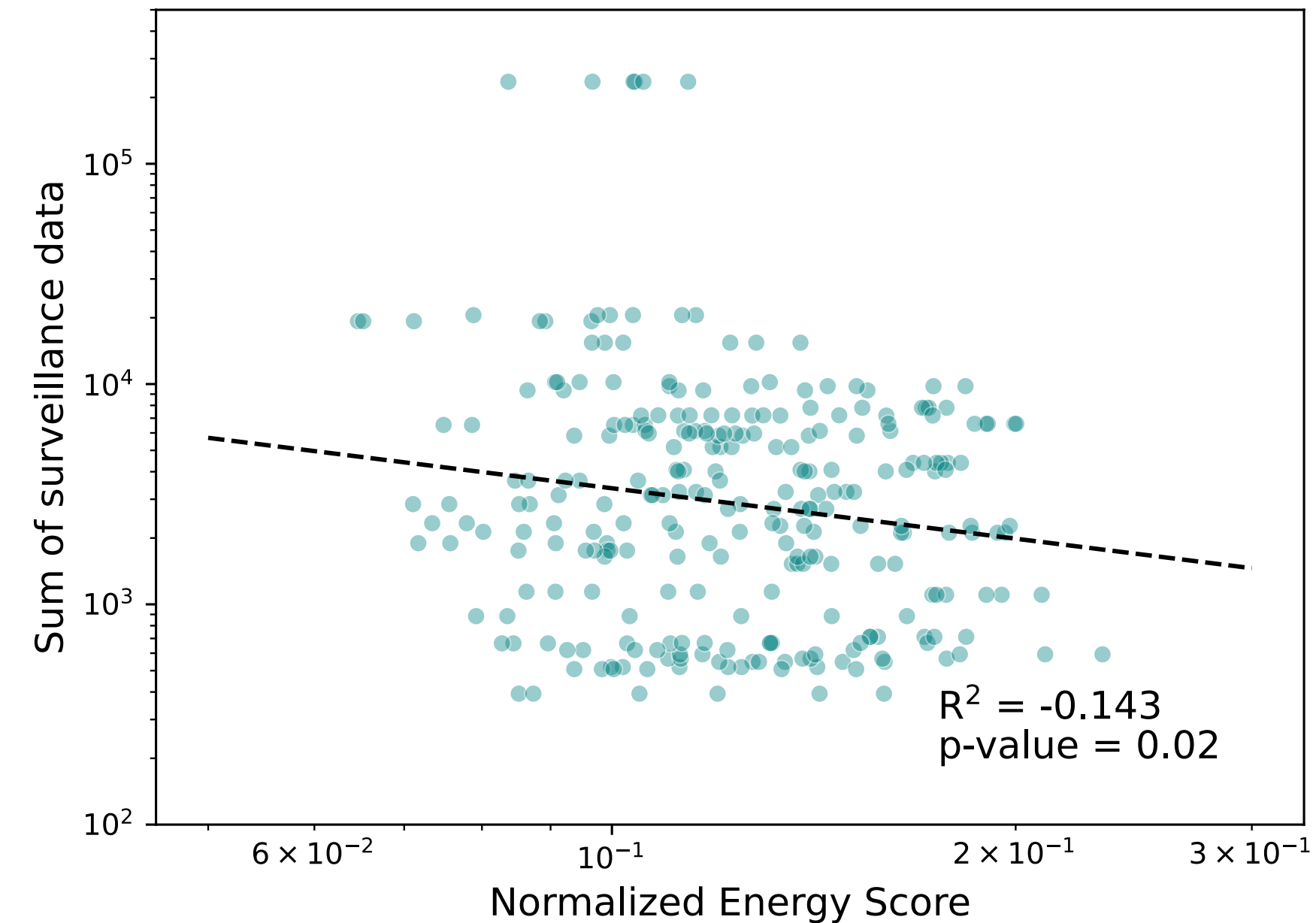
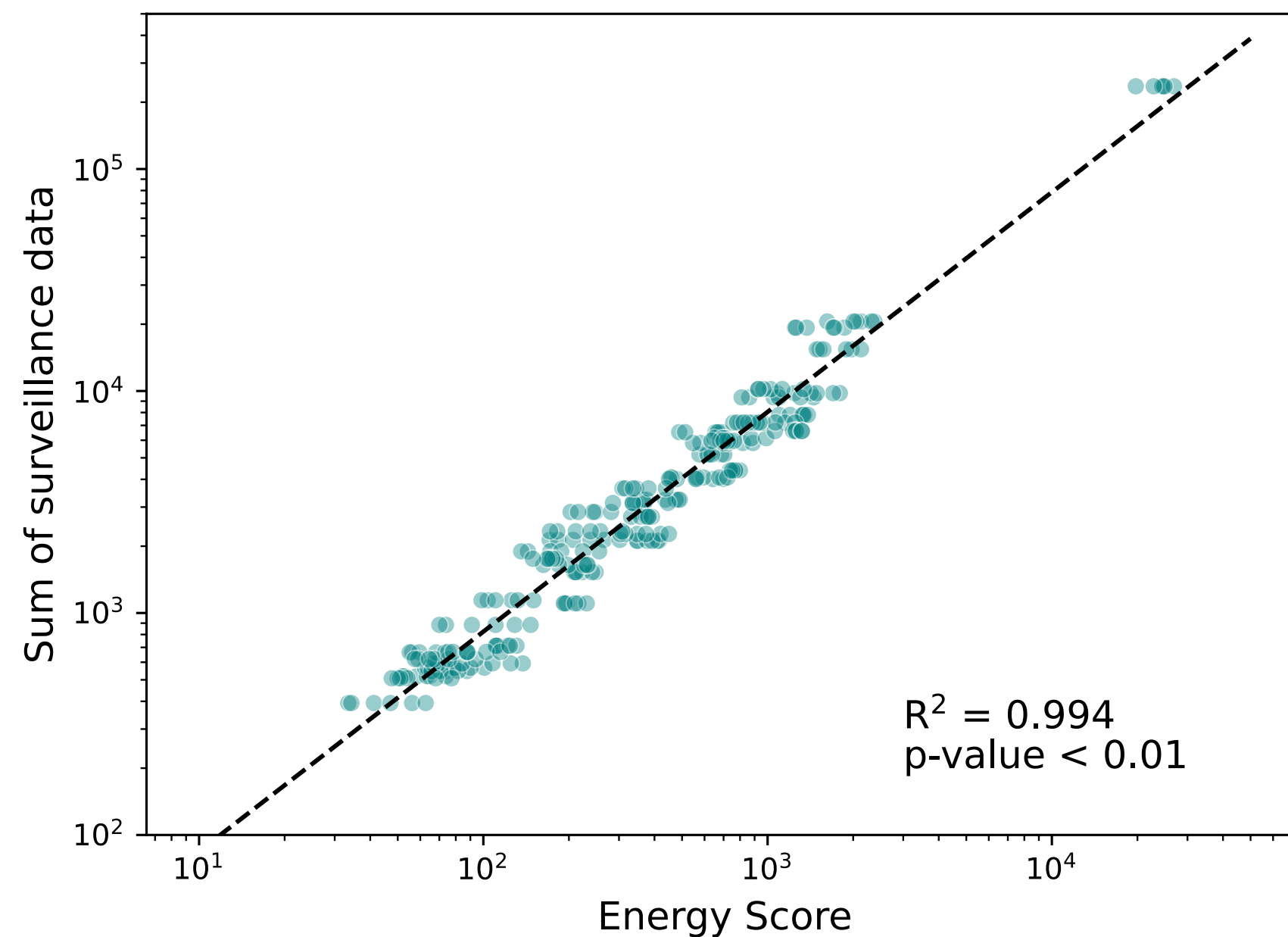
Energy score evaluates a synthetic predictive distribution similarly to the WIS at a single time point.



