plus migration intensity measured during the previous night. Our aim here was to determine how much additional explanatory power we could achieve with a model that took into account recently observed conditions and behavior.

Performance and importance

To assess performance of the final model using weather forecasts instead of reanalysis (i.e. NARR) data, we tested the model using archived NAM forecasts made 1-3 days in advance. We did the same for GFS forecasts made 1-7 days in advance. Because GFS does not contain a visibility variable, we first retrained the model without visibility included in order to conduct this evaluation.

To assess model performance at unobserved spatial locations, we performed a cross-validation where we randomly removed one station (out of 143 total) from the dataset, retrained the model on the remaining data, and tested its performance on the withheld station.

We identified the predictor variables that were most important for model predictions using gain, a measure of the variable's importance in making accurate predictions. We also generated partial dependence plots using the R package *mlr* (*35*) to explore how these variables influence predictions. Here, we used a learning rate *eta* of 0.05 instead of 0.01 to make computation tractable.

Prediction intervals

We constructed empirical prediction intervals using residuals from XGBoost predictions for the validation dataset. We fitted a generalized additive model (*36*) on squared XGBoost residuals against the XGBoost-predicted value to account for an error variance that increased with the magnitude of the predicted value. The generalized additive model produced an estimated error variance for each predicted value, which we used to construct 90% prediction intervals using 0.05/0.95 Gaussian quantiles. We constructed separate models for upper and lower limits to allow for asymmetry in the width of the interval, and we used the Gamma distribution family in the generalized additive model to constrain the predicted variances to be non-negative.

Forecast output and estimation of nightly migration magnitude

Using our validated migration forecast model, we made predictions across the entire 12-km NAM grid. For smooth presentation, we averaged predictions across 9×9 cell blocks. We also used our model to estimate the total number of birds migrating over the continental United States each night. For this we used the NARR dataset because it is the best retrospective estimate of occurred conditions. For each 32-km NARR grid cell covering the continental United States, we multiplied the bird density estimate by the area of the cell and summed totals across all grid cells for each night.

NEXRAD radars operate at slightly different carrier frequencies (and hence different wavelengths) to reduce interference from neighboring radars, and this variation may introduce noise into estimates of total bird numbers if radars differ substantially in wavelength (*37*). However, such noise is likely to be minor because (1) most radars operate at more similar wavelengths than the example presented in (*37*), (2) variation in carrier frequency is not correlated with geography (i.e. no consistent spatial bias would be introduced), and (3) including wavelength as a model predictor to account for any systematic difference in detected bird densities did not appreciably change model performance.