# Energy score vs. normalized energy score

Energy score is strongly correlated to the magnitude of the flu season, where the normalized energy score is not.

With the normalized energy score, we can compare scores across predictions with different signal magnitudes.



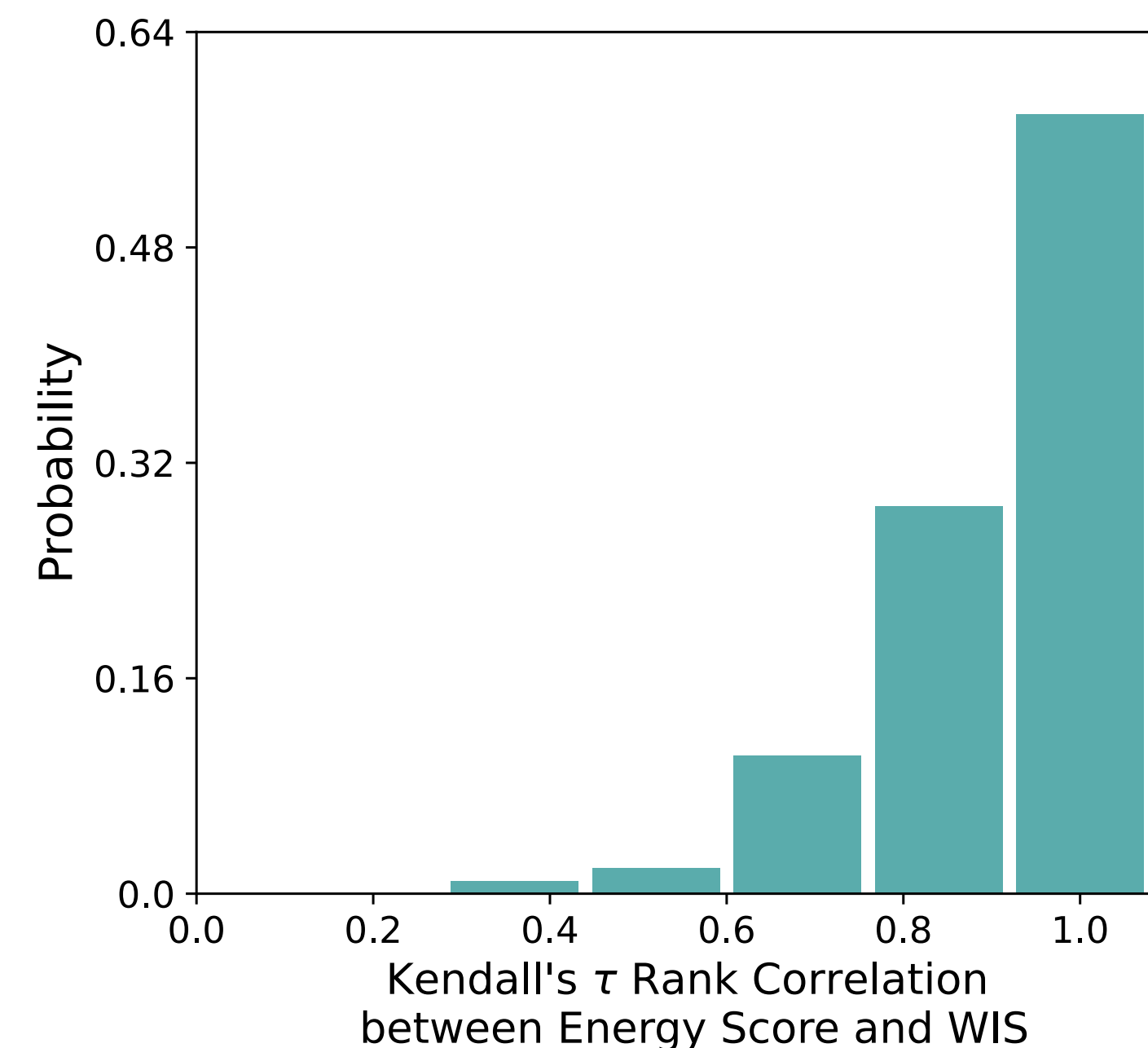
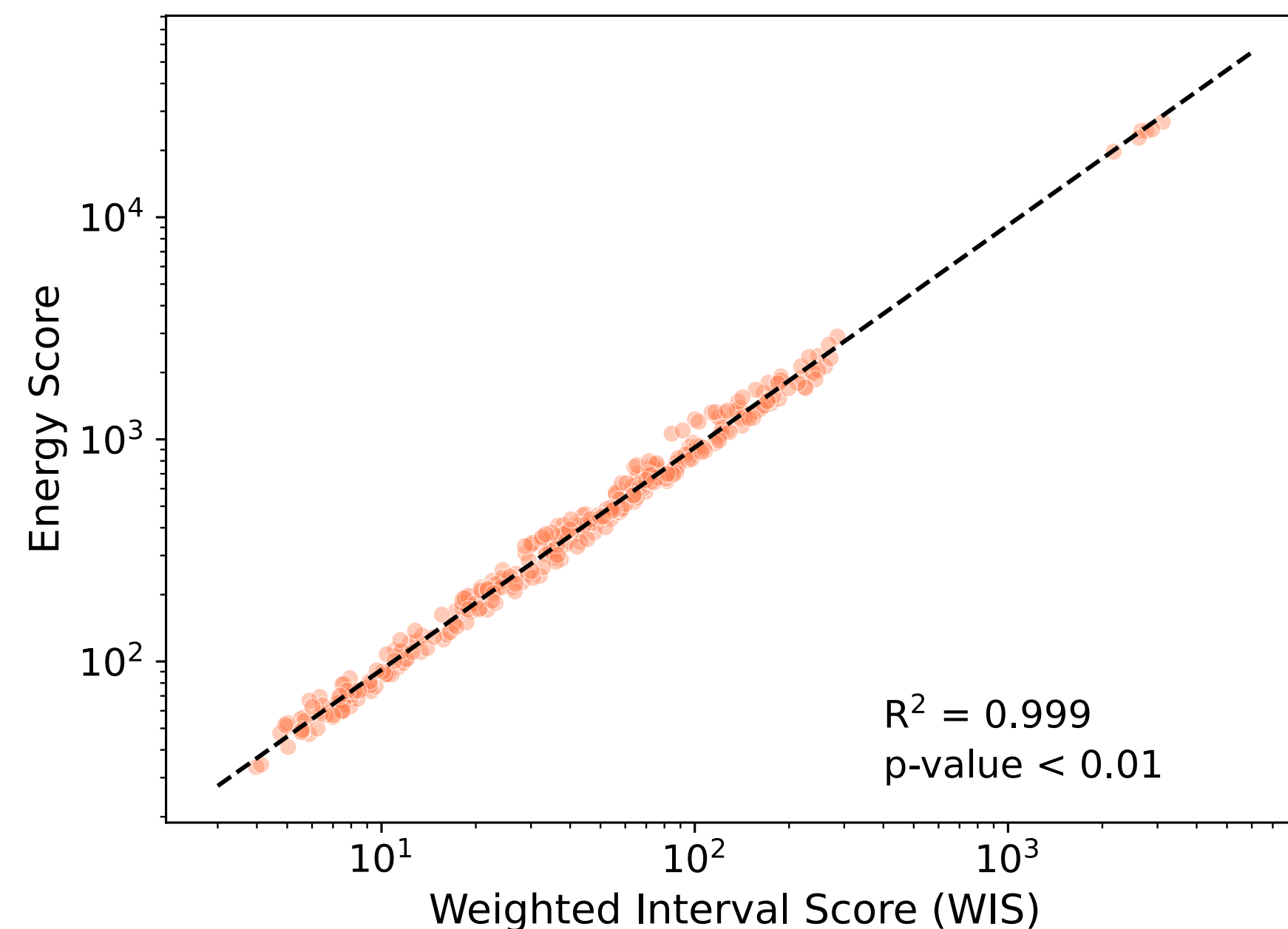


# Energy Score vs. WIS

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.

The two scores also rank models very similarly with respect to one another.

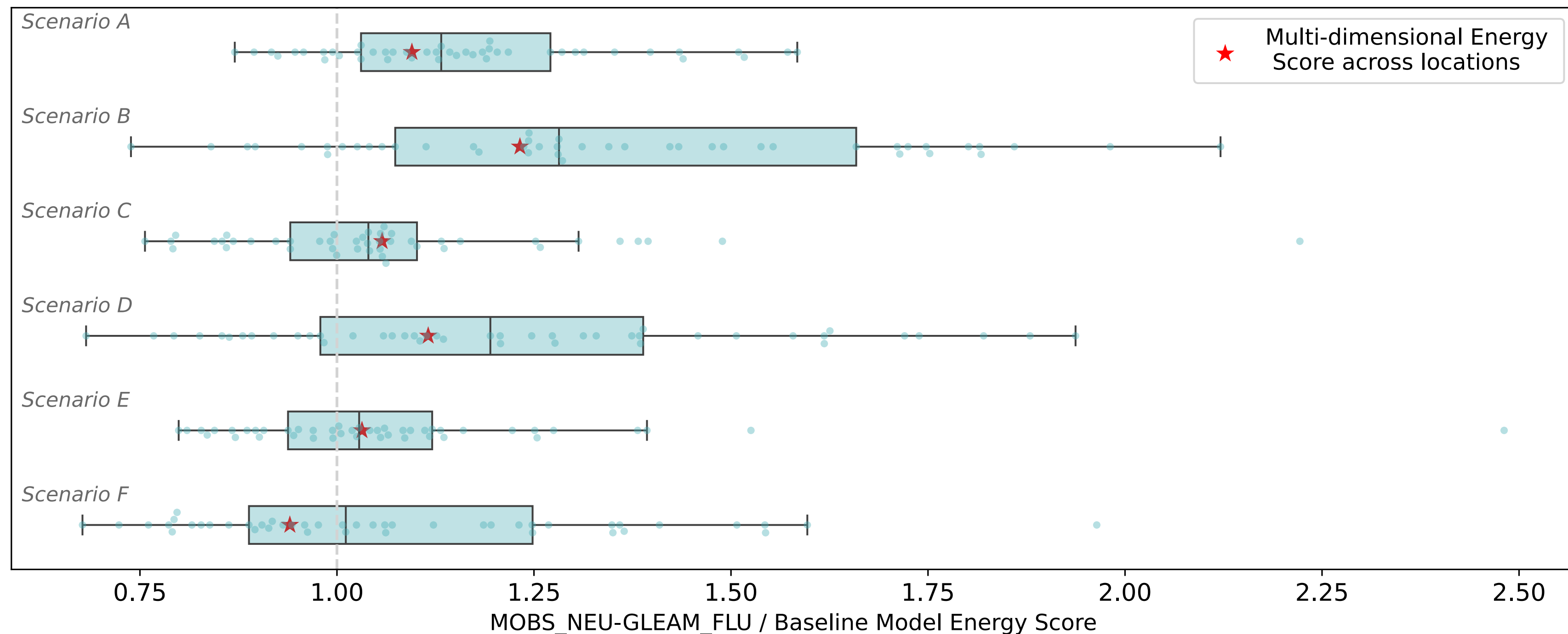




# Evaluation of MOBS model across locations

Using ratio comparing the MOBS predictions with a naive baseline model, we can assess performance across locations and scenarios.

Multi-dimensional energy score matches ranking of scenarios.



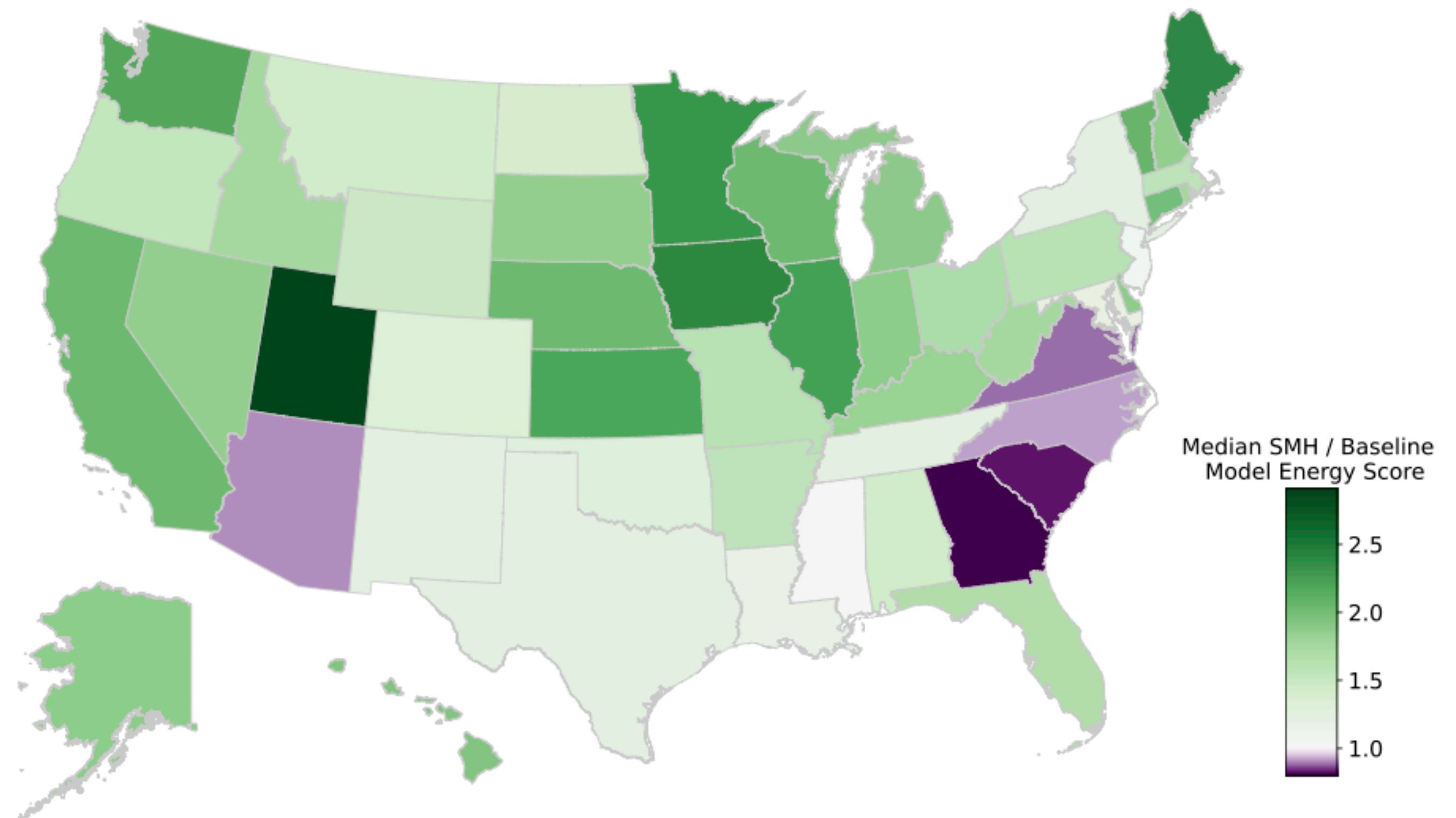
# Energy Score across locations

Median energy score ratio across SMH models for each US state.

Purple states: ratio $<1$ , SMH typically performs better than naive baseline.

Green states: ratio>1, SMH typically performs worse than naive baseline.

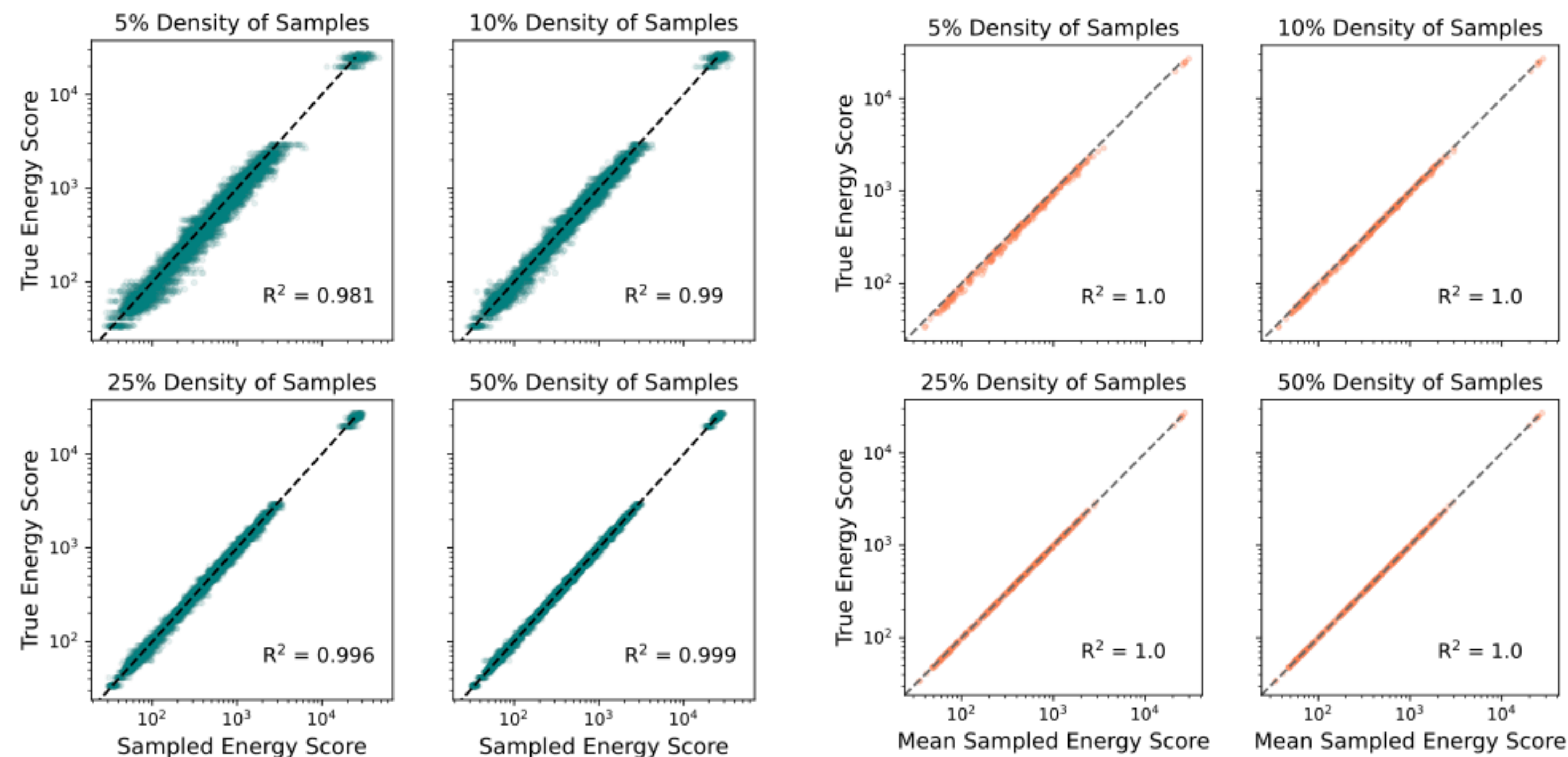
Most teams do well in southeastern US.



# 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 many iterations and compare with score found by using all trajectories.



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

# Conclusions

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 create a single score that spans multiple dimensions/types of target outcomes.

Possible extensions:

- Changing trajectory weights, forecasts, ensembling, multi-dimensional score